DeltaNet: Conditional Medical Report Generation for COVID-19 Diagnosis

Xian Wu[1], Shuxin Yang[2,5], Zhaopeng Qiu[1], Shen Ge[1], Yangtian Yan[1], Xingwang Wu[3]
Yefeng Zheng[1], S. Kevin Zhou[4], Li Xiao[2,5]*

[1]Tencent Jarvis Lab [2]Institute of Computing Technology of the Chinese Academy of Sciences
[3]The First Affiliated Hospital of Anhui Medical University
[4]University of Science and Technology of China
[5]University of Chinese Academy of Sciences

{kevinxwu, shuxinyang, zhaopengqiu, shenge, yangtianyan, yefengzheng}@tencent.com;
duobi2004@126.com; s.kevin.zhou@gmail.com;xiaoli@ict.ac.cn

Abstract

Fast screening and diagnosis are critical in COVID-19 patient treatment. In addition to the gold standard RT-PCR, radiological imaging like X-ray and CT also works as an important means in patient screening and follow-up. However, due to the excessive number of patients, writing reports becomes a heavy burden for radiologists. To reduce the workload of radiologists, we propose DeltaNet to generate medical reports automatically. Different from typical image captioning approaches that generate reports with an encoder and a decoder, DeltaNet applies a conditional generation process. In particular, given a medical image, DeltaNet employs three steps to generate a report: 1) first retrieving related medical reports, i.e., the historical reports from the same or similar patients; 2) then comparing retrieved images and current image to find the differences; 3) finally generating a new report to accommodate identified differences based on the conditional report. We evaluate DeltaNet on a COVID-19 dataset, where DeltaNet outperforms state-of-the-art approaches. Besides COVID-19, the proposed DeltaNet can be applied to other diseases as well. We validate its generalization capabilities on the public IU-Xray and MIMIC-CXR datasets for chest-related diseases.

1 Introduction

Since December 2019, the world has been suffering from a serious health crisis: the outbreak of COVID-19 (Wang et al., 2020). Fast screening and diagnosis is critical in COVID-19 patient treatment. In clinical practice, the Reverse Transcription Polymerase Chain Reaction (RT-PCR) is recognized as the golden standard (Zu et al., 2020). However, due to high false-negative rate and shortage of equipment (Fang et al., 2020; Ng et al., 2020), medical imaging like X-ray and Computed Tomography (CT) (Rubin et al., 2020) also works as

an alternative means in COVID-19 diagnosis and treatment which generates more timely results than RT-PCR and helps in evaluating the severity degree of COVID-19.

Given medical images of COVID-19 patients, radiologists need to write relatively long reports to address the impressions and findings. Considering the large volume of COVID-19 patients and potentially infected population, writing medical reports becomes a heavy burden for radiologists. Furthermore, due to the varied expertise of radiologists, some abnormalities in medical images may be ignored and thus not included in the final reports. To alleviate the heavy workload and aid less experienced radiologists, automatically generating medical reports becomes a critical task.

Due to its importance in clinical practice, automatically generating medical reports has attracted extensive research interests in recent years (Zhou et al., 2021). Existing works mainly follow the image captioning approaches and employ an encoder-decoder process. In the encoding stage, the visual features are extracted from medical images via a CNN; in the decoding stage, the reports are generated sequentially via an RNN. Such a two-step framework has been proven effective in generating general image captions, such as MS COCO. However, when applied to medical images, it may have following two problems: 1) visual bias: for most cases, the abnormal regions only occupy a small fraction of the entire medical images, thus visual features of normal regions dominate the extracted visual embedding, and abnormal regions are difficult to identify; 2) textual bias: in current medical reports, the majority of transcription focuses on describing the normal regions which distract the model from abnormal regions in training.

To address two problems mentioned above, we propose DeltaNet, which is customized for medical images. Different from encoder-decoder frameworks that generate reports from a single medical

---

*Corresponding author

image, DeltaNet introduces a retrieve-update process which consists of the following three steps: 1) Retrieval: DeltaNet firstly retrieves conditional medical images and reports from medical records. For patients who've already been examined before, we directly obtain his previous medical images and reports. For patients examined for the first time, we retrieve the medical images and reports with similar visual features from other patients; 2) Comparison: DeltaNet compares the embedding difference between two medical images to capture the visual difference; 3) Based on the identified visual difference and the retrieved reports, DeltaNet conditionally generates the final report for current medical image.

To prove the effectiveness of the proposed DeltaNet, we collect a dataset of COVID-19 patients. For each patient, we manage to collect all the historical medical images and the corresponding reports that cover the complete treatment process. The experimental results show that the proposed DeltaNet outperforms the baselines, including both the general image captioning approaches and the existing medical report generation works. Besides generating reports, DeltaNet can also highlight the differences between previous and current medical images and the corresponding updated reports, which provides a clear explanation for the generated report. In addition to COVID-19, the proposed DeltaNet can also be applied to medical report generation of other diseases. We evaluate the generalization ability of DeltaNet on IU-Xray and MIMIC-CXR datasets. DeltaNet consistently outperforms baselines.

