

# Automatic Radiology Report Generation by Learning with Increasingly Hard Negatives

Bhanu Prakash Voutharoja<sup>a</sup>, Lei Wang<sup>a,\*</sup> and Luping Zhou<sup>b</sup>

<sup>a</sup>School of Computing and IT, University of Wollongong, NSW, Australia

<sup>b</sup>School of Electrical and Information Engineering, University of Sydney, NSW, Australia

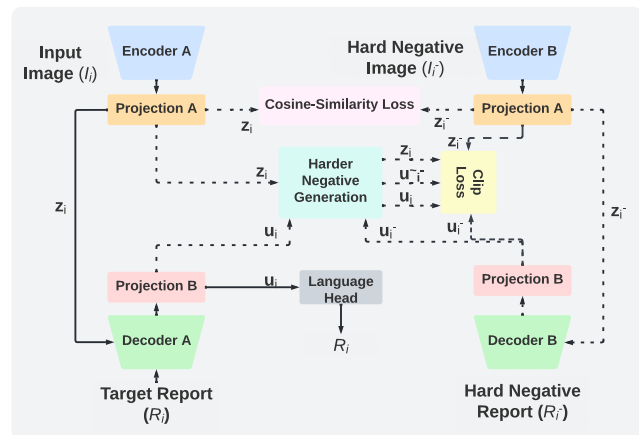
ORCID ID: Bhanu Prakash Voutharoja <https://orcid.org/0000-0002-1352-2858>

**Abstract.** Automatic radiology report generation is challenging as medical images or reports are usually similar to each other due to the common content of anatomy. This makes a model hard to capture the uniqueness of individual images and is prone to producing undesired generic or mismatched reports. This situation calls for learning more discriminative features that could capture even fine-grained mismatches between images and reports. To achieve this, this paper proposes a novel framework to learn discriminative image and report features by distinguishing them from their closest peers, i.e., hard negatives. Especially, to attain more discriminative features, we gradually raise the difficulty of such a learning task by creating increasingly hard negative reports for each image in the feature space during training, respectively. By treating the increasingly hard negatives as auxiliary variables, we formulate this process as a min-max alternating optimisation problem. At each iteration, conditioned on a given set of hard negative reports, image and report features are learned as usual by minimising the loss functions related to report generation. After that, a new set of harder negative reports will be created by maximising a loss reflecting image-report alignment. By solving this optimisation, we attain a model that can generate more specific and accurate reports. It is noteworthy that our framework enhances discriminative feature learning without introducing extra network weights. Also, in contrast to the existing way of generating hard negatives, our framework extends beyond the granularity of the dataset by generating harder samples out of the training set. Experimental study on benchmark datasets verifies the efficacy of our framework and shows that it can serve as a plug-in to readily improve existing medical report generation models. The code is publicly available at <https://github.com/Bhanu068/ITHN>

## 1 Introduction

A radiology report is a multi-sentence paragraph that precisely describes the normal and abnormal regions in a radiology image. Writing such reports requires proper experience and expertise [11]. Automating this process can reduce manual workload and speed up clinic procedure. Although radiology report generation is similar to image captioning, the former experiences more subtle correlation between medical images and corresponding reports. This could make image captioning models [32, 35, 2, 29, 23] fail when directly applied on medical datasets. Many works have been proposed for medical report generation recently [11, 10, 18, 5, 4, 34], with the focus

on improving the generated reports by using medical tags [11], large pretrain models [8, 7, 27], relational memory [5, 4], and so on.



**Figure 1.** Based on an encoder-decoder structure, our framework utilises increasingly hard negative reports to enforce the model to learn more distinctive features for both image and report so as to attain accurate and diverse medical reports. At the inference stage, only images are needed as the input for the model to output the generated reports. Detailed explanation is in Sec. 3.

Despite of the progress, the generated reports could still be heavily dominated by the terms describing the common contents of medical images [10], making the characteristics of individual reports submerged. Recently, some methods have been proposed to mitigate this issue, for example, by learning sample-specific features for image or report [21] or contrasting the given samples to their closer ones within the training set to guide the report generation [22]. Nevertheless, they 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 solely use either image or text modality to capture abnormality-related features, ignoring the crucial information from the other modality. For example, the former [21] relies on merely text modality, whereas the latter [22] solely utilizes image modality.

To address this situation, we argue that it is imperative to learn more discriminative features to represent image and report and con-

\* Corresponding Author. Email: leiw@uow.edu.au.

sider both visual and textual modalities so as to generate reports with sufficient individual characteristics. Following this idea, we propose a method that can better capture the unique features of a given image (or report) by differentiating it from its closest hard negative report (or image) in this paper. Particularly, to adequately capture the fine differences among similar images and reports, we innovatively synthesise *increasingly harder negative reports in the feature space* with the learning process. Specifically, we view the increasingly hard negatives as auxiliary variables and put forward a min-max alternating optimisation to synthesise them during training. At each iteration, by minimisation, we mean to reduce the loss on report generation to update the model weights; By maximisation, we mean to increase the loss on image-report alignment to create a new set of harder negatives to make the "minimisation" in the next iteration more difficult. Through this alternating optimisation, we keep challenging the model to push it to learn more discriminative features such that it can better characterise different medical images and reports and get the true pairs well aligned. After the model is trained, at the inference stage, we only need a medical image as the input to generate a report as usual.

