Cross-modal Prototype Driven Network for Radiology Report Generation

Jun Wang, Abhir Bhalerao, and Yulan He

Department of Computer Science, University of Warwick, UK
{jun.wang.3, abhir.bhalerao, yulan.he}@warwick.ac.uk

Abstract. Radiology report generation (RRG) aims to describe automatically a radiology image with human-like language and could potentially support the work of radiologists, reducing the burden of manual reporting. Previous approaches often adopt an encoder-decoder architecture and focus on single-modal feature learning, while few studies explore cross-modal feature interaction. Here we propose a Cross-modal PROtotype driven NETwork (XPRONET) to promote cross-modal pattern learning and exploit it to improve the task of radiology report generation. This is achieved by three well-designed, fully differentiable and complementary modules: a shared cross-modal prototype matrix to record the cross-modal prototypes; a cross-modal prototype network to learn the cross-modal prototypes and embed the cross-modal information into the visual and textual features; and an improved multi-label contrastive loss to enable and enhance multi-label prototype learning. XPRONET obtains substantial improvements on the IU-Xray and MIMIC-CXR benchmarks, where its performance exceeds recent state-of-the-art approaches by a large margin on IU-Xray and comparable performance on MIMIC-CXR.[1]

Keywords: Radiology Report Generation, Cross-Modal Pattern Learning, Prototype Learning, Transformers

1 Introduction

Radiology images, e.g., X-Ray and MRI, are widely used in medicine to support disease diagnosis. Nonetheless, traditional clinical practice is laborious since it requires the medical expert, such as a radiologist, to carefully analyze an image and then produce a medical report, which often takes more than five minutes [13]. This process could also be error-prone due to subjective factors, such as fatigue and distraction. Automatic radiology report generation, as an alternative to expert diagnosis, has therefore gained increasing attention from researchers. Automatic medical report generation has the potential to rapidly produce a report and assist a radiologist to make the final diagnosis significantly reducing the workload of radiologists and saving medical resources, especially in developing countries where well-trained radiologists can be in short supply.

[1] The code is publicly available at https://github.com/Markin-Wang/XProNet

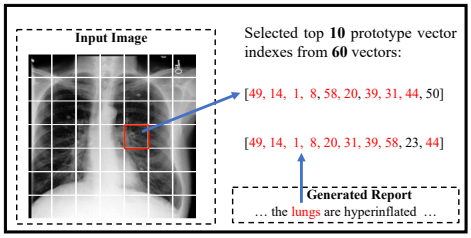

Fig. 1: An example generated report and the selected cross-modal prototype indices using XPRONET. The selected word "lungs" is marked as red and the associated image patch is highlighted in the red rectangle. The prototype indices selected both from the image patch and from the text instance are marked as red.

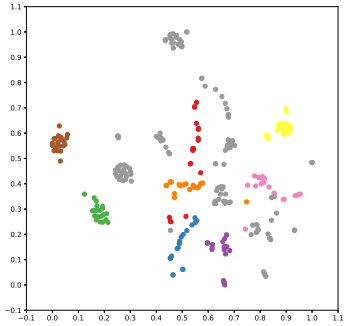

Fig. 2: A visualization of the cross-modal prototype matrix on the MIMIC-CXR dataset using T-SNE [25] . Points with the same colour come from the same prototype category.

Owing to developments in computer vision models for image captioning and availability of large-scale datasets, recently there have been significant advancements in automated radiology report generation [44,13,22]. Nevertheless, radiology report generation still remains a challenging task and is far from being solved. The reasons are three-fold. Firstly, unlike the traditional image captioning task which often produces only a single sentence, a medical report consists of several sentences and its length might be four-times longer than an image caption. Secondly, medical reports often exhibit more sophisticated linguistic and semantic patterns. Lastly, commonly used datasets suffer from notable data biases: the majority of the training samples are of normal cases, any abnormal regions often only exist in a small parts of an image, and even in pathological cases, most statements may be associated with a description of normal findings, e.g. see Figure 4. Overall, these problems present a substantial challenge to the modelling of cross-modal pattern interactions and learning informative features for accurate report generation.

Existing methods often focus on learning discriminative, *single*-modal features and ignore the importance of cross-modal interaction, essential for dealing with complex image and text semantic interrelationships. Thus, cross-modal interaction is of great importance as the model is required to generate a meaningful report only given the radiology image. Previous studies normally model cross-modal interaction by a self-attention mechanism on the extracted visual and textual features in an encoder-decoder architecture, which cannot adequately capture complex cross-modal patterns. Motivated by this, we propose a novel framework called *Cross-modal PROtotype driven NETwork* (XPRONET) which learns the cross-modal prototypes on the fly and utilizes them to embed cross-modal information into the single-model features. XPRONET regards the cross-modal prototypes as intermediate representations and explicitly establishes a cross-modal information flow to enrich single-modal features. Figure 1 shows

an example of the cross-modal information flow where the visual and textual features select almost the same (9 of top 10) cross-modal prototypes to perform interaction. These enriched features are more likely to capture the sophisticated patterns required for accurate report generation. Additionally, the imbalance problem is addressed by forcing single-model features to interact with their cross-modal prototypes via a class-related, cross-modal prototype querying and responding module. Our work makes three principal contributions:

1. We propose a novel end-to-end cross-modal prototype driven network where we utilize the cross-modal prototypes to enhance image and text pattern interactions. Leveraging cross-modal prototypes in this way for RRG has not been explicitly explored.
2. We employ a memory matrix to learn and record the cross-modal prototypes which are regarded as intermediate representations between the visual and textual features. A cross-modal prototype network is designed to embed cross-modal information into the single-modal features.
3. We propose an improved multi-label contrastive loss to learn cross-modal prototypes while simultaneously accommodating label differences via an adaptive controller term.