2 Related Works

2.1 Image Captioning

Image captioning aims to provide a short descriptive sentence for a given image, which has recently received extensive research interests (Vinyals et al., 2015; Xu et al., 2015; Anderson et al., 2018; Huang et al., 2019). Typical image captioning models, e.g., (Anderson et al., 2018; Huang et al., 2019), adopt the encoder-decoder framework to accomplish the image captioning task, where the encoder extracts the visual representations from the images, and the decoder transforms the acquired visual features to texts. Different from typical image captioning approaches, DeltaNet applies a retrieve-update process, consisting of retrieval, comparison, and generation steps to generate reliable and robust reports.

2.2 Medical Report Generation

Medical report generation generally produces longer texts than typical image captions (Zhou et al., 2019). Employing the typical image caption encoder-decoder framework, Jing et al. (2018) proposed to use a two-level hierarchical LSTM to deal with long reports, with the top-level handling topic generation and the bottom-level generating texts according to the currently selected topic; Wang et al. (2018); Xue and Huang (2019); Yuan et al. (2019) used attention mechanism to drive the encoder and the decoder to emphasize on more informative words or visual regions, resulting in improved performance; Li et al. (2018, 2019); Zhang et al. (2020); Liu et al. (2021) introduced external information such as template or knowledge graph to guide the generation of medical reports; Syeda-Mahmood et al. (2020) invited four clinicians to manually examine 220,000 reports and build a taxonomy of 11,000 unique terms for developing an automatic labeling algorithm. Then, they built and customized similar reports from a large report database by fine-grained labels as the generated report. However, for diseases not labeled in advance, this method may not fit well. Some other works (Wang et al., 2018) brought in auxiliary tasks to improve report generation, usually requiring extra expert labeling. Our work uses a new retrieve-update process to effectively generate reports automatically conditioned on historical reports. Besides, our method enhances the model by only using the information from the training dataset, avoiding any expert labeling or external information.

3 Model

In this section, we introduce the proposed DeltaNet. Firstly, we formulate the conditional medical report generation problem; secondly, we describe the basic encoder-decoder based generation model; then, we propose DeltaNet, a conditional medical report generation model; finally, we further extend DeltaNet to exploit multiple conditional reports.

3.1 Problem Formulation

We use a quadruple $q = \{I, I_c, R, R_c\}$ to refer to an input instance, in which I denotes the input medical image and $R = \{w_1, w_2, \ldots, w_N\}$ denotes the corresponding report to be generated. In this manner, medical report generation can be formulated as estimating the parameters of the con-
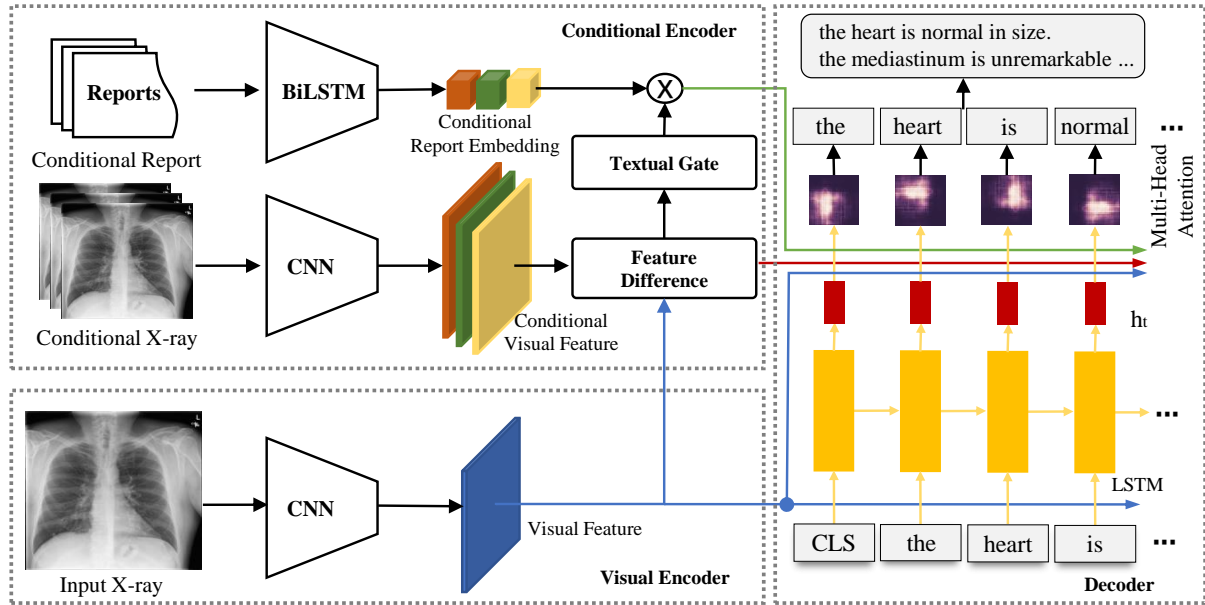
Figure 1: Framework of the proposed DeltaNet model including three major component: visual encoder, conditional encoder and decoder. The visual encoder extracts features from input X-ray image. The conditional encoder extracts features from conditional X-ray image, and embeds the conditional report by a BiLSTM, and acquires the difference between input x-ray and conditional X-ray. The decoder decodes all features, difference and embeddings to generate output report.