To realise and enrich the above methods, this work also considers the following issues. 1) When synthesising the increasingly hard negative reports in the learned feature space, we shall maintain their nature, i.e., restricting them to the manifold of "reports". We implement this by imposing a simple constraint during the maximisation process. 2) To contrast with our method, we also put forward another way to synthesise increasingly hard negative reports. It is a weighted linear combination of a given report and its initial hard negative in the feature space, which is inspired by a strategy called MoChI [13]. As will be seen by both theoretical analysis and experimental study, our method shows clear advantage thanks to the optimisation process. 3) We identify the initial set of hard negatives based on the report modality. This is because compared to images, textual reports more directly reflect high-level clinic-related semantics. An initial hard negative can readily be obtained for each report by extracting its features with an off-the-shelf transformer model and identifying its nearest neighbour report in the training set. The anchor report and this identified peer form a "hard negative report pair." 4) We observe that when two reports are nearest neighbours, their ground truth paired images will usually be visually similar. Based on this observation, we regard the two images as nearest neighbours too. In other words, based on the above "hard negative report pair," we form a "hard negative image pair" correspondingly. For the two images in a pair, a loss will be defined in our method to encourage their features to be different from each other. 5) Lastly, to fully utilise the synthesised harder negatives, we modify the contrastive loss (CLIP) [26] when using it in our method. A similarity matrix between the features of images and those of the synthesised harder negative reports is additionally computed. It is appended to the original similarity matrix in CLIP loss to form the final matrix to be used.

Our main contributions are summarised as follows.

- *First*, we propose a simple yet effective method that helps the model to capture discriminative features in both image and text modalities to generate diverse and accurate reports. Our method can readily work with existing encoder-decoder based medical report generation models with a minimum amount of alteration. Moreover, it does not introduce any additional network weights.
- *Second*, we propose an optimisation-based mechanism to utilise increasingly hard negatives to enhance the learning performance. A principled way is explored to synthesise a set of optimal harder

negative reports in the sense of loss function *maximisation*. By training our model with the increasingly harder negatives, the recognition resolution of our model can be increased. Furthermore, unlike the conventional approach of identifying hard negatives from within the data set, our framework extends beyond the granularity of training data by synthesising new, harder negatives.

- *Third*, we propose an improved variant of the contrastive loss to fully utilise the synthesised hard negative reports rather than merely relying on the negatives randomly sampled into a batch as in the original CLIP loss.

With three different backbones, we validate the proposed framework on two benchmarks IU-XRay [6] and MIMIC-CXR [12] under a variety of assessment criteria. As experimentally demonstrated, it outperforms the existing medical report generation methods and achieves the state-of-the-art performance.

## 2 Related Work

**Medical Report Generation.** Earlier works adopted a traditional encoder-decoder framework for report generation. To generate coherent reports, [11] proposed a co-attention mechanism to localize abnormal regions and a hierarchical 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, [10] proposed a multi-agent model to alleviate the data bias problem between normal and abnormal regions of the medical images. [5] 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, [4] used a relational memory matrix for cross-modal alignment of image and text features. After the success of GPT-2 [27] in text generation tasks, [1] used GPT-2 for report generation. First, they fine-tuned CheXNet [28] 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, [22] proposed contrastive attention mechanism focusing on abnormal regions of the image. It consists of two attentions - aggregate attention to summarise the information from all the reports in the normality pool and differentiate attention to remove common information between normal and abnormal images. 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. In other work, PPKED [21] used prior knowledge, posterior knowledge, and multi-domain knowledge distiller to generate the report. This method tries to mimic the behavior of a radiologist, by first predicting the disease tags focusing on the abnormal regions of the image (posterior knowledge), utilising 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, [34] used an additional image-text matching branch to align image and report features in the latent space, and utilised the SBert generated reports during training as the hard negative reports that are close to the ground-truth to supervise branch.

The main drawback of these methods is that they try to utilize distinctive information from just one modality. We propose an approach to utilise distinctive features from both modalities and closely align them in the latent space by evolving the hard negatives.

### 3 Proposed Method

#### 3.1 Overview

Given an image set  $\mathcal{I} = \{I_1, I_2, \dots, I_n\}$  and the corresponding ground-truth paired report set  $\mathcal{R} = \{R_1, R_2, \dots, R_n\}$ , our framework uses an encoder to extract image representation and passes it to a decoder to generate a report.

For each report  $R_i$ , we obtain its nearest neighbour, denoted by  $R_i^-$ , from the training set by using the RadBERT model [36] (Sec. 3.2). As mentioned in introduction, we also regard their paired images, denoted by  $I_i$  and  $I_i^-$ , as nearest neighbours to each other. Images  $I_i$  and  $I_i^-$  are fed into two encoders with shared parameters and reports  $R_i$  and  $R_i^-$  are fed into two decoders sharing parameters too. The features of the two images extracted by the encoders are denoted as  $\mathbf{z}_i$  and  $\mathbf{z}_i^-$ , respectively. To enforce the encoders to learn features that can differentiate the two neighbouring images, we define a loss to encourage the dissimilarity between  $\mathbf{z}_i$  and  $\mathbf{z}_i^-$  (Sec. 3.3). The two image features are further fed into the decoders to produce the corresponding report features, which are denoted by  $\mathbf{u}_i$  and  $\mathbf{u}_i^-$ , respectively. To further enhance the discriminative ability of report features  $\mathbf{u}$ , we synthesise an increasingly hard negative for each report  $\mathbf{u}_i$  during the training process, denoted by  $\tilde{\mathbf{u}}_i$ , in the feature space (Sec. 3.4). An improved variant of the CLIP loss is then defined to align report feature  $\mathbf{u}_i$  with the corresponding image feature  $\mathbf{z}_i$ , while pushing the synthesised harder negative features  $\tilde{\mathbf{u}}_i$  away from  $\mathbf{z}_i$  (Sec. 3.5). Our framework is illustrated in Fig. 1. We elaborate the framework in the following sections.

