Improving Radiology Summarization with Radiograph and Anatomy Prompts

Jinpeng Hu♡, Zhihong Chen♡, Yang Liu♡
Xiang Wan♢♦†, Tsung-Hui Chang♡†

♡Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, Guangdong, China
♢Pazhou Lab, Guangzhou, 510330, China

{jinpenghu, zhihongchen, yangliu5}@link.cuhk.edu.cn
wanxiang@sribd.cn changtsunghui@cuhk.edu.cn

Abstract

The impression is crucial for the referring physicians to grasp key information since it is concluded from the findings and reasoning of radiologists. To alleviate the workload of radiologists and reduce repetitive human labor in impression writing, many researchers have focused on automatic impression generation. However, recent works on this task mainly summarize the corresponding findings and pay less attention to the radiology images. In clinical, radiographs can provide more detailed valuable observations to enhance radiologists’ impression writing, especially for complicated cases. Besides, each sentence in findings usually focuses on single anatomy, such that they only need to be matched to corresponding anatomical regions instead of the whole image, which is beneficial for textual and visual features alignment. Therefore, we propose a novel anatomy-enhanced multimodal model to improve impression generation. In detail, we first construct a set of rules to extract anatomies and put these prompts into each sentence to highlight anatomy characteristics. Then, two separate encoders are applied to extract features from the radiograph and findings. Afterward, we apply a contrastive learning module to align these two representations at the overall level and use a co-attention to fuse them at the sentence level with the help of anatomy-enhanced sentence representation. The experimental results on two benchmark datasets confirm the effectiveness of the proposed method, which achieves state-of-the-art results.

1 Introduction

A radiology report of an examination is used to describe normal and abnormal conditions with one medical image and two important text sections: findings and impression. The findings section is a free-text description of a clinical radiograph (e.g., chest X-ray), providing the medical image’s detailed observations. Meanwhile, the impression is a more concise statement about critical observations summarized from the findings, images and the inference from radiologists and provides some clinical suggestions, such that in practice, clinicians prefer to read the impression to locate the prominent observations and evaluate their differential diagnoses. However, writing impressions is time-consuming and in high demand, which draws many researchers to focus on automatic impression generation (AIG) to alleviate the workload of radiologists (Gharebagh et al., 2020; Hu et al., 2021; Zhang et al., 2018, 2020c; Hu et al., 2022a; MacAvaney et al., 2019).

For example, (Gharebagh et al., 2020; Hu et al., 2021; Karn et al., 2022) propose to extract medical ontologies and entities from findings and then utilize graph neural networks (GNNs), dual encoder, or reinforcement learning to integrate this knowledge into general sequence-to-sequence models for promoting AIG. Yet, most existing studies mainly focus on fully using findings to produce impressions and pay rare attention to medical radiography. Owing to the fact that some diseases tend to have similar observations, they are difficult to get a clear diagnosis only depending on the textual statements. In this situation, most radiologists usually consider both the image and findings to make a more ac-
Figure 2: The overall architecture of our proposed model. The green box is used to provide sentence anatomy prompts. Besides, aligned contrastive learning and sentence-level co-attention fusion modules are shown in the purple and red boxes. 1, 2, 3 indicate different pairs (i.e., image and its corresponding findings).

Curate clinical suggestion in impressions. Besides, many approaches have been proposed for radiology report generation and have achieved considerable success (Chen et al., 2021; Zhang et al., 2020a), whose goal is to generate the findings based on a given medical image, further showing the value of knowledge in the medical image. In radiology reports, each findings can be regarded as a text representation of the corresponding medical image, and meanwhile, each image is a visual representation of the findings such that these two modal data can be effectively aligned.

