

# LEVERAGING HARD NEGATIVE PRIORS FOR AUTOMATIC MEDICAL REPORT GENERATION

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Recently, automatic medical report generation has become an active research topic in medical imaging field. It is imperative for the model to identify normal and abnormal regions in a medical image to generate a coherent and diverse report. However, medical datasets are highly biased towards normal regions. This makes most existing models tend to generate a generic report without sufficiently considering the uniqueness of individual images. In this paper, we propose a learning framework to extract distinctive image and report features for each sample by distinguishing it from its closest peer (denoted as hard negative in this paper) and gradually increasing the difficulty of such a task through synthesizing harder and harder negatives during training. Specifically, a prior hard negative report, which is the report closest to an anchor report in the dataset, is initially identified by using a pre-trained Sentence Transformer. To force our report decoder to capture highly distinctive and image-correlated text features, harder and harder negative reports keep being synthesized by gradually moving the prior hard negative report towards the anchor report in the latent space during training. The harder negative report is used to evaluate a triplet loss that is minimized to enforce the distance between the matched image and report to be smaller than the distance between an image and its synthesized harder negative report. Meanwhile, the associated images of the anchor report and its prior hard negative report form a hard negative image pair, and a cosine similarity loss is used to capture the distinctive features of the anchor image by pushing the hard negative image away. In this way, our model could achieve subtle representative resolution (i.e., the ability to distinguish two similar samples). As a general method, we demonstrate experimentally that our framework could be readily incorporated into a variety of existing medical report generation models, and significantly improve the corresponding baselines. Our code will be publicly released at

## 1 INTRODUCTION

Medical report is a multi-sentence paragraph that precisely describes the normal and abnormal regions in a medical image. Writing such reports requires proper experience and expertise (Jing et al., 2018). Using AI to automate this process can reduce manual workload and speed up clinic procedure. Automatic medical report generation is similar to image captioning but with more subtle correlation between medical images and corresponding reports. This could make image captioning models (Vinyals et al., 2015; Xu et al., 2015; Anderson et al., 2018; Rennie et al., 2017; Lu et al., 2017) fail when directly applied on medical datasets, making medical report generation more challenging (Wang et al., 2021b). Many works have been proposed for medical report generation via an encoder-decoder framework (Jing et al., 2018; 2019; Li et al., 2018; Chen et al., 2020; 2021; Wang et al., 2021b). They focus on improving the generated reports with the aid of medical tags (Jing et al., 2018), large pretrain models (Huang et al., 2017; He et al., 2016; Radford et al., 2019), and the use of relational memory (Chen et al., 2020; 2021). A detailed literature review is in Sec. 2.

Despite the effectiveness of these approaches, their generated reports could still be heavily dominated by the terms describing the common contents of medical images (Jing et al., 2019). As a result, the characteristics of an individual report could be submerged. It is therefore imperative to help the model to generate a report that covers the distinctive features of individual images. Recently, a few methods (Liu et al., 2021b;a) were proposed to learn sample-specific image or report features by

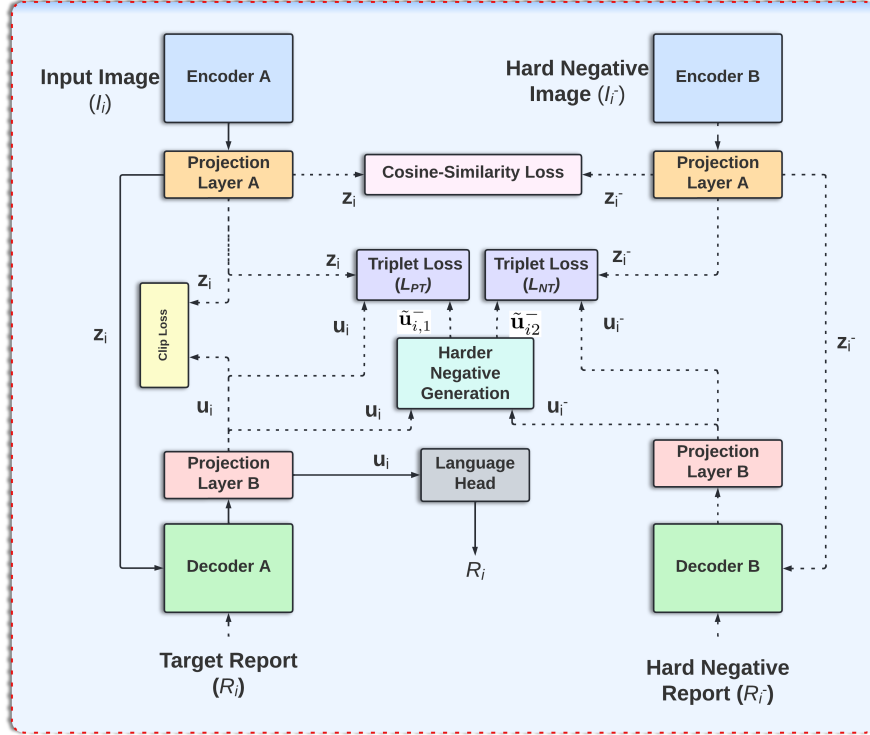


Figure 1: Illustration of our proposed framework, explained in Sec. 3.1. At the inference stage, only images are used as the input for the model to output the generated reports.

linking or contrasting the given sample to the closer ones in the training set to guide the report generation. For example, Posterior-and-Prior Knowledge Exploring and Distilling (PPKED) (Liu et al., 2021a) utilized the detected abnormal image regions to retrieve a report from the training set, and used this report as a reference to help generate the report of the given image. Another work (Liu et al., 2021b) proposed a contrastive attention mechanism (denoted as CA in this paper) to capture abnormal regions by comparing an input image with a pool of normal images while generating the report. An aggregation attention and a differentiation attention were used to remove the significant common information between an abnormal image and a set of normal images. Nevertheless, the two methods still focus on distinguishing abnormal samples from the common normal ones, which may not be able to sufficiently differentiate the subtle but critical changes across individual reports. Moreover, the two methods use only one of the two modalities to capture abnormality-related features, ignoring the crucial information from the other modality. For example, CA (Liu et al., 2021b) utilizes only image modality while PPKED (Liu et al., 2021a) relies on merely text modality.

To address this situation, we argue that it is imperative to extract distinctive features and consider both visual and text modalities so as to generate reports reflecting individual characteristics. We propose a method that captures the unique features of a given sample for both image and text modalities by differentiating it from its closest neighbour in the training set (called hard negative in this paper) and, more importantly, further gradually increase the hardness of discrimination via synthesizing **harder and harder** negatives to enforce even finer differences to be captured. The hard negatives are found from the perspective of reports since they reflect high-level clinic-related semantics.

