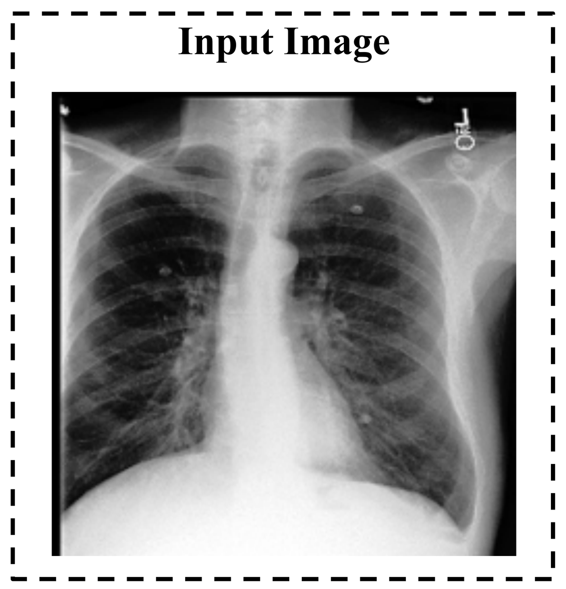
Cross-modal Prototype Driven Network for Radiology Report Generation

Jun Wang, Abhir Bhalerao, Yulan He

ECCV2022

Abstract

- **Radiology Report Generation** (RRG) aims to describe automatically a radiology image with human-like language.
- Approaches often use encoder-decoder architectures but focus on **single-modal feature learning.**
- We propose a CROSS-modal PROtotype driven NETwork (**XPRONET**) to promote **cross-modal pattern learning** to improve radiology report generation from X-rays.
- XPRONET obtains **substantial** improvements on the IU-Xray and MIMIC-CXR benchmarks, exceeding recent state-of-the-art approaches.



Ground Truth
frontal and lateral views of the chest are obtained . the lungs remain hyperinflated suggesting chronic obstructive pulmonary disease . no focal consolidation pleural effusion or evidence of pneumothorax is seen . the cardiac and mediastinal silhouettes are stable and unremarkable . hilar contours are also stable.

XPRONET: frontal and lateral views of the chest were obtained. the lungs are hyperinflated with flattening of the diaphragms suggesting chronic obstructive pulmonary disease . no focal consolidation pleural effusion or evidence of pneumothorax is seen. the cardiac and mediastinal silhouettes are unremarkable.

XPRONET w/o CMPNet: heart size is normal. the mediastinal and hilar contours are normal. the pulmonary vasculature is normal. lungs are hyperinflated but clear . no pleural effusion or pneumothorax is seen . there are no acute osseous abnormalities .

XPRONET w/o IMLCL: pa and lateral views of the chest provided. there is no focal consolidation effusion or pneumothorax. the cardiomediastinal silhouette is normal. imaged osseous structures are intact. no free air below the right hemidiaphragm is seen.

XPRONET w/o PI: pa and lateral views of the chest provided . there is no focal consolidation effusion or pneumothorax . the cardiomediastinal silhouette is normal. imaged osseous structures are intact. no free air below the right hemidiaphragm is seen.

Fig. 1: 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.

Challenges

- Radiology reports consist of several sentences and their length may be four-times longer than image captions.
- Medical reports often exhibit more sophisticated linguistic and semantic patterns.
- Datasets suffer from notable biases: (1) a majority of the training samples are of *normal* cases; (2) any abnormal regions often only exist in small parts of an image, even in pathological cases (3) most statements may be associated with a description of normal findings.