After a discussion of related work, our methods and implementation are described in detail in Section 3. Experimental results presented in Section 4 demonstrate that our approach outperforms a number of state-of-the-art methods over two widely-used benchmarks. We also undertake ablation studies to verify the effectiveness of individual components of our method. Discussion and proposals are given to inspire future work.

2 Related Work

Image Captioning Image captioning aims to generate human-like sentences to describe a given image. This task is considered as a high-level visual understanding problem which combines the research of computer vision and natural language processing. Recent state-of-the-art approaches [32,38,20,24,42,33] follow an encoder-decoder architecture and have demonstrated a great improvement in some traditional image captioning benchmarks. In particular, the most successful models [5,8,30,11] usually adopt the Transformer [36] as their backbone due to its self-attention mechanism and its impressive capability of extracting meaningful features for the task. However, these methods are designed for short textual description generation and are less capable for generating long reports. Though several works [16,26] have been proposed to deal with long text generation, they often cannot capture the specific medical observations and tend to produce reports ignoring abnormal regions in images, resulting in unsatisfactory performance.

Radiology Report Generation Inspired by the great success of encoder-decoder based frameworks in image captioning, recent radiology report generation methods have also employed similar architectures. Specifically, Jing et

al. [13] developed a hierarchical LSTM model to produce long reports and proposed a co-attention mechanism to detect abnormal patches. Liu et al. [22] proposed to firstly determine the topics of each report, which are then conditioned upon for report generation. Similarly, Zhang et al. [44] also ascertained the disease topics and utilized prior knowledge to assist report generation via a preconstructed knowledge graph. Liu et al. [21] extended this work by presenting a PPKED model which distills both the prior and posterior knowledge into report generation. A few works [27,29] have investigated reinforcement learning for improving the consistency of the generated reports. These encoder-decoder approaches often focus on extracting discriminative single-modal features (visual *or* textual), while few study explores the importance of the cross-modal pattern interactions.

The most similar work to ours is R2GenCMN [3] which utilizes an extra memory to learn the cross-modal patterns. Nonetheless, there are three main differences. First, we design a shared cross-modal prototype matrix to learn the class-related cross-modal patterns and propose an improved multi-label contrastive loss, while Chen et al. [3] randomly initialize a memory matrix and use a cross entropy loss. Additionally, our querying and responding process is class-related, that is, cross-modal pattern learning is only performed over the cross-modal prototypes sharing the *same* labels rather than on all cross-modal prototypes. Moreover, we adopt a more effective approach to distill the cross-modal information into the single-modal representations rather than the simple averaging function used in R2GenCMN. XPRONET is driven by the cross-modal *prototypes* which, to the best of our knowledge, has not been explored before in radiology report generation.

3 Methods

Our aim is to learn important informative cross-modal patterns and utilize them to explicitly model cross-modal feature interactions for radiology report generation, Figure 3 shows the overall architecture of XPRONET. The details of the main three modules, i.e., the image feature extractor, the cross-modal prototype network, and the encoder-decoder are described in the following subsections.

3.1 Image Feature Extractor

Given an input radiology image $\boldsymbol{I}$, a ResNet-101 [9] is utilized to extract the image features $\boldsymbol{v} \in \mathbb{R}^{H \times W \times C}$, shown in the blue-dashed rectangle in Figure 3. In particular, image features $\boldsymbol{v}$ are extracted from the last convolution layer, before the final average pooling operation. Here H, W *and* C are the height, width and the number of channels of an image, respectively. Once extracted, we linearize the image features $\boldsymbol{v}$ by concatenating the rows of the image features and regard each region (position) feature as a visual word token. The final feature representation sequence $\boldsymbol{v}_s \in \mathbb{R}^{HW \times C}$ is taken as the input for the subsequent

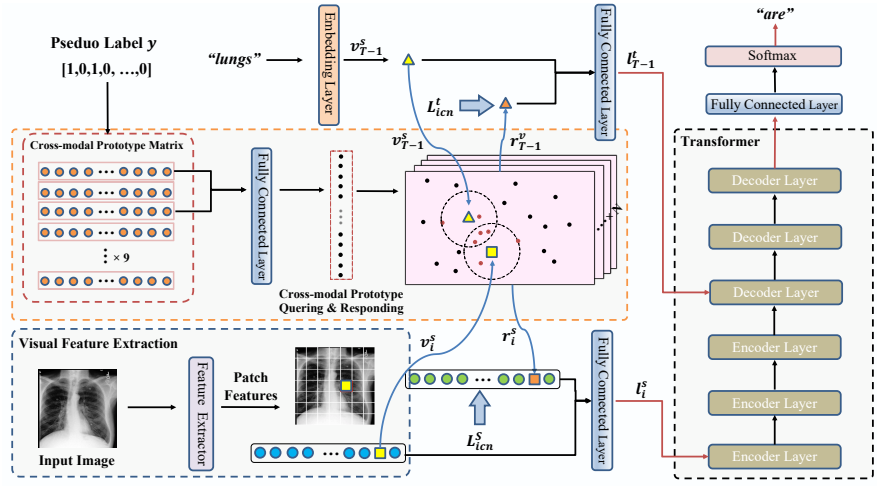

Fig. 3: The architecture of XPRONET: An image is fed into the Visual Feature Extractor to obtain patch features. A word at time step T (e.g. "lungs") is mapped onto a word embedding via an embedding layer. The visual and textual representations are then sent to the cross-modal prototype querying and responding module to perform cross-modal interaction on the selected cross-modal prototypes based on the associated pseudo label. Then the single-model feature are enriched by the generated responses through a linear layer and taken as the source inputs of the Transformer encoder-decoder to generate the report.

modules and is expressed as:

$$\{v_1^s, v_2^s, ..., v_i^s, ..., v_{N^s-1}^s, v_{N^s}^s\} = f_{ife}(\boldsymbol{I}), \tag{1}$$

where v_i denotes the region features in the i^{th} position of $\boldsymbol{v}_s$, $N^s = H \times W$, and $f_{ife}(\cdot)$ is the image feature extractor.