Specifically, for each report in the training set, we extract the report features using Sentence Transformer (SBert) (Reimers & Gurevych, 2019) and calculate its cosine-similarity scores with respect to other reports. The top-1 similar report is then picked as its prior hard negative. By observing if two reports are similar, their paired images are likely to be also similar, we simultaneously obtain the hard negatives for the associated images. Then we send the hard negative image pair to the two shared encoders to extract image representations and the hard negative report pair to the two shared decoders to extract report features. For the image modality, a cosine-similarity loss is minimized

to ensure that the learned image representations of the hard negative image pair are dissimilar from each other. For the text modality, in its feature space, we keep synthesizing harder and harder negative features through linearly combining the given report and its prior hard negative and making the synthesized ones gradually closer to the given report to increase their hardness to be differentiated. Moreover, we employ a triplet loss to explicitly align the image and the report features to be close to each other and at the same time push the synthesized harder negative features away from the image features in the feature embedding space. The importance of the synthesis is to have a much harder negative so that model can learn subtle relation between image and report features with the help of the triplet loss. For inference, only the text-aligned image features are used for report generation.

The main contributions of our paper are summarized as follows.

First, we propose a simple yet effective approach that helps the model to capture unique features in both the image and the text modalities to generate coherent and diverse reports. It can be easily incorporated with the existing encoder-decoder based models without modifying their core architecture or mechanism, as demonstrated in our experimental study.

Second, we propose a new mechanism to leverage hard negative priors for report generation, which gradually increases the hardness of the generated negative samples with the evolution of the training process. By training our model to differentiate these harder negatives, the recognition resolution of our model could be increased.

Third, we validate the effectiveness of our proposed framework upon five different backbones. On two radiology benchmarks IU-XRay (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019), Our model outperforms the state-of-the-art medical report generation methods and integrating our framework could consistently improve the corresponding backbones. In addition, we also show that incorporating our framework could also benefit the generic image captioning on COCO dataset (Lin et al., 2014).

## 2 RELATED WORK

**Image Captioning** The aim of an image captioning task is to automatically generate a single or multi-sentence description of an image using deep learning models. A basic image-captioning model follows an encoder-decoder framework where encoder is a convolutional neural network (CNN) and decoder is a recurrent neural network (RNN). The CNN extracts images features and passes it to the RNN for caption generation. Later, attention mechanisms were introduced to supervise the model to focus on the significant regions of an image while generating a caption. To perform attention on object level and regions level, Anderson et al. (2018) combined top-down and bottom-up attention mechanisms. To generate a multi-sentence description of an image, Krause et al. (2017) proposed a heirarchial RNN (HRNN). Recently, much stronger Transformer replaced RNN to improve the quality and diversity of the captions.

**Medical Report Generation** Earlier works adopted traditional encoder-decoder framework for report generation. To generate coherent reports, (Jing et al., 2018) proposed co-attention mechanism to localize abnormal regions and a heirarchial LSTM decoder to generate the reports. They also adopted disease tag classification task along with report generation, making it a multi-task learning framework. Later, (Jing et al., 2019) proposed a multi-agent model to alleviate data bias problem between normal and abnormal regions of the medical images. (Chen et al., 2020) incorporated a relational memory module into the vanilla Transformer to store the significant information related to earlier generated reports and utilized this information for effective report generation. Following this work, (Chen et al., 2021) used a relational memory matrix for cross-modal alignment of image and text features. After the success of GPT-2 (Radford et al., 2019) in text generation tasks, (Alfarghaly et al., 2021) used GPT-2 for report generation. First, they fine-tuned ChexNet (Rajpurkar et al., 2017) for disease tag prediction, then used the predicted tag’s embeddings to calculate weighted semantic features. Finally, they conditioned the GPT-2 model on semantic and visual features to generate a report.

To improve the diversity of the generated reports in describing abnormal diseases, (Liu et al., 2021b) proposed contrastive attention mechanism focusing on abnormal regions of the image. It consists of two attentions - aggregate attention to summarize the information from all the reports in normality

pool and differentiate attention to remove common information between normal and abnormal image. Each abnormal image is compared with a set of normal images to get contrastive information. This information is then used for report generation. They try to get the contrastive information from image-level, ignoring the crucial information from text-level. Other work, PPKED (Liu et al., 2021a) used prior knowledge, posterior knowledge and multi-domain knowledge distiller to generate the report. This method tries to mimic the behaviour of a radiologist, by first predicting the disease tags focusing on the abnormal regions of the image (posterior knowledge), utilizing the prior knowledge by retrieving the relevant reports from the corpus, and finally using a distiller module to incorporate the prior and posterior knowledge while doing report generation. Furthermore, (Wang et al., 2021b) used an additional image-text matching (ITM) branch to align image and report features in the latent space, and utilized the generated reports during training as the hard negative reports that are close to the ground-truth to supervise ITM branch.

The main drawback of these methods is that they try to utilize the distinctive information from just one modality. We propose an approach to utilize distinctive features from both modalities and closely align them in the latent space by evolving hard negatives. Unlike (Wang et al., 2021b), we achieve this without the need of an additional ITM branch, which significantly simplifies the model. More detailed discussion is available in Section 3.

### 3 PROPOSED METHOD

Our learning framework is model-agnostic and suits general encoder-decoder based methods. In this paper, we demonstrate its effectiveness on multiple models, including a basic baseline model using DenseNet-121(Huang et al., 2017) as the encoder and GPT-2(Radford et al., 2019) as the decoder and several recent models consisting of R2Gen (Chen et al. (2020)), R2GenCMN (Chen et al. (2021)), XproNet (Wang et al. (2022)), and VisualGPT (Chen et al. (2022)). The input to our model is a medical image and the output is the medical report describing the image. Each input image is resized to  $3 \times H \times W$  shape. We do not combine the frontal and lateral views of the image and rather pass them separately.

