Automated Generation of Accurate & Fluent Medical X-ray Reports

Hoang T.N. Nguyen*
University of Alberta

Dong Nie*
University of North Carolina at Chapel Hill

Taivanbat Badamdorj
University of Alberta

Yujie Liu
Guangzhou University of
Chinese Medicine

Yingying Zhu
University of Texas
at Arlington

Jason Truong
University of Alberta

Li Cheng
University of Alberta

Abstract

Our paper aims to automate the generation of medical reports from chest X-ray image inputs, a critical yet time-consuming task for radiologists. Existing medical report generation efforts emphasize producing human-readable reports, yet the generated text may not be well aligned to the clinical facts. Our generated medical reports, on the other hand, are fluent and, more importantly, *clinically accurate*. This is achieved by our fully differentiable and end-to-end paradigm that contains three complementary modules: taking the chest X-ray images and clinical history document of patients as inputs, our *classification* module produces an internal checklist of disease-related topics, referred to as enriched disease embedding; the embedding representation is then passed to our transformer-based *generator*, to produce the medical report; meanwhile, our generator also creates a weighted embedding representation, which is fed to our *interpreter* to ensure consistency with respect to disease-related topics. Empirical evaluations demonstrate very promising results achieved by our approach on commonly-used metrics concerning language fluency and clinical accuracy. Moreover, noticeable performance gains are consistently observed when additional input information is available, such as the clinical document and extra scans from different views.

1 Introduction

Medical reports are the primary medium, through which physicians communicate findings and diagnoses from the medical scans of patients. The process is usually laborious, where typing out a medical report takes on average five to ten minutes (Jing et al., 2018); it could also be error-prone. This has led to a surging need for automated generation of medical reports, to assist radiologists and physicians in making rapid and meaningful diagnoses.

Its potential efficiency and benefits could be enormous, especially during critical situations such as COVID or a similar pandemic. Clearly a successful medical report generation process is expected to possess two key properties: 1) clinical accuracy, to properly and correctly describe the disease and related symptoms; 2) language fluency, to produce realistic and human-readable text.

Fueled by recent progresses in the closely related computer vision problem of image-based captioning (Vinyals et al., 2015; Tran et al., 2020), there have been a number of research efforts in medical report generation in recent years (Jing et al., 2018, 2019; Li et al., 2018, 2019; Xue et al., 2018; Yuan et al., 2019; Wang et al., 2018; Yin et al., 2019; Lovelace and Mortazavi, 2020; Srinivasan et al., 2020). These methods often perform reasonably well in addressing the language fluency aspect; on the other hand, as is also evidenced in our empirical evaluation, their results are notably less satisfactory in terms of clinical accuracy. This we attribute to two reasons: one is closely tied to the textual characteristic of medical reports, which typically consists of many long sentences describing various disease related symptoms and related topics in precise and domain-specific terms. This clearly sets the medical report generation task apart from a typical image-to-text problem such as image-based captioning; another reason is related to the lack of full use of rich contextual information that encodes prior knowledge. These information include for example the patient's clinical document describing key clinical history and indication from doctors, and multiple scans from distinct 3D views – information that are typically existed in abundance in practical scenarios, as e.g. in the standard X-ray benchmarks of Open-I (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019).

The aforementioned observations motivate us to propose a categorize-generate-interpret framework that places specific emphasis on clinical accuracy

---

* indicates equal contribution. Code is available at https://github.com/ginobilinie/xray_report_generation
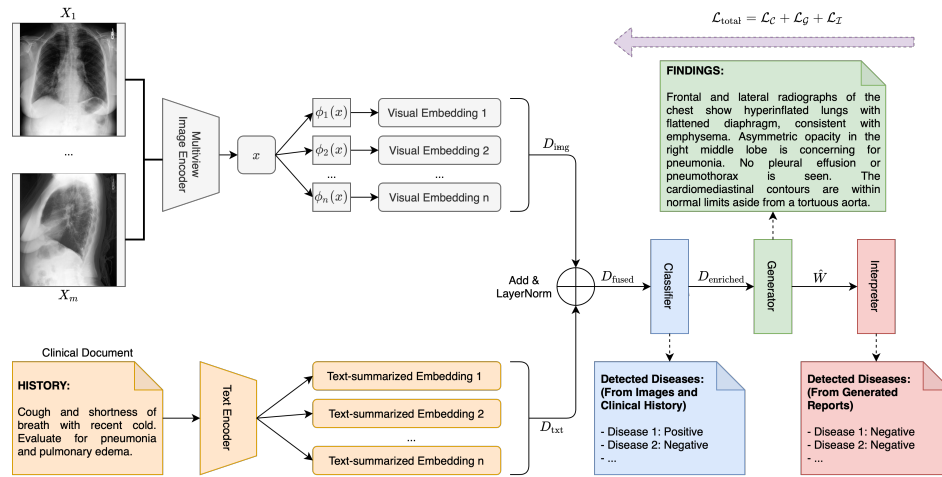
Figure 1: Our approach consists of three modules: a *classifier* that reads chest X-ray images and clinical history to produce an internal checklist of disease-related topics, a transformer-based *generator* to generate fluent text, and an *interpreter* to examine and fine-tune the generated text to be consistent with the disease-related topics.