3.2 Cross-modal Prototype Network

Learning complex related patterns between image features and related textual descriptions is challenging. But cross-modal learning enables jointly learn informative representations of image *and* text. Central to our network is a prototype matrix which contains image pseudo-labels, initialized using an approach described below.

Pseudo Label Generation Cross-modal prototypes require category information for each sample, which is however often not provided in the datasets. To address this problem for prototype learning, we utilize CheXbert [34], an automatic radiology report labeler, to generate a pseudo label for each image-text pair. We denote the report associated with image $\boldsymbol{I}$ as:

$$\boldsymbol{R} = \{w_1, w_2, ..., w_i, ..., w_{N^r-1}, w_{N^r}\}, \tag{2}$$

where w_i is the i^{th} word in the report and N^r is the number of words in the report. The labelling process can then be formulated as:

$$\{y_1, y_2, ..., y_i, ..., y_{N^l-1}, y_{N^l}\} = f_{al}(\boldsymbol{R}), \tag{3}$$

where the result is an one-hot vector and $y_i \in \{0, 1\}$ is the prediction result for i^{th} category. Note that the value of one indicates the existence of that category, N^l is the number of categories, and $f_{al}(\cdot)$ denotes the automatic radiology report labeler.

Prototype Matrix Initialization Existing methods often directly model the cross-modal information interactions using the encoded features and learn implicitly cross-modal patterns. The length of the report, the imbalanced distribution of text descriptions of normal and abnormal cases, and complex cross modal patterns, make it hard to capture cross-modal patterns effectively. For better cross-modal pattern learning, we design a shared cross-modal prototype matrix $\boldsymbol{PM} \in \mathbb{R}^{N^l \times N^p \times D}$ to learn and store the cross-modal patterns, which can be considered as intermediate representations. Here N^p and D are the number of learned cross-modal prototypes for each category and the dimension for each prototype, respectively. $\boldsymbol{PM}$ is updated and learned during training, and then utilized by the class-related prototype querying and responding modules to explicitly embed the cross-modal information to the single-modal features.

The initialization of the prototype matrix is critical. One way is to randomly initialize the matrix [3], but this does not capture any meaningful semantic information and hampers the subsequent prototype learning. Therefore, we propose to utilize prior information to initialize a semantic cross-modal prototype matrix. Specifically, for an image-text pair $< \boldsymbol{I}, \boldsymbol{R} >$ with the associated pseudo class labels $\boldsymbol{y}$, we employ a pretrained ResNet-101 and BERT [34] to extract the global visual and textual representations, $\boldsymbol{o}^i \in \mathbb{R}^{1 \times C_1}$ and $\boldsymbol{o}^t \in \mathbb{R}^{1 \times C_2}$, where C_1 and C_2 are the number of channels extracted of the visual and textual representations, respectively. To improve robustness, we also extract the flipped image features $\boldsymbol{o}^{if} \in \mathbb{R}^{1 \times C_1}$. By repeating this process on all the training samples, we can obtain a group of feature sets for each class, formulated as:

$$\boldsymbol{R}_k^I = \{o_u^{i(f)} | y_{u,k} = 1\}, \ \boldsymbol{R}_k^T = \{o_u^t | y_{u,k} = 1\}. \tag{4}$$

Here $\boldsymbol{R}_k^I$ and $\boldsymbol{R}_k^T$ are the visual and textual feature sets for category k, $i(f)$ means either the original image i or the flipped image if, and $y_{u,k}$ denotes the label of category k for sample u. After that, we concatenate the visual and textual representations to form the cross-modal features, $\boldsymbol{r} \in \mathbb{R}^{1 \times D}$. Note that $D = C_1 + C_2$. Finally, K-Means [23] is employed to cluster each feature set into N_p clusters and the average of features in each cluster is used as an initial

cross-modal prototype for $\boldsymbol{PM}$. This process can be summarised as:

$$o_u = Concat(o_u^{i(f)}, o_u^t), \tag{5}$$

$$\{\boldsymbol{g}_1^k, ..., \boldsymbol{g}_{N^p-1}^k, \boldsymbol{g}_{N^p}^k\} = f_{km}(\boldsymbol{R}_k), \qquad \boldsymbol{g}_i^k = \{o_1^{k,i}, ..., o_{N_{k,i}^d}^{k,i}\}, \tag{6}$$

$$\boldsymbol{PM}(k,i) = \frac{1}{N_{k,i}^s} \sum_{j=0}^{N} r_j^{k,i}, \tag{7}$$

Where o_u and $\boldsymbol{R}_k$ are the concatenated cross-modal representation for sample u and the cross-modal feature set for category k, $\boldsymbol{g}_i^k$ is the i^{th} grouped cluster for k^{th} category returned by the K-Mean algorithm f_{km}. $N_{k,i}^d$ is the number of samples in the i^{th} cluster for k^{th} category. $\boldsymbol{PM}(k,i)$ then represents the i^{th} vector in the cross-modal prototype set for the k^{th} category.

Cross-modal Prototype Querying After obtaining the prototype matrix, similar to [3], we adopt a querying and responding process to explicitly embed the cross-modal information into the single-modal features. Different from [3], given an image, our cross-modal prototype querying measures the similarity between its single-modal representation and the cross-modal prototype vectors under the same label as the image, and selects the top γ vectors having the highest similarity to interact with the single-model representations. This process is illustrated in the yellow-dashed rectangle in Figure 3.