#### 3.2 Identifying Prior Hard Negative Reports

Before training, all reports  $\{R_1, R_2, \dots, R_n\}$  in the training set are fed into RadBERT [36], a pre-trained transformer-based language model tailored to radiology, to extract their report features, denoted by  $\{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n\}$ . For each report  $\mathbf{r}_i$ , its nearest neighbour is identified by a cosine similarity score between  $\mathbf{r}_i$  and  $\mathbf{r}_j$ , where  $j = 1, \dots, n, j \neq i$ . By doing so, we create a hard negative report pair  $(R_i, R_i^-)$  for each report  $R_i$ . Accordingly, we regard their corresponding images as a hard negative image pair  $(I_i, I_i^-)$  too. Note that we create the hard negatives from the perspective of reports as they directly contain high-level clinic-related semantics. The hard negative reports obtained above are used as a prior in our method. Their hardness will be increased gradually during the training, as will be detailed in Sec. 3.4.

#### 3.3 Learning Distinctive Image Features

The pair of images  $I_i$  and  $I_i^-$  are sent through Encoder  $A$  and Encoder  $B$  in Fig. 1, respectively, to obtain image features  $\mathbf{z}_i$  and  $\mathbf{z}_i^-$ . We then define a loss  $\mathcal{L}_{CS}$  to be minimized to make each image  $I_i$  and its hard negative  $I_i^-$  dissimilar from each other. This aims to learn powerful encoders capable of differentiating the two closely similar images. A common cosine similarity score is used.

$$\mathcal{L}_{CS} = \frac{\mathbf{z}_i \cdot \mathbf{z}_i^-}{\|\mathbf{z}_i\| \cdot \|\mathbf{z}_i^-\|}. \quad (1)$$

In addition, a cross-entropy loss  $\mathcal{L}_{CE}$  used to generate reports adds another level of supervision on the encoder to learn image representations useful for report generation. Together, the losses  $\mathcal{L}_{CS}$  and  $\mathcal{L}_{CE}$  make the image representations from the encoder not only discriminative but also significant for report generation.

#### 3.4 Synthesising Increasingly Hard Negative Reports

The image features  $\mathbf{z}_i$  and  $\mathbf{z}_i^-$  along with the corresponding reports  $R_i$  and  $R_i^-$  are passed to Decoders  $A$  and  $B$ . The report features produced at the last hidden layer of the two decoders are denoted as  $\mathbf{u}_i$  and  $\mathbf{u}_i^-$ , respectively, where  $\mathbf{u}_i^-$  is the hard negative of  $\mathbf{u}_i$  as previously defined. The decoders should well differentiate  $\mathbf{u}_i$  from  $\mathbf{u}_i^-$  so as to achieve discriminative report features. To facilitate this, we argue that the efficacy of decoders could be enhanced by challenging it with increasingly hard  $\mathbf{u}_i^-$ , i.e., higher difficulty to differentiate  $\mathbf{u}_i^-$  from  $\mathbf{u}_i$ . To achieve this, two strategies to synthesise harder negative reports are put forward as follows.

The first strategy is inspired by a hard-negative mixing method MoChi [13]. MoChi designs harder negatives to facilitate better and faster contrastive learning in computer vision tasks such as image classification, object detection, and image segmentation. Following the spirit, we utilise it in our framework for medical report generation. In the feature space, given a report  $\mathbf{u}_i$  and its  $k$  nearest neighbours  $\{\mathbf{u}_{r_1}, \dots, \mathbf{u}_{r_k}\}$  identified via RadBERT before training starts, we synthesise the harder negative, denoted by  $\tilde{\mathbf{u}}_i$ , by linearly combining  $\mathbf{u}_i$  and a randomly sampled  $\mathbf{u}_{r_i}$ , where  $1 \leq i \leq k$ .

$$\tilde{\mathbf{u}}_i = (1 - \lambda)\mathbf{u}_{r_i} + \lambda\mathbf{u}_i \quad (2)$$

where  $\lambda$  is a parameter controlling the contribution of report  $\mathbf{u}_i$  to the synthesised harder negative.

As will be experimentally demonstrated, the harder negatives synthesised by our MoChi-inspired strategy can indeed improve the quality of the generated reports. However, note that MoChi [13] only considers the tasks of single modality (i.e., image-based classification, detection, and segmentation). This nature is also reflected in our MoChi-inspired strategy in which  $\mathbf{u}_{r_i}$  and  $\mathbf{u}_i$  are features from the same modality (i.e., report). Differently, medical report generation involves two modalities, i.e., image and text. Both of them shall be taken into account when synthesising the harder negatives. More importantly, the strategy in Eq.(2) is disconnected from the training process of report generation and its fitness to our task is not justified.

