Cross-modal Contrastive Attention Model for Medical Report Generation

Xiao Song [1], Xiaodan Zhang[1], Junzhong Ji[1,*], Ying Liu[2], Pengxu Wei [3]
[1] Beijing University of Technology
[2] Peking University Third Hospital
[3] Sun Yat-sen University
xiaos@emails.bjut.edu.cn; {zhangxiaodan, jjz01}@bjut.edu.cn;
lyyulia@163.com; weipx3@mail.sysu.edu.cn

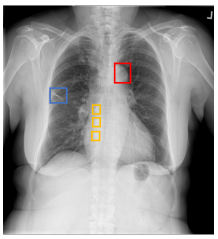
Abstract

Medical report automatic generation has gained increasing interest recently as a way to help radiologists write reports more efficiently. However, this image-to-text task is rather challenging due to the typical data biases: 1) Normal physiological structures dominate the images, with only tiny abnormalities; 2) Normal descriptions accordingly dominate the reports. Existing methods have attempted to solve these problems, but they neglect to exploit useful information from similar historical cases. In this paper, we propose a novel Cross-modal Contrastive Attention (CMCA) model to capture both visual and semantic information from similar cases, with mainly two modules: a Visual Contrastive Attention Module for refining the unique abnormal regions compared to the retrieved case images; a Cross-modal Attention Module for matching the positive semantic information from the case reports. Extensive experiments on two widely-used benchmarks, IU X-Ray and MIMIC-CXR, demonstrate that the proposed model outperforms the state-of-the-art methods on almost all metrics. Further analyses also validate that our proposed model is able to improve the reports with more accurate abnormal findings and richer descriptions.

1 Introduction

Medical report generation task in practice demands radiologists carefully examine details of images and write corresponding reports, which is time-consuming and technically rigorous. In addition, with the explosion of medical images, generating medical reports has increasingly become a tough burden for radiologists in clinical diagnosis and treatment. Thus, it is extremely desired for automatically generating medical reports from medical images, which has also attracted increasing attention especially in the Chest X-ray report generation.

Recently, the widely used Encoder-Decoder framework in image captioning task (Karpathy and Fei-Fei, 2015; You et al., 2016; Vaswani et al., 2017; Anderson et al., 2018) has been successfully inherited by medical report generation and has made great improvements (Jing et al., 2018; Zhang et al., 2017; Shin et al., 2016a; Wang et al., 2018; Li et al., 2019; Chen et al., 2020). Nevertheless, as shown in Figure 1, different from image captioning, medical report generation faces typical data biases which cause the failing of generating accurate descriptions: 1) the abnormal regions are usually tiny, rare and hard to be recognized in medical images with monotonous and homogeneous features (Guan et al., 2021; Li et al., 2018b; Guan et al., 2020); 2) the abnormal text descriptions are correspondingly rare in reports and the normal descriptions dominate the whole datasets (Shin et al., 2016b; Xue et al., 2018; Jing et al., 2019; Liu et al., 2021a,b).



Ground Truth: There is scarring in the right mid and upper lung zone with surgical clips identified as well. There is no pleural effusion or pneumothorax. The heart is not significantly enlarged. There are atherosclerotic changes of the aorta. Arthritic changes of the skeletal structures are noted.

Baseline: The lungs are clear. There is no pleural effusion or pneumothorax. The heart and mediastinum are normal. The skeletal structures are normal.

CMCA: The lungs are clear. There is no pleural effusion or pneumothorax. The heart is not significantly enlarged. There are calcified mediastinal lymph. There are atherosclerotic changes of the aorta. Arthritic changes of the skeletal structures are noted.

Figure 1: One example of Chest X-ray image with the corresponding ground truth, Baseline (Vaswani et al., 2017) and our model generated reports. The abnormal regions and their corresponding descriptions are marked in same colors, showing serious data biases of this task.

To tackle these issues, Jing et al. (Jing et al., 2019) employed an auxiliary detector to identify abnormality terms. Liu et al. (Liu et al., 2021b) compared the input image with normal samples to distill the visual abnormal information. However, these approaches mainly probed abnormalities from images themselves or comparing with manual selected normal samples, without considering the importance of exploiting abnormal information from historical similar cases and making use of their visual and semantic information. Based

*Corresponding Author.

Proceedings of the 29th International Conference on Computational Linguistics, pages 2388–2397
October 12–17, 2022.

on the observation that similar images are more likely to have similar reports (Ramos et al., 2014), we presume that taking the most similar historical case as a contrastive reference will make models able to relieve the data biases and capture more critical visual and semantic information. Unfortunately, exploring positive semantic information from cases faces another challenge on cross-modal matching (Xu et al., 2020; Liang et al., 2021): the retrieved report which contains useful but noisy semantic information is hard to be aligned with the input image solely across the unmatched visual-semantic modalities.