Given the image-text training pair $< \boldsymbol{I}, \boldsymbol{R} >$ and the associated pseudo label $\boldsymbol{y}$, the queried cross-modal prototype vectors for the sample are then generated. The queried cross-modal prototype vectors $\boldsymbol{pv} = \{\boldsymbol{PM}(k) |\ y_k = 1\}$, where $\boldsymbol{PM}(k)$ is the cross-modal prototype set for the k^{th} category generated by Equations (5) - (8). To filter out possible noise, a linear projection is applied to $\boldsymbol{pv}$ to map it to C_P dimensions before sending it into the querying process, as follows:

$$\boldsymbol{p} = \boldsymbol{pv} \cdot \boldsymbol{W}_{pv}, \tag{8}$$

where $W_{pv} \in \mathbb{R}^{D \times C_P}$ is a learnable weight matrix.

We denote the report representation output by the embedding layer as $\boldsymbol{v}_t = \{v_1^t, v_2^t, ..., v_i^t, ..., v_{N^t-1}^t, v_{N^t}^t\}$ and a cross-modal prototype vector as p_i, where $v_i^t \in \mathbb{R}^{1 \times C}$ is the i^{th} word embedding of the report. Before performing the querying, we linearly project the visual feature sequence $\boldsymbol{v}_s$, textual report embeddings $\boldsymbol{v}_t$ and the cross-modal prototype vector p_i into the same dimension d since they may have different dimensions:

$$v_i^{s*} = v_i^s \cdot \boldsymbol{W}_v, \quad v_i^{t*} = v_i^t \cdot \boldsymbol{W}_v, \quad p_i^* = p_i \cdot \boldsymbol{W}_p, \tag{9}$$

where $\boldsymbol{W}_v \in \mathbb{R}^{C \times d}$ and $\boldsymbol{W}_p \in \mathbb{R}^{C_P \times d}$ are two learnable weights. A similarity between each single-modal feature and cross-modal prototype vector pair is computed by:

$$D_{(i,u)}^s = \frac{v_j^{s*} \cdot p_u^*}{d}, \quad D_{(j,u)}^t = \frac{v_j^{t*} \cdot p_u^*}{d}. \tag{10}$$

Since the majority of the cross-modal prototypes might be irrelevant to the queried vectors, which may introduce noisy cross-modal patterns, we only select γ most similar vectors to respond to the query vectors. After that, we calculate the weights among these selected prototype vectors based on the similarities. This process between a cross-modal prototype p_u^*, a visual region representation v_i^{s*} and a textual word embedding v_i^{t*} is captured by:

$$w_{(i,u)}^s = \frac{D_{(i,u)}^s}{\sum_{j=1}^{\gamma} D_{(i,j)}^s}, \quad w_{(i,u)}^t = \frac{D_{(i,u)}^t}{\sum_{j=1}^{\gamma} D_{(i,j)}^t} \tag{11}$$

Cross-modal Prototype Responding After obtaining the top γ similar cross-modal prototype vectors and their weights, the next step is to generate the responses for the visual and textual features. In particular, we firstly transform the queried prototype vectors to the same representation space of the query vectors via a fully connected layer. The responses for the visual and textual features are created by taking the weighted sum over these transformed cross-modal prototype vectors:

$$e_{(i,j)}^s = p_{(i,j)}^{s*} \cdot \boldsymbol{W}_e, \qquad e_{(i,j)}^t = p_{(i,j)}^{t*} \cdot \boldsymbol{W}_e, \tag{12}$$

$$r_i^s = \sum_{j=1}^{\gamma} w_{(i,j)}^s \cdot e_{(i,j)}^s, \qquad r_i^t = \sum_{j=1}^{\gamma} w_{(i,j)}^t \cdot e_{(i,j)}^t, \tag{13}$$

where $p_{(i,j)}^{s*}$ and $p_{(i,j)}^{t*}$ are the j^{th} prototype vectors in the most similar cross-modal prototype sets for the i^{th} image patch and word, respectively. Similarly, the j^{th} transformed prototype vectors for i^{th} image patch and word are denoted as $e_{(i,j)}^s$ and $e_{(i,j)}^t$. We represent the responses for i^{th} image patch and word as r_i^s and r_i^t. The $w_{(i,j)}^s$ and $w_{(i,j)}^t$ are the weights obtained by Equations (11) to (12).

Feature Interaction Module The selected cross-modal prototype vectors contain class-related and cross-modal patterns. The last step is to introduce these informative patterns into the single-modal features via feature interaction. In [3], this is achieved by directly adding the single-model features and their associated responses, which pays the same attention to the responses and the single-modal features. However, this simple approach might be suboptimal given possibly noisy responses or non-discriminative single-model features. To mitigate this problem, we propose to automatically learn the importance difference and filter out noisy signals.

Specifically, we firstly concatenate the single-modal features with their associated responses. A linear layer is then applied to fuse the single-modal features and the cross-modal prototype vectors. Remember that the fused representations contain rich class-related features and cross-modal patterns. The process is:

$$\boldsymbol{l}^s = \boldsymbol{FCN}(Concat(\boldsymbol{v}^s, \boldsymbol{r}^s)), \quad \boldsymbol{l}^t = \boldsymbol{FCN}(Concat(\boldsymbol{v}^t, \boldsymbol{r}^t)), \tag{14}$$

where $\boldsymbol{FCN}$ denotes the fully connected layer and $Concat$ is the concatenating function. The outputs of the Feature Interaction Module are taken as the source inputs for the following Transformer module to generate the reports.