while maintaining adequate language fluency of the generated reports. It consists of a classifier module reads chest X-ray images (e.g., either single-view or multi-view images) and related documents to detect diseases and output enriched disease embedding, a transformer-based medical report generator, and a differentiable interpreter to evaluate and fine-tune the generated reports for factual correctness. The main contributions are two-fold:

- A differentiable end-to-end approach is proposed, consisting of three modules (classifier-generator-interpreter): the classifier module learns the disease feature representation via context modeling (section 3.1.3) and disease-state aware mechanism (section 3.1.4); the generator module transforms the disease embedding to medical report; the interpreter module reads and fine-tunes the generated reports, enhancing the consistency of the generated reports and the classifier's outputs.

- Empirically our approach is shown to outperform against a number of strong baselines over two widely-used benchmarks on an equal footing (i.e. without accessing to additional information). Moreover, empirical evidence demonstrates clinical patient history as well as additional scans may play a vital role in improving the quality of the generated reports.

2 Related Work

2.1 Image-based Captioning and Medical Report Generation

Apart from some familiar topics such as disease detection (Oh et al., 2020; Luo et al., 2020; Lu et al., 2020b; Rajpurkar et al., 2017; Lu et al., 2020a; Ranjan et al., 2018) and lung segmentation (Eslami et al., 2020), the most related computer vision task is the emerging topic of image-based captioning, which aims at generating realistic sentences or topic-related paragraphs to summarize visual contents from images or videos (Vinyals et al., 2015; Xu et al., 2015; Goyal et al., 2017; Rennie et al., 2017; Huang et al., 2019; Feng et al., 2019; Pei et al., 2019; Tran et al., 2020). Not surprisingly, the recent progresses in medical report generation (Jing et al., 2018, 2019; Li et al., 2018, 2019; Xue et al., 2018; Yuan et al., 2019; Wang et al., 2018; Yin et al., 2019; Lovelace and Mortazavi, 2020; Srinivasan et al., 2020; Zhang et al., 2020; Huang et al., 2021; Gasimova et al., 2020; Singh et al., 2019; Nishino et al., 2020) have been particularly influenced by the successes in image-based captioning.

The work of (Vinyals et al., 2015; Xu et al., 2015) is among the early approaches in medical report generation, where visual features are extracted by convolution neural networks (CNNs); they are subsequently fed into recurrent neural networks (RNNs) to generate textual descriptions. In remedying the issue of inaccurate textual descriptions, a secondary task is explicitly adopted by (Jing et al., 2018; Srinivasan et al., 2020) to select top-k most likely diseases to gauge report generation. The

methods of (Jing et al., 2019; Li et al., 2018), on the other hand, consider a reinforcement learning process to promote generating reports with correct contents. It has been noted by (Jing et al., 2018, 2019; Li et al., 2018) that traditional RNNs are not well suited in generating long sentences and paragraphs (Vaswani et al., 2017; Krause et al., 2017), which renders them insufficient in medical report generation task (Jing et al., 2018). This issue is relieved by either conceiving hierarchical RNN architectures (Krause et al., 2017) (Jing et al., 2018, 2019; Li et al., 2018; Xue et al., 2018; Yuan et al., 2019; Wang et al., 2018; Yin et al., 2019), or resorting to alternative techniques including in particular the recently developed transformer architectures (Vaswani et al., 2017) (Srinivasan et al., 2020; Lovelace and Mortazavi, 2020).

It is worth noting that most existing methods concentrate on the image-to-fluent-text aspect of the medical report generation problem; on the other hand, their results are considerably less well-versed at uncovering the intended disease and symptom related topics in the generated texts, the true gems where the physicians would base their decisions upon. To alleviate this issue, a graph-based approach is considered in (Li et al., 2019): it starts by compiling a list of common abnormalities, then transforms them into correlated disease graphs, and categorizes medical reports into templates for paraphrasing. Its practical performance is however less stellar, which may be credit to the fact that (Li et al., 2019) is fundamentally based on detecting abnormalities from medical images, thus may overlook other important information.

2.2 Transformers

The transformer technique (Vaswani et al., 2017) is first introduced in the context of machine translation with the purpose of expediting training and improving long-range dependency modeling. They are achieved by processing sequential data in parallel with an attention mechanism, consisting of a multi-head self-attention module and a feedforward layer. By considering multi-head self-attention mechanisms, including e.g. a graph attention network (Velickovic et al., 2018), recent transformer-based models have shown considerable advancement in many difficult tasks, such as image generation (Chen et al., 2020), story generation (Radford et al., 2018), question answering, and language inference (Devlin et al., 2019).

