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. $\gamma$ is set to $15$ which means we only select the top $15$ cross-modal prototype vectors to respond the single-modal representations. The term $\theta$ 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. $\lambda$ and $\epsilon$ 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 visualization of prediction results by XPRONET. GT is the abbreviation of the Ground Truth.
References