#### 3.1 OVERVIEW

Given an image set  $\mathcal{I} = \{I_1, I_2, \dots, I_n\}$  and the corresponding ground-truth report set  $\mathcal{R} = \{R_1, R_2, \dots, R_n\}$ , the encoder extracts the image representation  $\mathbf{z}_i = f(I_i) \in \mathbb{R}^d$  and passes it to the decoder to generate a predicted report  $\hat{R}_i$ . To extract distinctive and significant features of images and reports, treating each image-report pair  $(I_i, R_i)$  as an anchor, we compute its hard negative image-report pair, denoted by  $(I_i^-, R_i^-)$ , using SBert (Reimers & Gurevych, 2019) based on the cosine-similarity score (Sec. 3.2). The resulting two images  $(I_i, I_i^-)$  are fed as the input to two encoders whose parameters are shared while the resulting two reports  $(R_i, R_i^-)$  are fed to two decoders sharing parameters also. The features extracted by the two encoders are denoted as  $(\mathbf{z}_i, \mathbf{z}_i^-)$ . To enforce the encoder to extract distinctive features, we use a cosine-similarity loss to encourage the dissimilarity between the image features  $\mathbf{z}_i$  and  $\mathbf{z}_i^-$  (Sec. 3.3). Next, we pass these extracted image features to the decoders to extract the corresponding report features, denoted by  $(\mathbf{u}_i, \mathbf{u}_i^-)$ . To further enhance the discrimination of the report features  $\mathbf{u}$ , we synthesize a series of harder negative reports, denoted by  $\tilde{\mathbf{u}}_i$ , in the embedding feature space by gradually moving the hard negative report feature  $\mathbf{u}_i^-$  towards the report feature  $\mathbf{u}_i$  via using a convex linear combination of them (Sec. 3.4). We then employ a triplet loss and a clip loss to align the report features  $\mathbf{u}_i$  with the corresponding image features  $\mathbf{z}_i$ , while pushing the synthesized harder negative features  $\tilde{\mathbf{u}}_i$  away from the image features  $\mathbf{z}_i$  in the embedding feature space (Sec. 3.5). Our framework is illustrated in Fig. 1.

#### 3.2 HARD NEGATIVE PRIORS USING SBERT

Before starting the training, for each report in the training set its closest report (referred to as the hard negative in this paper) is identified based on the cosine similarity score. Specifically, the reports  $\mathcal{R} = \{R_1, R_2, \dots, R_n\}$  are passed to the SBert (Reimers & Gurevych, 2019) to extract semantically meaningful report features  $\{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n\} \in \mathbb{R}^{d \times n}$ . For each report feature  $\mathbf{r}_i$ , its cosine similarity score is computed against all the other report features  $\mathbf{r}_j$  in the training set, and the one with the

highest score is considered as the hard negative for that report. Formally,

$$\cos(\mathbf{r}_i, \mathbf{r}_j) = \frac{\mathbf{r}_i \cdot \mathbf{r}_j}{\|\mathbf{r}_i\| \cdot \|\mathbf{r}_j\|} \quad (1)$$

where  $i, j = 1, \dots, n, i \neq j$ . By doing so, we can create a hard negative report pair  $(R_i, R_i^-)$  for each report  $R_i$ . We assume if  $(R_i, R_i^-)$  forms a hard negative report pair, then their corresponding images  $(I_i, I_i^-)$  usually also form a hard negative image pair. In this way, we obtain a pair of hard negatives of images and their corresponding reports  $(I_i, I_i^-, R_i, R_i^-)$  (where  $i = 1, \dots, n$ ) for all the samples. It is noted that we create the hard negatives from the perspective of reports as they directly contain the high-level clinic-related semantics.

The computed hard negative reports are used as a prior or a starting point. The hardness of the negative reports is further increased gradually by moving the features of the hard negative reports towards the anchor report features during the training, as discussed in detail in Sec. 3.4.

### 3.3 SIGNIFICANT IMAGE FEATURES

Once the hard negative pairs are obtained, the images  $I_i$  and  $I_i^-$  are sent through Encoder A and Encoder B, respectively, to extract features  $\mathbf{z}_i \in \mathbb{R}^d$  and  $\mathbf{z}_i^- \in \mathbb{R}^d$ . As aforementioned, the parameters of the two encoders are shared. A cosine-similarity loss  $\mathcal{L}_{CS}$  is then minimized to make the representations of an image and its hard negative dissimilar to each other in order to learn powerful encoders capable of differentiating two closely similar images. The loss  $\mathcal{L}_{CS}$  is defined as:

$$\mathcal{L}_{CS} = \frac{\mathbf{z}_i \cdot \mathbf{z}_i^-}{\|\mathbf{z}_i\| \cdot \|\mathbf{z}_i^-\|}. \quad (2)$$

$\mathcal{L}_{CS}$  would be small if  $\mathbf{z}_i$  is very different from  $\mathbf{z}_i^-$  when Encoder A and Encoder B extract distinctive features for the input image  $I_i$  and its hard negative  $I_i^-$ , respectively. Moreover, a cross-entropy loss used to generate reports adds another level of supervision on the encoder to extract image representations useful for report generation. Hence, the cosine-similarity loss  $\mathcal{L}_{CS}$  and the cross-entropy loss  $\mathcal{L}_{CE}$  together make the image representations from the encoder not only discriminative but also significant for report generation. More details regarding the cross-entropy loss are in Sec. 3.5.

### 3.4 SYNTHETIC HARDER NEGATIVES

The image representations  $\mathbf{z}_i$  and  $\mathbf{z}_i^-$  output from the Encoders A and B along with their corresponding reports  $R_i$  and  $R_i^-$  are passed to two decoders which also share parameters. The report representations from the last hidden layer of the Decoders A and B are denoted as  $\mathbf{u}_i \in \mathbb{R}^d$  and  $\mathbf{u}_i^- \in \mathbb{R}^d$ , respectively, where  $\mathbf{u}_i^-$  is the hard negative of  $\mathbf{u}_i$  as previously defined. The decoders should be trained to be able to differentiate  $\mathbf{u}_i$  and  $\mathbf{u}_i^-$  to achieve discriminative report representations. Moreover, we argue that the efficacy of decoders could be further enhanced by increasing the hardness of  $\mathbf{u}_i^-$ , i.e., the difficulty to differentiate  $\mathbf{u}_i^-$  from  $\mathbf{u}_i$ . We therefore introduce the hard negative mixing strategy (Kalantidis et al., 2020) into this process.