2.3 CheXpert Labeler

The CheXpert labeler (Irvin et al., 2019) is a rule-based system that extracts and classifies medical reports into 14 common diseases. Each disease label is either positive, negative, uncertain, or unmentioned. This is a crucial part in building large-scale chest X-ray datasets, such as (Irvin et al., 2019; Johnson et al., 2019), where an alternative manual labeling process may take years of effort. It could also be used to evaluate the clinical accuracy of a generated medical report (Liu et al., 2019). Another important use of the CheXpert labeler is to facilitate the generation of medical reports. Since the rule-based CheXpert labeler is not differentiable, it is regarded as a score function estimator for reinforcement learning models (Liu et al., 2019) to fine-tune the generated texts. However, the reinforcement learning methods are often computationally expensive and practically difficult to convergence. As an alternative, Lovelace et al. (Lovelace and Mortazavi, 2020) propose an attention LSTM model and fine-tune the generated report via a differentiable Gumbel random sampling trick, with promising results.

3 Our Approach

Our framework consists of a *classification* module, a *generation* module, and an *interpretation* module, as illustrated in Fig. 1. The classification module reads multiple chest X-ray images and extracts the global visual feature representation via a multiview image encoder. They are then disentangled into multiple low-dimensional visual embedding. Meanwhile, the text encoder reads clinical documents, including, e.g., doctor indication, and summarizes the content into text-summarized embedding. The visual and text-summarized embeddings are entangled via an "add & layerNorm" operation to form contextualized embedding in terms of disease-related topics. The generation module takes our enriched disease embedding as initial input and generates text word-by-word, as shown in Fig. 2. Finally, the generated text is fed to the interpretation module for fine-tuning to align to the checklist of disease-related topics from the classification module. In what follows, we are to elaborate on these three modules in detail.
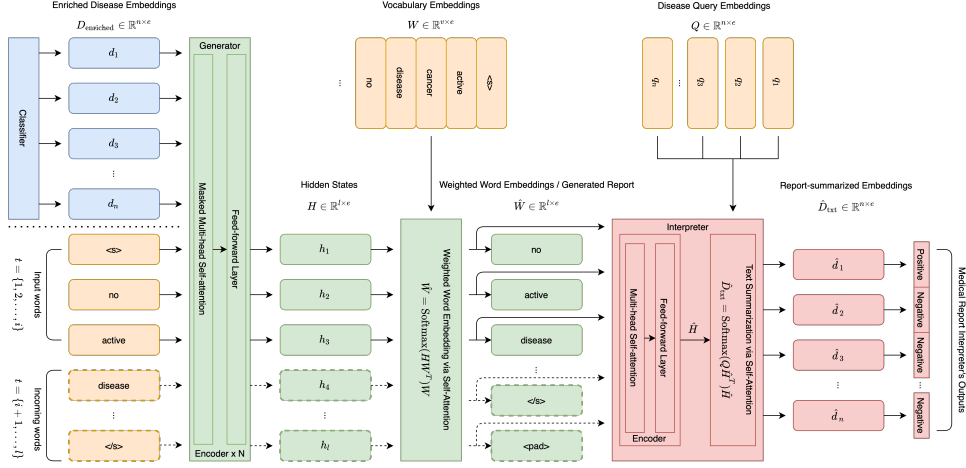
Figure 2: An exemplar illustration of our approach in action. Specifically, the enriched disease embedding produced from the classification module are fed into the generation module as initial inputs. Then, at each time step, the hidden state h_i is obtained to predict the next output word. Finally, the interpretation module takes as input all predicted outputs $\hat{W}$ to produce a checklist of disease-related topics, which are to be gauged with the same topics output from the classification module for consistency verification.

3.1 The Classification Module

3.1.1 Multi-view Image Encoder

For each medical study which consists of m chest X-ray images $\{X_i\}_{i=1}^m$, we extract the corresponding latent features $\{x_i\}_{i=1}^m \in \mathbb{R}^c$, where c is the number of features, via a shared DenseNet-121 image encoder (Huang et al., 2017). Then, the multi-view latent features $x \in \mathbb{R}^c$ can be obtained by max-pooling across the set of m latent features $\{x_i\}_{i=1}^m$, as proposed in (Su et al., 2015). When $m = 1$, the multi-view encoder boils down to a single-image encoder.

3.1.2 Text Encoder

Let T be a text document with length l consisting of word embeddings $\{w_1, w_2, ..., w_l\}$, where $w_i \in \mathbb{R}^e$ embodies the i-th word in the text and e is the embedding dimension. We use the transformer encoder (Vaswani et al., 2017) as our text feature extractor to retrieve a set of hidden states $H = \{h_1, h_2, ..., h_l\}$, where $h_i \in \mathbb{R}^e$ is the attended features of the i-th word to other words in the text,

$$h_i = \text{Encoder}(w_i|w_1, w_2, ..., w_l). \quad (1)$$

The entire document T is then summarized by $Q = \{q_1, q_2, ..., q_n\}$, representing n disease-related topics (e.g., pneumonia or atelectasis) to be queried from the document. We refer to this retrieval process as *text-summarized embedding* $D_{\text{txt}} \in \mathbb{R}^{n \times e}$,

$$D_{\text{txt}} = \text{Softmax}\left(QH^\mathsf{T}\right)H. \quad (2)$$

Here matrix $Q \in \mathbb{R}^{n \times e}$ is formed by stacking the set of vectors $\{q_1, q_2, ..., q_n\}$ where $q_i \in \mathbb{R}^e$ is randomly initialized, then learned via the attention process. Similarly, the matrix $H \in \mathbb{R}^{l \times e}$ is formed by $\{h_1, h_2, ..., h_l\}$ from Eq. (1). The term $\text{Softmax}(QH^\mathsf{T})$ is the word attention heat-map for the n queried diseases in the document. The intuition here is for each disease (e.g., pneumonia) to be queried from the text document T. We only pay attention to the most relevant words (e.g., *cough* or *shortness of breath*) in the text that associates with that disease, also known as a vector similarity dot product. This way, the weighted sum of these words by Eq. (2) gives the feature that summarizes the document w.r.t. the queried disease.

