Cross-modal Prototype Driven Network for Radiology Report Generation (Supplementary Material)

Jun Wang, Abhir Bhalerao, and Yulan He

Department of Computer Science, University of Warwick, UK

A.1 Implementation Details

Following the same strategy of previous work, e.g. [5,1], both images of a patient are utilized on IU-XRay and one image for MIMIC-CXR. In the training phase, images are first resized to $(256, 256)$ followed by a random cropping with the size of $(224, 224)$ before being fed into the model, while they are directly resized to $(224, 224)$ during the testing phase. We select the ResNet-101 [3] pretrained on ImageNet [2] as our visual extractor both in the prototype initialization module and our main task. Specifically, ResNet-101 produces patch features with 512 dimensions for each one in the main task. In the prototype initialization module, ResNet-101 extracts global visual representation with 2048 dimensions, and the global textual representation is obtain by a pretrained BERT [7] with 768 dimensions.

We utilize a randomly initialized Transformer as the backbone for the encoder-decoder module with 3 layers, 8 attention heads and 512 dimensions for the hidden states. The cross-modal prototype querying and responding follow a multi-head paradigm where each head has the same procedure as described in Section 3. The number of clusters N^P in equation (6) is set to 20. The pseudo label has 14 categories, hence the cross-modal prototype matrix contain $14 \times 20 = 280$ vectors. γ is set to 15 which means we only select the top 15 cross-modal prototype vectors to respond the single-modal representations. The term θ in the improved multi-label contrastive loss are 1.5 and 1.75 for the IU-Xray and MIMIC-CXR datasets respectively.

We use Adam as the optimizer [4] to optimize XPRONET under the cross entropy loss and our improved multi-label contrastive loss. λ and ϵ in equation (21) are 1 and 0.1. The learning rates are set to $1e-3$ and $2e-3$ for the visual extractor and encoder-decoder on IU-Xray, while MIMIC-CXR has a smaller learning rate with $5e-5$ and $1e-4$ respectively. The learning rates are decayed by 0.8 per epoch and the bath sizes are 16 for all the datasets. The same as most promising studies, we adopt a beam size of three in the report generation to balance the effectiveness and efficiency. Note that the optimal hyper-parameters are determined by estimating the models on the validation sets. We implement our model via the PyTorch [6] deep learning framework.

A.2 More Example Visualizations

This section demonstrates more visualization results predicted by XPRONet.

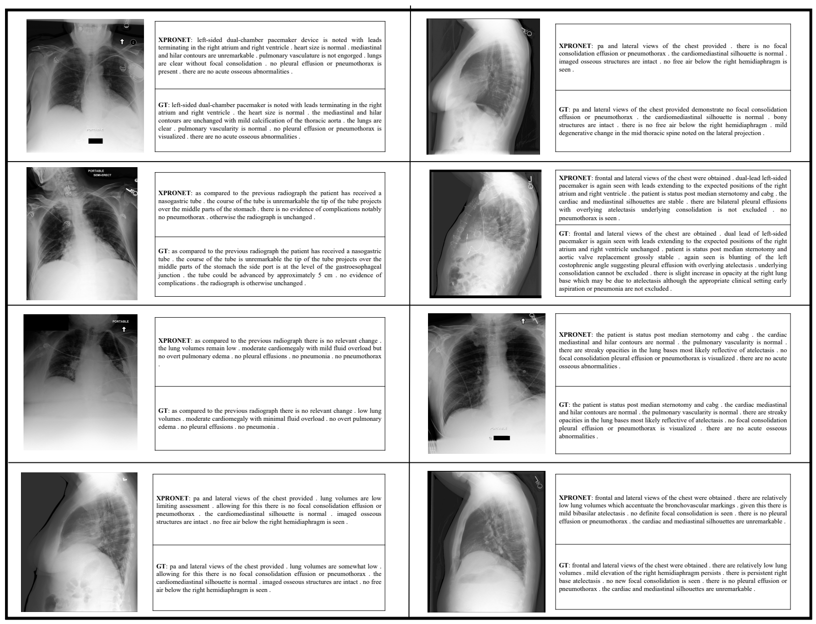

Fig. 1: The visulization of prediction results by XPRONET. GT is the abbreviation of the Ground Truth.

References

1. Z. Chen, Y. Shen, Y. Song, and X. Wan. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5904–5914, 2021.
2. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.
3. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
4. D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations, (Poster)*, 2015.
5. Y. Li, X. Liang, Z. Hu, and E. P. Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. *Advances in Neural Information Processing Systems*, 31, 2018.
6. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.
7. A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, 2020.