Therefore, we propose a task that integrates the images and anatomy-enhanced findings for impression generation. According to communication with radiologists, each sentence in the findings focuses on single anatomy, so the sentence-level representation should be easier to align to a certain anatomical region of the image. To enhance such a process, we first construct some rules under the guidance of radiologists and utilize these rules to extract the main anatomies from each sentence. Then we put these anatomies at the beginning of the sentence to emphasize anatomy information. Next, we use a visual extractor to extract visual features from the radiology image and apply a Transformer-based text encoder to embed the corresponding findings. Afterward, an extra encoder is used to further model visual features, whose output will be aligned to the textual representation at the document level by a contrastive learning module. Finally, we employ a co-attention to integrate the visual and text features at the sentence level to obtain the final fused representation, which is then input to the decoder to generate the impressions. Experimental results on two benchmark datasets, MIMIC-CXR and OpenI, demonstrate the effectiveness of our proposed model, which achieves better performance than most existing studies. Furthermore, analysis of impression length shows that our proposed multimodal model is better at long impression generation, where our model obtains significant improvements when the impression is longer than 20.

2 Method

We follow existing studies on report generation (Chen et al., 2020; Zhou et al., 2021) and impression generation (Zhang et al., 2018; Gharabaghi et al., 2020; Hu et al., 2021) and utilize the standard sequence-to-sequence paradigm for this task. In doing so, we regard patch features extracted from radiology image $X_I$ as one of the source inputs. In addition, the other input is the findings sequence $X_F = s_1, s_2, \ldots, s_M$, where $M$ is the number of sentence and $s_i = [CLS]_i, x_{i,1}, x_{i,2}, \ldots, x_{i,N_i}, [SEP]_i$ with external [CLS] token. The goal is to utilize $X_T$ and $X_F$ to find a target impression $Y = [y_1, \ldots, y_i, \ldots, y_L]$ that summarizes the most critical observations, where $L$ is the number of tokens and $y_i \in V$ is the generated token and $V$ is the vocabulary of all possible tokens. The impression generation process
We employ a pre-trained convolutional neural network (e.g., ResNet (He et al., 2016)) to extract features from $X_I$. We follow Chen et al. (2020) to decompose the image into multiple regions with equal size and then expand these patch features into a sequence:

$$[\text{im}_1, \text{im}_2, \cdots, \text{im}_p] = f_{ve}(X_I), \tag{3}$$

where $f_{ve}$ refers to the visual extractor and $\text{im}_i$ is the patch feature.

### 2.2 Sentence Anatomy Prompts

It is known that each sentence in findings usually focuses on describing observations in single anatomies, such as lung, heart, etc., instead of stating multiple anatomy observations in one sentence. This might be because many radiologists usually draw on radiology report templates when writing findings, and most templates follow this characteristic, which describes medical observations anatomy by anatomy. For example, radiology report templates in the radreport website\(^1\) mainly divide the radiology findings into six sections: Lungs, Pleural Spaces, Heart, Mediastinum, Osseous Structures, and Additional Findings, respectively. Motivated by this, we manually construct a rule lexicon under the guidance of radiologists to extract anatomy information from the sentence, with the details shown in Table 1. After that, we use the following ways to deal with different types of sentences:

- **Type I**: For the sentence that only describes observation in single anatomy, we assign the sentence to the corresponding anatomy type. For example, the sentence “The lungs are hyperexpanded and mild interstitial opacities” only contains one anatomy (i.e., lungs), and thus, we assign type **lungs** to this sentence.
- **Type II**: Although most sentences focus on single anatomy, there are still some with multiple anatomies. For these sentences, we follow the priority ranking from normal observations to comparisons, as shown in Table 1. For instance, although both lung and pleural spaces are in the sentence “lungs are grossly clear, and there are no pleural effusions”, we distribute this sentence into type **normal observations**.
- **Type III**: For the remaining sentences, we use a particular type **other observations** to mark.

Next, we plan anatomy type into the corresponding sentence and modify the original sentence as “anatomy: sentence”. For instance, the type **lungs** is inserted into “The lungs are hyperexpanded and mild interstitial opacities” as “lungs: The lungs are hyperexpanded and mild interstitial opacities”. In this way, the original findings $X_F$ is updated as an anatomy-enhanced one $X'_F$.