3.1.3 Contextualized Disease Embedding

The latent visual features $x \in \mathbb{R}^c$ are subsequently decoupled into low-dimensional disease representations, as illustrated in Fig. 1. They are regarded as the *visual embedding* $D_{\text{img}} \in \mathbb{R}^{n \times e}$, where each row is a vector $\phi_j(x) \in \mathbb{R}^e, j = 1, \ldots, n$ defined as follows:

$$\phi_j(x) = A_j^\mathsf{T} x + b_j. \quad (3)$$

Here $A_j \in \mathbb{R}^{c \times e}$ and $b_j \in \mathbb{R}^e$ are learnable parameters of the j-th disease representation. n is the number of disease representations, and e is the embedding dimension. Now, together with the available clinical documents, the visual embedding D_{img} and the text-summarized embedding D_{txt} are entangled to form *contextualized disease represen-*

tations $D_{\text{fused}} \in \mathbb{R}^{n \times e}$ as

$$D_{\text{fused}} = \text{LayerNorm}(D_{\text{img}} + D_{\text{txt}}). \quad (4)$$

Intuitively, the entanglement of visual and textual information allows our model to mimic the hospital workflow, to screen the disease's visual representations conditioned on the patients' clinical history or doctors' indication. For example, the doctor's indication in Fig. 1 shows *cough* and *shortness of breath* symptoms. It is reasonable for a medical doctor to request a follow-up check of the *pneumonia* disease. As for the radiologists receiving the doctors' indication, they may prioritize diagnosing the presence of *pneumonia* and related diseases based on X-ray scans and look for specific abnormalities. As empirically shown in Table 4, the proposed contextualized disease representations bring a significant performance boost in the medical report generation task. Meanwhile, our current embedding is basically a plain mingling of heterogeneous sources of information such as disease type (i.e., disease name) and disease state (e.g., positive or negative). As shown by the ablation study in Table 4, this embedding by itself is insufficient for generating accurate medical reports. This leads us to conceive a follow-up enriched representation below.

3.1.4 Enriched Disease Embedding

The main idea behind *enriched disease embedding* is to further encode informative attributes about disease states, such as *positive*, *negative*, *uncertain*, or *unmentioned*. Formally, let k be the number of states and $S \in \mathbb{R}^{k \times e}$ the state embedding. Then the confidence of classifying each disease into one of the k disease states is

$$p = \text{Softmax}(D_{\text{fused}} S^{\mathsf{T}}). \quad (5)$$

$S \in \mathbb{R}^{k \times e}$ is randomly initialized, then learned via the classification of D_{fused}. D_{fused} acts as features for the multi-label classification, and the classification loss is computed as

$$\mathcal{L}_{\mathcal{C}} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} y_{ij} \log(p_{ij}), \quad (6)$$

where $y_{ij} \in \{0, 1\}$ and $p_{ij} \in (0, 1)$ are the j-th ground-truth and predicted values for the disease i-th, respectively. The state-aware embedding $D_{\text{states}} \in \mathbb{R}^{n \times e}$ are then computed as

$$D_{\text{states}} = \begin{cases} yS, & \text{if training phase} \\ pS, & \text{otherwise.} \end{cases} \quad (7)$$

$y \in \{0, 1\}^{n \times k}$ is the one-hot ground-truth labels about the disease-related topics, whereas $p \in (0, 1)^{n \times k}$ is the predicted values. During training, the ground-truth disease states facilitate our generator in describing the diseases & related symptoms based on accurate information (teacher forcing). At test time, our generator then furnishes its recount based on the predicted states.

Finally, the *enriched disease embedding* $D_{\text{enriched}} \in \mathbb{R}^{n \times e}$ is the composition of state-aware disease embedding D_{states} (i.e., good or bad), disease names D_{topics} (i.e., which disease/topic), and the disease representations D_{fused} (i.e., severity and details of the diseases),

$$D_{\text{enriched}} = D_{\text{states}} + D_{\text{topics}} + D_{\text{fused}}. \quad (8)$$

Like the disease queries Q, $D_{\text{topics}} \in \mathbb{R}^{n \times e}$ is randomly initialized, representing diseases or topics to be generated. It is then learned in training through the medical report generation pipeline. The enriched disease embedding provides explicit and precise disease descriptions, and endows our follow-up generation module with a powerful data representation.