The second strategy, which is developed by this work, addresses the above issues by proposing an optimisation-based approach. Let  $\mathcal{L}_F$  denote our final loss function used to train the model for report generation. We use  $\theta$  to denote the set of network weights to be learned and recall that  $\tilde{\mathbf{u}}_i$  stands for the harder negative to be synthesised for a report  $\mathbf{u}_i$ , where  $i = 1, \dots, n$ . By treating the harder negatives as auxiliary variables, we can explicitly express  $\mathcal{L}_F$  as  $\mathcal{L}_F(\theta; \tilde{\mathbf{u}}_1, \dots, \tilde{\mathbf{u}}_n)$ . To synthesise the harder negatives with the training process, we propose a min-max optimisation problem

$$\min_{\theta} \max_{\{\tilde{\mathbf{u}}_1, \dots, \tilde{\mathbf{u}}_n\}} \mathcal{L}_F(\theta; \tilde{\mathbf{u}}_1, \dots, \tilde{\mathbf{u}}_n). \quad (3)$$

This optimisation is solved in an alternating manner. At each training iteration, given the current set of  $\{\tilde{\mathbf{u}}_1, \dots, \tilde{\mathbf{u}}_n\}$ , the above problem reduces to  $\min_{\theta} \mathcal{L}_F(\theta)$ , which corresponds to the common case of network weights updating, say, via the back-propagation process. After that, given the updated  $\theta$ , the above problem reduces to  $\max_{\{\tilde{\mathbf{u}}_1, \dots, \tilde{\mathbf{u}}_n\}} \mathcal{L}_F(\tilde{\mathbf{u}}_1, \dots, \tilde{\mathbf{u}}_n)$ . It means to find a new set of harder negatives that can maximally increase the current loss to make the subsequent learning more challenging. In report generation tasks, the final loss  $\mathcal{L}_F$  is usually a linear combination of multiple sub-loss functions, in which a triplet loss or a contrastive loss depends on the hard negatives while the remaining losses are not.

In the following, we demonstrate this maximisation problem via a triplet loss. We use this loss by considering that 1) triplet loss will

lead to a simpler optimisation for each  $\mathbf{u}_i^-$ ; 2) triplet loss makes the comparison of this strategy with the above MoChi-inspired strategy more clear; and 3) triplet loss shares the similar spirit with the contrastive loss. Let  $(\mathbf{z}_i, \mathbf{u}_i, \tilde{\mathbf{u}}_i^-)$  denote a triplet, where we recall that  $\mathbf{z}_i$  is image feature,  $\mathbf{u}_i$  the feature of paired report, and  $\tilde{\mathbf{u}}_i^-$  the hard negative to be synthesised. A common triplet loss is expressed as

$$\mathcal{L}(\tilde{\mathbf{u}}_i^-) = \max(\|\mathbf{z}_i - \mathbf{u}_i\|^2 - \|\mathbf{z}_i - \tilde{\mathbf{u}}_i^-\|^2 + m_0, 0), \quad (4)$$

where  $m_0$  is a preset constant of margin. When the first term in the bracket is greater than zero, a loss will be incurred. In this case, the gradient of the loss with respect to  $\tilde{\mathbf{u}}_i^-$  can be easily shown as  $\frac{\partial \mathcal{L}(\tilde{\mathbf{u}}_i^-)}{\partial \tilde{\mathbf{u}}_i^-} = 2(\mathbf{z}_i - \tilde{\mathbf{u}}_i^-)$ . This indicates that to *increase* the loss, the harder negative shall be synthesised by moving the current  $\tilde{\mathbf{u}}_i^-$  towards the image feature  $\mathbf{z}_i$ .

Nevertheless, we observe that the harder negatives synthesised in this way degrade the quality of report generation. Our explanation is that this way does not maintain the nature of the synthesised harder negative, i.e., "being a report." To handle this issue, we constrain the synthesised hard negative on the manifold of "report" in the feature space. Since the current  $\tilde{\mathbf{u}}_i^-$  is the hard negative of  $\mathbf{u}_i$ , they are close in the feature space. Therefore, the line segment between the current  $\tilde{\mathbf{u}}_i^-$  and  $\mathbf{u}_i$ , denoted by  $\tilde{\mathbf{u}}_i^- \mathbf{u}_i$ , will be a good local approximation to the "report" manifold near  $\tilde{\mathbf{u}}_i^-$ . Therefore, we impose the constrain that the synthesised harder negative must reside on the line  $\tilde{\mathbf{u}}_i^- \mathbf{u}_i$ . In the presence of such a constraint, we can utilise projected gradient descent to obtain the synthesised harder negative as

$$\begin{aligned} \tilde{\mathbf{u}}_i^- &\triangleq \mathcal{P}(\tilde{\mathbf{u}}_i^- + \frac{1}{2}\lambda \frac{\partial \mathcal{L}(\tilde{\mathbf{u}}_i^-)}{\partial \tilde{\mathbf{u}}_i^-}) \\ &= \mathcal{P}(\tilde{\mathbf{u}}_i^- + \lambda(\mathbf{z}_i - \tilde{\mathbf{u}}_i^-)) = \mathcal{P}((1-\lambda)\tilde{\mathbf{u}}_i^- + \lambda\mathbf{z}_i), \end{aligned} \quad (5)$$

where  $\mathcal{P}(\mathbf{x})$  denotes the operator that projects a point  $\mathbf{x}$  in the feature space onto the line  $\tilde{\mathbf{u}}_i^- \mathbf{u}_i$ . It is not difficult to derive the projection of  $\mathbf{z}_i$  onto this line as

$$\mathbf{p}_i = \tilde{\mathbf{u}}_i^- + \frac{(\mathbf{z}_i - \tilde{\mathbf{u}}_i^-) \cdot (\mathbf{u}_i - \tilde{\mathbf{u}}_i^-)}{\|\mathbf{u}_i - \tilde{\mathbf{u}}_i^-\|^2} \cdot (\mathbf{u}_i - \tilde{\mathbf{u}}_i^-). \quad (6)$$

Formally, the synthesised harder negative can be expressed as

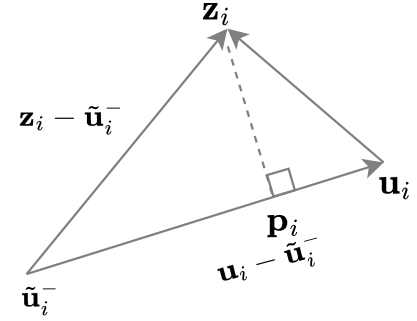
$$\tilde{\mathbf{u}}_i^- \triangleq (1-\lambda)\tilde{\mathbf{u}}_i^- + \lambda\mathbf{p}_i. \quad (7)$$

The above process is illustrated in Fig. 2.