### 2.3 Text Encoder

Pre-trained language models have achieved great success in many NLP tasks (Hu et al., 2022b,c; Zhong and Chen, 2021; Xu et al., 2021b; Fang et al., 2023a,b; Hu et al., 2023). Therefore, we

---

\(^1\)https://radreport.org/
employ a pre-trained model BioBERT (Lee et al., 2020) as our text encoder to extract features from the findings:

$$[h_1, h_2, \cdots, h_n] = f_{te}(\chi^t),$$

where $f_{te}(\cdot)$ refers to the text encoder, and $h_i$ is a high dimensional vector for representing tokens $x_i$. We regard the representation of $[CLS]_i$ in $s_i$ (i.e., $h_{CLS,i}$) as the $i$th sentence representation.

### 2.4 Document-level Cross-Modal Alignment

In radiology reports, findings and radiology images usually describe the same medical observations by using different media (i.e., vision and text, respectively). To pull the image representation close to the output of the text encoder, we first utilize an extra Transformer encoder to further model the visual features $\chi^t_x$, computed by:

$$[c_1, c_2, \cdots, c_p] = f_{te}(im).$$

Herein the outputs are the hidden states $c_i$ encoded from the input visual features in subsection 2.1 and $f_{te}$ refers to the Transformer image encoder. Afterward, we use the mean pooling to obtain the overall representation with respect to the findings and the corresponding image, formalized as:

$$z_f = Mean(c_1, c_2, \cdots, c_p),$$

$$z_F = Mean(h_{CLS,1}, h_{CLS,2}, \cdots, h_{CLS,n}).$$

Owing to the characteristic of the radiology report, $z_f$ and $z_F$ should be close to each other if the image and findings are from the same examination. On the contrary, radiology images and reports from different tests tend to have distinct medical observations and further should be different from each other. Therefore, we introduce a contrastive learning module to map positive samples closer and push apart negative ones, where the positive indicates that $z_f$ and $z_F$ are from the same pair (i.e., the same examination) and the negative refers to the samples from different pairs. For example, we assume there are two tests, $\{findings_1, images_1\}$ and $\{findings_2, image_2\}$, and thus, in this case, for $findings_1$, the $image_2^*$ is a positive sample while the $image_1^*$ is a negative instance. We follow Gao et al. (2021) to compute the cosine similarity between the original representation and its positive and negative examples. Then, for a batch of $2Q$ examples $z \in \{z_f \cup z_F\}$, we compute the contrastive loss for each $z_m$ as:

$$L^\text{con}_m = -\log \frac{e^{\text{sim}(z_m, z_m^+)/\tau}}{\sum_{z \in \{z\}} (e^{\text{sim}(z, z^+)/\tau})},$$

where $\text{sim}(\cdot, \cdot)$ is the cosine similarity, and $\tau$ is a temperature hyperparameter. The total contrastive loss is the mean loss of all examples:

$$L^\text{con}_m = \frac{1}{2Q} \sum_{m=1}^{2Q} L^\text{con}_m.$$ (8)

### 2.5 Sentence-Level Co-Attention Fusion

As mentioned in subsection 2.2, each sentence in the findings usually focuses on single anatomy, meaning that sentence-level textual information can be mapped to corresponding anatomy regions in images. Therefore, we propose to utilize the anatomy-enhanced sentence representation to align with the image. In detail, as introduced in 2.3, we extract anatomy-enhanced sentence representations from the text encoder $h_{CLS} = [h_{CLS,1}, h_{CLS,2}, \cdots, h_{CLS,n}]$, which are then used to perform co-attention to fuse two modal knowledge. We first treat $h_{CLS}$ as query and the corresponding image representations $c$ as key and value matrix and compute the attention weight with the softmax function:

$$a^b_i = \text{Softmax}(h_{CLS}, c^T),$$

where $a^b_i$ can be viewed as a probability distribution over the image features, which is then used to compute a weighted sum:

$$c^b_i = \sum_k a^b_i,k c_k.$$ (10)