Specifically, in the feature embedding space, given a report sample  $\mathbf{u}_i$ , our main motive is to start from its prior hard negative report  $\mathbf{u}_i^-$  within the training set (as precomputed using SBert in Sec. 3.2) and gradually increase the hardness of  $\mathbf{u}_i^-$  by moving it closer to the anchor  $\mathbf{u}_i$ . This is achieved by synthesizing a series of harder negatives via the convex linear combinations of  $\mathbf{u}_i$  and  $\mathbf{u}_i^-$ , i.e.,

$$\tilde{\mathbf{u}}_{i1}^- = \lambda \times \mathbf{u}_i^- + (1 - \lambda) \times \mathbf{u}_i, \quad (3)$$

where  $\tilde{\mathbf{u}}_{i1}^-$  denotes the synthetic harder negative of  $\mathbf{u}_i$ . The hyper-parameter  $\lambda$  is used to control the rate at which the hard negative moves towards the anchor report in the feature embedding space. The value of  $\lambda$  is defined as  $\lambda = e^{-\alpha}$  and it changes over the epochs so that the hardness of the synthetic harder negative gradually increases rather than staying constant. Also, we cap the value of  $\lambda$  from bottom say 0.5, so that  $\mathbf{u}_i$  does not dominate  $\mathbf{u}_i^-$  in Eq. 3. A decay parameter  $\alpha$  is used to control the rate of change in hardness.

Meanwhile, since  $\mathbf{u}_i$  could reciprocally be regarded as the hard negative of  $\mathbf{u}_i^-$ , we could also compute a synthetic harder negative for  $\mathbf{u}_i^-$  in a similar way and denote it as  $\tilde{\mathbf{u}}_{i2}^-$ ,

$$\tilde{\mathbf{u}}_{i2}^- = \lambda \times \mathbf{u}_i + (1 - \lambda) \times \mathbf{u}_i^-. \quad (4)$$

### 3.5 TRAINING PROCESS

To this end, we could form two triplets  $(\mathbf{z}_i, \mathbf{u}_i, \tilde{\mathbf{u}}_{i1}^-)$  and  $(\mathbf{z}_i^-, \mathbf{u}_i^-, \tilde{\mathbf{u}}_{i2}^-)$ . In order to narrate finer visual findings in a given medical image, highly correlated image and report features are desired. Two triplet losses in Eq. 5 and Eq. 6 are applied to align image-report features. Specifically, for the triplet  $(\mathbf{z}_i, \mathbf{u}_i, \tilde{\mathbf{u}}_{i1}^-)$ , we require that in the feature embedding space, the distance between the image  $\mathbf{z}_i$  and its corresponding report  $\mathbf{u}_i$  is smaller than the distance between the image  $\mathbf{z}_i$  and its harder negative report  $\tilde{\mathbf{u}}_{i1}^-$ . This applies to the triplet  $(\mathbf{z}_i^-, \mathbf{u}_i^-, \tilde{\mathbf{u}}_{i2}^-)$  in a similar way. In other words, the triplet loss in Eq. 5 pulls the true match  $\mathbf{u}_i$  closer to the image  $\mathbf{z}_i$  and also pushes the harder negative report  $\tilde{\mathbf{u}}_{i1}^-$  away from  $\mathbf{z}_i$ . The triplet loss in Eq. 6 functions similarly.

$$\mathcal{L}_{PT} = \max\{d(\mathbf{z}_i, \mathbf{u}_i) - d(\mathbf{z}_i, \tilde{\mathbf{u}}_{i1}^-) + \epsilon, 0\}, \quad (5)$$

where  $\epsilon$  is a small positive value, denoting the margin of distances, and

$$\mathcal{L}_{NT} = \max\{d(\mathbf{z}_i^-, \mathbf{u}_i^-) - d(\mathbf{z}_i^-, \tilde{\mathbf{u}}_{i2}^-) + \epsilon, 0\}. \quad (6)$$

Moreover, a triplet loss considers only one positive report (the true match) and one negative report for an anchor image. To well align the positive with the anchor and at the same time push more negatives away from the anchor, it is essential to consider additional negatives. This helps to prevent the second, third or  $k$ -th closest unmatched report to the anchor from being close to the anchor, and therefore create a strong boundary around the anchor against potential negatives in the feature embedding space. To achieve this, we further employ a contrastive image-text pre-training (CLIP) loss (Radford et al., 2021). Given a batch of  $B$  images and their truly matched reports, in the feature embedding space, we could form  $B$  matched image-report pairs as well as  $B^2 - B$  unmatched image-report pairs. The encoder and decoder are then jointly trained to maximize the cosine similarity of the matched pairs while minimizing that of the unmatched pairings. We then optimize a symmetric cross entropy loss over these similarity scores. The clip loss is denoted as  $\mathcal{L}_{CP}$ .

For report generation, we use the standard cross-entropy loss  $\mathcal{L}_{CE}$ . Our final learning objective is

$$\mathcal{L}_F = \delta \times \mathcal{L}_{CE} + \beta \times (\mathcal{L}_{CP} + \mathcal{L}_{CS}) + \gamma \times (\mathcal{L}_{PT} + \mathcal{L}_{NT}), \quad (7)$$

The hyper-parameters  $\delta$ ,  $\beta$ , and  $\gamma$  are simply set to balance different objective terms.

## 4 EXPERIMENTAL RESULTS

### 4.1 DATASETS

We evaluate the performance of our proposed learning framework on two radiology benchmarks IU-XRay (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019) using four encoder-decoder based backbones. In addition, although focusing on medical report generation, we also show the generality of our framework on the image captioning benchmark COCO (Lin et al., 2014).

**IU-XRay** (Demner-Fushman et al., 2016) is a classic radiology dataset from Indiana University with 7,470 frontal and/or lateral X-ray images and 3,955 radiology reports. Each report consists of impression, findings and indication sections. The findings section contains multi-sentence paragraphs describing the image, and is used as the ground-truth, following the previous works (Li et al., 2018; Jing et al., 2019; Chen et al., 2020; 2021).

**MIMIC-CXR** (Johnson et al., 2019) is the largest radiology dataset consisting of 377,110 images with 227,835 reports from 64,588 patients. We use the official split that has 368,960 training samples, 2,991 validation samples and 5,159 test samples.

**COCO** (Lin et al., 2014) is the most widely used and standard dataset for image captioning. It comprises of 120,000 images, each with 5 different captions. We used the split provided by Karapathy (Karapathy & Fei-Fei, 2017), where 5000 images are used for validation, 5000 images for testing and the rest of the images for training. We download and use the splitted COCO dataset from the github repository of the meshed-memory-transformer<sup>1</sup>. For IU-XRay, samples without complete

<sup>1</sup><https://github.com/aimagelab/meshed-memory-transformer>

Table 1: Comparison of our proposed approach and state-of-the-art models on IU-XRay, MIMIC-CXR, and COCO datasets. \* indicates the results quoted from Chen et al. (2020) for IU X-Ray and MIMIC-CXR datasets. \*\* indicates the results taken directly from the respective paper. \*\*\* indicates the results obtained by running the author-released codes using the same dataset split as our approach. Basic baseline refers to the model using DenseNet-121 as encoder and GPT-2 as decoder. The metrics B is for BLEU, R-L for ROUGE-L, M for METEOR, C for CIDER.

