

Replace and Report: NLP Assisted Radiology Report Generation

Kaveri Kale, Pushpak Bhattacharyya and Kshitij Jadhav

Indian Institute of Technology Bombay

{kaverikale,pb}@cse.iitb.ac.in and kshitij.jadhav@iitb.ac.in

Abstract

Clinical practice frequently uses medical imaging for diagnosis and treatment. A significant challenge for automatic radiology report generation is that the radiology reports are long narratives consisting of multiple sentences for both abnormal and normal findings. Therefore, applying conventional image captioning approaches to generate the whole report proves to be insufficient, as these are designed to briefly describe images with short sentences. We propose a template-based approach to generate radiology reports from radiographs. Our approach involves the following: i) using a multi-label image classifier, produce the tags for the input radiograph; ii) using a transformer-based model, generate pathological descriptions (a description of abnormal findings seen on radiographs) from the tags generated in step (i); iii) using a BERT-based multi-label text classifier, find the spans in the normal report template to replace with the generated pathological descriptions; and iv) using a rule-based system, replace the identified span with the generated pathological description. We performed experiments with the two most popular radiology report datasets, IU Chest X-ray and MIMIC-CXR and demonstrated that the BLEU-1, ROUGE-L, METEOR, and CIDEr scores are better than the State-of-the-Art models by 25%, 36%, 44% and 48% respectively, on the IU X-RAY dataset. To the best of our knowledge, this is the first attempt to generate chest X-ray radiology reports by first creating small sentences for abnormal findings and then replacing them in the normal report template.

1 Introduction

Radiology report generation, which aims to automatically generate a free-text description of a clinical radiograph (like a chest X-ray), has become an important and interesting area of research in both clinical medicine and artificial intelligence. Natural Language Processing (NLP) can speed up

report generation and improve healthcare quality and standardization. Thus, recently, several methods have been proposed in this area (Jing et al., 2017; Li et al., 2018; Yuan et al., 2019; Chen et al., 2020; Alfarghaly et al., 2021). Radiology reports are lengthy narratives, which makes report generation difficult. Therefore, applying conventional image captioning approaches to generate the whole report (Vinyals et al., 2015; Anderson et al., 2018) proves to be insufficient, as such approaches are designed to briefly describe images with short sentences. Further, even if benchmark datasets are balanced between normal and abnormal studies, multiple organs' findings are included in a single report, and if at least one organ is abnormal, the report is classified as abnormal, but it still contains more normal sentences than abnormal. As a result, existing text generation models may be overly focused on widely reported normal findings. Hence, we propose an approach to generating radiology reports by generating pathological descriptions, i.e., descriptions of abnormal findings, and then replacing corresponding normal descriptions from the normal report template with a generated pathological description. Experimental results on the IU Chest X-ray (referred to as IU X-RAY) and the MIMIC-CXR benchmark datasets confirm the validity and effectiveness of our approach and demonstrate that our approach achieves better results than the State-of-the-Art methods. Two experts in the field were involved in this work. One is a radiologist with 30 years of experience, and the other is a doctor with an MBBS, MD, and Ph.D. and 2 years of experience in medicine.

Our contributions are:

1. A new approach to generating radiology reports: i) generating the tags for the input radiograph; ii) generating pathological descriptions from the tags generated in step (i); iii) identifying the spans in the normal report template to replace with the generated pathologi-

cal descriptions; and iv) replacing the identified span with the generated pathological description—improves the quality and accuracy of the radiology reports. Compared to the previous State-of-the-Art models, the BLEU-1, ROUGE-L, METEOR, and CIDEr scores of the pathological descriptions generated by our approach are raised by 25%, 36%, 44% and 48% respectively, on the IU X-RAY dataset.

2. A dataset of tags (tags are the disease keywords and radiological concepts associated with X-ray images) and their corresponding pathological descriptions. (Derived from IU X-RAY and MIMIC-CXR datasets containing 3827 and 44578 data points respectively.)
3. A dataset of the pathological descriptions and the corresponding normal sentences from the normal report template to replace with pathological descriptions (6501 data points).

2 Related Work

The topic of automatic report generation was researched by [Jing et al. \(2018\)](#); [Zhang et al. \(2017\)](#); [Yuan et al. \(2019\)](#). Several attempts have been made in the medical field to create medical reports from the corresponding images. Most researchers use multilabel image captioning to produce X-ray reports, and they subsequently use those captions as textual features. The IU X-ray dataset was created by [Demner-Fushman et al. \(2016\)](#) to generate radiology reports automatically. The IU X-RAY dataset's chest X-ray images were used to generate the first structured report using tags predicted by a CNN-RNN model ([Shin et al., 2016](#)). A system for generating natural reports for the Chest-Xray14 dataset, employing private reports, was presented by [Wang et al. \(2017\)](#). This framework used a non-hierarchical CNN-LSTM architecture and focused on semantic and visual aspects. Visual attention given to recurrent decoders and convolution-recurrent architectures (CNN-RNN) was first introduced by [Vinyals et al. \(2015\)](#) on image captioning.