Afterward, on the contrary, $c$ is regarded as the key and value matrix, and $h_{CLS}$ is represented as the query. We then adopt a similar method to obtain another fusion representation:

$$h^r_i = \sum_k a^r_i,k h_{CLS,k}, a^r_i = \text{Softmax}(c_i h_{CLS}).$$ (11)

After that, we obtain the updated image and sentence representation by adding the fusion vectors to the original ones:

$$c = c + c^b_i, h_{CLS} = h_{CLS} + h^r.$$ (12)

### 2.6 Decoder

The backbone decoder in our model is the one from the standard Transformer, where $e = [c, h_{CLS}, h]$ is functionalized as the input of the decoder so as to improve the decoding process. Then, the decoding process at time step $t$ can be formulated as the function of a combination of previous output (i.e., $y_1, \cdots, y_{t-1}$) and the feature input (i.e., $e$):

$$y_t = f_{de}(e, y_1, \cdots, y_{t-1}),$$ (13)
where $f_{de}(\cdot)$ refers to the Transformer-based decoder, and this process will generate a complete impression. We define the final loss function as the linear combination of impression generation loss and contrastive objectives:

$$L = L_{\text{generator}} + \lambda L_{\text{con}},$$

(14)

where $\lambda$ is the tuned hyper-parameter controlling the weight of the contrastive loss.

### 3 Experimental Setting

#### 3.1 Dataset

Our experiments are conducted on two benchmark datasets: OpenI (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019), respectively, which are described as follows:

- **OpenI**: it is a public dataset containing 7,470 chest X-ray images and 3,955 corresponding reports collected by Indiana University.
- **MIMIC-CXR**: it is a large-scale radiography dataset with 473,057 chest X-ray images and 206,563 report.

We follow Hu et al. (2021) to remove the following cases: (a) incomplete reports without findings or impressions; (b) reports whose findings have fewer than ten words or impression has fewer than two words. Besides, since some reports have multiple radiology images from different views, such as posteroanterior, anteroposterior and lateral, we only select one image from posteroanterior or anteroposterior. As for partition, we follow Chen et al. (2020) to split OpenI and MIMIC-CXR, where the former is split as 70%/10%/20% for train/validation/test, and the latter follows its official split.

#### 3.2 Baseline and Evaluation Metrics

To illustrate the validity of our proposed model, we use the following models as our main baselines:

- **BASE-FINDINGS** and **BASE-IMAGE**: They are unimodal models, where the former utilizes a pre-trained text encoder and a randomly initialized Transformer-based decoder, and the latter replaces the text encoder with image encoders.
- **BASE**: This is the base backbone multimodal summarization model with pre-trained image and text encoders and a Transformer-based decoder, which utilizes both findings and images to generate impressions.
- **BASE+DCA** and **BASE+AP**: They are the multimodal summarization models. The former utilizes document-level representations to align findings and images, and the latter utilizes the rules to enhance anatomy prompts for each sentence.

We follow Zhang et al. (2020c) to utilize summarization and factual consistency (FC) metrics to examine the model performance. Specially, we use ROUGE (Lin, 2004) and report $F_1$ scores of ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L) for summarization metrics. Meanwhile, a pre-trained CheXbert (Smit et al., 2020) is used to recognize 14 types of observation from reference and generated impression, respectively, whose detected results are used to calculate the precision.
Table 3: Comparisons of our proposed models with the previous studies on the test sets of OPENI and MIMIC-CXR with respect to the ROUGE metric. CHESTXRAYBERT is regarded as a weak reference since their data processing method was not public.