ditional probability $P(R|I)$. In this paper, we introduce a conditional image I_c and a conditional report $R_c = \{w_1, w_2, \ldots, w_{N_c}\}$. For patients who have historical medical images, we select their previous medical image and report as I_c and R_c respectively; as for patients without historical reports, we retrieve the most similar medical image I_c and report R_c from the pre-built medical report repository. Then the conditional medical report generation problem can be formulated as $P(R|I, I_c, R_c)$.

3.2 Basic Model

Many existing medical report generation works follow the image captioning approaches and employ the encoder-decoder two-step manner. Typically, a Convolution Neural Network (CNN) is introduced as the encoder, which extracts visual features from the input medical image. In this paper, we extract the output of the last convolution layer of the visual encoder following a linear projection layer as visual features:

$$V = \text{CNN}(I), \tag{1}$$

where $V \in \mathbb{R}^{K \times D}$, K denotes the size of feature maps and D denotes the number of feature maps; in this paper, K is set to 49 and D is set to 512.

After acquiring the visual features V, the next step is to generate medical reports which we refer

as the decoding process. In the decoding process, we use LSTM to generate medical report sequentially. At each timestamp, we firstly acquire current hidden state:

$$h_t = \text{LSTM}(w_{t-1}, h_{t-1}), \tag{2}$$

where $h_0 = \mathbf{0}$ and $h_t \in \mathbb{R}^{1 \times D}$.

After acquiring h_t, we attend it to the extracted visual features. In Eq.(3), we use h_t as query and V as both the key and value:

$$a_t = \text{MHA}(h_t, V, V), \tag{3}$$

where $\text{MHA}(\cdot, \cdot, \cdot)$ refers to the multi-head attention function proposed in (Vaswani et al., 2017). The attention embedding $a_t \in \mathbb{R}^{1 \times D}$ can be regarded as attended visual features given the current hidden state.

Finally, we combine the hidden state h_t and the attention embedding a_t to estimate the probability distribution of generating w_t. The probability distribution can be formulated as:

$$P(w_t|w_1, \ldots, w_{t-1}, V) = \sigma([h_t; a_t]W_p), \tag{4}$$

where $W_p \in \mathbb{R}^{2D \times E}$ is a learnable linear projection and E is vocabulary size.

2954

3.3 Conditional Generation Model

In Section 3.2, we adopt the two-step encoder-decoder framework to generate medical reports. Although such a two-step approach has been proven effective in general image captioning, e.g., MSCOCO, due to the unavailability of large-scale labeled data, it is difficult to generate an accurate yet fluent report when applying to the medical report generation problem.

Observing the fact that during a complete treatment process, one patient usually have been examined for multiple times, therefore he may have multiple medical images. For example, the COVID-19 patients may first take a chest X-ray examination for diagnosis and take several more X-ray to track the severity progresses of COVID-19 during the treatment. We review the medical reports from the same patient and find that the consecutive reports share the majority of the content and only differ in the disease progresses. Therefore, a natural thought is to generate the medical report in a conditional manner, not only from the current input medical image but also from the historical images and reports. For first time patients without previous reports, we can also select reports with similar visual appearances from other patients. For all medical images, we extract their visual features with Eq.(5). Then given an input image, we retrieve the most similar images according to the cosine distance of visual features. Let I_c and R_c denote the conditional image and report respectively, then conditional medical report generation problem can be formulated as $P(R|I, I_c, R_c)$. Here we propose DeltaNet as the medical report's conditional generation model.

DeltaNet model first extracts the visual features V_c and textual features T_c from the conditional image and report.

$$V_c = \text{CNN}(I_c), \tag{5}$$
$$T_c = \text{BiLSTM}(R_c), \tag{6}$$

.

As demonstrated in (Wu et al., 2018), the subtraction of visual features is an efficient operation to acquire the difference between features. Thus DeltaNet acquires the varied visual features between V and V_c using:

$$\Delta V_c = V - V_c, \tag{7}$$

In the decoding process, at each timestamp, after generating the hidden state h_t with Eq.(2), we attend h_t to both the varied visual features of medical images and the textual features from the conditional report as follows:

$$s_t = \text{MHA}(h_t, \Delta V_c, \Delta V_c), \tag{8}$$
$$c_t = \text{MHA}(h_t, T_c, T_c), \tag{9}$$

Finally, besides the h_t and a_t, we utilize s_t and c_t to estimate the probability distribution of generating w_t. The probability distribution can be formulated as:

$$P(w_t|w_1, \ldots, w_{t-1}, V) \tag{10}$$
$$= \sigma([h_t; a_t; s_t; c_t]W_p),$$

where W_p is a learnable linear projection in the shape of $4D \times E$ and E is vocabulary size.