3.2 The Generation Module

Our report generator is derived from the transformer encoder of (Vaswani et al., 2017). The network is formed by sandwiching & stacking a masked multi-head self-attention component and a feed-forward layer being on top of each other for N times, as illustrated in Fig. 2. The hidden state for each word position $h_i \in \mathbb{R}^e$ in the medical report is then computed based on previous words and disease embedding, as $D_{\text{enriched}} = \{d_i\}_{i=1}^{n}$,

$$h_i = \text{Encoder}(w_i | w_1, w_2, ..., w_{i-1}, d_1, d_2, ..., d_n). \quad (9)$$

This is followed by predicting future words based on the hidden states $H = \{h_i\}_{i=1}^{l} \in \mathbb{R}^{l \times e}$, as

$$p_{\text{word}} = \text{Softmax}(HW^{\mathsf{T}}). \quad (10)$$

Here $W \in \mathbb{R}^{v \times e}$ is the entire vocabulary embedding, v the vocabulary size, and l the document length. Let $p_{\text{word},ij}$ denote the confidence of selecting the j-th word in the vocabulary W for the i-th position in the generated medical report. The generator loss is defined as a cross entropy of the ground-truth words y_{word} and predicted words p_{word},

$$\mathcal{L}_{\mathcal{G}} = -\frac{1}{l} \sum_{i=1}^{l} \sum_{j=1}^{v} y_{\text{word},ij} \log(p_{\text{word},ij}). \quad (11)$$

3556

Finally, the *weighted word embedding* $\hat{W} \in \mathbb{R}^{l \times e}$, also known as the *generated report*, are:

$$\hat{W} = p_{\text{word}} W. \tag{12}$$

It is worth noting that this set-up facilitate the back-propagation of errors from the follow-up interpretation module.

3.3 The Interpretation Module

It is observed from empirical evaluations that the generated reports are often distorted in the process, such that they become inconsistent with the original output of the classification module – the enriched disease embedding that encodes the disease and symptom related topics. Inspired by the CycleGAN idea of (Zhu et al., 2017), we consider a fully differentiable network module to estimate the checklist of disease-related topics based on the generator's output, and to compare with the original output of the classification module. This provides a meaningful feedback loop to regulate the generated reports, which is used to fine-tune the generated report through the word representation outputs $\hat{W}$.

Specifically, we build on top of the proposed text encoder (described in section 3.1.2) a classification network that classifies disease-related topics, as follows. First, the text encoder summarizes the current medical report $\hat{W}$, and outputs the report-summarized embedding of the queried diseases Q,

$$\hat{D}_{\text{txt}} = \text{Softmax}(Q\hat{H}^{\mathsf{T}})\hat{H} \in \mathbb{R}^{n \times e}. \tag{13}$$

Here $\hat{H}$ is computed from the generated medical reports $\hat{W}$ using Eq. (1). Second, each of the report-summarized embedding $\hat{d}_i \in \mathbb{R}^e$ (i.e., each row of the matrix $\hat{D}_{\text{txt}} \in \mathbb{R}^{n \times e}$) is classified into one of the k disease-related states (i.e., positive or negative), as

$$p_{\text{int}} = \text{Softmax}(\hat{D}_{\text{txt}} S^{\mathsf{T}}) \in \mathbb{R}^{n \times k}. \tag{14}$$

Finally, the interpreter is trained to minimize the subsequent multi-label classification loss,

$$\mathcal{L}_{\mathcal{I}} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} y_{ij} \log(p_{\text{int},ij}). \tag{15}$$

here $y_{ij} \in \{0, 1\}$ is the ground-truth disease label and $p_{\text{int},ij} \in (0, 1)$ is the predicted disease label of the interpreter.

In fine-tuning the generated medical reports $\hat{W}$, all interpreter parameters are frozen, which acts as a guide to force the word representations $\hat{W}$ being close to what the interpreter has learned from the ground-truth medical reports. If the weighted word embedding $\hat{W}$ is different from the learned representation – which leads to incorrect classification – a large loss value will be imposed in the interpretation module. This thus forces the generator to move toward producing a correct word representation.

Collectively our model is trained in an end-to-end manner by jointly minimizing the total loss,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\mathcal{C}} + \mathcal{L}_{\mathcal{G}} + \mathcal{L}_{\mathcal{I}}. \tag{16}$$

4 Experiments

This section evaluates the medical report generation task on two fronts: the language performance and the clinical accuracy performance. Empirical evaluations are carried out on two widely-used chest X-ray datasets, MIMIC-CXR (Johnson et al., 2019) and Open-I (Demner-Fushman et al., 2016).

4.1 Datasets

4.1.1 MIMIC-CXR Dataset

The MIMIC-CXR dataset (Johnson et al., 2019) is a large-scale dataset with 227,835 medical reports of 65,379 patients, associated with 377,110 images from multiple views: anterior-posterior (AP), posterior-anterior (PA), lateral (LA). Each study comprises multiple sections, including *comparison*, *clinical history*, *indication*, *reasons for examination*, *impressions*, and *findings*. Here we utilize the multi-view images of AP/PA/LA views, and adopt as contextual information the concatenation of the *clinical history*, *reason for examination*, and *indication* sections. For consistency, we follow the experimental set-up of (Lovelace and Mortazavi, 2020) to focus on generating text in the "findings" section as the corresponding medical report.