Radiology report generation has recently shifted to transformer-based models ([Vaswani et al., 2017](#); [Devlin et al., 2018](#)). Knowledge-driven Encode, Retrieve, and Paraphrase (KERP) ([Li et al., 2019](#)) is used for accurate and reliable medical report generation. [Yuan et al. \(2019\)](#) suggests pretraining the encoder with several chest X-ray images to properly recognise 14 typical radiographic observations. According to [Chen et al. \(2020\)](#) proposals,

radiology reports can be generated using a memory-driven transformer, while [Pino et al. \(2021\)](#) suggests a template-based X-ray report generation approach. [Pino et al. \(2021\)](#) clinically defines the templates for each abnormality to indicate its presence or absence. If the generated tags indicate any disease, then the system will choose the corresponding abnormal sentence template. This method cannot generate patient-specific data like anatomical location, size, *etc.* [Wang et al. \(2021\)](#) proposed a template-based multi-attention report generation model (TMRGM) for normal and abnormal studies. This work uses template-based techniques exclusively for normal reports.

The differences between our work and previous work are as follows: i) Instead of generating the whole report at once, we generate smaller sentences for only abnormal findings and then replace them in the normal report template. ii) Unlike other state-of-the-art models, our methodology does not put excessive emphasis on normal sentences. iii) If the report's findings are normal, we use a standard template for normal reports.

3 Methodology

The approach that we follow to generate the radiology reports from the radiographs is as follows:

- Generate the tags for the input chest radiograph.
- Generate the pathological description from generated tags.
- Replace appropriate normal sentences (referred to as **normal description**) from the normal report template with the generated pathological descriptions.
 - Identify the span in the normal report template to replace.
 - Replace the identified span with the generated pathological description.

3.1 Model Architecture

Figure 1 illustrates the four main components of our model, which are the image tagger, text-generator (i.e., transformer), span identifier, and replacement module. The overall description of the four parts is provided below.

3.1.1 Image Tagger

Tagging X-ray images with multiple tags is a multi-label image classification problem. We build our multi-label classifier on top of convolutional neural

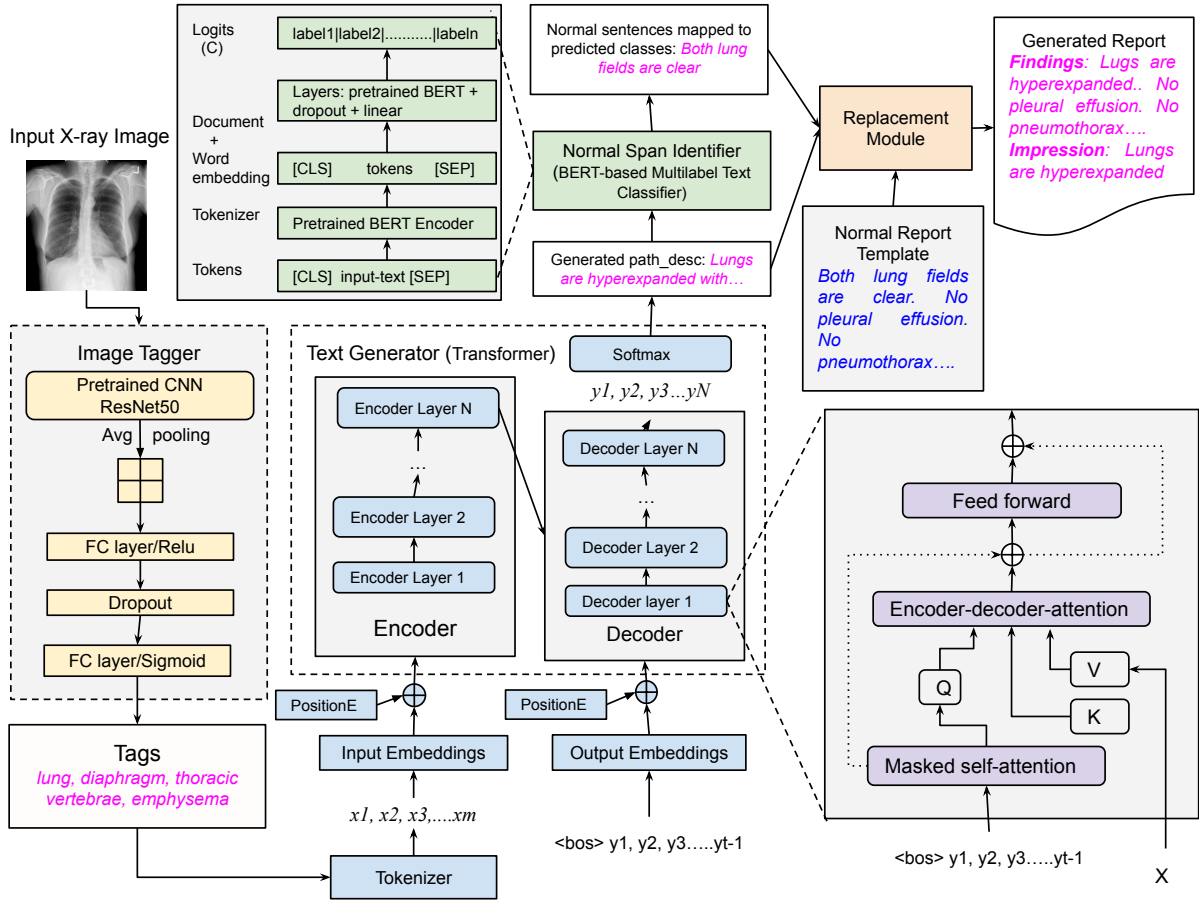


Figure 1: The architecture of our proposed model: Our model has four important components: an image tagger, a text generator, a span identifier, and a replacement module. The image tagger module produces tags for the input X-ray image. The text-generator module generates the pathological description for input tags. The span identifier module identifies the normal sentences that need to be replaced by generated pathological descriptions in the normal report template. The replacement module replaces identified normal spans with generated pathological descriptions.