In this paper, we propose a novel Cross-modal Contrastive Attention (CMCA) model to tackle the aforementioned problems. CMCA first retrieves the most similar case for the input image from a historical database, then generates a contrastive feature by enlarging the differences and maintaining the commons between the input image and the retrieved image. Subsequently, the contrastive feature is used to extract visually abnormal and semantically matched information through two modules: a Visual Contrastive Attention Module (VCAM) and a Cross-modal Attention Module (CAM). Specifically, VCAM extracts discriminative abnormal visual information from the contrastive feature, where the unique abnormal regions of input image are enhanced and similar regions are retained. CAM matches the positive semantic information from the retrieved report by aligning it with the contrastive feature, which builds interactions across the unmatched visual and semantic modalities. Finally, we propose a Parallel Attention Module (PAM) to further enhance the feature representation for generating accurate report. Extensive experiments on the widely-used benchmark IU X-Ray (Demner-Fushman et al., 2016) dataset and the largest public MIMIC-CXR (Johnson et al., 2019) dataset demonstrate that our model outperforms the state-of-the-art methods.

Overall, our contributions are as follows:

- We propose to take the most similar historical case as a contrastive reference to relieve the data biases for medical report generation.
- We propose a novel Cross-modal Contrastive Attention model to distill unique abnormal features for input image and match positive words from the case report.
- Extensive experimental results on the public IU X-Ray and MIMIC-CXR datasets demonstrate the effectiveness of the proposed model.
- We conduct analyses to validate the hypothesis that historical similar cases can significantly assist this task, and our CMCA model is able to generate reports with more accurate abnormal findings and richer descriptions.

2 Method

In this section, we introduce the background and the details of the proposed CMCA model in order.

2.1 Background

2.1.1 Overall Framework

Our CMCA model follows the Encoder-Decoder pipeline, as shown in Figure 2.

In Encoder: Firstly, the spatial visual feature X_I of input image I can be extracted through a DenseNet (Huang et al., 2017) network:

$$X_I = DenseNet(I), \quad (1)$$

Then, we introduce a Visual Attention Module (VAM), a Visual Contrastive Attention Module (VCAM) and a Cross-modal Attention Module (CAM) to respectively extract: 1) the visual attention feature V^a of X_I, 2) the visual contrastive attention feature V_c^a between X_I and the spatial visual feature of the retrieved similar case X_d, 3) the cross-modal attention feature Cr^a which matches the useful semantic information from the report of retrieved similar case R_d using the contrastive feature compared with X_I and X_d:

$$V^a = VAM(X_I), \quad (2)$$
$$V_c^a = VCAM(X_I, X_d), \quad (3)$$
$$Cr^a = CAM(X_I, X_d, R_d). \quad (4)$$

In Decoder: we propose a Parallel Attention Module (PAM) that allows the decoder model parallel grasping encoder features:

$$\tilde{Y} \leftarrow PAM(V^a, V_c^a, Cr^a). \quad (5)$$

where $\tilde{Y} = \{\tilde{y_1}, \tilde{y_2}, ..., \tilde{y_T}\}$ are word tokens of the generated report.

2.1.2 Basic Modules

Our proposed method is accomplished on stacks of identical Multi-head Attention (MHA) layers and Position-wised Feed-Forward Network (FFN) layers. The given input feature X are firstly converted into queries Q, keys K and values V:

$$Q = XW^Q, K = XW^K, V = XW^V, \quad (6)$$

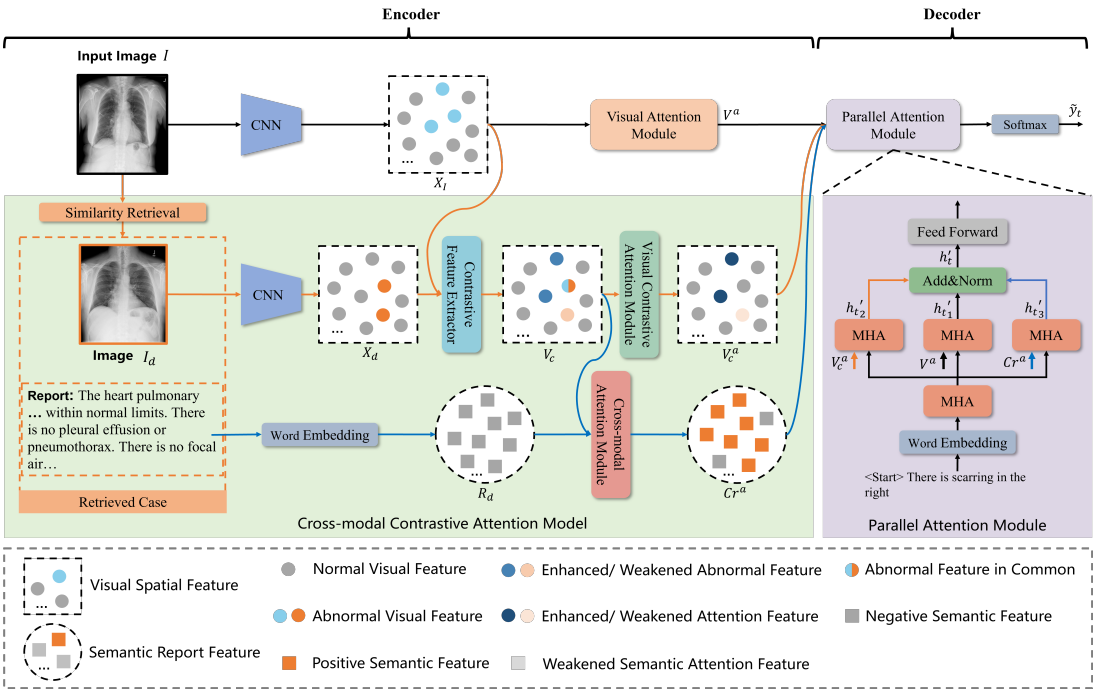


Figure 2: Overview of our proposed Cross-modal Contrastive Attention (CMCA) model for Medical Report Generation. In encoder, CMCA first retrieves the most similar image I_d, and generates a contrastive feature V_c through the contrastive feature extractor. Then, the Visual Contrastive Attention Module distills the unique abnormal features of I (blue circles), and the Cross-modal Attention Module matches the positive semantic information from the case report R_d. In decoder, the encoded features are integrated through a stack of Parallel Attention Modules to more effectively generate the final word $\tilde{y}_t$.

where $W^Q, W^K, W^V \in \mathbb{R}^{d\times d}$ are learned weights. In each MHA layer, the inputs are divided into h parallel attention heads, which allows the model to focus on the different representation sub-spaces of different positions jointly. And then the attention features of all heads can be calculated and concatenated as follows:

$$MHA(Q,K,V) = [head_1, ..., head_h]W^O, \quad (7)$$

$$head_i = Attn(QW_i^Q, KW_i^K, VW_i^V), \quad (8)$$

$$Attn(q,k,v) = softmax(\frac{qk^T}{\sqrt{d_k}})v. \quad (9)$$

where $W^O \in \mathbb{R}^{d\times d}$ and $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d\times \frac{d}{h}}$ are learned parameters, and $[\cdot]$ indicates the concatenation operation.

Then, the FFN layer is applied as follows:

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2. \quad (10)$$

where W_1, W_2 are learned parameters and b_1, b_2 are biases. There is also a residual connection around each of the MHA and FFN layers, followed by layer normalization.

2.1.3 Retrieval Case Database

For further retrieving the most similar historical cases of the input images, we create a Retrieval Case Database $\hat{DB}$. All of the records in $\hat{DB}$ are derived from the training sets, and each of them is a triplet consisting the global visual feature v_d of the image to be used for retrieval, the spatial visual feature X_d of the image to be used for calculating the contrastive feature compared to the input image, and the associated report R_d to be used for cross-modal alignment, which can be denoted by $< v_d, X_d, R_d >$. It can be noted that the spatial visual feature X_d is also extracted by DenseNet in Eq. 1, and the global visual feature v_d can be obtained by a average-pooling operation on X_d.

2.2 Cross-modal Contrastive Attention Model

To effectively extract critical abnormal visual features and positive semantic features from the input image and the most similar case, our CMCA model is proposed with mainly two modules: VCAM and CAM, both of which are applied based on the proposed simple-yet-effective contrastive feature, as shown in Figure 2.

Contrastive feature: To obtain the contrastive feature, we firstly retrieve the most similar historical case of the input image based on the historical Case Database $\hat{DB}$ established in Sec. 2.1.3 which includes the global and spatial visual features of case images and their corresponding reports, we

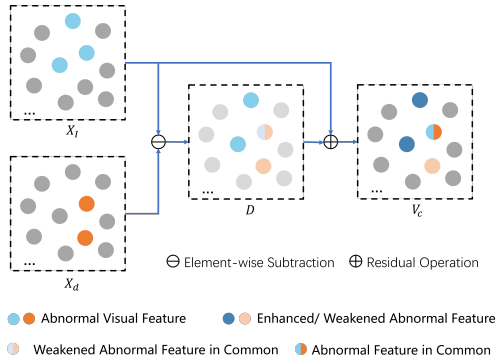


Figure 3: Structure of the proposed contrastive feature extractor. The circles in blue and orange correspond to the abnormal parts of the input image feature X_I and the retrieved image feature X_d, respectively.

retrieve the most similar case from $\hat{DB}$ by computing the highest cosine similarity between the global visual feature of input image v_I and historical image global features in case database $v_{\hat{DB}}$:

$$Case^{N_K} \leftarrow max(cosine(v_I, v_{\hat{DB}})). \quad (11)$$

where $Case^{N_K}$ denotes the retrieved N_K historical cases with the highest cosine similarity. It can be noted that we utilize each of the retrieved cases independently for the remaining task.

As shown in Figure 3, for one retrieved case $Case_d$, we calculate the contrastive feature V_c between the spatial visual features X_I and $X_d \in \mathbb{R}^{7\times7\times d}$ of the input image I and the case image I_d, the contrastive feature can be obtained through the following operations:

$$D = X_I - X_d, \quad (12)$$

$$V_c = D + X_I. \quad (13)$$

where D and $V_c \in \mathbb{R}^{7\times7\times d}$ denote the difference value of the spatial features and the final visual contrastive feature. The residual operation aims to tackle the problem of zero or negative value in D, and keep more representation of the input visual feature X_I for further cross-modal matching.