Now we can contrast Eq.(7) with the MoChi-inspired strategy in Eq.(2). It can be seen that both strategies are based on linear combination. Their first terms in the right hand side share the similar nature in that either  $\tilde{\mathbf{u}}_i^-$  or  $\mathbf{u}_{r_i}^-$  is a sample close (as hard negative or nearest neighbour) to  $\mathbf{u}_i$ . Meanwhile, through the projection  $\mathbf{p}_i$ , Eq.(7) additionally considers the image modality  $\mathbf{z}_i$ , which is missing in the MoChi-inspired one. Finally, it is worth noting that in real implementation, we empirically set  $\tilde{\mathbf{u}}_i^-$  in the right hand side of Eq.(7) as the prior hard negative  $\mathbf{u}_i^-$  instead of the  $\tilde{\mathbf{u}}_i^-$  synthesised at the last iteration to avoid being too aggressive. In this setting, the improvement of our strategy to the MoChi-inspired one will more directly reflect the benefit of considering both image and text modalities when synthesising harder negatives.

### 3.5 Training Process

The above process forms an updated triplet  $(\mathbf{z}_i, \mathbf{u}_i, \tilde{\mathbf{u}}_i^-)$ . To capture finer visual findings, highly correlated image and report features are



**Figure 2.** Illustration of our optimisation-based strategy to synthesize a harder negative report.  $\mathbf{z}_i$  denotes the feature of image  $I_i$ ,  $\mathbf{u}_i$  denotes the feature of the paired report  $R_i$ ,  $\tilde{\mathbf{u}}_i^-$  is the synthesised harder negative in the last iteration.  $\mathbf{p}_i$  is the projection of  $\mathbf{z}_i$  onto the line  $\tilde{\mathbf{u}}_i^- \mathbf{u}_i$ . The synthesised harder negative is a linear combination of  $\tilde{\mathbf{u}}_i^-$  and  $\mathbf{p}_i$ , as shown in Eq.(7).

desired. We explicitly align image ( $\mathbf{z}_i$ ) and the truly matched report ( $\mathbf{u}_i$ ) to be close and push the synthesised harder negative report ( $\tilde{\mathbf{u}}_i^-$ ) away from the image. To achieve stronger alignment, we use multiple negatives and therefore the contrastive loss, CLIP [26], which has a similar spirit to Eq. (4) in terms of aligning matched pairs together and pushing unmatched ones away from each other. The CLIP loss for a batch  $B$  containing paired sample features  $\mathbf{a}$  and  $\mathbf{b}$  is

$$L_{con} = -\log \sum_{i=1}^B \frac{\exp\{\text{sim}(\mathbf{a}_i, \mathbf{b}_i)/\tau\}}{\sum_{k=1; [k \neq i]}^B \exp\{\text{sim}(\mathbf{a}_i, \mathbf{b}_k)/\tau\}} \quad (8)$$

The loss depends on the similarity matrix of sample features in a batch. The diagonal elements of the matrix are considered to be paired and pulled close to each other, while the off-diagonal elements are unpaired samples and pushed away.

However, in the CLIP loss, negative samples are usually randomly sampled from a batch of paired samples during training. Due to this, all the negatives within the batch might not really be hard negatives for each positive sample. To address this situation, we propose to compute an additional similarity matrix between image features and the synthesised harder negative report features and then average the two similarity matrices to obtain a final similarity matrix to be used in our CLIP loss.

Specifically, we first calculate a cosine similarity matrix for samples  $(\mathbf{z}_i, \mathbf{u}_j)$ , where  $i, j = 1, \dots, B$  as

$$\mathbf{S}_1(\mathbf{z}_i, \mathbf{u}_j) = \frac{\mathbf{z}_i \cdot \mathbf{u}_j}{\|\mathbf{z}_i\| \cdot \|\mathbf{u}_j\|}. \quad (9)$$

A similar matrix  $\mathbf{S}_2$  can be obtained in the same way for samples  $(\mathbf{z}_i^-, \mathbf{u}_j^-)$ . After this, we calculate one more similarity matrix with synthesised harder negative  $(\mathbf{z}_i, \tilde{\mathbf{u}}_j^-)$ , where  $i, j = 1, \dots, B$  as

$$\mathbf{S}_3(\mathbf{z}_i, \tilde{\mathbf{u}}_j^-) = \frac{\mathbf{z}_i \cdot \tilde{\mathbf{u}}_j^-}{\|\mathbf{z}_i\| \cdot \|\tilde{\mathbf{u}}_j^-\|} \quad (10)$$

Meanwhile, since  $\tilde{\mathbf{u}}_i^-$  is a harder negative to  $\mathbf{z}_i$ , we need to push the diagonal elements of this matrix to be small as the off-diagonal ones do. To implement this, we multiply each of the diagonal elements of this matrix with  $-1$  and denote the result matrix as  $\mathbf{S}_3'$ .