4.1.2 Open-I Dataset

The Open-I dataset (Demner-Fushman et al., 2016) collected by the Indiana University hospital network contains 3,955 radiology studies that correspond to 7,470 frontal and lateral chest X-rays. Some radiology studies are associated with more than one chest X-ray image. Each study typically consists of *impression*, *findings*, *comparison*, and *indication* sections. Similar to the MIMIC-CXR dataset, we utilized both the multi-view chest X-ray images (frontal and lateral) and the *indication* section as our contextual inputs. For generating medi-

Datasets	Methods	B-1	B-2	B-3	B-4	MTR	RG-L	SV	MV	AI	FT
Open-I	S&T (Vinyals et al., 2015)	0.316	0.211	0.140	0.095	0.159	0.267	x			
	LRCN (Donahue et al., 2015)	0.369	0.229	0.149	0.099	0.155	0.278	x			
	SA&T (Xu et al., 2015)	0.399	0.251	0.168	0.118	0.167	0.323	x			
	Att-RK (You et al., 2016)	0.369	0.226	0.151	0.108	0.171	0.323	x			
	HRNN (Yin et al., 2019)	0.445	0.292	0.201	0.154	0.175	0.344	x			
	1-NN (Boag et al., 2020)	0.232	0.116	0.051	0.018	N/A	0.201	x			
	TieNet (Wang et al., 2018)	0.330	0.194	0.124	0.081	N/A	0.311	x			
	Liu et. al. (Liu et al., 2019)	0.359	0.237	0.164	0.113	N/A	0.354	x			x
	CoAtt (Jing et al., 2018)	0.455	0.288	0.205	0.154	N/A	0.369	x			
	HRGR-Agent (Li et al., 2018)	0.438	0.298	0.208	0.151	N/A	0.322		x		x
	KERP (Li et al., 2019)	0.482	0.325	0.226	0.162	N/A	0.339		x	x	
	ReinforcedTransformer (Xiong et al., 2019)	0.350	0.234	0.143	0.096	N/A	N/A	x			x
	HRG-Transformer (Srinivasan et al., 2020)	0.464	0.301	0.212	0.158	N/A	N/A		x		
	SD&C (Jing et al., 2019)	0.464	0.301	0.210	0.154	N/A	0.362	x			x
	Ours (SV)	0.463	0.310	0.215	0.151	0.186	0.377	x			
	Ours (MV)	0.476	0.324	0.228	0.164	0.192	0.379		x		
	Ours (MV+T)	0.485	0.355	0.273	0.217	0.205	0.422		x	x	
	Ours (MV+T+I)	**0.515**	**0.378**	**0.293**	**0.235**	**0.219**	**0.436**		x	x	x
MIMIC	1-NN (Boag et al., 2020)	0.367	0.215	0.138	0.095	0.139	0.228	x			
	SA&T (Xu et al., 2015)	0.370	0.240	0.170	0.128	0.141	0.310	x			
	AdpAtt (Lu et al., 2017)	0.384	0.251	0.178	0.134	0.148	0.314	x			
	Liu et. al. (Liu et al., 2019)	0.313	0.206	0.146	0.103	N/A	0.306	x			x
	Transformer (Vaswani et al., 2017)	0.409	0.268	0.191	0.144	0.157	0.318	x			
	GumbelTransformer (Lovelace and Mortazavi, 2020)	0.415	0.272	0.193	0.146	0.159	0.318	x			x
	Ours (SV)	0.447	0.290	0.200	0.144	0.186	0.317	x			
	Ours (MV)	0.451	0.292	0.201	0.144	0.185	0.320		x		
	Ours (MV+T)	0.491	0.357	0.276	0.223	0.213	0.389		x	x	
	Ours (MV+T+I)	**0.495**	**0.360**	**0.278**	**0.224**	**0.222**	**0.390**		x	x	x

Table 1: Quantitative comparison of our approach and a number of recent works. Since these works are evaluated under different setups of Single-view (SV), Multi-view (MV), w/ clinical text (T), and interpreter (I), for a fair comparison, all methods are categorized based on the following four aspects: Single-View (SV), Multi-view (MV), Additional Information (AI), and Fine-tuning of the generated reports (FT). Best results are highlighted in **bold face**. Different language metrics are employed, including BLEU-1 to BLEU-4 (B-1 to B-4), METEOR (MTR), and ROUGE-L (RG-L).

cal reports, we follow the existing literature (Jing et al., 2018; Srinivasan et al., 2020) by concatenating the *impression* and the *findings* sections as the target output.

An important note: the implementation details, dataset splits, preprocessing steps, generated examples, and qualitative analysis are described in the supplementary materials.

4.2 Experimental Results

4.2.1 Language Generation Performance

A comprehensive quantitative comparison of our approach and many baselines as shown in Table 1 on the two benchmarks using the widely-used language evaluation metrics: BLEU-1 to BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) scores. Since all comparison methods have their own experiment setups, for a fair comparison, we further categorize these methods into four aspects: single-view (SV), multi-view (MV), accessing to additional information (AI) such as clinical document, and applying fine-tuning (FT) to the generated medical reports. Experiments in Table 1 show that our models outperform the baselines in most language metrics.