<table>
<thead>
<tr>
<th>Model</th>
<th>OPENI</th>
<th>MIMIC-CXR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R-1</td>
<td>R-2</td>
</tr>
<tr>
<td>R2GEN (Chen et al., 2020)</td>
<td>50.68</td>
<td>38.02</td>
</tr>
<tr>
<td>R2GENCMN (Chen et al., 2021)</td>
<td>51.30</td>
<td>34.35</td>
</tr>
<tr>
<td>TRANSABS (Liu and Lapata, 2019)</td>
<td>62.90</td>
<td>53.51</td>
</tr>
<tr>
<td>CHESTXRAYBERT (Cai et al., 2021)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>WGSUM (Hu et al., 2021)</td>
<td>63.90</td>
<td>54.49</td>
</tr>
<tr>
<td>AIG_CL (Hu et al., 2022a)</td>
<td>64.97</td>
<td>54.26</td>
</tr>
<tr>
<td>CLIPABS (Radford et al., 2021)</td>
<td>53.13</td>
<td>39.69</td>
</tr>
<tr>
<td>OURS</td>
<td><strong>68.00</strong></td>
<td><strong>59.89</strong></td>
</tr>
</tbody>
</table>

recall, and F1 score for measuring FC.

3.3 Implementation Details

In our experiments, we select biobert-base-cased-v1.1⁴ as our text encoder and follow its default model settings which are 12 layers of self-attention with 768-dimensional embeddings. Besides, for the visual extractor, we select the ResNet101 pre-trained on the ImageNet to extract patch features with the dimension 2048. For the Transformer image encoder, we use a 6-layer Transformer with 768 hidden sizes and 2048 feed-forward filter sizes. The decoder has a similar structure: 6-layer Transformer with 768 dimensions, 8 attention heads, and 2048 feed-forward filter sizes. As for training, we use Adam (Kingma and Ba, 2014) to optimize the trainable parameters in our model.

4 Experimental Results

4.1 Overall Results

To explore the effect of integrating image and text to generate impressions, we compare our model to corresponding single modal summarization baselines in Table 2. We can observe that compared to BASE-FINDINGS and BASE-IMAGE, all other models (except BASE) obtain better results with respect to ROUGE scores, which shows the value of multimodal information fusion. The main reason might be that findings can provide key and accurate information, and the image can present detailed and rich features, such that these two different types of features can complement each other to enhance impression generation. Besides, BASE-FINDINGS outperforms BASE-IMAGE, illustrating that textual features are more valuable than visual ones because the gap between two related texts is smaller than that between vision and text.

Moreover, we conduct experiments on the different models, and the results are reported in Table 2 where BASE+AP+DCA indicates our full model. There are several observations drawn from different aspects. First, the comparisons between BASE+DCA, BASE+AP, and BASE illustrate the effectiveness of each component in our proposed model (i.e., contrastive learning and lexicon matching). Second, our full model (i.e., BASE+AP+DCA) achieves the best results among these baselines, which confirms the validity of our design that combines contrastive learning and anatomy information planning. Contrastive learning can map the image closer to the corresponding findings if they are in the same pair and push them apart if they are not, which can effectively align these two modalities at the document level. For another, highlighting anatomy characteristics can potentially help the model align the sentence feature to the corresponding organ or body part position in the images, further improving feature fusion between different modalities. Third, in terms of FC metrics on the MIMIC-CXR dataset, our proposed model outperforms all baselines and achieves higher F1 scores, indicating that our model is able to generate more accurate impressions. This is because our model can enhance feature matching between findings and images to facilitate critical information extraction, contributing to better impression generation with the help of such information.

⁴https://github.com/dmis-lab/biobert
Comparison Metric Win Tie Lose

Ours vs. Base
READ 8% 88% 4%
ACC. 25% 58% 17%
COMP. 13% 80% 7%

Ours vs. Ref
READ 4% 77% 9%
ACC. 12% 70% 18%
COMP. 5% 85% 10%

Table 4: Results of the human evaluation. The top three give results for comparison between BASE+AP+DCA and BASE. The bottom three are results for BASE+AP+DCA versus the reference impressions.