3.3 Reports Generation with Transformer

Transformers have been shown to be quite potent for NLP tasks, e.g., sentiment analysis [40,4,41], machine translation [43,39,2] and question answering [28,15,45]. Consequently, we adopt a transformer to generate the final reports. Generally, the Transformer consists of the Encoder and Decoder. At the first step, the responded visual features $\boldsymbol{l}^s$ are fed into the Encoder to obtain the intermediate representations. Combined with the current fused textual representation sequence $\boldsymbol{l}^t = \{l_1^t, l_2^t, ..., l_i^t, ..., l_{T-1}^s\}$, these intermediate representations are then taken as the source inputs for the Decoder to predict the current output. In general, the encoding and decoding processes can be expressed as:

$$\{m_1, m_2, ..., m_{N^s}\} = \boldsymbol{Encoder}(l_1^s, l_2^s, ..., l_{N^s}^s), \tag{15}$$

$$p_T = \boldsymbol{Decoder}(m_1, m_2, ..., m_{N^s}; l_1^t, l_2^t, ..., l_{T-1}^t), \tag{16}$$

where p_T denotes the word prediction for time step T. The complete report is obtained by repeating the above process.

3.4 Improved Multi-Label Contrastive Loss

Though the cross-modal prototype matrix is determistically initialized, further learning is required to learn class-related and informative cross-modal patterns, since the cross-modal patterns are actually far more sophisticated than the simple concatenation of the visual and textual representations in the Prototype Initialization module. Moreover, the cross-modal prototype features extractor (pretrained ResNet-101 and BERT) are not trained on our target benchmarks, leading to potentially noisy signals. Therefore, online cross-modal prototype learning becomes of greater significance.

A simple way is to utilize the widely used contrastive loss to supervise the learning of the cross-modal prototypes. Nonetheless, the vanilla contrastive loss is designed for the single-label prototype learning, while each training sample can belong to multiple categories in our task. Therefore, we modify the contrastive loss into a multi-label scenario by regarding the samples having at least one common label (excluding label 0) as positive pairs. If two samples do not share any common label, they form a negative pair. Instead of employing the contrastive loss on the responded features, we propose applying the loss on the responses since the fused features are used for medical report generation rather than for classification.

Given the visual responses $\boldsymbol{r}^s = \{r_1^s, r_2^s, ..., r_i^s, ..., r_{N^s-1}^s, r_{N^s}^s\}$ and textual responses $\boldsymbol{r}^t = \{r_1^t, r_2^t, ..., r_i^t, ..., r_{N^t-1}^t, r_{N^t}^s\}$, our modified multi-label contrastive

loss is formulated as:

$$\boldsymbol{L}_{icn}^{s} = \frac{1}{B^2}\sum_{i=1}^{B}\sum_{j:\boldsymbol{y}_i\otimes\boldsymbol{y}_j\neq 0}^{B}(\boldsymbol{\theta}^{-\frac{\boldsymbol{h_d}}{\boldsymbol{h_t}}} - Sim(\sigma(r_i^s, r_j^s)))+ \sum_{j:\boldsymbol{y}_i\otimes\boldsymbol{y}_j=0}^{B}\max(Sim(\sigma(r_i^s, r_j^s)) - \alpha, 0) \tag{17}$$

Here B denotes the number of training samples in one batch and $\otimes$ is the dot product operation. $\boldsymbol{y}_i \otimes \boldsymbol{y}_j \neq 0$ ensures that the responses r_i^s and r_j^s have at least one common label (excluding 0). $\sigma(\cdot)$ and $Sim(\cdot)$ are the the average function over all the image patch responses followed by the L_2 normalization and the cosine similarity function, respectively. Only negative pairs with similarity larger than a constant margin α can make a contribution to $\boldsymbol{L}_{cn}^{s}$.

Note that different from a standard contrastive loss, the maximum positive similarity (or one) is replaced with a label difference term, $\theta^{(\cdot)}$. In this way, the model can tolerate some dissimilarity between the positive pairs in terms of the label difference, instead of forcing them to be the same which is unreasonable under a multi-label setting:

$$h_d = \epsilon(abs(\boldsymbol{y}_i - \boldsymbol{y}_j)), \quad h_t = \epsilon(\boldsymbol{y}_i + \boldsymbol{y}_j), \tag{18}$$

where abs and ϵ are the absolute value and the summary functions, respectively. h_d calculates the number of different labels and h_t denotes the number of total labels of two training samples (excluding zero). Thus θ controls the relative tolerance where a smaller value represents less tolerance given the same label difference. An improved contrastive loss for textual responses $\boldsymbol{L}_{icn}^{t}$ is obtained in a similar way.

Objective Function Given the entire predicted report sequence $\{p_i\}$ and the associated ground truth report $\{w_i\}$, XPRONET is jointly optimized with a cross-entropy loss and our improved multi-label contrastive loss by:

$$\boldsymbol{L}_{ce} = -\frac{1}{N^r}\sum_{i=1}^{N^r} w_i \cdot log(p_i), \tag{19}$$

$$\boldsymbol{L}_{fnl} = \boldsymbol{L}_{ce} + \lambda\boldsymbol{L}_{icn}^{s} + \delta\boldsymbol{L}_{icn}^{t}, \tag{20}$$

Here λ an δ are two hyper-parameters which balance the loss contributions.

4 Experiments

We verify the effectiveness of XPRONET on two widely used medical report generation benchmarks, i.e., IU-Xray and MIMIC-CXR. Four common natural language processing evaluation metrics: BLEU{1-4} [31], ROUGE-L [19], METEOR [7] and CIDEr [37], are utilized to gauge performance. The implementation details are given in Appendix A.1.

Table 1: Comparative results of XPRONET with previous studies. The best values are highlighted in bold and the second best are underlined. BL, RG and MTOR are the abbreviations of BLEU, ROUGE and METEOR. The symbol $*$ denotes our replicated results with the official codes.