3.4 Multiple Conditional Generation Model

In the previous subsection, we mainly deal with the cases that the patient only has one conditional report. However, some patients have multiple historical reports. For first time patients without previous reports, we can also retrieve multiple reports according to the visual similarity from other patients. In this subsection, we extend the proposed DeltaNet to fit for multiple conditional reports.

For each input image, we assume that it has L conditional images and reports, we denote them with two sets $\{I_c^{(1)}, I_c^{(2)}, \ldots, I_c^{(L)}\}$ and $\{R_c^{(1)}, R_c^{(2)}, \ldots, R_c^{(L)}\}$. DeltaNet firstly extracts the visual features and textual features from conditional images and reports by two encoders:

$$V_c^{(i)} = \text{CNN}(I_c^{(i)}), \tag{11}$$
$$T_c^{(i)} = \text{BiLSTM}(R_c^{(i)}), \tag{12}$$

Then DeltaNet acquires varied visual features:

$$\Delta V_c^{(i)} = V - V_c^{(i)}. \tag{13}$$

As shown in Figure 1, for each textual feature $T_c^{(i)}$ extracted from the conditional report, we add a textual gate to control its contribution since the contribution of each report is different. Here we introduce a gate weight g_i as follows:

$$g^{(i)} = \sigma(W_v V + W_c V_c^{(i)} + b_i), \tag{14}$$

where W_v and W_c are learnable linear projections. The $g^{(i)}$ takes visual features of both input image and conditional image into consideration which is used to re-weight the textual features:

$$\hat{T}_c^{(i)} = g^{(i)} * T_c^{(i)}. \tag{15}$$

DeltaNet further concatenate varied visual features and weighted textual features in a row-wise manner:

$$\Delta V_c = [\Delta V_c^{(1)}; \ldots; \Delta V_c^{(L)}], \quad (16)$$

$$T_c = [\hat{T}_c^{(1)}; \ldots; \hat{T}_c^{(L)}], \quad (17)$$

where $\Delta V_c \in \mathbb{R}^{KL \times D}$ and $T_c \in \mathbb{R}^{N_c L \times D}$.

At each timestamp, DeltaNet use multi-head attention to generate the subtractive visual attend feature s_t and the context embedding c_t.

$$s_t = \text{MHA}(h_t, \Delta V_c, \Delta V_c), \quad (18)$$

$$c_t = \text{MHA}(h_t, T_c, T_c), \quad (19)$$

where both s_t and c_t are in the shape of $1 \times D$.

Then in each time step, the word w_t is generated according to the probability in Eq.(20).

$$P(w_t|w_1, \ldots, w_{t-1}, V) \quad (20)$$
$$= \sigma([h_t; a_t; s_t; c_t]W_p).$$

4 Experiments

In this section, we evaluate the proposed DeltaNet from three perspectives: (1) Whether incorporating conditional reports can bring in performance gain in medical report generation; (2) Whether increasing the number of conditional reports can further improve the performance; (3) Which will perform better, the conditional reports that are acquired from the historical reports of patients themselves or retrieved from other patients according to visual similarity.

4.1 Implementation Details

For first time patients without historical reports, we retrieve conditional reports via the embedding similarity of their corresponding images. The embeddings are acquired from an encoder pre-trained on the ChestX-Ray14 dataset (Wang et al., 2017). Since the lengths of conditional reports are different, we use zero-padding for each conditional case to pad each report to a fixed length. We adopt ResNet-152 as our visual encoder for both current input image and conditional images, a two-layer bidirectional LSTM as conditional report encoder, and a single-layer LSTM as medical report decoder. The dimension of the visual feature and hidden states are set to 512. We adopt the Adam optimizer with an initial learning rate of 5e-4 and a mini-batch size of 32. We train the model with cross entropy loss for 100 epochs and early stop strategy is adopted.

4.2 Datasets and Settings

We introduce three data sets COVID-19, IU-Xray (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019) to evaluate the proposed DeltaNet.

The COVID-19 dataset includes 1,261 exams (including both images and reports) from 1,085 patients from mobile field hospitals used in the COVID-19 pandemic. The private information has been removed during the data collection process. Each case includes both the X-ray images and the corresponding reports. The max, median, and mean length of the reports are 180, 69, and 72 words, respectively. Thus the report to be generated are relatively lengthy. The ratio of patients with only one visit is 57.48%. Among these 1,261 cases, in terms of disease severity, 112 cases are severe, 1,113 cases are general, 30 cases are mild, and 6 cases are not labeled. More than 166 patients have more than 2 reports; we select the latest one as the prediction target and select the second latest one as the conditional report. We split these 166 patients in the ratio of 7:1:2, that is 116 for training, 16 for validating, and 34 for testing. These three groups have no overlap in patients.