Dataset	Methods	B-1	B-2	B-3	B-4	R-L	M	C
IU-XRay	ST*	0.216	0.124	0.087	0.066	0.306	-	-
	ATT2IN*	0.224	0.129	0.089	0.068	0.308	-	-
	ADAATT*	0.220	0.127	0.089	0.068	0.308	-	-
	CO-ATT*	0.455	0.288	0.205	0.154	0.369	-	-
	HRGR*	0.438	0.298	0.208	0.151	0.322	-	-
	CMAS-RL*	0.464	0.301	0.210	0.154	0.362	-	-
	MedSkip**	0.467	0.297	0.214	0.162	0.355	0.187	-
	CA**	0.492	0.314	0.222	0.169	0.381	0.193	-
	KERP**	0.470	0.304	0.219	0.165	0.371	0.187	0.280
	PPKED**	0.483	0.315	0.224	0.168	0.376	0.190	0.351
	R2Gen***	0.438	0.283	0.205	0.152	0.347	0.176	0.408
	<b>R2Gen + Ours</b>	0.471	0.299	0.212	0.156	0.366	0.178	0.423
	R2GenCMN***	0.467	0.300	0.210	0.156	0.369	0.185	0.401
	<b>R2GenCMN + Ours</b>	<b>0.505</b>	0.318	0.219	0.159	0.374	0.190	<b>0.440</b>
	Basic baseline	0.467	0.300	0.215	0.161	0.379	0.192	0.369
	<b>Basic baseline + Ours</b>	0.482	<b>0.318</b>	<b>0.229</b>	<b>0.171</b>	<b>0.385</b>	0.194	0.392
MIMIC-CXR	XproNet***	0.454	0.288	0.199	0.143	0.362	0.179	0.380
	<b>XproNet + Ours</b>	0.455	0.306	0.224	0.169	0.382	<b>0.196</b>	0.410
	ST*	0.299	0.184	0.121	0.084	0.263	0.124	-
	ATT2IN*	0.325	0.203	0.136	0.096	0.276	0.134	-
	ADAATT*	0.299	0.185	0.124	0.088	0.266	0.118	-
	CA**	0.350	0.219	0.148	0.106	0.278	0.142	-
	PPKED**	0.360	0.224	0.149	0.106	0.284	0.149	0.237
	R2Gen***	0.363	0.216	0.143	0.101	0.269	0.135	0.141
	<b>R2Gen + Ours</b>	0.360	0.218	0.146	0.105	0.270	0.139	0.253
	R2GenCMN***	<b>0.369</b>	0.223	0.148	0.105	0.270	0.136	0.143
	<b>R2GenCMN + Ours</b>	0.364	<b>0.225</b>	<b>0.152</b>	0.109	0.276	0.139	0.237
	Basic baseline	0.268	0.146	0.083	0.047	0.210	0.114	0.107
	<b>Basic baseline + Ours</b>	0.327	0.194	0.127	0.090	0.225	0.125	0.227
	XproNet	0.344	0.215	0.146	0.105	0.279	0.138	0.359
	<b>XproNet + Ours</b>	0.354	0.220	0.150	<b>0.112</b>	<b>0.283</b>	<b>0.138</b>	<b>0.410</b>
COCO	VisualGPT	0.677	-	-	0.236	0.486	0.221	0.768
	<b>VisualGPT + ours</b>	<b>0.694</b>	-	-	<b>0.250</b>	<b>0.493</b>	<b>0.226</b>	<b>0.830</b>

findings sections are removed, following (Li et al., 2018). The filtered images and reports for IU-XRay and MIMIC-CXR are publicly available in this repository<sup>2</sup> by (Chen et al., 2020), and they are directly downloaded and used in our experiments.

#### 4.2 IMPLEMENTATION DETAILS

We evaluate our framework upon four encoder-decoder backbones, including a basic baseline, R2Gen (Chen et al., 2020), R2GenCMN (Chen et al., 2021), and XproNet (Wang et al., 2022). While the last three are the existing state-of-the-arts (SOTA) medical report generation methods, our basic baseline uses DenseNet-121 (Huang et al., 2017) pretrained on ImageNet as the encoder and GPT-2 (Radford et al., 2019) as the decoder. Our model is trained using the Adam (Kingma & Ba, 2015) optimizer with a weight decay of  $5 \times 10^{-5}$  for 50 epochs, and a batch size of 64. The loss

<sup>2</sup><https://github.com/cuhksz-nlp/R2Gen>

Table 2: Ablation studies on component contributions (left) and the impact of the hyper-parameter  $\alpha$ . The metrics B is for BLEU, R-L for ROUGE-L, C for CIDER.

Model	B-4	R-L	C	Dataset	$\alpha$	B-4	R-L	C
Basic Baseline	0.161	0.371	0.379	IU-XRay	0.1	0.165	0.372	0.379
+ $\mathcal{L}_{CP}$	0.161	0.371	0.382		0.01	<b>0.171</b>	<b>0.385</b>	<b>0.392</b>
+ $\mathcal{L}_{CP} + \mathcal{L}_{CS}$	0.164	0.375	0.383		0.001	0.165	0.374	0.388
+ $\mathcal{L}_{CP} + \mathcal{L}_{CS} + \mathcal{L}_{PT}$	0.167	0.378	0.388	COCO	0.1	<b>0.250</b>	<b>0.493</b>	<b>0.830</b>
+ $\mathcal{L}_{CP} + \mathcal{L}_{CS} + \mathcal{L}_{PT} + \mathcal{L}_{NT}$	<b>0.171</b>	<b>0.380</b>	<b>0.392</b>		0.01	0.239	0.485	0.766
					0.001	0.235	0.481	0.760

weights in Eq. 7 are set as 0.30, 0.05, and 0.65 for  $\delta$ ,  $\beta$  and  $\gamma$  for all datasets. The initial learning rate is set as  $1 \times 10^{-4}$  and further reduced by 10 times if there is no improvement in either BLEU-3 or BLEU-4 score on the validation set. When evaluating the existing SOTA models with or without adding our losses, we use the default hyperparameters, optimizer, scheduler, and experimental settings of the original methods for fair comparison. Following the previous works in medical report generation and image captioning, we evaluate our model on widely used Natural Language Generation (NLG) metrics: BLEU (Papineni et al., 2002), CIDER (Vedantam et al., 2015), METEOR (Lavie & Agarwal, 2007) and ROUGE-L (Lin, 2004).