network (CNN) architectures such as ResNet-50. Our model takes an X-ray image as input, calculates a score for each of the L labels, and then uses a cutoff threshold to decide which labels to keep. Target in each sample presented a binary vector $[y_1^j, y_2^j, \dots, y_L^j]$, $j = 1 \dots n$, $y \in Y$ and indicates whether a label is present (1) or absent (0). We use the sigmoid activation function in the output layer, and the binary cross-entropy loss function is used to fit the model. The loss is calculated as:

$$Loss = \frac{1}{L} \sum_{i=1}^n [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (1)$$

where \hat{y}_i is the i th predicted value by model, y_i is the corresponding target value, and L is the number of scalar values in the model output.

3.1.2 Tags Embedding

Input tag set X is tokenized as $\{x_1, x_2, \dots, x_{|s|}\}$. Input token embeddings is given by, $e_x = \{e_{x_1}, e_{x_2}, \dots, e_{x_{|n|}}\}$. For a token x_i , its embedding is $e_{x_i} \in R^d$, where d is the dimension of the token embeddings. The positional encoding is added to the input embeddings before passing it to the transformer layer. The text embeddings are the sum of the token embeddings and the positional embeddings, i.e., $e_{xp} = e_x + e_p$, where e_p is the positional embeddings.

3.1.3 Encoder-decoder Architecture

Generating pathological descriptions from tags can be looked at as a text generation problem. Given the input tag sequence $X = (x_1, x_2, \dots, x_s)$, the goal of general text generation model is to generate the pathological description sequence $Y =$

(y_1, y_2, \dots, y_T) , where S and T are the length of the input and output sequence respectively. The text generation model can be defined as follows:

$$p(Y|X) = \prod_{t=1}^T p(y_t|X, Y_{<t}) \quad (2)$$

Finally, it generates a pathological description with the highest probability. If the image is labeled as *normal* then the text generation module generates the sentence *No acute abnormality found*. Table 3 shows the samples from the dataset that include tags and their corresponding pathological descriptions.

3.1.4 Span Identification and Replacement Module

Span Identification: The span identification module identifies the span from the normal report template that would be replaced with a generated pathological description. To create a standard normal report template, we manually curated a set of sentences per abnormality indicating absence of abnormality (that means normal findings), totaling 23 sentences. With the expert’s opinion, we built the template by examining the reports and picking existing sentences or creating new ones. Due to the fixed nature of sentences, we treat each as a separate label. Figure 2 shows an example of the normal report template with a label for each sentence mentioned in brackets. We create a dataset of pathological description sentences, which we extract from the findings and impressions section of the original dataset, and annotate it with corresponding normal sentence labels (i.e., *lung1*, *lung2*, etc.). If a pathological description cannot be replaced with any of the normal sentences, then we annotate that sentence with the *extra* label.

This normal sentence identification problem can be formulated as a multilabel text classification problem. Consider input sentence as series of words: $X = \{w_1, w_2, \dots, w_n\}$ and for span identification, we need to predict the sentence category $Y = \{lung1, lung2, \dots, extra\}$. We use a BERT-based multilabel text classifier to identify the normal sentences. The last layer uses a sigmoid activation function to generate the probability of a sample belonging to the corresponding classes. The loss function is the same as equation 1.

Replacement: Once we obtain the list of normal sentences to replace with the generated pathological description, we submit the normal sen-

tences, the normal report template, and the generated pathological description to the replacement module. If a sentence is labeled as *extra* then we do not replace any normal sentences, but instead add pathological description sentences as an extra part in the report. For example, if the generated pathological description sentence is, *Multiple surgical clips are noted.*, then we add this as an extra part in the report without removing any normal sentences. Replacement module finds the exact match of the identified normal span sentences with normal report sentences and replaces it with generated pathological description. If the generated description is *No acute abnormality found*. then our replacement algorithm returns the standard normal report template as the generated report. If the span identifier gives multiple sentences to replace by a single sentence of pathological description, then the replacement module replaces the first sentence from the span and removes the remaining span sentences from the template. For example, if generated pathological description is *Stable cardiomegaly with large hiatal hernia.*, then we have to replace *No evidence of hernia (lung14)* and *Heart size is within normal limits. (heart1)*. Here, our replacement algorithm replaces *lung14* sentence with a generated pathological description and removes the *heart1* from the template.

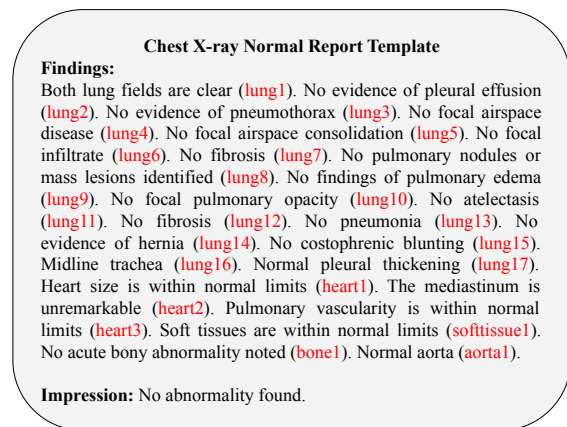


Figure 2: Chest X-ray normal report template. Each sentence is mapped with a unique label. Labels are highlighted in red.