4.2 Comparison with Previous Studies

We further compare our model with existing methods, with the results reported in Table 3. We can observe that our model outperforms other methods, although those studies utilize complicated structures to enhance the generation, e.g., WGSUM utilizes a complicated graph structure, and R2GEN uses a recurrent relational memory. In addition, it is surprising that CLIPABs achieve worse performance than text-based models (i.e., TRANSABS, WGSUM and AIG_CL). This might be because CLIP pays more attention to the images and is less powerful in encoding text, while textual features are more important in this task.

4.3 Human Evaluation

We also conduct a human evaluation to evaluate the quality of the generated impressions with respect to three metrics: Readability, Accuracy, and Completeness (Gharebagh et al., 2020). In detail, we randomly select 100 chest X-ray images and their findings and impressions from the test set of MIMIC-CXR, as well as impressions generated from different models. Afterward, three experts who are familiar with radiology reports are invited to evaluate the generated impression with the results shown in Table 4. We can observe that our model is better than BASE, where more impressions from our model have higher quality than those from BASE, further confirming the effectiveness of our model. Meanwhile, when comparing our model against references, we find that although some cases are worse than ground truth (9%, 18%, and 10%), most of the impressions from our model are at least as good as the reference impressions.

Figure 3: R-1 score of generated impressions from different models, where OURS represents the BASE+AP+DCA. Note that when the word-based impression length is longer than 20, the \( p \)-value is less than 0.05.

5 Analyses

5.1 Impression Length

To test the effect of the length of impressions in AIG, we categorize the generated impressions on the MIMIC-CXR test set into several groups according to the length of reference impression, with the R-1 scores shown in Figure 3. Note that the average impression length for MIMIC-CXR is 17. We can observe that these models tend to have worse performance with increasing impression length, especially in the last group, where all obtain the worst R-1 scores. Our proposed model achieves more promising results in most groups, except the first group where the BASE-FINDINGS achieves the best results, which illustrates that our model is better at generating longer impressions. The main reason is that short impressions are usually normal observations without complicated abnormalities so that findings are enough to describe such information, and images may lead to some redundant noise due to their being too detailed. In contrast, for the long impression, detailed information can complement textual features to help the model accurately grasp complex observations.

5.2 Case Study

To further qualitatively investigate the effectiveness of our proposed model, we conduct a case study on the generated impressions from different models whose inputs are X-ray images and corresponding findings. The results are shown in Figure 4, and different colors represent the observations found in different locations. It is observed that OURS is able to produce better impressions than the BASE.
model, where impressions from our models can almost cover all the key points in these two examples with the help of the corresponding regions in images. On the contrary, the BASE model ignores some critical observations written in reference impressions, such as “right basilar loculated hydropneumothorax.” in the first example and “Stable mild cardiomegaly” in the second example, and even generates some unrelated information (e.g., “No pneumonia” in the second case).

6 Related Work

6.1 Multimodal Summarization

With the increase of multimedia data, multimodal summarization has recently become a hot topic, and many works have focused on this area, whose goal is to generate a summary from multimodal data, such as textual and visual (Zhu et al., 2018; Li et al., 2018; Zhu et al., 2020; Li et al., 2020; Im et al., 2021; Atri et al., 2021; Delbrouck et al., 2021). For example, Li et al. (2017) proposed to generate a textual summary from a set of asynchronous documents, images, audios and videos by a budgeted maximization of submodular functions.

6.2 Radiology report generation

Image captioning is a traditional task and has received extensive research interest (You et al., 2016; Aneja et al., 2018; Xu et al., 2021a). Radiology report generation can be treated as an extension of image captioning tasks to the medical domain, aiming to describe radiology images in the text (i.e., findings), and has achieved considerable improvements in recent years (Chen et al., 2020; Zhang et al., 2020a; Liu et al., 2019b, 2021b; Zhou et al., 2021; Boag et al., 2020; Pahwa et al., 2021; Jing et al., 2019; Zhang et al., 2020b; You et al., 2021; Liu et al., 2019a). Liu et al. (2021a) employed competence-based curriculum learning to improve report generation, which started from simple reports and then attempted to consume harder reports.