Our similarity matrix is therefore

$$\mathbf{S} = \frac{1}{3}(\mathbf{S}_1 + \mathbf{S}_2 + \mathbf{S}_3'). \quad (11)$$



and it defines the  $\text{sim}(\cdot, \cdot)$  in Eq. (8). The CLIP loss in our case is

$$\mathcal{L}_{CP} = \frac{1}{2} \left( \mathcal{L}_{\text{con}}(\text{sim}(\cdot, \cdot)) + \mathcal{L}_{\text{con}}(\text{sim}^\top(\cdot, \cdot)) \right). \quad (12)$$

For report generation, we use the standard cross-entropy loss  $\mathcal{L}_{CE}$ . By also considering the loss  $\mathcal{L}_{CS}$  in Eq. (1), our final learning objective is defined as

$$\mathcal{L}_F = \mathcal{L}_{CE} + \beta \mathcal{L}_{CP} + \gamma \mathcal{L}_{CS}. \quad (13)$$

The hyper-parameters  $\beta$  and  $\gamma$  balance different objective terms.

## 4 Experimental Results

### 4.1 Datasets

We evaluate our framework on two radiology benchmarks IU-XRay [6] and MIMIC-CXR [12] for radiology report generation. In addition, we show the generality of our framework on the image captioning benchmark COCO [20].

**IU-XRay** [6] 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 impressions, 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 [18, 10, 5, 4].

**MIMIC-CXR** [12] is the large-scale radiology dataset having 377,110 images with 227,835 reports from 64,588 patients. We use the official data split with 368,960 training samples, 2,991 validation samples, and 5,159 test samples.

**COCO** [20] is the most widely used and standard dataset for image captioning. It comprises 120,000 images, each with 5 different captions. We used the split provided by Karapathy [14], where 5,000 images are used for validation, 5,000 images for testing, and the rest of the images for training. We download the split COCO dataset from the GitHub repository of the meshed-memory-transformer<sup>1</sup>.

For IU-XRay, samples without complete findings sections are removed, following [18]. The filtered images and reports for IU-XRay and MIMIC-CXR are publicly available in this repository<sup>2</sup> by [5], and they are directly downloaded and used in all our experiments.

### 4.2 Implementation Details

We test our framework with different encoder-decoder backbones, including R2GenCMN [4], which is our baseline model, and XproNet [33], a SOTA model for generating medical reports. Additionally, we integrate our framework into VisualGPT [3], a SOTA image captioning model. We use PyTorch to build our framework. Each input image is resized to  $3 \times 256 \times 256$  and later randomly cropped to the size of  $3 \times 224 \times 224$ . We do not combine the frontal and lateral views of the image and rather pass them separately. The baseline model is trained using Adam [15] optimizer with a weight decay of  $5 \times 10^{-5}$  for 50 epochs, and a batch size of 16. The hidden dimensions for the image and report features are 2048 and 512, respectively. To achieve alignment between the image and report features in a shared space, we employ a linear layer with a hidden size of 512, followed by an l2-normalization layer for both the image and report features. This ensures that both sets of features are projected into the common embedding space. The loss

weights in Eq.(13) are set as 0.1 and 0.2 for  $\beta$  and  $\gamma$ , respectively, 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. The value of  $\lambda$ , in Eq. (2) and Eq. (7) is equally set as  $1 - e^{-\alpha \cdot \text{epoch}}$ , where  $\alpha$  is a hyperparameter and  $\text{epoch}$  denotes the current epoch number. As the epoch number increases, the value of  $\lambda$  gradually increases from 0. Our experimental evaluations reveal that the value of  $\alpha = 0.01$  leads to superior performance, as shown in Table 1. When evaluating the existing SOTA models with or without adding our proposed framework, we use default hyper-parameters, optimizer, scheduler, and experimental settings of their methods for a fair comparison. We performed hyperparameter analysis only for  $\alpha$ ,  $\beta$ , and  $\gamma$  values. We have chosen the best values based on the higher BLEU-4 score on the validation set. We use one NVIDIA V100 32GB GPU for all our experiments. The training took nearly 3 hours to complete for IU X-Ray and 3 days for MIMIC-CXR.

**Evaluation Metrics:** Following the previous works in medical report generation and image captioning, we evaluate our model on widely used Natural Language Generation (NLG) metrics: BLEU [25], CIDER [31], METEOR [16] and ROUGE-L [19]. The evaluation metrics are readily available in pycocotools<sup>3</sup> package and we directly use it for our experiments. In addition, following the previous works [5, 4, 33], we also evaluate our framework on clinical efficacy (CE) metrics. To calculate CE metrics, the model trained on CheXpert [9] is used to label the generated reports and compare the findings to ground-truth reports in 14 distinct categories pertaining to thoracic disorders and support equipment. To assess model performance for CE metrics, we employ *precision*, *recall*, and *F1*. Furthermore, to measure the uniqueness of the generated reports, we use a metric called *diversity*, proposed by a recent work SimCTG [30]

### 4.3 Results

We first compare our model with the SOTA medical report generation models CO-ATT [11], CMAS-RL [10], HRGR [18], R2Gen [5], R2GenCMN [4], PPKED [21], KERP [17], XproNet [33], MedSkip [24], and CA [22]. 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[5], R2GenCMN [4], and XproNet [33], we run their codes with their default experimental settings. Moreover, to further verify if our framework also benefits generic image captioning tasks, we also incorporate our framework into the image captioning model VisualGPT [3] and test it on the image captioning benchmark COCO.