4 Experiments

In this section we cover the datasets, evaluation metrics and baselines used for the training and evaluation of our model in detail.

4.1 Datasets

We conduct our experiments on two datasets, which are described as follows:

- **IU X-RAY** (Demner-Fushman et al., 2016) : a public radiography dataset compiled by Indiana University with 7,470 chest X-ray images and 3,955 reports. Each report has three parts: an impression, which is a title or summary of the report; findings, which contain the report in detail; and manual tags.
- **MIMIC-CXR** (Johnson et al., 2019) : the largest publicly available radiology dataset that consists of 473,057 chest X-ray images and 227,943 reports. For the purpose of our experiments we utilized 44578 reports.

Figure 3 shows the data samples from the IU-Chest X-ray dataset and the MIMIC-CXR dataset.

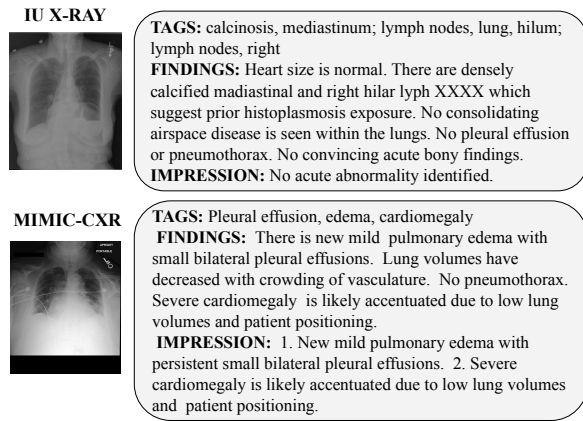


Figure 3: The data samples are from the IU X-ray and MIMIC-CXR datasets.

We use distinct fields from the original datasets to train different modules. In this section, we give details about experimental settings for each module separately.

4.1.1 Image Tagger Dataset

We consider frontal chest-radiographs as input and target as tags for the IU X-RAY dataset and CheXpert labels for MIMIC-CXR dataset. We build a model using Convolutional Neural Networks (CNN) to analyze each image and classify it with one or more of the 189 labels for IU X-Ray dataset and 14 labels for MIMIC-CXR dataset. Figure 4 and figure 5 shows the distribution of important classes in IU X-RAY dataset and MIMIC-CXR datasets, respectively. For the IU X-RAY dataset out of 3827 samples train, validation and test split

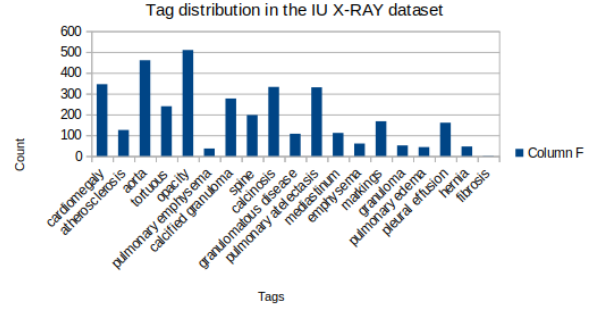


Figure 4: IU X-RAY dataset tag distribution. There are a total of 189 unique tags but here we represent only 19 important tags.

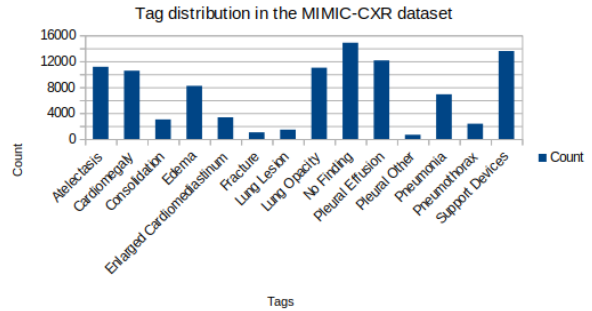


Figure 5: MIMIC-CXR dataset tag distribution with all 14 tags.

is 3000, 327 and 500 respectively. For the MIMIC-CXR dataset out of 44578 samples train, validation and test split is 40000, 2000 and 2578 respectively.

4.1.2 Pathological Description Generator Dataset

To train a model we derive two separate datasets from the original IU X-RAY and MIMIC-CXR datasets. Two tasks are involved in creating this dataset:

- We constructed one more dataset that includes unique sentences from the IU X-RAY dataset, and we annotated them as *normal* or *abnormal*. The binary classifier dataset contains 5000 samples, each labeled as *normal* or *abnormal* by domain experts. We then train a BERT-based binary classifier to classify each sentence as *normal* or *abnormal*.
- We use the trained model to classify each sentence from findings and impressions into *normal* and *abnormal* classes. We remove normal sentences for each report and consider only abnormal sentences as our pathological description.