For further illustrating the ability of contrastive feature to distill the abnormal visual information, our explanations are as follows: As shown in Figure 3, we suppose the gray circles in X_I and X_d represent the normal portions in input and case images, respectively, while the colored circles represent the abnormal regions. Accordingly, there are four statuses in the contrastive feature V_c:

- The blue circles denote the regions of the input image I that are abnormal but normal in the case image I_d. As a result, the contrastive features of these regions can be expressed as $V_c^{blue} = X_I^{blue} + D^{blue}$, thus the distinctive abnormal regions of image I are reinforced.

- The orange circles denote the regions where I are abnormal but I_d are normal. The contrastive features decrease the distinctive abnormal regions of case I_d in orange, since $V_c^{orange} = X_I^{orange} + D^{orange}$ and the value of D^{orange} is negative.

- Identical regions are indicated by gray circles in two input spatial features. As a result, the differences of these parts are zero: $D^{gray} = 0$. The contrastive features of these regions are equivalent to the original spatial visual feature of I, which is indicated by $V_c^{gray} = X_I^{gray}$.

- The abnormal regions which occur in both I and I_d are the mixed color circles. As a result, the difference of these parts are also zero: $D^{mix} = 0$. And the contrastive features of these portions are also the original visual features of I, $V_c^{mix} = X_I^{blue}$.

In summary, the contrastive feature $V_c = \{V_c^{blue}, V_c^{orange}, V_c^{gray}, V_c^{mix}\}$ enhances the unique abnormal regions of the input image I, retains the identical regions, and weakens the unique abnormal regions of the case image I_d. Base on V_c, the following two modules respectively explain how to distill the unique abnormal visual representation of I and how to match the positive semantic features from the case report R_d.

2.2.1 Visual Contrastive Attention Module

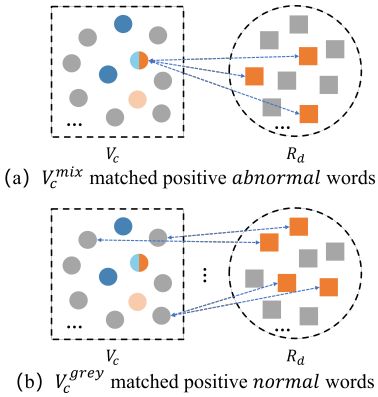
The proposed VCAM aims to extract the unique abnormal regions of the input image I based on the calculated contrastive feature V_c.

Given the contrastive feature V_c, VCAM uses it to reconstruct an attention feature as follows:

$$V_c^a = FFN(MHA(V_c, V_c, V_c)). \quad (14)$$

In V_c, the unique abnormal regions (V_c^{blue}) of I are enhanced. Thus through the attention module, the generated visual contrastive attention feature V_c^a is able to more effectively represent the unique abnormal regions of the input image.

VCAM attends to represent the unique abnormal regions in the input images. However, there are still some important abnormal participants, such as the mixed region V_c^{mix} in Figure 3, ignored by VCAM. In addition, it is equally meaningful to match both normal and abnormal positive semantic words in



(a) V_c^{mix} matched positive *abnormal* words

(b) V_c^{grey} matched positive *normal* words

Figure 4: The cross-modal matching for the visual contrastive feature V_c and semantic feature of the retrieved report R_d. The mixed circles in V_c is the same abnormal regions of image I and I_d. The orange squares are positive semantic features.

the retrieved case report. Thus, the following module is proposed to solve these problems.

2.2.2 Cross-modal Attention Module

We propose CAM to align the contrastive feature V_c with the retrieved report R_d for exploring the positive semantic information from the retrieved case report R_d.

As shown in Figure 4, for better illustrating the ability of CAM, we hypothesize splitting case report features into positive and negative parts denoted by orange blocks and gray blocks in $R_d = \{R_d^{orange}, R_d^{gray}\}$. Given the contrastive features V_c and the case reports R_d, CAM feeds them into the following layers:

$$Cr^a = FFN(MHA(V_c, R_d, R_d)). \quad (15)$$

The negative words R_d^{gray} are unmatched because they are corresponding to the unique abnormal regions ($V_c{}^{blue}$ and $V_c{}^{orange}$) which has been enhanced or weakened in V_c. The positive words R_d^{orange} contain two statuses as follows:

- As shown in Figure 4 (a), the first status is that the abnormal positive words are corresponding to the same abnormal parts ($V_c{}^{mix}$) of I and I_d, which are retained in V_c and can be matched with CAM.
- As shown in Figure 4 (b), the second status is that the normal positive words are corresponding to the normal visual features ($V_c{}^{grey}$) which also retained in V_c and can be matched with CAM.

In short, CAM extracts both normal and abnormal positive semantic words from the retrieved semantic report R_d by building their cross-modal interactions with the visual contrastive feature V_c.