#### 4.3.1 Quantitative Analysis

Table 1 demonstrates the effectiveness of our approach compared with the SOTA methods. We call our approach ITHN, short for image-text hard negatives. All the results with ITHN in Table 1 are averages of three runs with three random seed values. After integrating our framework into the baseline and the SOTA models in radiology report generation and generic image captioning, we consistently observe a significant boost in the scores of evaluation. 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

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

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

<sup>3</sup> <https://github.com/tylin/coco-caption>

Dataset	Methods	B-1	B-2	B-3	B-4	R-L	M	C
IU-XRay[6]	HRGR*[18]	0.438	0.298	0.208	0.151	0.322	-	-
	CO-ATT*[11]	0.455	0.288	0.205	0.154	0.369	-	-
	CMAS-RL*[10]	0.464	0.301	0.210	0.154	0.362	-	-
	R2Gen***[5]	0.458	0.295	0.210	0.159	0.375	0.176	0.408
	MedSkip**[24]	0.467	0.297	0.214	0.162	0.355	0.187	-
	KERP**[17]	0.470	0.304	0.219	0.165	0.371	0.187	0.280
	PPKED**[21]	0.483	0.315	0.224	0.168	0.376	0.190	0.351
	CA**[22]	0.492	0.314	0.222	0.169	0.381	0.193	-
	R2GenCMN***[4]	0.467	0.300	0.210	0.156	0.369	0.185	0.401
	R2GenCMN + <b>MoChi</b>	0.489	0.310	0.211	0.151	0.374	0.190	0.440
	R2GenCMN + <b>ITHN (ours)</b>	<b>0.496</b>	0.319	0.220	0.160	0.378	0.194	0.461
	XproNet***[33]	0.468	0.299	0.205	0.158	0.367	0.181	0.375
MIMIC-CXR[12]	XproNet + <b>MoChi</b>	0.455	0.306	0.224	0.169	0.382	0.196	0.410
	XproNet + <b>ITHN (ours)</b>	0.491	<b>0.328</b>	<b>0.231</b>	<b>0.183</b>	<b>0.387</b>	<b>0.210</b>	<b>0.495</b>
	CA**[22]	0.350	0.219	0.148	0.106	0.278	0.142	-
	R2Gen***[5]	0.363	0.216	0.143	0.101	0.269	0.135	0.141
	PPKED**[21]	0.360	0.224	0.149	0.106	0.284	0.149	0.237
	R2GenCMN***[4]	0.369	0.223	0.148	0.105	0.270	0.136	0.143
	R2GenCMN + <b>MoChi</b>	0.364	0.225	0.152	0.109	0.276	0.139	0.237
	R2GenCMN + <b>ITHN (ours)</b>	<b>0.375</b>	0.226	<b>0.155</b>	0.112	0.284	0.142	0.300
COCO[20]	XproNet[33]	0.344	0.215	0.146	0.105	0.279	0.138	0.359
	XproNet + <b>MoChi</b>	0.354	0.220	0.150	0.112	0.283	0.138	0.371
	XproNet + <b>ITHN (ours)</b>	0.370	<b>0.230</b>	0.154	<b>0.114</b>	<b>0.294</b>	<b>0.145</b>	<b>0.387</b>
	VisualGPT***[3]	0.677	-	-	0.236	0.486	0.221	0.768
	VisualGPT + <b>MoChi</b>	0.694	-	-	0.250	0.493	0.226	0.830
	VisualGPT + <b>ITHN (ours)</b>	<b>0.701</b>	-	-	<b>0.259</b>	<b>0.516</b>	<b>0.235</b>	<b>0.852</b>

**Table 1.** Comparison of our approach and SOTA models on IU-XRay, MIMIC-CXR, and COCO datasets. \* indicates the results quoted from [5] 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. The metrics B-k (k = 1 to 4) is for BLEU, R-L for ROUGE-L, M for METEOR, and C for CIDER.

Model	P	R	F1	D
R2Gen[5]	0.288	0.215	0.246	0.50
R2GenCMN[4]	0.295	0.221	0.252	0.55
XproNet[33]	0.302	0.222	0.255	0.64
R2GenCMN + <b>MoChi</b>	0.321	0.252	0.282	0.63
R2GenCMN + <b>ITHN</b>	0.355	0.281	0.313	0.69
XproNet + <b>MoChi</b>	0.333	0.265	0.295	0.71
XproNet + <b>ITHN</b>	<b>0.362</b>	<b>0.283</b>	<b>0.317</b>	<b>0.82</b>

**Table 2.** Comparison of clinical metrics and diversity scores on MIMIC-CXR. P is precision, R is recall, F1 is F1 score, and D is diversity score.

the literature, and could further benefit the latter. On comparing the results of our approach ITHN with that of MoChi, we could observe an average relative increase of +6.6% in BLUE, +7.1% in METEOR, and +20% in CIDER for the best performing model XproNet on IU-XRay dataset. These scores indicate that utilising both image and text modalities is an effective approach to synthesising harder negatives, as opposed to relying solely on one modality as in the case of MoChi. Moreover, mining the hard negatives in our way could also enhance the standard image captioning on the COCO dataset. Our approach significantly boosts the performance of the 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. Moreover, we observe consistent improvement over SOTA models even in clinical and diversity metrics shown in Table 2.

#### 4.3.2 Qualitative Analysis

In Fig. 3, we compare the reports generated by the SOTA models with and without using our learning framework. The generated sentences are differently colored, with green indicating semantically correct generation, red indicating incorrect generation, and blue indicating the text present in the ground-truth but not generated by any models under comparison. As can be seen, comparing the reports generated by the baseline R2GenCMN and XproNet, incorporating our framework could reduce the incorrect sentences generated in the reports. For instance, in the example from row 3 of Fig. 3, the SOTA models trained with our framework can identify and generate the sentences - "hilar contours are stable" and "lungs are hyperinflated", which are missing in the generated reports without using our framework. This is consistent with the quantitative analysis.