Table 1 shows the examples of findings from the

Tags	Findings from the Original Dataset	Extracted Pathological Descriptions
calcinosis, abdomen, left, severe	The heart size and cardiomediastinal silhouette are stable and within normal limits. Pulmonary vasculature appears normal. There is no focal air space consolidation. No pleural effusion or pneumothorax. Extensive left upper quadrant splenic calcification may reflect old granulomatous disease	Extensive left upper quadrant splenic calcification may reflect old granulomatous disease
nodule, lung, base, calcinosis, lung, hilum, lymph nodes, right, granuloma, right	There is a 1 cm nodule within one of the lung bases, seen only on the lateral view. There is a calcified right hilar lymph node and right granuloma. Heart size is normal. No pneumothorax.	There is a 1 cm nodule within one of the lung bases, seen only on the lateral view. There is a calcified right hilar lymph node and right granuloma.
opacity, lung, apex, right, focal, opacity, lung, base, left, mild, spine, degenerative, cicatrix, lung, base, left, mild, pulmonary atelectasis, base, left, mild	The heart, pulmonary XXXX and mediastinum are within normal limits. There is no pleural effusion or pneumothorax. There is no focal air space opacity to suggest a pneumonia. There is a 1 cm focal opacity in the right lung apex incompletely evaluated by this exam. There is minimal left basilar XXXX opacity compatible with scarring or atelectasis. There are degenerative changes of the spine.	There is a 1 cm focal opacity in the right lung apex incompletely evaluated by this exam. There is minimal left basilar XXXX opacity compatible with scarring or atelectasis. There are degenerative changes of the spine.

Table 1: The samples are from the IU X-ray dataset, including the findings and the pathological descriptions extracted from it.

IU X-RAY dataset and extracted pathological descriptions from it. The train, validation, and test split is the same as the image tagger dataset.

4.1.3 Normal Span Identifier Dataset

Concept	Annotation labels	Concept	Annotation labels
cardiomegaly	heart1	airspace disease	lung1, lung4
heart size	heart1	infiltrate	lung1, lung6
hilar	heart2	nodule	lung1, lung8
sternotomy	bone1	pulmonary edema	lung1, lung9
kyphosis	bone1	clavicle	bone1
scoliosis	bone1	shoulder	bone1
pleural fluid	lung2	humeral	bone1
atelectasis	lung1, lung11	sternotomy	bone1
consolidation	lung1, lung5, lung13	spine	bone1
fibrosis	lung1, lung7	bronchial cuffing	lung1, lung9
pneumonia	lung1, lung13	bronchovascular crowding	lung1, lung11
costophrenic blunting	lung2, lung15	degenerative changes	bone1
bronchial	lung1	CABG	heart1
granuloma	lung1	scarring	lung1, lung6, lung7
COPD	lung1, lung4	interstitial prominence	lung1, lung6, lung7
interstitial marking	lung1		

Table 2: Any mention of a concept given in column one in a pathological sentence should be labeled with the labels given in column two.

We construct a dataset to identify the sentences to replace from the normal report template with the generated pathological description. The constructed dataset contains sentences from the findings and impressions of the IU X-RAY dataset and their corresponding list of normal sentences to replace. Annotation guidelines are provided by the domain expert. Table 2 shows the guidelines provided by experts. Using those guidelines, we annotate the pathological description sentences. Data annotated by us is verified by a domain expert and corrected if necessary. Table 3 shows the samples from the span identification dataset. Out of total

6500 samples, the train, validation, and test splits are 5000, 500, and 1000, respectively. Table 3 shows the samples from the multilabel text classification dataset that we have constructed.

Pathological Description	Labels
The thoracic aorta is tortuous and calcified.	aorta1
XXXX right pleural opacity along the lower chest wall compatible with thickening and/or some loculated effusion, accompanied with some adjacent atelectasis / airspace disease within the right lung base.	lung2, lung4, lung10, lung11, lung17
Stable cardiomegaly with large hiatal hernia.	lung14, heart1
Left greater than right basilar opacity, probable atelectasis and/or scarring.	lung10, lung11

Table 3: The samples are from the span identifier dataset. It includes the pathological descriptions and labels of the corresponding normal sentences to replace.

4.2 Training Details

We adopted the ResNet50 pretrained on Imagenet as the image tagger to produce tags for radiographs. We used frontal radiographs for both the IU X-RAY and the MIMIC-dataset. The IU X-RAY dataset contains 189 labels, and the MIMIC-CXR dataset contains 14 labels. According to the number of labels, the last layer of the image tagger module is different for both modules. We resize the images from both datasets to 224 x 224. We trained our model up to 20 epochs for both datasets. DGX A100-SXM-80GB GPU server takes 100 minutes for a single epoch for the MIMIC-CXR dataset and approximately 3 minutes for the IU X-RAY dataset. The transformer in our proposed model was initialized with pretrained T5-large model weights. The