6.3 Radiology impression generation

Summarization is a fundamental text generation task in natural language processing (NLP), drawing sustained attention over the past decades (See et al., 2017; Liu and Lapata, 2019; Duan et al., 2019; Chen and Bansal, 2018; Lebanoff et al., 2019). General Impression generation can be regarded as a special type of summarization task in the medical domain, aiming to summarize findings and generate impressions. There are many methods proposed for this area (Gharebagh et al., 2020; Hu et al., 2021; Zhang et al., 2018; Hu et al., 2022a; Karn et al., 2022; MacAvaney et al., 2019; Zhang et al., 2020c; Delbrouck et al., 2022). MacAvaney et al. (2019); Gharebagh et al. (2020) proposed to extract medical ontologies and then utilize a separate encoder to extract features from such critical words for improving the decoding process and thus promoting AIG. Hu et al. (2021) further constructed a word graph by medical entities and dependence tree and then utilized the GNN to extract features from such graph for guiding the generation process. However, recent works in this area mainly focus on the text section while failing to fully explore the valuable information in corresponding radiology images.

7 Conclusion

This paper proposes an anatomy-enhanced multimodal summarization framework to integrate radiology images and text for facilitating impression generation. In detail, for radiology images, we use a visual extractor to extract detailed visual features. For radiology findings, we first plan anatomical
prompts into each sentence by keywords and rules and then apply a pre-trained encoder to distillate features from modified findings. Afterward, we employ a contrastive learning module to align the visual and textual features at the document level and use a co-attention to fuse these two features at the sentence level, which are then input to the decoder to improve impression generation. Furthermore, experimental results on two benchmark datasets illustrate the effectiveness of our model, especially for long impression generation, where our model achieves significant improvements.

8 Limitations

Although our model has achieved considerable improvements, as shown in Figure 3, our model tends to have a slight decrease in short impression generation, which need to be further solved in the future. In this paper, we follow previous studies and only utilize English radiology report datasets to verify the effectiveness of our proposed model, which is limited in verification in other languages. The main reason is that most publicly available radiology report datasets center on English. In addition, our model needs relatively more parameters than the models only using findings to generate impressions.

9 Acknowledgments

This work is supported by Chinese Key-Area Research and Development Program of Guangdong Province (2020B0101350001), and the Shenzhen Science and Technology Program (JCYJ20220818103001002), and the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen.

References


<table>
<thead>
<tr>
<th>MODEL</th>
<th>HYPER-PARAMETER</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIMIC-CXR</td>
<td>BATCH SIZE</td>
<td>640, 1024, 2048, 3072</td>
</tr>
<tr>
<td></td>
<td>LEARNING RATE</td>
<td>6e-5, 5e-4, 1e-3</td>
</tr>
<tr>
<td></td>
<td>TRAINING STEPS</td>
<td>200000</td>
</tr>
<tr>
<td></td>
<td>λ</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>τ</td>
<td>0.5</td>
</tr>
<tr>
<td>OpenI</td>
<td>BATCH SIZE</td>
<td>640, 1024, 2048, 3072</td>
</tr>
<tr>
<td></td>
<td>LEARNING RATE</td>
<td>6e-5, 5e-4, 1e-3</td>
</tr>
<tr>
<td></td>
<td>TRAINING STEPS</td>
<td>30000</td>
</tr>
<tr>
<td></td>
<td>λ</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>τ</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Table 5: The hyper-parameters that we have experimented on the datasets. The bold values illustrate the best configurations of different models.