VCAM and CAM complement each other in representing visual and semantic information as well as settling the problems caused by data biases.

2.3 Parallel Attention Module

As shown in Figure 2, for each time step t, the decoder layer first takes the embedded previous words $y_{1:t-1}$ as the input of the MHA layers:

$$h_t = MHA(y_{1:t-1}, y_{1:t-1}, y_{1:t-1}), \quad (16)$$

Then, the obtained visual attention feature V^a, visual contrastive attention feature V_c^a and cross-modal attention feature Cr^a are fed into three MHA layers separately, which calculates the attention features in parallel:

$$h_t^{'1} = MHA(h_t, V^a, V^a), \quad (17)$$
$$h_t^{'2} = MHA(h_t, V_c^a, V_c^a,), \quad (18)$$
$$h_t^{'3} = MHA(h_t, Cr^a, Cr^a), \quad (19)$$

The parallel operation would further enhance the decoding features for the three encoded attention feature. Thereafter, the three attention features are gathered as follows:

$$h_t^{'} = h_t^{'1}W_t^1 + h_t^{'2}W_t^2 + h_t^{'3}W_t^3, \quad (20)$$

where W_t^1, W_t^2 and W_t^3 are learned parameters.

Finally, the $h_t^{'}$ goes through a FFN layer and a linear layer followed by softmax activation to predict the current word:

$$\tilde{y_t} \leftarrow p_t = softmax(FFN(h_t^{'})W_y + b_y). \quad (21)$$

where $\tilde{y_t}$ is the predicted word at current timestep, and W_y, b_y are learnable weight and bias.

3 Experiments

3.1 Datasets

We conduct experiments on two datasets to evaluate the effectiveness of our proposed model.

3.1.1 IU X-Ray

Indiana University Chest X-Ray Collection (IU X-Ray) (Demner-Fushman et al., 2016) is a widely used public radiography dataset which totally contains 7,470 Chest X-ray images and 3,955 reports. Following (Chen et al., 2020), we randomly split the dataset into train/validation/test sets by 7:1:2.

3.1.2 MIMIC-CXR

The new released MIMIC-CXR (Johnson et al., 2019) dataset is the largest dataset so far. It con-

Table 1: Comparison of the proposed Cross-modal Contrastive model with other state-of-the-art methods on the IU X-Ray dataset and MIMIC-CXR dataset. BLEU-n denotes the BLEU scores using n-grams.

Datasets	Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
IU X-Ray	CNN-RNN (Vinyals et al., 2015)	0.216	0.124	0.087	0.066	-	0.306
	AdaAtt (Lu et al., 2017)	0.220	0.127	0.089	0.068	-	0.308
	Att2in (Rennie et al., 2017)	0.224	0.129	0.089	0.068		0.308
	HRNN (Krause et al., 2017)	0.439	0.281	0.190	0.133	-	0.342
	CoAtt (Jing et al., 2018)	0.455	0.288	0.205	0.154	-	0.369
	HRGR-Agent (Li et al., 2018a)	0.438	0.298	0.208	0.151	-	0.322
	CMAS-RL (Jing et al., 2019)	0.464	0.301	0.210	0.154	-	0.362
	KERP (Li et al., 2019)	0.482	0.325	0.226	0.162	-	0.339
	R2Gen (Chen et al., 2020)	0.470	0.304	0.219	0.165	0.187	0.371
	CMN (Chen et al., 2021)	0.475	0.309	0.222	0.170	0.191	0.375
	CA (Liu et al., 2021b)	0.492	0.314	0.222	0.169	0.193	0.381
	Transformer (Vaswani et al., 2017)	0.396	0.254	0.179	0.135	0.164	0.342
	CMCA	**0.497**	**0.349**	**0.268**	**0.215**	**0.209**	**0.392**
MIMIC-CXR	CNN-RNN (Vinyals et al., 2015)	0.299	0.184	0.121	0.084	0.124	0.263
	AdaAtt (Lu et al., 2017)	0.299	0.185	0.124	0.088	0.118	0.266
	Att2in (Rennie et al., 2017)	0.325	0.203	0.136	0.096	0.134	0.276
	Top-Down (Anderson et al., 2018)	0.317	0.195	0.130	0.092	0.128	0.267
	R2Gen (Chen et al., 2020)	0.353	0.218	0.145	0.103	0.142	0.277
	CMN (Chen et al., 2021)	0.353	0.218	0.148	0.106	0.142	0.278
	CA (Liu et al., 2021b)	0.350	0.219	0.152	0.109	**0.151**	0.283
	Transformer (Vaswani et al., 2017)	0.314	0.192	0 .127	0.090	0.125	0.265
	CMCA	0.**360**	**0.227**	**0.156**	**0.117**	0.148	**0.287**

tains 473,057 Chest X-ray images and 206,563 reports. For fair comparison, we adopt the official split with 368,960 images and 222,758 reports for training, 2,991 images and 1,808 reports for validation, 5,159 images and 3,269 reports for testing.