Dataset	Method	NLG Metrics							CE Metrics		
		Bleu-1	Bleu-2	Bleu-3	Bleu-4	Rouge-L	Meteor	CIDEr	P	R	F
IU X-RAY	CDGPT2 (Alfarghaly et al., 2021)	0.387	0.245	0.166	0.111	0.289	0.164	-	0.0	0.0	0.0
	Visual Transformer (Chen et al., 2020)	0.470	0.304	0.219	0.165	0.371	0.187	-	-	-	-
	MvH+AttL+MC (Yuan et al., 2019)	<u>0.529</u>	<u>0.372</u>	<u>0.315</u>	<u>0.255</u>	0.453	0.343	-	-	-	-
	SentSAT+KG (Zhang et al., 2020)	0.441	0.291	0.203	0.147	0.367	-	0.304	0.483	0.490	0.478
	CoAttn (Jing et al., 2018)	0.517	0.386	0.306	0.247	0.447	0.217	0.327	0.491	0.503	0.491
	CNN-TRG (Pino et al., 2021)	0.273	-	-	-	0.352	-	0.249	<u>0.529</u>	<u>0.534</u>	<u>0.540</u>
	TMRGM (Wang et al., 2021)	0.419	0.281	0.201	0.145	0.280	0.183	0.359	-	-	-
	Ours (pathological description)	0.402	0.322	0.285	0.180	<u>0.567</u>	<u>0.455</u>	<u>0.473</u>	0.892	0.890	0.889
	Ours (full report)	0.775	0.699	0.658	0.627	0.817	0.782	0.835	0.533	0.874	0.648
MIMIC-CXR	Visual Transformer (Chen et al., 2020)	0.353	0.218	0.145	0.103	0.277	0.142	-	0.333	0.273	0.276
	SentSAT+KG (Zhang et al., 2020)	<u>0.441</u>	<u>0.291</u>	<u>0.203</u>	<u>0.147</u>	<u>0.367</u>	-	0.304	-	-	-
	CNN-TRG (Pino et al., 2021)	0.080	-	-	-	0.151	-	0.026	<u>0.668</u>	<u>0.749</u>	<u>0.640</u>
	Ours (pathological description)	0.253	0.188	0.169	<u>0.163</u>	0.348	<u>0.268</u>	<u>0.331</u>	0.769	0.771	0.765
	Ours (full report)	0.833	0.807	0.794	0.785	0.833	0.861	0.861	0.488	0.863	0.606

Table 4: The NLG metrics and CE metrics score of generated X-ray reports by previous methods and our approach vs. gold standard X-ray reports. The best results are in bold font, and the second best is underlined.

model was trained under cross entropy loss with the ADAM optimizer. We set the learning rate to $1e-4$. We decayed such a rate by a factor of 0.8 per epoch for each dataset and set the beam size to 5 to balance the generation’s effectiveness and efficiency. Maximum input and output sequence lengths were set to 100. We train the transformer model up to 20 epochs for IU X-RAY dataset and up to 15 epochs for the MIMIC-CXR dataset. For one epoch it takes approximately 10 minutes and 60 minutes on a single DGX A100-SXM-80GB GPU for the IU X-RAY and MIMIC-CXR datasets, respectively. The last module was a BERT-based multilabel classifier, which identifies the normal span to replace. We used pretrained BERT weights to initialize our model. There are a total of 24 labels, according to the number of nodes in the last layer change. We train all the models on a DGX A100-SXM-80GB GPU server. For all transformer based models we use hugging face transformer libraries.¹

4.3 Evaluation

We evaluate the results of different modules separately. For the image tagger and span identifier modules, we compare the accuracy, AUC score, precision, recall, and F1 scores. The first slot in the table 5 shows the results for the image tagger module, and the second slot shows the results for the span identifier module on the IU X-RAY and MIMIC-CXR datasets.

We compare the performance of our model with previous State-of-the-Art image captioning based

¹<https://huggingface.co/docs/transformers/index>

Module	Dataset	Acc	auROC	F1 score	Prec	Recall
Image Tagger	MIMIC-CXR	0.71	0.82	0.68	0.80	0.64
	IU X-RAY	0.75	0.79	0.61	0.71	0.58
Span Identifier	MIMIC-CXR	0.94	0.95	0.96	0.95	0.96
	IU X-RAY	0.96	0.95	0.96	0.96	0.97

Table 5: Results of the image tagger module and span identification module on the IU X-RAY and MIMIC-CXR datasets given by our model. First slot shows the results for image tagger and second slot shows the results for span identifier.

methods like CNN-RNN (Vinyals et al., 2015), CDGPT2 (Alfarghaly et al., 2021) and Visual Transformer (Chen et al., 2020) and template based methods such as CNN-TRG (Pino et al., 2021) and TMRGM (Wang et al., 2021). To evaluate the generated pathological descriptions, we consider the pathological descriptions that we extract from original reports as ground truth. To evaluate the generated full reports, we generate templated reports by replacing ground truth pathological description in normal report template and consider it as ground truth. The performance of the aforementioned models is evaluated by conventional natural language generation (NLG) metrics and clinical efficacy (CE) metrics. Clinical Efficacy (CE) metrics provides a quantitative assessment of the quality of generated radiology reports. Clinical efficacy (CE) metrics are calculated by comparing the critical radiology terminology extracted from the generated and reference reports. We use MIRQI² tool to calculate the precision, recall, and F1 scores to evaluate the model performance for these metrics. NLG metrics such as BLEU (Papineni et al.,

²<https://github.com/xiaosongwang/MIRQI>

2002), ROUGE (Lin, 2004), CIDEr and METEOR (Banerjee and Lavie, 2005), which primarily focus on measuring n-gram similarities. For CE metrics, precision, recall, and F1-score are used to evaluate model performance. We use the CheXpert (Irvin et al., 2019) to label the generated reports and compare the results with ground truths in 14 different categories of thoracic diseases and support devices. Table 4 shows NLG and CE metrics of the generated pathological description as well as full reports by our model and baseline models compared to gold standard reports.