Architecture

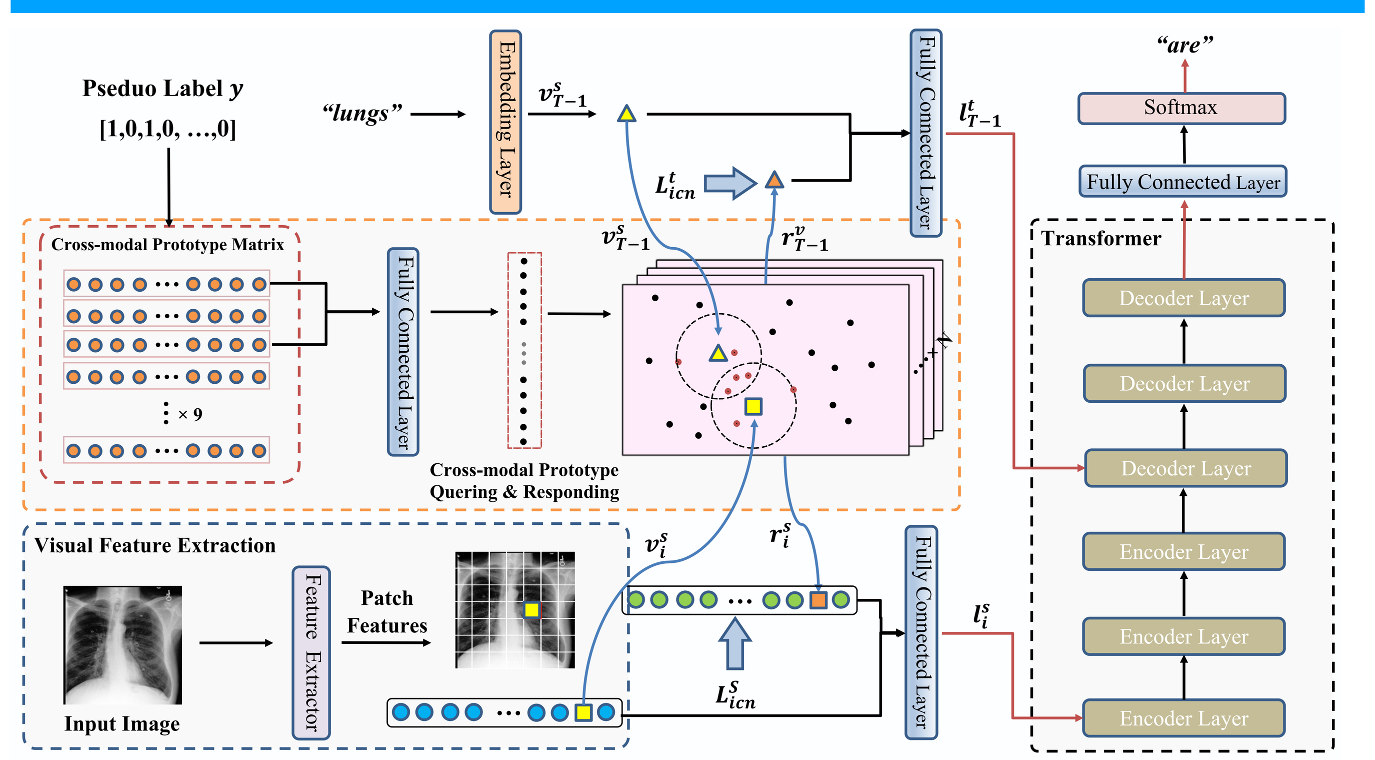


Fig. 2: The architecture of XPRONET: An image is fed into the Visual Feature Extractor to obtain patch features. A word at time step T (e.g. "lungs") is mapped onto a word embedding via an embedding layer. The visual and textual representations are then sent to the cross-modal prototype querying and 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.

PRESENTER
Jun Wang
@markin-wang
THE UNIVERSITY OF WARWICK
Department of Computer Science

Cross-modal prototypes increase performance of radiology report generation

1. **A novel end-to-end cross-modal prototype driven network is developed utilising cross-modal prototypes to enhance image and text pattern interactions**
2. **A memory matrix is used to learn and record the cross-modal prototypes which are regarded as intermediate representations between the visual and textual features**
3. **An improved multi-label contrastive loss learns cross-modal prototypes while simultaneously accommodating label differences via an adaptive controller term**

Scan for our paper

We are very grateful to the support of the UKRI Tier 2 Sulis HPC Platform, enabling the efficient training on the MIMIC-CXR dataset.

Method

Aim: To learn important informative cross-modal patterns and utilize them to explicitly model cross-modal feature interactions for RRG.

A. **Image Feature Extractor**: Given an input radiology image and its report , we firstly extract the visual feature sequences (tokens) and pseudo labels.

B. **Prototype Matrix Initialisation**: We design a shared cross-modal prototype matrix $PM \in \mathbb{R}^{N^l \times N^p \times D}$ to learn and store the cross-modal patterns, which can be considered as **intermediate** representations. It is initialized from the clustered, concatenated features (visual+textual) via K-Mean algorithm.

C. **Cross-modal Prototype Querying&Responding**: To generate the responses containing the most related cross-modal patterns to visual/textual features:

- Measure the similarity (weight) between its **single**-modal representation and the **cross**-modal prototype vectors.
- Select the top γ vectors having the highest similarity to interact with the single-model representations.
- Generate the responses $\boldsymbol{r}^s$ and $\boldsymbol{r}^t$ for the visual and textual features by taking the weighted sum over these transformed cross-modal prototype vectors.

D. **Feature Interaction:** The last step is to introduce these informative patterns (selected cross-modal vectors) into the single-modal features by a linear layer which takes the concatenated single-model features and responses as input and outputs the fused features $\boldsymbol{l}^s$ and $\boldsymbol{l}^t$.