For both datasets, we adopt and tokenize the *findings* section which has long sentences as the target reports and convert words into lower-cases.

3.2 Experimental Settings

3.2.1 Evaluation Metrics

We evaluate our proposed approach on the widely used metrics: BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004) and METEOR (Banerjee and Lavie, 2005). The metric scores are calculated by the standard image caption evaluation tool [1].

For clinical efficacy estimate, we employ the CheXpert labeling tool [2] proposed in (Irvin et al., 2019) to label our generated reports and the ground-truth reports in 14 different categories related to thoracic diseases and support devices. *Precision*, *Recall* and *F1* are taken as the evaluation metrics.

3.2.2 Implementation Details

We use the DenseNet-121 (Huang et al., 2017) pre-tained on ImageNet (Deng et al., 2009) and fine-tuned on CheXpert (Irvin et al., 2019) dataset to extract visual features of images in this paper. The dimension of each extracted visual feature map is set to 1024, and we then converted it to 512. In addition, following the previous works (Chen et al., 2020), we use the frontal and lateral view images as input and concatenate the features of two view images together for IU X-Ray dataset, and use single image as input for MIMIC-CXR dataset. For the proposed method, the dimension of our multi-head attention model is set to 512, and the number of heads is set to 8. And we set the number of layers to 3 for all modules. Moreover, the model is trained for 100 epochs under ADAM optimizer (Kingma and Ba, 2014). We set the initial learning rate to 1e-4 decaying by 0.99 per epoch. The beam size is set to 3 and we select top 5 similar cases for each input image.

Table 2: The comparison of the clinical efficacy metrics on MIMIC-CXR dataset, which measures the Precision, Recall and F1-score of the clinical abnormalities for the generated reports.

Methods	CE Metrics		
	Precision	Recall	F1-score
CNN-RNN (Vinyals et al., 2015)	0.249	0.203	0.204
AdaAtt (Lu et al., 2017)	0.268	0.186	0.181
Att2in (Rennie et al., 2017)	0.322	0.239	0.249
Top-Down (Anderson et al., 2018)	0.320	0.231	0.238
R2Gen (Chen et al., 2020)	0.333	0.273	0.276
CMN (Chen et al., 2021)	0.334	0.275	0.278
CA (Liu et al., 2021b)	0.352	**0.298**	0.303
CMCA	**0.444**	0.297	**0.356**

[1] https://github.com/tylin/coco-caption

[2] https://github.com/stanfordmlgroup/chexpert-labeler

Table 3: The comparison on the IU X-Ray dataset of the Baseline model (Vaswani et al., 2017) with the different components of our proposed method: VCAM, CAM, and PAM.

Methods	VCAM	CAM	PAM	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
Baseline				0.396	0.254	0.179	0.135	0.164	0.342
(a)	✓			0.481	0.328	0.242	0.187	0.202	0.380
(b)		✓		0.470	0.305	0.215	0.160	0.186	0.384
(c)	✓	✓		0.484	0.335	0.248	0.194	0.203	0.381
CMCA	✓	✓	✓	**0.497**	**0.349**	**0.268**	**0.215**	**0.209**	**0.392**

3.3 Results

We take Transformer (Vaswani et al., 2017) with 3 layers for both encoder and decoder modules as the Baseline model. In addition, we compare our approach with the state-of-the-art medical report generation models, i.e., CoAtt (Jing et al., 2018), HRGR-Agent (Li et al., 2018a), CMAS-RL (Jing et al., 2019), KERP (Li et al., 2019), R2Gen (Chen et al., 2020), CMN (Chen et al., 2021) and CA (Liu et al., 2021b). And we also adopt image captioning methods, i.e., CNN-RNN (Vinyals et al., 2015), AdaAtt (Lu et al., 2017), Att2in (Rennie et al., 2017), Top-Down (Anderson et al., 2018), and the model designed for long sentence generation task: HRNN (Krause et al., 2017). We directly quote the results from the original papers for the comparison methods.

As shown in Table 1, our CMCA outperforms on almost all metrics on both datasets, which validates our hypothesis that historical similar cases can significantly assist medical report generation task, and CMCA is able to exploit useful visual and semantic information from similar cases for generating more accurate reports. In addition, the clinical efficacy metrics in Table 2 show that CMCA outperforms the state-of-the-art methods on almost all metrics especially on *precision* and *F1-score*, which indicates that more abnormal findings identified by CMCA are exact and our model greatly boosts the comprehensive performance on clinical efficacy.

4 Analysis

4.1 Quantitative Analysis

4.1.1 Effect of the Visual Contrastive Attention Module

VCAM is used to calculate the visual contrastive attention feature, which makes the model focus on the unique abnormal regions of the input image in the contrastive feature. Comparing with Baseline and (a) in Table 3, we can find that VCAM boosts the performance of Baseline model on all evaluation metrics. More encouragingly, comparing with the state-of-the-art methods in Table 1, VCAM achieves comparable performance on most of the metrics. We hypothesize that these performance gains may due to the contrastive feature which enhances the unique abnormal regions of the input image, and the following VCAM makes the model focus on the abnormal regions.