4.4 Qualitative Evaluation and Error Analysis

This section provides a qualitative analysis performed by a domain expert. The domain expert classified the generated reports into three categories: **accurate** (reports with most of the vital information), **missing details** (reports with no false information but missing some vital details), and **misleading** (reports with false information and an overall incorrect diagnosis).

Method	Samples	Accurate	Missing Details	Misleading
Ours	All(500)	78.00%	12.00%	10.00%
	Normal(183)	99.95%	00.00%	00.05%
	Abnormal(317)	63.13%	22.00%	14.05%
CDGPT2	All(500)	61.60%	28.20%	10.20%
	Normal(201)	99.00%	00.00%	01.00%
	Abnormal(299)	36.50%	47.10%	16.40%

Table 6: Results of generated reports, manually evaluated by radiologist. Manual evaluation is done on the IU X-RAY dataset. Best results are shown in a bold face.

We provide 500 test samples from the test dataset and their corresponding generated reports to the domain expert for qualitative analysis. The model generated 78% accurate reports, 12% reports with missing information, and 10% with misleading predictions. Further, these random samples were classified into *normal* and *abnormal* reports. Out of 183 normal reports, the model generated 99.95% correct reports, 0.0% reports with missing details, and 0.05% misleading reports. Out of 317 abnormal reports, the model could produce 64% accurate reports, 22% of them with missing details, and 14.5% with false reports. Table 6 contains the results of the qualitative analysis. Figure 6 shows the case studies of ground truth reports and generated pathological descriptions and full reports by our method for the above mentioned categories.

5 Summary, Conclusion and Future Work

We present a template-based approach for generating X-ray reports from radiographs. Our model generates small sentences exclusively for abnormalities, which are then substituted in the normal report template to produce a high-quality radiology report. We create a replacement dataset that contains pathological descriptions and their corresponding normal sentences from the normal report template. Our experimental results demonstrate that, compared to the State-of-the-Art models, the BLEU-1, ROUGE-L, METEOR, and CIDEr scores of the full reports generated by our approach are raised by 25%, 36%, 44% and 48%, respectively. Also, clinical evaluation metrics show that our method generates more clinically accurate reports than the State-of-the-Art methods. Unlike other State-of-the-Art models, our methodology does not put excessive emphasis on normal sentences. In the future, we plan to apply the proposed method to generate radiology reports for CT, MRI, etc. For our experiments, we have used all samples from the IU X-RAY dataset. But we have used only 44578 reports out of 227827 reports for the MIMIC-CXR dataset. Our immediate plan is to perform experiments on the whole MIMIC-CXR dataset. Take away from our work is that creating smaller sentences of pathological descriptions and replacing them in the normal template produces better quality reports than generating the whole report at once.

Limitations

Data unbalancing is one of the limitations of our work. In the future, we would like to address this problem by data oversampling or undersampling. For our experiments, we have used all samples from the IU X-ray dataset. But from the MIMIC-CXR dataset, we have used only 44578 reports out of 227827 reports. Our results for the MIMIC-CXR dataset might differ when we use the whole dataset. To evaluate the generated pathological descriptions, we consider the pathological descriptions that we extract from original reports as ground truth. To evaluate the generated full reports, we generate templated reports by replacing ground truth pathological description in normal report template and consider it as ground truth. So it considers the abnormalities from the original reports and the normal sentences from the normal report template. Automatic generation of chest X-ray reports will make it easier for radiologists to diagnose and write reports.

Ground Truth	Generated Pathological Description	Generated Full Report
The heart is normal in size. Atherosclerotic calcifications of the aorta . The mediastinum is stable. There is again soft tissue density projected over the right mid chest, XXXX patient's known large breast mass. The appearance is grossly stable to decreased from prior study. The lateral projection is suboptimal as patient could not raise XXXX. There is no pleural effusion.	Atherosclerotic calcifications of the aortic XXXX are noted.	Both lung fields are clear. No evidence of pleural effusion. No evidence of pneumothorax. No focal airspace disease. No focal airspace consolidation. No focal infiltrate. No fibrosis. No pulmonary nodules or mass lesions identified. No findings of pulmonary edema. No focal pulmonary opacity. No atelectasis. No fibrosis. No pneumonia. No evidence of hernia. No costophrenic blunting. Midline trachea. Normal pleural thickening. Heart size is within normal limits. The mediastinum is unremarkable. Pulmonary vascularity is within normal limits. Soft tissues are within normal limits. No bony abnormalities. Atherosclerotic calcifications of the aortic XXXX are noted.
The heart size is normal. There is minimal fibronodular scarring right apex, otherwise lungs clear.	There is a 1.2 cm nodule in the right upper lobe. There is a 1.2 cm nodule in the right apex. There is a 1.3 cm nodule in the right upper lobe. There is XXXX scarring.	There is a 1.2 cm nodule in the right upper lobe. There is a 1.2 cm nodule in the right apex. There is a 1.3 cm nodule in the right upper lobe. No evidence of pleural effusion. No evidence of pneumothorax. No focal airspace disease. No focal airspace consolidation. There is XXXX scarring. No findings of pulmonary edema. No focal pulmonary opacity. No atelectasis. No fibrosis. No pneumonia. No evidence of hernia. No costophrenic blunting. Midline trachea. Normal pleural thickening. Heart size is within normal limits. The mediastinum is unremarkable. Pulmonary vascularity is within normal limits. Soft tissues are within normal limits. No bony abnormalities. Normal aorta.
Heart size is normal. Lungs are clear. Calcified 5 mm granuloma in the left midlung. Status post resection of left upper lobe. No adenopathy, nodules, or masses. No effusion.	Calcified left upper lobe granuloma.	Calcified left upper lobe granuloma. No evidence of pleural effusion. No evidence of pneumothorax. No focal airspace disease. No focal airspace consolidation. No focal infiltrate. No fibrosis. No pulmonary nodules or mass lesions identified. No findings of pulmonary edema. No focal pulmonary opacity. No atelectasis. No fibrosis. No pneumonia. No evidence of hernia. No costophrenic blunting. Midline trachea. Normal pleural thickening. Heart size is within normal limits. The mediastinum is unremarkable. Pulmonary vascularity is within normal limits. Soft tissues are within normal limits. No bony abnormalities. Normal aorta.
Cardiomediastinal silhouette is within normal limits for size and contour. Lungs are hyperinflated with flattening of the diaphragms consistent with emphysematous change. No evidence of focal airspace disease, pleural effusion, or pneumothorax. Multilevel degenerative changes of the spine are noted.	Hyperinflated lungs with flattened diaphragms consistent with emphysema. Degenerative changes in the spine.	Hyperinflated lungs with flattened diaphragms consistent with emphysema. No evidence of pleural effusion. No evidence of pneumothorax. No focal airspace disease. No focal airspace consolidation. No focal infiltrate. No fibrosis. No pulmonary nodules or mass lesions identified. No findings of pulmonary edema. No focal pulmonary opacity. No atelectasis. No fibrosis. No pneumonia. No evidence of hernia. No costophrenic blunting. Midline trachea. Normal pleural thickening. Heart size is within normal limits. The mediastinum is unremarkable. Pulmonary vascularity is within normal limits. Soft tissues are within normal limits. Degenerative changes in the spine. Normal aorta.