### 4.3 RESULTS

To validate our approach on medical report generation, we first compare our model with the SOTA image captioners ST (Vinyals et al., 2015), ATT2IN (Rennie et al., 2017), ADAATT (Lu et al., 2017), and medical report generation models CO-ATT (Jing et al., 2018), CMAS-RL (Jing et al., 2019), HRGR (Li et al., 2018), R2Gen (Chen et al., 2020), R2GenCMN (Chen et al., 2021), PPKED (Liu et al., 2021a), KERP (Li et al., 2019), XproNet (Wang et al., 2022), MedSkip (Pahwa et al., 2021), and CA (Liu et al., 2021b). Since the models PPKED, KERP, MedSkip, and CA are not open-sourced, we directly quote the results published in their literature. For the open-sourced models R2Gen(Chen et al., 2020), R2GenCMN (Chen et al., 2021), and XproNet (Wang et al., 2022), we run their codes with their default experimental settings. Moreover, to further verify if our framework also benefits generic image captioning tasks, we also build our framework upon the image captioning model VisualGPT Chen et al. (2022) and test it on the image captioning benchmark COCO.

#### 4.3.1 QUANTITATIVE ANALYSIS


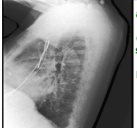
The results in Table 1 demonstrate the effectiveness of our approach compared with the SOTA methods. After integrating our proposed learning framework into the basic baseline and the SOTA models in medical report generation and generic image captioning, we consistently observe a significant boost in the scores of evaluation metrics, especially in CIDER. The best performance is mostly achieved when incorporating our framework into the strong backbones of the SOTA methods, showing that our way of improvement is orthogonal to the current efforts seen in the literature, and could further benefit the latter. Moreover, mining the hard negatives in our way could also enhance the standard image captioning on COCO dataset. Our approach significantly boosts the performance of the very recently proposed VisualGPT across all NLG metrics. It is noteworthy that out of all NLG metrics, our increase in CIDER scores is the most significant across all the models. CIDER in general could reflect the diversity of the generated text as it down-weights the common content across reports. This verifies the effectiveness of our proposed learning framework through leveraging harder and harder negatives during training.

#### 4.3.2 QUALITATIVE ANALYSIS

In Figure 2, we compare the example reports generated by the SOTA models with and without our learning framework. The generated sentences are differently colored, with green indicating semantically correct generation and red indicating incorrect generation. As can be seen, comparing with the reports generated by XproNet, R2Gen and R2GenCMN, incorporating our framework could



Figure 2: Examples of reports generated by various models. From top to bottom, images are taken from the IU-XRay and MIMIC-CXR. Green colour represents the text semantically similar to the ground-truth. Red colour represents the text incorrectly generated. Blue colour in the ground-truth report represents the text not generated by any of the models under comparison.

Medical Image	GT	XproNet	XproNet + Ours	R2Gen	R2Gen + Ours	R2GenCMN	R2GenCMN + Ours
	The heart size and mediastinal contours appear within normal limits. No focal airspace consolidation, pleural effusion or pneumothorax. No acute bony abnormalities.	The heart size and mediastinal contours are within normal limits. The lungs are clear. No pneumothorax or pleural effusion.	The heart size and mediastinal contours are within normal limits. The lungs are clear. There is no pneumothorax or large pleural effusion. There is no acute bony abnormalities.	heart size normal . lungs are clear . xxxx are normal . no pneumonia effusions edema pneumothorax adenopathy nodules or masses .	the cardiomedastinal silhouette is normal in size and contour . no focal consolidation pneumothorax or large pleural effusion . negative for acute bone abnormality	heart size normal . lungs are clear . xxxx are normal . no pneumonia effusions edema pneumothorax adenopathy nodules or masses .	the lungs are clear . there is no pleural effusion or pneumothorax . the heart and mediastinum are normal . the skeletal structures are normal .
	ap upright and lateral views of the chest were provided . left chest wall pacer pack is again seen with leads extending into the right heart . abandoned pacing leads are also noted in the right chest wall extending into the right heart . the heart remains moderately enlarged . lung volumes are low with equivocal ground-glass opacity	ap upright and lateral views of the chest provided . a left chest wall is seen right atrium bony structures are intact	ap upright and lateral views of the chest provided . a left chest wall pacer device is again seen with leads extending into the region of the right atrium and right ventricle . the heart remains mildly enlarged .	ap upright and lateral views of the chest were provided . left chest extending into the right heart . enlarged heart	ap upright and lateral views of the chest are given . left chest wall pacer pack is extending into the right heart . abandoned pacing leads are also noted in the right chest wall extending into the right heart . the heart is enlarged .	frontal and lateral views of the chest provided . a left chest wall is seen right atrium bony structures are intact	ap upright and lateral views of the chest provided . a left chest wall is seen right atrium bony structures are intact . abandoned pacing noted in the chest wall right heart .

reduce the incorrect sentences generated in the reports. For instance, in the example from MIMIC-CXR, the SOTA models trained with our framework are able to identify and generate the sentence - “abandoned pacing leads are also noted in the right chest wall extending into the right heart”, which is missing in the generated reports without using our approach. This is consistent with the quantitative analysis. More extensive analysis with multiple example images from all three datasets is available in Appendix A.

#### 4.4 ABLATION STUDY

To understand the contribution of each component of our proposed framework, we conduct an ablation study on three evaluation metrics - BLEU-4, ROUGE-L and CIDER on IU-XRay. We use the basic baseline as the backbone and add the losses of  $\mathcal{L}_{CP}$ ,  $\mathcal{L}_{CS}$ ,  $\mathcal{L}_{PT}$ , and  $\mathcal{L}_{NT}$  one by one. The results are shown in Table 2 Left. By adding the clip loss  $\mathcal{L}_{CP}$  to the basic baseline, performance increase is observed for all the four metrics, showing the effectiveness of using contrastive based learning process with more hard negatives. After adding the cosine-similarity loss  $\mathcal{L}_{CS}$  to supervise the encoder to capture distinctive image features, further performance improvement could be achieved. Finally, by adding the two triplet losses  $\mathcal{L}_{PT}$  and  $\mathcal{L}_{NT}$ , we achieve the optimal performance compared to the baseline. Moreover, to show the influence of  $\alpha$  which controls the rate of hardness increase of the synthesized harder negative features computed using Eq. 3 and Eq. 4, we investigate the performance of the model with different values of  $\alpha$  as shown in Table 2 Right. The optimal value of  $\alpha$  is depends on the nature of dataset. The higher the  $\alpha$  value, the faster will be the rate of increase in the hardness of negative reports. For medical datasets such as IU-Xray and MIMIC-CXR, where most of the medical images have the reports of similar nature, having a lower  $\alpha$  value gives better performance. On the other hand, for a diverse dataset such as COCO, slightly higher  $\alpha$  value gives better results.