4.1.2 Effect of the Cross-modal Attention Module

To make full use of the semantic information of the retrieved case, we propose to align the contrastive feature with the retrieved report to match the positive words by CAM. Comparing the Baseline and (b) in Table 3, we can find that CAM also greatly improve the performance of baseline model on all evaluation metrics. In addition, it can be noted from (a) and (b) that CAM performs better on the ROUGE-L metric, which indicates that the longest common sub-sequence between the generated report and ground truth is longer. We assume this performance gain might because CAM catches more common semantic information between the input image and retrieved case.

4.1.3 Effect of the Parallel Attention Module

Based on the standard Transformer decoder layers, PAM decodes the attention features obtained from the Encoder in parallel, and then the parallel features are fused together to generate the final word. Comparing (c) with the full CMCA model in Table 3, we find that the decoder layers with parallel attention perform better than the normal ones. It verifies that the parallel attention can further extract discrimitive information from decoding the attention features separately.

In summary, both VCAM and CAM modules can boost the performance respectively. And the incorporation of the two modules can generate better results than each. Moreover, the combination of VCAM, CAM and PAM achieves the state-of-the-art performances on both neural language and clinical efficacy metrics.

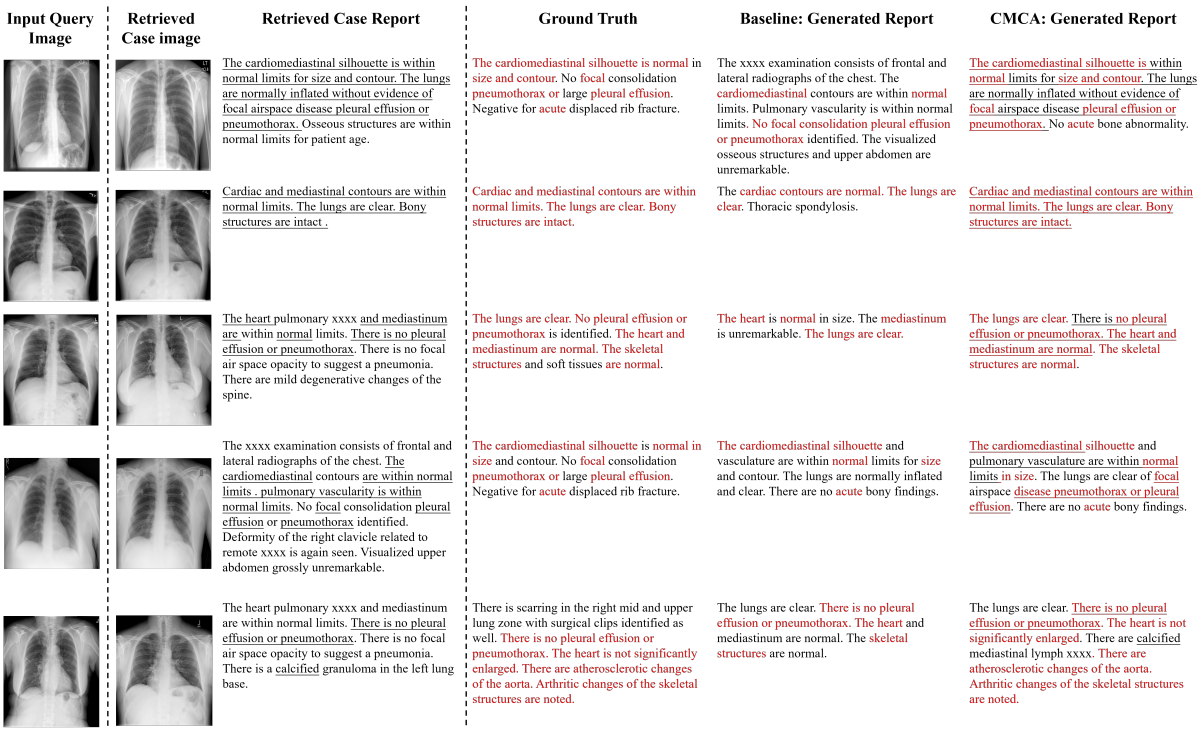


Figure 5: Visualization results of our proposed CMCA model on IU X-Ray dataset. The red fonts mark that our generated words match the ground truth, and the underlines mean that our generated words match the case reports.


Ground Truth: The lungs and pleural spaces show no acute abnormality. Heart size and pulmonary vascularity within normal limits.

Generated Report: The cardiomediastinal silhouette and vasculature are within normal limits for size and contour. The lungs are normally inflated and clear. Osseous structures are within normal limits for patient age.

Ground Truth: The cardiomediastinal silhouette is normal in size and contour. No focal consolidation pneumothorax or large pleural effusion. Negative for acute displaced rib fracture.

Generated Report: The cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size. The lungs are clear of focal airspace disease pneumothorax or pleural effusion. There are no acute bony findings.

Ground Truth: The lungs are clear without evidence of focal airspace disease. There is no evidence of pneumothorax or large pleural effusion. The cardiac and mediastinal contours are within normal limits. The xxxx are unremarkable.