#### 4.4 Ablation Study

To understand the contribution of each component of our framework, we conduct an ablation study on three evaluation metrics - BLEU-4, ROUGE-L, and CIDER on IU-XRay. We use R2GenCMN as our baseline. The results are shown in Table 3 (left). Firstly, to show the impact of contrastive loss on model performance, we add  $\mathcal{L}_{con}$  to XproNet. To show the importance of increasing the hardness of reports during training, we compare the results of "XproNet w/o harder negative synthesis" with "XproNet + **MoChi**", and "XproNet + **ITHN**". We observe that synthesising harder negatives can improve the results by +14.3% on BLEU, +4.3% on Rouge-L, and +24.3%

Model (tested on IU X-Ray)	B-4	R-L	C	Dataset	$\alpha$	B-4	R-L	C
XproNet	0.158	0.367	0.375	IU-XRay	0.1	0.182	0.384	0.488
XproNet + $\mathcal{L}_{con}$	0.160	0.371	0.398		0.01	<b>0.183</b>	<b>0.387</b>	<b>0.495</b>
XproNet w/o harder negative synthesis	0.165	0.375	0.426		0.001	0.177	0.379	0.482
XproNet + <b>MoChi</b>	0.169	0.382	0.410	COCO	0.1	<b>0.259</b>	<b>0.516</b>	<b>0.852</b>
XproNet + <b>ITHN</b>	<b>0.183</b>	<b>0.387</b>	<b>0.495</b>		0.01	0.256	0.499	0.849
XproNet + <b>ITHN</b> w/o $\mathcal{L}_{CS}$	0.178	0.384	0.477		0.001	0.252	0.492	0.838

**Table 3.** Ablation studies on component contributions (left) and the impact of hyper-parameter  $\alpha$ . B is for BLEU, R-L for ROUGE-L, and C for CIDER.



**Figure 3.** Examples of reports generated by various models. From top to bottom, images are taken from the IU-XRay. The green colour represents the text semantically similar to the ground-truth. The red colour represents the text incorrectly generated. The blue colour in the ground-truth report represents the text not generated by any of the models under comparison.

on CIDER. These improvements indicate that the model learns more fine-grained characteristics between the image and report by making the prior hard negative reports much harder during training. Finally, after removing the cosine-similarity loss  $\mathcal{L}_{CS}$  in Eq. (13), which supervises the encoder to capture distinctive image features, we could see a drop in performance. This supports our motivation of learning distinctive features from both modalities to achieve better performance. Moreover, to show the impact of  $\alpha$  which controls the rate of hardness increase of the synthesised harder negative features computed using Eq. (7), we investigate the performance of the model with different values of  $\alpha$  as in Table 3 (right). The optimal value of  $\alpha$  depends on the nature of the dataset. The higher the  $\alpha$  value, the faster the rate of increase in the hardness of negative reports will be. For medical datasets such as IU-Xray and MIMIC-CXR, where most medical images have reports of similar nature, having a lower  $\alpha$  value gives better performance. On the other hand, for a diverse dataset such as COCO, a slightly higher  $\alpha$  value gives better results.

## 5 Conclusion

This paper proposes a framework to improve medical report generation by learning significant and unique features for images and reports. We demonstrate that our framework can readily work 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 framework although this is not the focus of this paper. Meanwhile, in terms of potential risk, the medical reports generated by current AI have not achieved human-level performance and therefore shall not be used for practical diagnosis at this stage.

## References

- [1] 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).
- [2] 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).
- [3] 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).
- [4] 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.
- [5] 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.
  - [6] 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*, 304–10, (2016).
  - [7] 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).
  - [8] 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).
  - [9] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Illcus, Chris Chute, Henrik Marklund, Behzad Haghighi, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng, ‘Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison’, in *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pp. 590–597. AAAI Press, (2019).
  - [10] 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.
  - [11] 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.
  - [12] 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*, (2019).
  - [13] 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).
  - [14] 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).
  - [15] Diederick P Kingma and Jimmy Ba, ‘Adam: A method for stochastic optimization’, in *International Conference on Learning Representations (ICLR)*, (2015).
  - [16] 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.
  - [17] 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).
  - [18] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing, ‘Hybrid retrieval-generation reinforced agent for medical image report generation’, in *Advances in Neural Information Processing Systems*, eds., S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, volume 31. Curran Associates, Inc., (2018).
  - [19] 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.
  - [20] 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 *Computer Vision – ECCV 2014*, eds., David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, pp. 740–755, Cham, (2014). Springer International Publishing.
  - [21] 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 2021).
  - [22] 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 2021). Association for Computational Linguistics.
  - [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).
  - [24] 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).
  - [25] 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.
  - [26] 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 *Proceedings of the 38th International Conference on Machine Learning*, eds., Marina Meila and Tong Zhang, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, (18–24 Jul 2021).
  - [27] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, ‘Language models are unsupervised multitask learners’, (2019).
  - [28] 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*, arXiv:1711.05225, (November 2017).
  - [29] 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).
  - [30] Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier, ‘A contrastive framework for neural text generation, 2022.
  - [31] 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).
  - [32] 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)*, 3156–3164, (2015).
  - [33] Jun Wang, Abhir Bhalerao, and Yulan He, ‘Cross-modal prototype driven network for radiology report generation’, in *ECCV*, (2022).
  - [34] 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, (2021).
  - [35] 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 *Proceedings of the 32nd International Conference on Machine Learning*, eds., Francis Bach and David Blei, volume 37 of *Proceedings of Machine Learning Research*, pp. 2048–2057, Lille, France, (07–09 Jul 2015). PMLR.
  - [36] An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu, ‘Radbert: Adapting transformer-based language models to radiology’, *Radiology: Artificial Intelligence*, **4**(4), e210258, (2022).