Dataset	Method	BL-1	BL-2	BL-3	BL-4	RG-L	MTOR	CIDEr
IU-Xray	*ST* [35]	0.216	0.124	0.087	0.066	0.306	-	-
	ADAATT [24]	0.220	0.127	0.089	0.068	0.308	-	0.295
	ATT2IN [33]	0.224	0.129	0.089	0.068	0.308	-	0.220
	$SentSAT + KG$ [44]	0.441	0.291	0.203	0.147	0.304	-	0.304
	HRGR [18]	0.438	0.298	0.208	0.151	0.322	-	<u>0.343</u>
	CoAT[13]	0.455	0.288	0.205	0.154	0.369	-	0.277
	$CMAS - RL$ [12]	0.464	0.301	0.210	0.154	0.362	-	0.275
	KERP [17]	<u>0.482</u>	<u>0.325</u>	<u>0.226</u>	0.162	0.339	-	0.280
	R2GenCMN* [3]	0.474	0.302	0.220	<u>0.168</u>	<u>0.370</u>	<u>0.198</u>	-
	XPRONET(*Ours*)	**0.525**	**0.357**	**0.262**	**0.199**	**0.411**	**0.220**	**0.359**
MIMIC -CXR	*RATCHET* [10]	0.232	-	-	-	0.240	0.101	-
	ST [35]	0.299	0.184	0.121	0.084	0.263	0.124	-
	ADAATT [24]	0.299	0.185	0.124	0.088	0.266	0.118	-
	ATT2IN [33]	0.325	0.203	0.136	0.096	<u>0.276</u>	0.134	-
	TopDown[1]	0.317	0.195	0.130	0.092	0.267	0.128	-
	R2GenCMN* [3]	**0.354**	<u>0.212</u>	<u>0.139</u>	<u>0.097</u>	0.271	<u>0.137</u>	-
	XPRONET(*Ours*)	<u>0.344</u>	**0.215**	**0.146**	**0.105**	**0.279**	**0.138**	-

Datasets IU-Xray [6] is a widely used benchmark which contains 7,470 X-ray images and 3,955 corresponding reports established by Indiana University. The majority of patients provided both the frontal and lateral radiology images. MIMIC-CXR [14] is a recently released large chest X-ray dataset with 473,057 X-ray images and 206,563 reports provided by the Beth Israel Deaconess Medical Center. Both of these two datasets are publicly available [2]. We follow the same data splits proportions as [18] to divide the IU-Xray dataset into train (70%), validation (10%) and test (20%) sets, while the official data split is adopted for the MIMIC-CXR dataset.

Comparisons with Previous Studies Here, we compare the experimental results with previous studies on the IU-Xray and MIMIC-CXR datasets. As shown in Table 1, ours (XPRONET) outperforms the previous best SOTA method of R2GenCMN by a noteable margin on the IU-Xray dataset. In particular, XPRONET surpasses the second best-performing method by 4.3%, 3.1% and 4.1% on BLEU-1, BLEU-4 and RG-L scores respectively. A similar pattern can be seen on the MIMIC-CXR benchmark where XPRONET achieves the best performance on all the evaluation metrics except BLEU-1 in which it is slightly inferior to R2GenCMN. We mainly attribute the improved performance to the enriched single-modal feature representation via the cross-modal prototype learning. The superiority of XPRONET on IU-Xray is more obvious than MIMIC-CXR. This could be partly explained by the data size differences as the number of samples in MIMIC-CXR is almost 50 times larger than IU-Xray, hence

[2] https://openi.nlm.nih.gov/
https://physionet.org/content/MIMIC-cxr-jpg/2.0.0/

Table 2: The experimental results of ablation studies on the IU-Xray and MIMIC-CXR datasets. The best values are highlighted in bold. BL and RG are the abbreviations of BLEU and ROUGE.

IU-Xray	**BL-1**	**BL-2**	**BL-3**	**BL-4**	**RG-L**	**METEOR**
XPRONET	**0.525**	**0.357**	**0.262**	**0.199**	**0.411**	**0.220**
w/o PI	0.476	0.307	0.218	0.160	0.371	0.196
w/o IMLCS	0.471	0.307	0.215	0.159	0.377	0.196
w/o CMPNet	0.467	0.303	0.210	0.155	0.367	0.197
MIMIC-CXR	**BL-1**	**BL-2**	**BL-3**	**BL-4**	**RG-L**	**METEOR**
XPRONET	**0.344**	**0.215**	**0.146**	**0.105**	**0.279**	**0.138**
w/o PI	0.329	0.205	0.139	0.100	0.275	0.133
w/o IMLCS	0.336	0.204	0.137	0.098	0.269	0.135
w/o CMPNet	0.321	0.198	0.133	0.095	0.273	0.131

it is more difficult to learn informative and class-related cross-modal prototypes. We present a visual example in Figure 4 and give a further analysis below.

Ablation Analysis Ablation studies were conducted to further explore the impact of each component of XPRONET on report generation performance. We investigated the following variants:

- **XPRONET w/o CMPNet**: the base model which only consists of the visual extractor (ResNet-101) and the encoder-decoder (Transformer) without other extensions.
- **XPRONET w/o PI**: XPRONET without the cross-modal Prototype Initialization (PI), i.e., the cross-modal prototype matrix is randomly initialized.
- **XPRONET w/o IMLCS**: XPRONET without the improved multi-label contrastive loss (IMLCS). We replace the adaptable maximum similarity $\theta^{-\frac{h_d}{h_t}}$ in Equation (17) with one to switch it back to the standard multi-label contrastive loss.