Generated Report: The cardiomediastinal silhouette is within normal limits for appearance. No focal areas of pulmonary consolidation. No pneumothorax . no pleural effusion. The thoracic spine appears intact. No acute displaced rib fractures.

Figure 6: Visualization of CMCA generated reports and the ground truth. The underlined texts indicate the words which correctly describe medical image information but absent in the ground truth reports.

4.2 Qualitative Analysis

In Figure 5, we visualize five examples in row to illustrate the effectiveness of our proposed model. For each row, the first column denotes the input image. The second and third columns show the retrieved case, which contains a visual similar image and the corresponding report. The fourth column is the ground truth report of the input image. The fifth column is the report generated by Baseline model and the last column is the generated report of our proposed CMCA. The red word indicates that the generated report matches the ground truth, and the underlined word means that the generated result matches the retrieved case report.

According to the underlined texts of the first two normal examples in Figure 5, our model adopts positive words from case report, and generates more accurate reports. The third and fourth examples are normal cases, but the retrieved cases are abnormal. It can be seen that CMCA employs the normal components and eliminates the abnormal ones in case reports, then generates other sentences according to the input images like *"the lungs are clear"* and *"the skeletal structures are normal"*. The fifth example and its retrieved case are both abnormal. Our result contains useful sentences from the case report, such as *"there is no pleural effusion or pneumothorax"*, and *"calcified"*. The visualization of the examples again verifies the effectiveness of our proposed CMCA model, which can select the useful semantic words from the retrieved reports.

In addition, as shown in Figure 6, the underlined texts show that our generated reports contain richer information than the ground truth. For example, in the first image, our model generates *"the cardiomediastinal silhouette and vasculature are within normal limits for size and contour"*, which is absent in the ground truth. In practice, radiologist might only write the most significant findings according to the images, while other pathological information might be neglected or incompletely recorded. Our proposed CMCA model can mitigate this problem and generate much richer reports.

5 Conclusion

In this paper, we propose a novel Cross-modal Contrastive Attention (CMCA) model to exploit the contrastive information from historical similar cases to tackle the data biases for medical report generation. CMCA contains two modules: the Visual Contrastive Attention Module that distills abnormal information of the input images, and the Cross-modal Attention Module that builds interactions of the unmatched cross-modalities. Extensive experimental results show that CMCA outperforms the state-of-the-art methods on almost all metrics. Further analyses verify the ability of CMCA in generating reports with more accurate abnormal findings and richer descriptions.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 61906007 and 62276010, in part by the R&D Program of Beijing Municipal Education Commission under Grant KM202110005022 and KZ202210005009.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. 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*, pages 6077–6086.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. 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*, pages 5904–5914.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1439–1449.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Qingji Guan, Yaping Huang, Yawei Luo, Ping Liu, Mingliang Xu, and Yi Yang. 2021. Discriminative feature learning for thorax disease classification in chest x-ray images. *IEEE Transactions on Image Processing*, 30:2476–2487.

Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, and Yi Yang. 2020. Thorax disease classification with attention guided convolutional neural network. *Pattern Recognition Letters*, 131:38–45.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*.

Baoyu Jing, Zeya Wang, and Eric Xing. 2019. 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*, pages 6570–6580.

Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2577–2586.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

D. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 317–325.

Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2018a. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Advances in neural information processing systems*, pages 1537–1547.

Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2019. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6666–6673.

Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. 2018b. Thoracic disease identification and localization with limited supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8290–8299.

Gongbo Liang, Connor Greenwell, Yu Zhang, Xin Xing, Xiaoqin Wang, Ramakanth Kavuluru, and Nathan Jacobs. 2021. Contrastive cross-modal pre-training: A general strategy for small sample medical imaging. *IEEE Journal of Biomedical and Health Informatics*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021a. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13753–13762.

Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. 2021b. Contrastive attention for automatic chest x-ray report generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 269–280.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. 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*, pages 375–383.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

José Ramos, Thessa TJP Kockelkorn, Isabel Ramos, Rui Ramos, Jan Grutters, Max A Viergever, Bram van Ginneken, and Aurélio Campilho. 2014. Content-based image retrieval by metric learning from radiology reports: application to interstitial lung diseases. *IEEE Journal of Biomedical and Health Informatics*, 20(1):281–292.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024.

Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. 2016a. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2497–2506.

Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. 2016b. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2497–2506.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3156 – 3164.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9049–9058.

Xing Xu, Tan Wang, Yang Yang, Lin Zuo, Fumin Shen, and Heng Tao Shen. 2020. Cross-modal attention with semantic consistence for image-text matching. *IEEE Transactions on Neural Networks and Learning Systems*, 31(12):5412–5425.

Yuan Xue, Tao Xu, L Rodney Long, Zhiyun Xue, Sameer Antani, George R Thoma, and Xiaolei Huang. 2018. Multimodal recurrent model with attention for automated radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 457–466. Springer.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4659.

Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. 2017. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6428–6436.