IU-Xray (Demner-Fushman et al., 2016) is a public data set which includes 3,955 radiology reports and 7,470 frontal- and lateral-view chest X-ray images. We follow the split of (Chen et al., 2020) which divides the entire data set into train/validation/test in the ratio 7:1:2.

MIMIC-CXR (Johnson et al., 2019) is the largest dataset for the medical report generation, which contains 377,110 chest X-ray images and 227,827 free-text radiology reports. The dataset contains multi-view images, and we select frontal view images in this work.

For the COVID-19 data set, the conditional report is selected from the previous one of the same patient. As to the IU-Xray data set, we are unable to acquire the historical reports for a patient. Therefore we first acquire the visual feature for each image via a CNN based model pre-trained on the ChestX-Ray14 dataset. Then for each image, we select the conditional reports via cosine similarity of visual feature. For the case in the train/validate/test groups, we select their conditional reports only from the training group. In this case, we avoid the situation that the report to be generated appears as the conditional reports in the training data set. As a result, the label leakage is prevented.

Table 1: The performance of baselines and the proposed DeltaNet on IU-Xray and MIMIC dataset. The conditional reports are extracted from reports of other patients according to visual similarity. For the baseline methods, we cite their performance reported in their papers and reported in (Li et al., 2018) and (Jing et al., 2018). The *Basic* refers to the one introduced in Section 3.2.

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE-L
IU-Xray	S&T (Vinyals et al., 2015)	0.216	0.124	0.087	0.066	0.294	0.306
	SA&T (Xu et al., 2015)	0.399	0.251	0.168	0.118	0.302	0.323
	AdaAtt (Lu et al., 2017)	0.220	0.127	0.089	0.068	0.295	0.308
	TieNet (Wang et al., 2018)	0.286	0.160	0.104	0.074	/	0.226
	CoAtt (Jing et al., 2018)	0.455	0.288	0.205	0.154	0.277	0.369
	R2Gen (Chen et al., 2020)	0.470	0.304	0.219	0.165	/	0.371
	PPKED (Liu et al., 2021)	0.483	0.315	0.224	0.168	0.351	0.376
	Basic	0.417	0.264	0.184	0.138	0.467	0.343
	DeltaNet-1C	0.470	0.307	0.224	0.175	**0.853**	0.369
	DeltaNet-3C	**0.485**	**0.324**	**0.238**	**0.184**	0.802	**0.379**
MIMIC	S&T (Vinyals et al., 2015)	0.256	0.157	0.102	0.070	0.063	0.249
	SA&T (Xu et al., 2015)	0.304	0.177	0.112	0.077	0.083	0.249
	AdaAtt (Lu et al., 2017)	0.311	0.178	0.111	0.075	0.084	0.246
	BU&TD (Anderson et al., 2018)	0.280	0.169	0.108	0.074	0.073	0.250
	R2Gen (Chen et al., 2020)	0.353	0.218	0.145	0.103	/	0.277
	PPKED (Liu et al., 2021)	0.360	0.224	0.149	0.106	/	**0.284**
	Basic	0.335	0.206	0.138	0.100	0.156	0.263
	DeltaNet-1C	0.355	0.221	0.152	0.113	0.220	0.279
	DeltaNet-3C	**0.361**	**0.225**	**0.154**	**0.114**	**0.281**	0.277

For the MIMIC-CXR data set, we exclude the patients without reports. Then we generate two data sets to evaluate two types of conditional report respectively: 1) MIMIC: this data set is the same as original MIMIC-CXR dataset. We follow standard split provided by (Johnson et al., 2019) to divide the entire data set into train/validate/test and the conditional cases are retrieved from the training set using visual feature similarity; 2) MIMIC-Multi-Visit: we first select the patients with more than three reports as candidates. We choose their latest reports as the prediction target and select three most recent previous reports as the conditional reports. The offical training/validation/testing sets split of MIMIC-CXR is 222,758/1,808/3,269, respectively. From the orignial MIMIC-CXR data set, we select the patients with more than 3 visits, resulting the data set MIMIC-Multi-Visit of 11,978/83/165, respectively. To conduct a fair comparison, during the training of the basic model and existing models, we append the conditional reports into their training set. In this manner, the proposed DeltaNet will not include extra labeled data in training.

We select the popular metrics for natural language generation (NLG) tasks and a specific clinical metric for evaluation. The NLG metrics include BLEU-n (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), and ROUGE-L (Lin, 2004) score. The results are computed by MS-COCO caption evaluation tool [1] automatically. The clinical efficacy (CE) metric is proposed by R2gen (Chen et al., 2020) to quantify the precision, recall, and f1 score of medical terminology described in reference and generated reports. Because the IU-Xray dataset does not provide such labels, we only report CE metric on the MIMIC-CXR dataset. For consistency, we employ the CheXpert (Irvin et al., 2019) to extract the labels from generated reports.