With a single input X-ray image as the sole input, ours (SV) outperforms by a noticeable margin the

best SOTA methods of CoAtt on Open-I and Transformer on MIMIC, respectively. This we mainly attribute to the utilization of the enriched disease embedding that explicitly incorporates the disease-related topics. With multiple X-ray images as input, Ours (MV) again outperforms the best comparison methods of HRG-Transformer on Open-I. With multiple X-ray images and additional clinical document information as input, ours (MV+T) outperforms the comparison methods of KERP on Open-I. Finally, with the complete contextual information available as input, ours (MV+T+I) outperforms all the comparison methods available in both Open-I and MIMIC datasets.

4.2.2 Clinical Accuracy Performance

To evaluate the clinical accuracy of the generated reports, we use the LSTM CheXpert labeler (Lovelace and Mortazavi, 2020) as a universal measurement. We compare different methods based on accuracy, F-1, precision (prec.), and recall (rec.) metrics on 14 common diseases. Since there are 14 independent diseases, we also report the macro and micro scores. Intuitively, a high macro score means the detection of all 14 diseases is improved. Meanwhile, a high micro score implies the dominant diseases are improved (i.e., some diseases appear more frequently than others). As

Datasets	Methods	Acc.	Macro scores				Micro scores			
			AUC	F-1	Prec.	Rec.	AUC	F-1	Prec.	Rec.
Open-I	1-NN (Boag et al., 2020)	0.911	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	S&T (Vinyals et al., 2015)	0.915	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	SA&T (Xu et al., 2015)	0.908	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	TieNet (Wang et al., 2018)	0.902	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Liu et. al. (Liu et al., 2019)	0.918	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Ours (SV)	0.944	0.595	0.118	0.125	0.136	0.857	0.657	0.651	**0.663**
	Ours (MV)	0.943	0.626	0.144	0.149	0.150	**0.878**	0.648	0.647	0.649
	Ours (MV+T)	**0.947**	0.671	0.130	**0.192**	0.124	0.873	**0.659**	**0.687**	0.634
	Ours (MV+T+I)	0.937	**0.702**	**0.152**	0.142	**0.173**	0.877	0.626	0.604	0.649
MIMIC	1-NN (Boag et al., 2020)	N/A	N/A	0.206	0.213	0.200	N/A	0.335	0.346	0.324
	SA&T (Xu et al., 2015)	N/A	N/A	0.101	0.247	0.119	N/A	0.282	0.364	0.230
	AdpAtt (Lu et al., 2017)	N/A	N/A	0.163	0.341	0.166	N/A	0.347	0.417	0.298
	Liu et. al. (Liu et al., 2019)	0.867	N/A	N/A	0.309	0.134	N/A	N/A	**0.586**	0.237
	Transformer (Vaswani et al., 2017)	N/A	N/A	0.214	0.327	0.204	N/A	0.398	0.461	0.350
	GumbelTransformer (Lovelace and Mortazavi, 2020)	N/A	N/A	0.228	0.333	0.217	N/A	0.411	0.475	0.361
	Ours (SV)	0.877	0.743	0.342	0.357	0.347	0.857	0.530	0.533	0.528
	Ours (MV)	0.880	0.752	0.347	0.385	0.347	0.862	0.533	0.545	0.522
	Ours (MV+T)	**0.890**	0.778	0.407	**0.448**	0.399	0.872	**0.578**	0.583	0.574
	Ours (MV+T+I)	0.887	**0.784**	**0.412**	0.432	**0.418**	**0.874**	0.576	0.567	**0.585**

Table 2: Quantitative comparison of clinical accuracy from the generated reports of a number of recent methods, evaluated on the 14 common CheXpert's diseases. The best results are highlighted in **bold face**.

observed in Table 2, our clinical performance increased significantly compared to the baselines in both macro and micro scores.

Among our ablation models in Table 2, the precision and accuracy scores of our contextualized variant (MV+T) tend to be higher, whereas other scores are lower than the one with the interpreter (MV+T+I). This opposite behavior is due to the interpreter, which encourages detecting diseases, thus increases False Positives (FP). Note in the medical context, it is usually critically important to lower the False Negatives (FN) rate, thus a high recall score with a slight decrease in precision is more preferred.

4.2.3 Human Evaluation

In addition to the automated evaluations, we ask an experienced medical doctor to evaluate our generated medical reports. Specifically, the chest X-ray images and ground-truth medical reports are given to the doctor. Then, the doctor evaluates the quality of the generated reports by assigning a score from 0 (totally disagree) to 10 (totally agree). The final score for each model is computed by averaging all scores (97 test samples for each proposed model).

It can be inferred from Table. 3 that the MV+T+I gives more accurate medical reports and using the interpreter to fine-tune the outputs is indeed improving the reports' quality. Additionally, it is also clear from the human evaluation that incorporating clinical history information positively affects the final performance. Moreover, the human evaluation shows that most generated examples are good (8.031 on average), indicating the proposed model's effectiveness in terms of clinical accuracy.