The main results of the ablation studies of XPRONET are shown in Table 2. First, all the three components, i.e., prototype initialization, improved multi-label contrastive loss and the whole cross-modal prototype network architecture, significantly boost the performance as a notable drop can be seen when any of them is removed. For instance, the BLEU-4 score decreases from 0.199 to 0.160 and 0.105 to 0.100 on the IU-Xray and MIMIC-CXR datasets when the prototype initialization is removed. Similarly, removing the improved multi-label contrastive loss lead to lower scores on BLEU-2 and ROUGE-L. These results verify the importance of informatively initializing the cross-modal prototype and allowing some dissimilarity between positive pairs under the multi-label, cross-modal prototype learning settings.

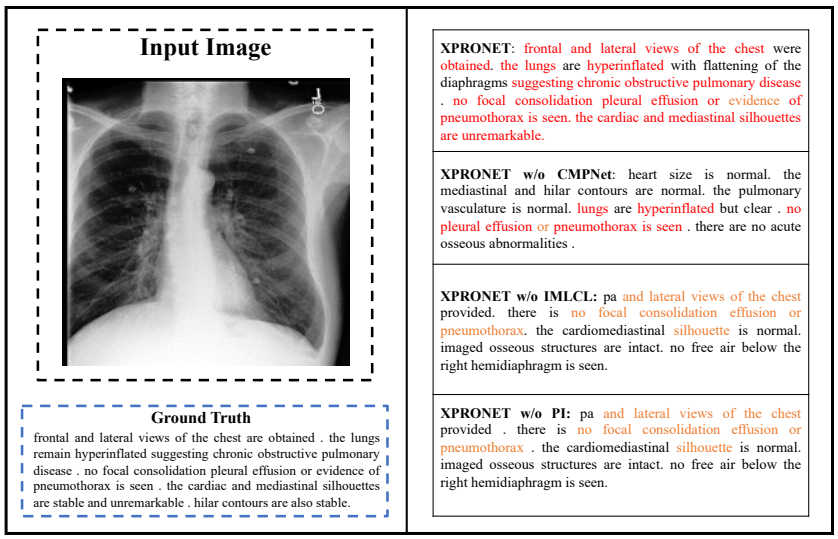

Fig. 4: An example of the report generated by different models. The ground truth report is shown in the blue dashed rectangle. Words that occurred in the ground truth are marked as red.

Moreover, the biggest performance drop can be seen on the model without the whole cross-modal prototype network on all the evaluation metrics, e.g., 0.525 to 0.467 and 0.344 to 0.321 of BL-1 on IU-Xray and MIMIC-CXR dataset respectively. An example visualization is shown in Figure 4 to illustrate the strength of XPRONET. More example visualizations are given in Appendix A.2. As we can see, XPRONET can capture the abnormal information and generate a better report, while the reaming models tend to produce sentences ignoring the abnormal patterns observed in images. This could be attributed to the well-learned cross-modal prototypes and the class-related querying and responding module which better capture the cross-modal flow and embed the prototype information into the feature learning procedure. We illustrate the cross-modal prototype matrix extracted from the linear projection (Equation (8)) in Figure 2. It can be seen that there is an obvious clustering pattern shown in the cross-modal prototype matrix. It should be mentioned that XPRONET can tolerate some dissimilarities between positive pairs, hence a category always occurring with other categories may lead to the associated prototypes being scattered with others (e.g., the orange), which is an expected outcome. To further explore the effectiveness of the XPRONET, we show an example of the generated report and the selected cross-modal prototype indices in Figure 1. For the word "lungs" and its corresponding image patch, the majority (nine of ten) of their selected responding cross-modal prototypes are the same, indicating that they learn the same cross-modal patterns and establish the cross-modal information flow via XPRONET, which is the expected behavior.

The sensitivity of XPRONET to the number of responding prototype vectors γ is shown in Figure 5. The BL-4 score reduces modestly when γ increases from 13 to 14, and then culminates at (0.199) at 15, after which the score decreases steadily to 0.171 with the γ increasing to 17 on the IU-Xray dataset. Generally, excessive or less responding prototype vectors can lead to notable performance drop. The reason being that excessive cross-modal prototype vectors may introduce noisy information, while inadequate numbers cannot provide sufficient cross-modal and class-related patterns. Figure 6 illustrates the influence of the tolerance rate controller term θ of XPRONET on the MIMIC-CXR benchmark. As we can see, the best performance is achieved with a θ value of 1.750, and performance drops at other values. A smaller θ represents a larger maximum similarity which forces the positive pairs to be more similar, causing a performance drop given dissimilar positive pairs. In contrast, XPRONET cannot learn useful cross-modal prototypes with a large θ which leads to a small maximum similarity. Therefore, it appears important to strike a good balance between the cross-modal prototype learning and dissimilarity tolerance.

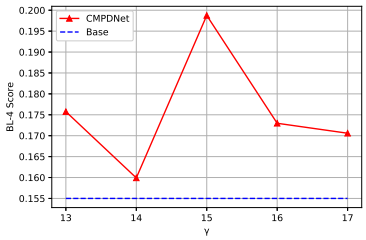

Fig. 5: Effect of varying γ, number of responding prototype vectors on (BLEU-4 score).

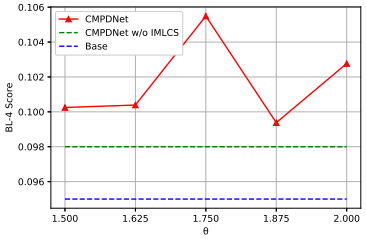

Fig. 6: Effect of varying θ, tolerance rate control on (BLEU-4 score).