4.3 Quantitative Results

Baselines. The Tables display the comparison results between existing works and the proposed DeltaNet. Among the baselines, S&T (Vinyals et al., 2015), SA&T (Xu et al., 2015), AdaAtt (Lu et al., 2017) and BU&TD (Anderson et al., 2018) belong to general image captioning approaches; while CoAtt (Jing et al., 2018), TieNet (Wang et al., 2018), R2Gen (Chen et al., 2020) and PPKED (Liu et al., 2021) focus on the specific medical report generation task. Since MIMIC-Multi-Visit only includes patients with more than three images, and COVID-19 is a private data set, we cannot directly cite the results of existing models. Therefore, we re-train existing models on these two data sets. For R2Gen, we directly use the implementation [2] provided by its authors; for CoAtt, we use the third party implementation [3]. For other works, we re-implement them ourselves; The *Basis Model* refers to the one described in *Basic Model* section; The

---

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

[2] https://github.com/cuhksz-nlp/R2Gen

[3] https://github.com/ZexinYan/Medical-Report-Generation

Table 2: The performance of baselines and the proposed DeltaNet on COVID-19 and MIMIC-Multi-Visit data sets. The conditional reports are extracted from patients' own historical reports.

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	ROUGE-L
COVID-19	S&T (Vinyals et al., 2015)	0.604	0.573	0.546	0.523	0.148	0.639
	SA&T (Xu et al., 2015)	0.619	0.586	0.557	0.534	0.167	**0.642**
	AdaAtt (Lu et al., 2017)	0.617	0.583	0.553	0.529	0.150	0.625
	BU&TD (Anderson et al., 2018)	0.600	0.563	0.531	0.504	0.120	0.603
	R2Gen (Chen et al., 2020)	0.610	0.576	0.547	0.523	0.176	0.624
	Basic	0.614	0.566	0.530	0.503	0.256	0.610
	DeltaNet	**0.664**	**0.622**	**0.588**	**0.561**	**0.273**	0.635
MIMIC-Multi-Visit	S&T (Vinyals et al., 2015)	0.300	0.176	0.111	0.074	0.066	0.246
	SA&T (Xu et al., 2015)	0.284	0.170	0.114	0.080	0.046	0.234
	AdaAtt (Lu et al., 2017)	0.310	0.180	0.114	0.075	0.068	0.240
	BU&TD (Anderson et al., 2018)	0.284	0.170	0.111	0.079	0.089	0.253
	R2Gen (Chen et al., 2020)	0.303	0.193	0.131	0.095	0.148	0.266
	Basic	0.325	0.196	0.129	0.091	0.102	0.248
	DeltaNet-1C	0.358	0.217	0.144	0.103	0.248	**0.270**
	DeltaNet-3C	**0.371**	**0.228**	**0.152**	**0.107**	**0.301**	0.264

DeltaNet and *DeltaNet-1C* refers to the one introduced in *Conditional Generation Model* section; The *DeltaNet-3C* refers to the one introduced in *Multiple Conditional Generation Model* section with three conditional reports.

Conditional Reports from Other Patients. We evaluate DeltaNet with conditional reports retrieved by the visual feature similarity from other patients. Here we conduct experiments on IU-Xray and MIMIC data sets. Since these two data sets are public data sets and the split of train/validate/test is in the standard manner, for existing works, we directly cite the reported performance. As shown in Table 1, *DeltaNet* consistently outperforms the *Basic Model*. This demonstrates that besides using the historical reports of the same patient as the conditional report, selecting similar reports as conditional reports can also boost the performance. As shown in Table 1 and Table 3, for the existing methods, the proposed DeltaNet with three conditional reports outperforms almost all baselines on both NLG and clinical efficiency metrics. As to the number of conditional reports, we evaluate DeltaNet with 1 and 3 conditional reports. As shown in both Table 2 and Table 1, the inclusion of more conditional reports generally increases the performance.

Conditional Reports from the Patients Themselves. As shown in Table 2, *DeltaNet* significantly outperforms the *Basic Model* which proves the effectiveness of introducing conditional reports in medical report generation. Furthermore, the proposed *DeltaNet* outperforms state-of-the-art approaches almost all the metrics.

Table 3: The performance of clinical efficiency.

Dataset	Model	Precision	Recall	F1
MIMIC	S&T	0.084	0.066	0.072
	SA&T	0.181	0.134	0.144
	AdaAtt	0.265	0.178	0.197
	BU&TD	0.166	0.121	0.133
	R2Gen	0.333	0.273	0.276
	DeltaNet-1C	0.460	0.353	0.376
	DeltaNet-3C	**0.470**	**0.399**	**0.406**

4.4 Ablation Study

Self vs. Other Patients. Here we conduct an experiment to compare the performance of two types of conditional reports. We use MIMIC-CXR dataset and select the patients with ≥ 2 reports. For each patient, we select two most recent reports, that is 49,180 images from 24,590 patients. The method *Self* denotes that the conditional report is acquired from the same patient; the method *Others* denotes that the conditional report is retrieved via visual feature similarity. To be fair, the conditional reports of the *Self* methods are added into the training set of the *Others* method. As shown in Table 4, the BLEU-4 is 0.114 vs. 0.101. The conditional report from the same patient outperforms the one from other patients.