## 5 CONCLUSION AND FUTURE WORK

In this paper, we propose a general framework to improve medical report generation through learning significant and unique image and report features. We demonstrate that our proposed framework can be readily incorporated with a variety of SOTA medical report generation models, and consistently boost the performance of the latter. We also show that even generic image captioners could benefit from our proposed framework although this is not the focus of this paper. On the other hand, currently, we compute the hard negative priors for each report in the dataset in a brute force manner. Although this is an offline one-off expense, the time for this pre-computation would increase significantly for larger datasets ( $> 10M$  samples). In future, we would explore effective ways such as approximate nearest neighbour search to compute hard negative priors more efficiently.

## REFERENCES

- Omar Alfarghaly, Rana Khaled, Abeer Elkorany, Maha Helal, and Aly Fahmy. Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, 24:100557, 2021. ISSN 2352-9148. doi: <https://doi.org/10.1016/j.imu.2021.100557>. URL <https://www.sciencedirect.com/science/article/pii/S2352914821000472>.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, 2018. doi: 10.1109/CVPR.2018.00636.
- Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18030–18040, June 2022.
- Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1439–1449, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.112. URL <https://aclanthology.org/2020.emnlp-main.112>.
- Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5904–5914, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.459. URL <https://aclanthology.org/2021.acl-long.459>.
- Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza M. Rodriguez, S. Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association : JAMIA*, 23 2:304–10, 2016.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.
- Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017. doi: 10.1109/CVPR.2017.243.
- Baoyu Jing, Pengtao Xie, and Eric Xing. 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, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1240. URL <https://aclanthology.org/P18-1240>.
- Baoyu Jing, Zeya Wang, and Eric Xing. 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, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1657. URL <https://aclanthology.org/P19-1657>.
- Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *CoRR*, abs/1901.07042, 2019. URL <http://arxiv.org/abs/1901.07042>.
- Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In *Neural Information Processing Systems (NeurIPS)*, 2020.

- Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:664–676, 2017.
- Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.
- Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3337–3345, 2017.
- Alon Lavie and Abhaya Agarwal. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 228–231, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL <https://aclanthology.org/W07-0734>.
- Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6666–6673, Jul. 2019. doi: 10.1609/aaai.v33i01.33016666. URL <https://ojs.aaai.org/index.php/AAAI/article/view/4637>.
- Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL <https://proceedings.neurips.cc/paper/2018/file/e07413354875be01a996dc560274708e-Paper.pdf>.
- Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL <https://aclanthology.org/W04-1013>.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.
- Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13753–13762, June 2021a.
- Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. Contrastive attention for automatic chest X-ray report generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 269–280, Online, August 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.23. URL <https://aclanthology.org/2021.findings-acl.23>.
- Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3242–3250, 2017. doi: 10.1109/CVPR.2017.345.
- Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. Progressive transformer-based generation of radiology reports. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2824–2832, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.241. URL <https://aclanthology.org/2021.findings-emnlp.241>.
- Esha Pahwa, Dwij Mehta, Sanjeet Kapadia, Devansh Jain, and Achleshwar Luthra. Medskip: Medical report generation using skip connections and integrated attention. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3402–3408, 2021. doi: 10.1109/ICCVW54120.2021.00380.

- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL <https://aclanthology.org/P02-1040>.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/radford21a.html>.
- Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv e-prints*, art. arXiv:1711.05225, November 2017.
- Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL <https://aclanthology.org/D19-1410>.
- Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1179–1195, 2017. doi: 10.1109/CVPR.2017.131.
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, 2015. doi: 10.1109/CVPR.2015.7299087.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and D. Erhan. Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3156–3164, 2015.
- Jun Wang, Abhir Bhalerao, and Yulan He. Cross-modal prototype driven network for radiology report generation. In *ECCV*, 2022.
- Yixin Wang, Zihao Lin, Jiang Tian, Zhongchao Shi, Yang Zhang, Jianping Fan, and Zhiqiang He. Confidence-guided radiology report generation. *ArXiv*, abs/2106.10887, 2021a.
- Zhanyu Wang, Luping Zhou, Lei Wang, and Xiu Li. A self-boosting framework for automated radiographic report generation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2433–2442, 2021b. doi: 10.1109/CVPR46437.2021.00246.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2048–2057, Lille, France, 07–09 Jul 2015. PMLR. URL <https://proceedings.mlr.press/v37/xuc15.html>.
- Yuan Xue, Tao Xu, L. Rodney Long, Zhiyun Xue, Sameer Antani, George R. Thoma, and Xiaolei Huang. Multimodal recurrent model with attention for automated radiology report generation. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pp. 457–466, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00928-1.

Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pp. 721–729, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32226-7.

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan L. Yuille, and Daguang Xu. When radiology report generation meets knowledge graph. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 12910–12917. AAAI Press, 2020. URL <https://ojs.aaai.org/index.php/AAAI/article/view/6989>.

## A MORE QUALITATIVE RESULT

In Figure 3 and Figure 4, we provide additional images from COCO, IU-XRay, and MIMIC-CXR datasets for comparing the reports/captions generated by the SOTA models with and without our framework.

Figure 3: Comparison of captions generated by VisualGPT on COCO images with and without our framework. Green colour represents the text semantically similar to the ground-truth. Red colour represents the text incorrectly generated. Blue colour in the ground-truth caption represents the text not generated by any of the models under comparison.








COCO	Ground Truth	VisualGPT	VisualGPT + Ours
	The lady has her tennis racket aimed at the ball in the air.	A woman holding a tennis racket on a court	A woman swining the tennis racket on a tennis court
	A living area with table, chairs and a sofa	A living area with table, and sofa	A living area with table chairs, and a sofa
	a large clock tower is on top of a building	a building with the clock tower on the road	a large building with a large clock tower on the top.