<table>
<thead>
<tr>
<th>DATA</th>
<th>TYPE</th>
<th>TRAIN</th>
<th>DEV</th>
<th>TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenI</td>
<td>REPORT #</td>
<td>2.4K</td>
<td>0.3K</td>
<td>0.6K</td>
</tr>
<tr>
<td></td>
<td>AVG. WF</td>
<td>37.9</td>
<td>37.8</td>
<td>30.0</td>
</tr>
<tr>
<td></td>
<td>AVG. SF</td>
<td>5.75</td>
<td>5.68</td>
<td>5.77</td>
</tr>
<tr>
<td></td>
<td>AVG. WI</td>
<td>10.4</td>
<td>11.2</td>
<td>10.6</td>
</tr>
<tr>
<td></td>
<td>AVG. SI</td>
<td>2.86</td>
<td>2.94</td>
<td>2.82</td>
</tr>
<tr>
<td>MIMIC-CXR</td>
<td>REPORT #</td>
<td>117.7K</td>
<td>0.9K</td>
<td>1.5K</td>
</tr>
<tr>
<td></td>
<td>IMAGE #</td>
<td>117.7K</td>
<td>0.9K</td>
<td>1.5K</td>
</tr>
<tr>
<td></td>
<td>AVG. WF</td>
<td>55.4</td>
<td>56.3</td>
<td>70.0</td>
</tr>
<tr>
<td></td>
<td>AVG. SF</td>
<td>5.49</td>
<td>5.51</td>
<td>6.24</td>
</tr>
<tr>
<td></td>
<td>AVG. WI</td>
<td>16.4</td>
<td>16.26</td>
<td>21.1</td>
</tr>
<tr>
<td></td>
<td>AVG. SI</td>
<td>1.66</td>
<td>1.65</td>
<td>1.87</td>
</tr>
</tbody>
</table>

Table 6: The statistics of the two benchmark datasets with random split for OpenI and official split for MIMIC-CXR, including the numbers of report, the averaged sentence-based length (AVG. SF, AVG. SI), the averaged word-based length (AVG. WF, AVG. WI) of both IMPRESSION and FINDINGS.

A Appendix

A.1 Hyper-parameter Settings

Table 5 reports the hyper-parameters tested in tuning our models on MIMIC-CXR and OpenI. For each dataset, we try combinations of the hyper-parameters and use the one achieving the highest R-L for MIMIC-CXR and OpenI.

A.2 Dataset

We present the statistics of these two datasets in Table 6.

A.3 Model Size

Table 7 reports the number of trainable parameters (PARA.) of the baselines and our proposed model on MIMIC-CXR dataset when the hyper-parameters use the best configuration.

<table>
<thead>
<tr>
<th>MODEL</th>
<th>PARA.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASE-FINDING</td>
<td>177.87M</td>
</tr>
<tr>
<td>BERT+AP+CL</td>
<td>255.03M</td>
</tr>
</tbody>
</table>

Table 7: The parameter size of the methods in the experiments.
ACL 2023 Responsible NLP Checklist

A For every submission:

✓ A1. Did you describe the limitations of your work?  
Limitations

✓ A2. Did you discuss any potential risks of your work?  
section 5.1

✓ A3. Do the abstract and introduction summarize the paper’s main claims?  
abstract section 1

☐ A4. Have you used AI writing assistants when working on this paper?  
Not applicable. Left blank.

B ☐ Did you use or create scientific artifacts?  
Not applicable. Left blank.

✓ B1. Did you cite the creators of artifacts you used?  
section 3.1

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?  
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?  
Not applicable. Left blank.

✓ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?  
section 3.1

✓ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?  
section 3.1

✓ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.  
section 3.1 Appendix A.2

C ✓ Did you run computational experiments?  
section 4.1 4.2

✓ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?  
Appendix A.3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.
✓ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
  *Appendix A.1 section 3.3*

✓ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
  *section 4.1 section 4.2 section 5.1*

✓ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
  *Appendix A.1 Section 3.3*

D ✓ Did you use human annotators (e.g., crowdworkers) or research with human participants?
  *Section 4.3*

  □ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
    *Not applicable. Left blank.*

  □ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants’ demographic (e.g., country of residence)?
    *Not applicable. Left blank.*

  □ D3. Did you discuss whether and how consent was obtained from people whose data you’re using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
    *Not applicable. Left blank.*

  □ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
    *Not applicable. Left blank.*

  □ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
    *Not applicable. Left blank.*