Table 4: The performance of two types of conditional reports.

Conditional Reports	BLEU-3	BLEU-4	CIDE-r	ROUGE-L
Others	0.140	0.101	0.179	0.278
Self	**0.158**	**0.114**	**0.231**	**0.289**

Effectiveness of Conditional Image and Report. Compared with the basic model, our model utilizes two additional features I_c and R_c. To analyze the
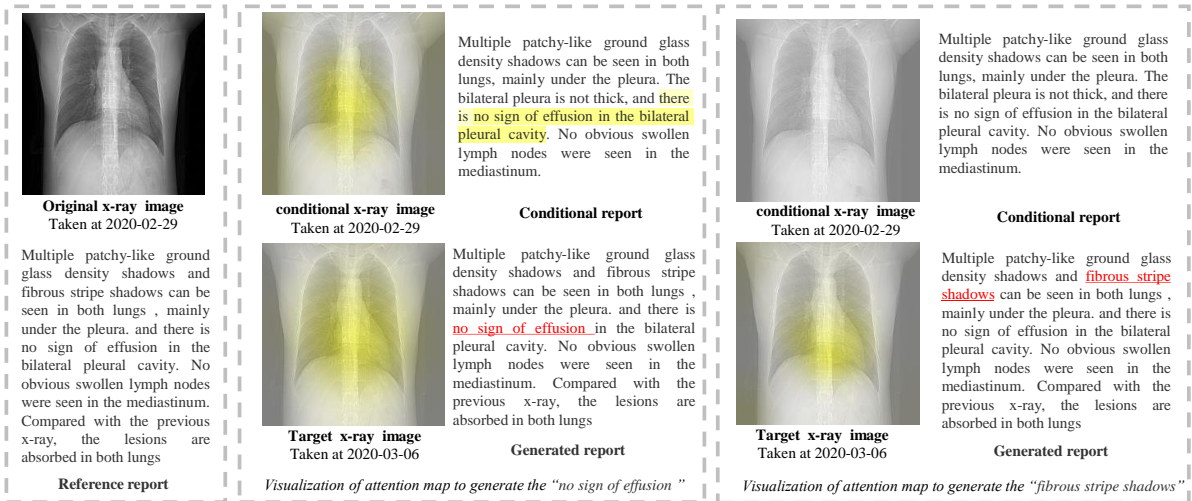
Figure 2: Visualizations of attention map on conditional/target image and report when DeltaNet generates the highlight words.

contribution of each feature, we design two models which separately concatenate I_c or R_c with visual features. As shown in first two rows of Table 5, the model with conditional reports outperforms the model with conditional images, which shows that the conditional reports are more effective than conditional images. To analyze the effectiveness of image feature subtraction in Eq.(7), we design a model which directly concatenates both with visual features. As shown in "$I_c + R_c$" and ours rows of Table 5, the model directly concatenates with both features is lower than DeltaNet, which shows the benefit by combining current input and conditional input by feature subtraction and cross attention in DeltaNet.

Table 5: The contribution of conditional image and report. I_c and R_c refer to the features of conditional images and reports. $I_c + R_c$ refers to feature concatenation.

Dataset	Model	BLEU-3	BLEU-4	ROUGE-L
IU-Xray	I_c	0.137	0.103	0.281
	R_c	0.176	0.134	0.310
	I_c+R_c	0.208	0.162	0.329
	Ours	**0.224**	**0.175**	**0.369**
MIMIC-Multi-Visit	I_c	0.108	0.075	0.222
	R_c	0.126	0.090	0.238
	I_c+R_c	0.127	0.091	0.242
	Ours	**0.144**	**0.103**	**0.270**
COVID-19	I_c	0.481	0.456	0.572
	R_c	0.507	0.482	0.595
	I_c+R_c	0.531	0.508	0.619
	Ours	**0.588**	**0.561**	**0.635**

4.5 Case Study

In this section, we demonstrate the effectiveness of DeltaNet with a case study of a COVID-19 patient. Here we select a patient who has two consecutive X-ray examinations during COVID treatment. This patient is in mild COVID-19 severity. We use the report taken on Feb 29, 2020 as the conditional report to generate the report taken on March 6, 2020.

On the left part of Figure 2, we visualize the attention heat map to generate the phrase "no sign of effusion". The finding of "no sign of effusion" remains the same in both conditional and current reports. Therefore, generating this phrase attends to both current and conditional medical images. It also attends to the corresponding phrase in the conditional reports. As a result, the combination of the conditional report and both images enable the correct generation of this phrase.

On the right of Figure 2, we visualize the attention heat map to generate the phrase "fibrous stripe shadows". Since this is a new finding which does not exist in a conditional report, it only attends to the current medical image. The proposed DeltaNet can learn the difference between conditional and current images and correspondingly generate this new finding in the target report.