E. **Report Generation with Transformer**: Given the fused visual and textual representation sequences $\boldsymbol{l}^s$ and $\boldsymbol{l}^t = \{l^t_1, l^t_2, \ldots, l^t_{T-1}\}$, the report is generated by the encoder-decoder through a repeating process:

$$\{m_1, m_2, \ldots, m_{N^s}\} = \boldsymbol{Encoder}(l^s_1, l^s_2, \ldots, l^s_{N^s}) \quad (1)$$

$$p_T = \boldsymbol{Decoder}(m_1, m_2, \ldots, m_{N^s}; l^t_1, l^t_2, \ldots, l^t_{T-1}) \quad (2)$$

F. **Prototype Learning**: An improved multi-label contrastive loss is proposed to further supervise the learning of the crocs-modal prototypes, where the maximum positive similarity is replaced with a label difference term, $\theta^{(\cdot)}$:

$$L^s_{icn} = \frac{1}{B^2}\sum_{i=1}^{B}\sum_{j:y_i\otimes y_j\neq 0}^{B}(\theta^{-\frac{h_d}{h_t}} - Sim(\sigma(r^s_i, r^s_j))) + \sum_{j:y_i\otimes y_j=0}^{B}\max(Sim(\sigma(r^s_i, r^s_j)) - \alpha, 0) \quad (3)$$

$$h_d = \epsilon(abs(\boldsymbol{y}_i - \boldsymbol{y}_j)), \quad h_t = \epsilon(\boldsymbol{y}_i + \boldsymbol{y}_j) \quad (4)$$

Results

Dataset	Method	BL-1	BL-2	BL-3	BL-4	RG-L	MTOR	CIDEr
IU-Xray	*ST* [38]	0.216	0.124	0.087	0.066	0.306	-	-
	ADAATT [26]	0.220	0.127	0.089	0.068	0.308	-	0.295
	ATT2IN [36]	0.224	0.129	0.089	0.068	0.308	-	0.220
	SentSAT + KG [47]	0.441	0.291	0.203	0.147	0.304	-	0.304
	HRGR [20]	0.438	0.298	0.208	0.151	0.322	-	<u>0.343</u>
	CoAT[14]	0.455	0.288	0.205	0.154	0.369	-	0.277
	CMAS − RL [13]	0.464	0.301	0.210	0.154	0.362	-	0.275
	KERP [19]	<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* [11]	0.232	-	-	-	0.240	0.101	-
	ST [38]	0.299	0.184	0.121	0.084	0.263	0.124	-
	ADAATT [26]	0.299	0.185	0.124	0.088	0.266	0.118	-
	ATT2IN [36]	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**	-

Table 1: Comparative results of XPRONET with previous studies. The best values are highlighted in bold and the second best are underlined.

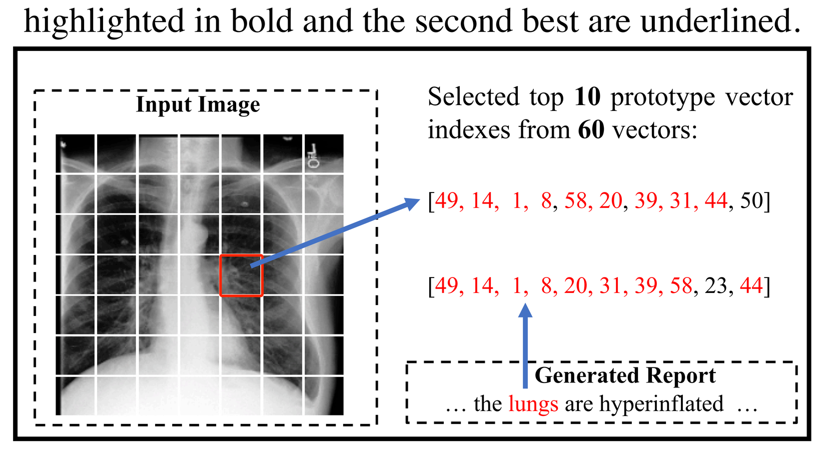


Fig. 3: An example generated report and the selected cross-modal prototype indices.The prototype indices selected both from the image patch and from the text instance are marked as red.

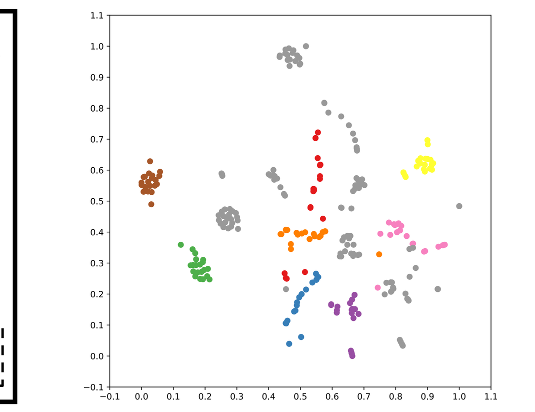


Fig. 4: T-SNE visualization of the cross-modal prototype matrix on the MIMIC-CXR. Points with the same colour come from the same prototype category.