Figure 4: Comparison of reports generated by the SOTA models on IU-XRay and MIMIC-CXR images with and without our framework. Green colour represents the text semantically similar to the ground-truth. Red colour represents the text incorrectly generated. Blue colour in the ground-truth report represents the text not generated by any of the models under comparison.

IU XRAY		<b>Baseline</b> Heart size and mediastinal contours are within normal limits. The lungs are clear. No pneumothorax or pleural effusion. No pneumothorax. No pleural effusion.	<b>R2Gen</b> the heart is normal in size . the mediastinum is unremarkable . the lungs are clear .	<b>R2GenCMN</b> heart size normal . lungs are clear . xxxx are normal . no pneumonia effusions edema pneumothorax adenopathy nodules or masses .	<b>XproNet</b> the lungs are clear . heart size is normal . no pneumothorax .
	<b>Ground Truth</b> Lungs are clear bilaterally. Cardiac and mediastinal silhouettes are normal. Pulmonary vasculature is normal. No pneumothorax or pleural effusion. No acute bony abnormality.	<b>Baseline + Ours</b> The cardiomeastinal silhouette is within normal limits. The lungs are clear. There is no pneumothorax or pleural effusion. No acute bony abnormalities.	<b>R2Gen + Ours</b> heart size is normal . the lungs are clear . there are no focal air space consolidations . no pleural effusions or pneumothoraces . the hilar and mediastinal contours are normal . normal pulmonary vascularity .	<b>R2GenCMN + Ours</b> The cardiomeastinal silhouette is within normal limits. The lungs are clear. There is no pneumothorax or pleural effusion. No acute bony abnormalities.	<b>XproNet + Ours</b> the heart and mediastinal contours are within normal limits . the lungs are clear . the lungs are clear . there is no pneumothorax or pleural effusion .
IU XRAY		<b>Baseline</b> The cardiomeastinal silhouette is within normal limits for size and contour. The lungs are clear. No pneumothorax or pleural effusion. There are no focal airspace consolidation.	<b>R2Gen</b> there is a left base opacity . the right lung is grossly clear . heart size is normal . left venous catheter with tip in the right atrium . there is no pneumothorax .	<b>R2GenCMN</b> heart size is normal . no mediastinal silhouette . no pneumothorax . lungs are clear .	<b>XproNet</b> there is a left base opacity . the right lung is grossly clear . heart size is normal . there is no pneumothorax .
	<b>Ground Truth</b> The heart size and pulmonary vascularity appear within normal limits. Right pleural effusion is present and appears increased. No pneumothorax is identified. Some scattered XXXX of right base atelectasis are seen. Surgical XXXX remain in XXXX. The left lung appears clear.	<b>Baseline + Ours</b> The cardiomeastinal silhouette is normal. The lungs are clear. No pneumothorax or pleural effusion. Surgical XXXX remain XXXX.	<b>R2Gen + Ours</b> the heart size and pulmonary vascularity appear within normal limits . the lungs are free of focal airspace disease . no pleural effusion or pneumothorax is seen . a few bandlike opacities are present which are xxxx to represent small areas of scarring or atelectasis .	<b>R2GenCMN + Ours</b> the heart size and pulmonary vascularity appear within normal limits . the lungs are free of focal airspace disease . no pleural effusion or pneumothorax is seen . surgical xxxx remain in xxxx . increased pleural effusion	<b>XproNet + Ours</b> the cardiomeastinal silhouette is normal . the heart size and pulmonary vascularity appear within normal limits . the lungs are free . no pleural effusion or pneumothorax is seen . increased pleural effusion
MIMIC-CXR		<b>Baseline</b> the lungs are clear. no acute osseous abnormalities identified degenerative changes are seen in the thoracic spine no acute osseous abnormalities identified surgical clips are noted in the right upper quadrant	<b>R2Gen</b> frontal and lateral views of the chest are obtained . no focal consolidation pleural effusion or evidence of pneumothorax is seen	<b>R2GenCMN</b> pa and lateral views of the chest provided . there is no focal consolidation effusion or pneumothorax . the cardiomeastinal silhouette is normal . imaged osseous structures are intact .	<b>XproNet</b> pa and lateral views of the chest provided . there is no focal consolidation effusion or pneumothorax . the cardiomeastinal silhouette is normal . imaged osseous structures are intact . imaged osseous structures are intact
	<b>Ground Truth</b> 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 .	<b>Baseline + Ours</b> frontal and lateral views of the chest are obtained . lungs are hyperinflated . no pleural effusion pneumothorax .	<b>R2Gen + Ours</b> frontal and lateral views of the chest are obtained . no focal consolidation pleural effusion or pneumothorax . no focal consolidation	<b>R2GenCMN + Ours</b> pa and lateral views of the chest provided . there is no focal consolidation effusion or pneumothorax . the lungs remain hyperinflated the cardiomeastinal silhouette is normal .	<b>XproNet + Ours</b> pa and lateral views of the chest provided . lungs are hyperinflated . there is no focal consolidation effusion or pneumothorax . the cardiomeastinal silhouette is normal . imaged osseous structures are intact . stable hilar contours
MIMIC-CXR		<b>Baseline</b> the lungs are clear. no acute osseous abnormalities identified degenerative changes are seen in the thoracic spine no acute osseous abnormalities identified surgical clips are noted in the right upper quadrant	<b>R2Gen</b> frontal and lateral views of the chest are obtained . no focal consolidation pleural effusion or evidence of pneumothorax is seen	<b>R2GenCMN</b> pa and lateral views of the chest provided . there is no focal consolidation effusion or pneumothorax . the cardiomeastinal silhouette is normal . imaged osseous structures are intact .	<b>XproNet</b> pa and lateral views of the chest provided . there is no focal consolidation effusion or pneumothorax . the cardiomeastinal silhouette is normal . imaged osseous structures are intact . imaged osseous structures are intact
	<b>Ground Truth</b> 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 .	<b>Baseline + Ours</b> frontal and lateral views of the chest are obtained . lungs are hyperinflated . no pleural effusion pneumothorax .	<b>R2Gen + Ours</b> frontal and lateral views of the chest are obtained . no focal consolidation pleural effusion or pneumothorax . no focal consolidation	<b>R2GenCMN + Ours</b> pa and lateral views of the chest provided . there is no focal consolidation effusion or pneumothorax . the lungs remain hyperinflated the cardiomeastinal silhouette is normal .	<b>XproNet + Ours</b> pa and lateral views of the chest provided . lungs are hyperinflated . there is no focal consolidation effusion or pneumothorax . the cardiomeastinal silhouette is normal . imaged osseous structures are intact . stable hilar contours