5 Conclusions

We propose a novel cross-modal prototype driven framework for medical report generation, XPRONET, which aims to explicitly model cross-modal pattern learning via a cross-modal prototype network. The class-related cross-modal prototype querying and responding module distills the cross-modal information into the single-model features and addresses the data bias problem. An improved multi-label contrastive loss is designed to better learn the cross-modal prototypes and can be easily incorporated into existing works. Experimental results on two publicly available benchmark datasets verify the superiority of XPRONET. We also provide ablation studies to demonstrate the effectiveness of the proposed component parts. A potential way to improve XPRONET is to increase the number of cross-modal prototypes, especially for larger datasets. In addition, we speculate that a more effective clustering approach in cross-modal prototype matrix initialization could bring further improvements.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6077–6086 (2018)
2. Bao, G., Zhang, Y., Teng, Z., Chen, B., Luo, W.: G-transformer for document-level machine translation. 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). pp. 3442–3455 (2021)
3. Chen, Z., Shen, Y., Song, Y., Wan, X.: 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). pp. 5904–5914 (2021)
4. Cheng, J., Fostiropoulos, I., Boehm, B., Soleymani, M.: Multimodal phased transformer for sentiment analysis. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 2447–2458 (2021)
5. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10578–10587 (2020)
6. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
7. Denkowski, M., Lavie, A.: Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In: Proceedings of the sixth workshop on Statistical Machine Translation. pp. 85–91 (2011)
8. Guo, L., Liu, J., Zhu, X., Yao, P., Lu, S., Lu, H.: Normalized and geometry-aware self-attention network for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10327–10336 (2020)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
10. Hou, B., Kaissis, G., Summers, R.M., Kainz, B.: Ratchet: Medical transformer for chest x-ray diagnosis and reporting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 293–303. Springer (2021)
11. Ji, J., Luo, Y., Sun, X., Chen, F., Luo, G., Wu, Y., Gao, Y., Ji, R.: Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1655–1663 (2021)
12. Jing, B., Wang, Z., Xing, E.: Show, describe and conclude: On exploiting the structure information of chest x-ray reports. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 6570–6580 (2019)
13. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2577–2586 (2018)
14. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)

15. Kacupaj, E., Plepi, J., Singh, K., Thakkar, H., Lehmann, J., Maleshkova, M.: Conversational question answering over knowledge graphs with transformer and graph attention networks. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 850–862 (2021)
16. Krause, J., Johnson, J., Krishna, R., Fei-Fei, L.: A hierarchical approach for generating descriptive image paragraphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 317–325 (2017)
17. Li, C.Y., Liang, X., Hu, Z., Xing, E.P.: Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 6666–6673 (2019)
18. Li, Y., Liang, X., Hu, Z., Xing, E.P.: Hybrid retrieval-generation reinforced agent for medical image report generation. Advances in Neural Information Processing Systems **31** (2018)
19. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81 (2004)
20. Liu, F., Ren, X., Liu, Y., Wang, H., Sun, X.: simnet: Stepwise image-topic merging network for generating detailed and comprehensive image captions. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 137–149 (2018)
21. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 13753–13762 (2021)
22. Liu, G., Hsu, T.M.H., McDermott, M., Boag, W., Weng, W.H., Szolovits, P., Ghassemi, M.: Clinically accurate chest x-ray report generation. In: Machine Learning for Healthcare Conference. pp. 249–269. PMLR (2019)
23. Lloyd, S.: Least squares quantization in pcm. IEEE Transactions on Information Theory **28**(2), 129–137 (1982)
24. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 375–383 (2017)
25. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(11) (2008)
26. Melas-Kyriazi, L., Rush, A.M., Han, G.: Training for diversity in image paragraph captioning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 757–761 (2018)
27. Miura, Y., Zhang, Y., Tsai, E., Langlotz, C., Jurafsky, D.: Improving factual completeness and consistency of image-to-text radiology report generation. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5288–5304 (2021)
28. Naseem, T., Ravishankar, S., Mihindukulasooriya, N., Abdelaziz, I., Lee, Y.S., Kapanipathi, P., Roukos, S., Gliozzo, A., Gray, A.: A semantics-aware transformer model of relation linking for knowledge base question answering. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 256–262 (2021)
29. Nishino, T., Ozaki, R., Momoki, Y., Taniguchi, T., Kano, R., Nakano, N., Tagawa, Y., Taniguchi, M., Ohkuma, T., Nakamura, K.: Reinforcement learning with imbalanced dataset for data-to-text medical report generation. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 2223–2236 (2020)

30. Pan, Y., Yao, T., Li, Y., Mei, T.: X-linear attention networks for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10971–10980 (2020)
31. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
32. Pei, W., Zhang, J., Wang, X., Ke, L., Shen, X., Tai, Y.W.: Memory-attended recurrent network for video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8347–8356 (2019)
33. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7008–7024 (2017)
34. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A.Y., Lungren, M.: 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). pp. 1500–1519 (2020)
35. Sukhbaatar, S., Weston, J., Fergus, R., et al.: End-to-end memory networks. Advances in Neural Information Processing Systems **28** (2015)
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
37. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4566–4575 (2015)
38. Wang, J., Tang, J., Luo, J.: Multimodal attention with image text spatial relationship for ocr-based image captioning. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 4337–4345 (2020)
39. Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D.F., Chao, L.S.: Learning deep transformer models for machine translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 1810–1822 (2019)
40. Wang, Z., Wan, Z., Wan, X.: Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In: Proceedings of The Web Conference 2020. pp. 2514–2520 (2020)
41. Yang, K., Xu, H., Gao, K.: Cm-bert: Cross-modal bert for text-audio sentiment analysis. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 521–528 (2020)
42. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4651–4659 (2016)
43. Zhang, J., Luan, H., Sun, M., Zhai, F., Xu, J., Zhang, M., Liu, Y.: Improving the transformer translation model with document-level context. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 533–542 (2018)
44. Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., Xu, D.: When radiology report generation meets knowledge graph. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12910–12917 (2020)
45. Zhao, X., Xiao, F., Zhong, H., Yao, J., Chen, H.: Condition aware and revise transformer for question answering. In: Proceedings of The Web Conference 2020. pp. 2377–2387 (2020)