5 Conclusion and Future Works

In this paper, we targeted to automate the medical report generation. Different from typical encoder-decoder framework, we proposed a conditional generation model DeltaNet. For patients with historical

reports, we combined the input image with the historical report to generate reports. For first time patients without historical reports, we retrieved visual similar reports from other patients as conditional reports. We proved the advantage of the proposed DeltaNet over state-of-the-art approaches on the IU-Xray, MIMIC-CXR and COVID-19 datasets.

In this paper, for new patients without historical reports, we retrieve the conditional reports only via visual similarity. However, more features like age, gender and diseases could be useful in selecting conditional reports. In addition, we may attempt to combine historical reports from the same patient and simliar reports from other patients together to futher improve the performance.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and VQA. In *CVPR*.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating Radiology Reports via Memory-driven Transformer. In *EMNLP*.

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

Yicheng Fang, Huangqi Zhang, Jicheng Xie, Minjie Lin, Lingjun Ying, Peipei Pang, and Wenbin Ji. 2020. Sensitivity of Chest CT for COVID-19: Comparison to RT-PCR. *Radiology*, pages 200–432.

Lun Huang, Wenmin Wang, Jie Chen, and Xiaoyong Wei. 2019. Attention on attention for image captioning. In *ICCV*.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. 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. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*.

Baoyu Jing, Pengtao Xie, and Eric P. Xing. 2018. On the automatic generation of medical imaging reports. In *ACL*.

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.

2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317.

Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2019. Knowledge-Driven Encode, Retrieve, Paraphrase for Medical Image Report Generation. In *AAAI*.

Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2018. Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation. In *NeurIPS*.

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

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

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*.

Ming-Yen Ng, Elaine YP Lee, Jin Yang, Fangfang Yang, Xia Li, Hongxia Wang, Macy Mei-sze Lui, Christine Shing-Yen Lo, Barry Leung, Pek-Lan Khong, et al. 2020. Imaging profile of the covid-19 infection: radiologic findings and literature review. *Radiology: Cardiothoracic Imaging*, 2(1):e200034.

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

Geoffrey D. Rubin, Linda B. Haramati, Jeffrey P. Kanne, Neil W. Schluger, and Athol U. Wells. 2020. The role of chest imaging in patient management during the covid-19 pandemic: A multinational consensus statement from the fleischner society. *Radiology*, 296(1):201365.

Tanveer F. Syeda-Mahmood, Ken C. L. Wong, Yaniv Gur, Joy T. Wu, Ashutosh Jadhav, Satyananda Kashyap, Alexandros Karargyris, Anup Pillai, Arjun Sharma, Ali Bin Syed, Orest B. Boyko, and Mehdi Moradi. 2020. Chest X-Ray Report Generation Through Fine-Grained Label Learning. In *MICCAI*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE*

conference on computer vision and pattern recognition, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*.

C. Wang, P. W. Horby, F. G. Hayden, and G. F. Gao. 2020. A novel coronavirus outbreak of global health concern. *The Lancet*, 395(10223):470–473.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. Chest x-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *CVPR*.

Chenfei Wu, Jinlai Liu, Xiaojie Wang, and Xuan Dong. 2018. Object-difference attention: A simple relational attention for visual question answering. In *MM 2018 - Proceedings of the 2018 ACM Multimedia Conference*, volume 1, pages 519–527.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Yuan Xue and Xiaolei Huang. 2019. Improved disease classification in chest x-rays with transferred features from report generation. In *IPMI*.

Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. 2019. Automatic Radiology Report Generation Based on Multi-view Image Fusion and Medical Concept Enrichment. In *MICCAI 2019*, volume 11769 LNCS, pages 721–729.

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. 2020. When Radiology Report Generation Meets Knowledge Graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12910–12917.

S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. 2021. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*.

S Kevin Zhou, Daniel Rueckert, and Gabor Fichtinger. 2019. *Handbook of medical image computing and computer assisted intervention*. Academic Press.

Zi Yue Zu, Meng Di Jiang, Peng Peng Xu, Wen Chen, and Long Jiang Zhang. 2020. Coronavirus disease 2019 (covid-19): A perspective from china. *Radiology*, 296(2):200490.

A Appendix

A.1 Ethical Impact

This work aims to provide efficient and accurate radiology reports to assist radiologists rather than replace radiologists. It benefits the development of human health. We conduct experiments on COVID-19, IU-Xray, and MIMIC-CXR datasets. All personal information was de-identified. We have removed the information related to the data collection. All necessary permissions have been obtained and the appropriate institutional forms have been archived.

A.2 Limitation

In this paper, our method applies a conditional generation process that generates medical reports based on historical medical records. There are still some limitations. First, the generated medical report only contains the diseases in the historical medical records. Second, our method only contains visual and textual information and lacks other examination results (e.g., blood examination and indication) and medical knowledge. We will explore how to involve more features to improve the performance of the proposed method.

A.3 Potential Risks

Given a lot of medical images, it can automatically generate radiology reports. The radiologists only need to make revisions rather than write a new report from scratch. However, inexperienced radiologists may rely on it. Therefore, it is necessary to take additional measures to avoid the abuse of our model.