Figure 6: Examples of the ground truth and the pathological descriptions and full reports generated using our approach. The first column shows the findings from the IU X-RAY dataset. Abnormal findings in the original report are highlighted in magenta. The second column shows the pathological description generated by our system. The third column shows the full report generated by our method. For both second and third column, correctly generated sentences are highlighted in green, partially correct sentences are highlighted in blue and misleading sentences are highlighted in red. Example 1 shows that the generated report is correct but missing some important information. Example 2 shows that the generated report is misleading. Example 3 shows that the generated report is correct but missing the measurements. Example 4 shows that the generated report is correct and reports all findings.

Our model achieved comparable performance with State-of-the-Art models on chest X-ray report generation. In realistic scenarios, it is still a long way from being used clinically.

Ethics Statement

The IU Chest X-ray dataset's authors used appropriate techniques to de-identify the text reports. Data is anonymized; hence, our model will not disclose information about the patient's identity. The MIMIC-CXR dataset does not contain any information related to the patient's identity, like name, age, or address. These reports are also anonymized. Data itself does not reveal the patient's identity; hence, our model also does not reveal the patient's

identity.

Acknowledgements

We thank the radiologist, Dr. Milind Gune, for his helpful suggestions and feedback.

References

- Omar Alfarghaly, Rana Khaled, Abeer Elkorany, Maha Helal, and Aly Fahmy. 2021. Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, 24:100557.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image

- captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.
- Satanjeev Banerjee and Alon Lavie. 2005. **METEOR: An automatic metric for MT evaluation with improved correlation with human judgments**. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*.
- 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.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.
- Baoyu Jing, Pengtao Xie, and Eric Xing. 2017. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*.
- Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. **On the automatic generation of medical imaging reports**. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2577–2586, Melbourne, Australia. Association for Computational Linguistics.
- Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.
- Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2019. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6666–6673.
- Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. *Advances in neural information processing systems*, 31.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- 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.
- Pablo Pino, Denis Parra, Cecilia Besa, and Claudio Lagos. 2021. Clinically correct report generation from chest x-rays using templates. In *International Workshop on Machine Learning in Medical Imaging*, pages 654–663. Springer.
- Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. 2016. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2497–2506.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.
- Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106.
- Xuwen Wang, Yu Zhang, Zhen Guo, and Jiao Li. 2021. Tmrgm: A template-based multi-attention model for x-ray imaging report generation. *Journal of Artificial Intelligence for Medical Sciences*, 2(1-2):21–32.
- 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 721–729. Springer.
- Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. 2020. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12910–12917.
- Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. 2017. Mdnnet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6428–6436.

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
Section: Limitations
- A2. Did you discuss any potential risks of your work?
Limitations
- A3. Do the abstract and introduction summarize the paper’s main claims?
Abstract, 1, 5
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

4.1, 4.1.2, 4.1.3

- B1. Did you cite the creators of artifacts you used?
4.1, 4.1.1, 4.1.2, 4.1.3
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
4.1
- 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)?
4.1
- 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?
Provided data is already in anonymized form.
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
4.1
- 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.
4.1

C Did you run computational experiments?

4

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4.2

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?
4.2
- 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?
4.3, 4.4
- 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.)?
4
- D** **Did you use human annotators (e.g., crowdworkers) or research with human participants?**
4
- 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.?
4.1.3
- 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)?
Annotator is researcher and co-author
- 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?
4.1
- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Section: Ethic statement
- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Annotator is researcher and co-author