Methods	Average	Min	Max	Median	Q1	Q3
Ours (SV)	7.000	**3**	**10**	7	6	8
Ours (MV)	7.237	**3**	**10**	7	**7**	8
Ours (MV+T)	7.794	**3**	**10**	8	**7**	9
Ours (MV+T+I)	**8.031**	**3**	**10**	8	**7**	9

Table 3: The human evaluation scores for the generated reports from an experienced medical doctor. For each model, we take the average, min, max, median, first, and the third quartile of all ratings given by the doctor. The score is in the range of 0 (totally disagree) to 10 (totally agree).

Methods	B-1	B-2	B-3	B-4	MTR	RG-L
R w/o D_{states}	0.400	0.253	0.175	0.127	0.166	0.362
R w/o D_{topics}	0.453	0.300	0.206	0.142	0.183	0.366
R w/o D_{fused}	0.468	0.310	0.215	0.151	0.189	0.373
R with D_{enriched}	0.463	0.310	0.215	0.151	0.186	0.377
R + Interpreter	0.470	0.314	0.220	0.158	0.192	0.375
C w/o D_{states}	0.404	0.286	0.215	0.169	0.183	0.396
C w/o D_{topics}	0.474	0.329	0.244	0.187	0.194	0.401
C w/o D_{fused}	0.470	0.337	0.257	0.204	0.212	0.408
C with D_{enriched}	0.485	0.355	0.273	0.217	0.205	0.422
C + Interpreter	**0.515**	**0.378**	**0.293**	**0.235**	**0.219**	**0.436**

Table 4: The table compares a regular image-to-text version (R) and a contextualized version (C) of our proposed method that utilizes clinical history on the Open-I dataset. For each version, we evaluate the importance of each component D_{states}, D_{topics}, and D_{fused} in the proposed enriched disease embedding D_{enriched} by removing one component at a time.

4.3 Ablation studies

4.3.1 Enriched disease embedding

We observe that the latent features D_{fused} extracted from the classifier are insufficient to generate robust medical reports, as shown in Table 4. Based on our human languages, a meaningful story needs three factors: the topic (i.e., what disease), the tone (i.e., is it negative or positive), and the details (i.e., the severity). However, there is no guarantee that the learned latent features D_{fused} has all three required elements. On the other hand, with the the

3559

explicit representations (i.e., D_{fused}, D_{topics}, and D_{states}), all three factors are preserved. Therefore, the enriched disease embedding D_{enriched} can generate precise and complete medical reports, leading to the language metrics' substantial improvement.

4.3.2 Contextualized embedding

Table 4 also shows that our proposed "contextualized" version can improve the language scores over the "regular" version, which reads only images. Notably, the contextualized version is the entanglement of the chest X-ray images and the clinical history, which is crucial to improve the generated report's quality and accommodate doctors' practical needs. It mimics how radiologists receive requests from medical doctors and write reports to answer their questions. Hence, the generated reports are believed to be more "on point" and receives higher language scores than the regular "image-to-text" setting.

5 Limitations and Future Work

Our work has several limitations that future works can take into consideration for further improvement. Firstly, our model does not explicitly consider disease orientation or direction (e.g., left or right, top or bottom). For example, future works can include visual-semantic embedding (direction/orientation/location) to learn and localize diseases during generating medical reports. Secondly, our work does not support time-series relationships between different studies of a patient. This information is vital to analyze existing diseases by comparing their size or structure to determine if a disease is getting worse or not. If these limitations can be addressed, the medical report system can be much more reliable for real-world applications.

Noticeably, we observe some hallucination facts (False Positives) where some diseases are mistakenly described as positive in the medical reports. For example, some images with "pneumonia" are wrongly described as "pulmonary edema". In fact, human radiologists often mistakenly classify some diseases (Satia et al., 2013). For example, (Satia et al., 2013) shows that human radiologists or physicians can accurately detect normal lung X-ray images almost all of the time; but, for abnormal lung X-ray images, the correctness of diagnosis drops to only 50% (Satia et al., 2013). For this reason, it is challenging to generate accurate medical reports even for experienced radiologists.

In the future, we will expand our work to related medical applications such as retinal and brain medical report generation on X-ray/MRI/CT scans. We believe that our model can be generalized to a wide range of medical report generation problems where common symptoms or disease labels and medical reports are available in most medical scan datasets. Moreover, extending the current work to incorporate tabular data inputs could be another exciting direction because some clinical information is in the form of tabular structure such as patient's age, heart pressure, or temperature (Cohen et al., 2020). In some cases, physicians must include this information in medical reports, which cannot be inferred from only reading medical scans.

6 Conclusion

This paper introduces a novel three-module approach for generating medical reports from X-ray scans. Superior performance of our approach over state-of-the-art methods has been empirically demonstrated on widely-used benchmarks with a range of evaluation metrics. Our approach is also flexible and can work with additional input information, where consistent performance gains are observed. For future work, we plan to apply our approach to related medical report generation tasks that go beyond X-rays.

Acknowledgement

References

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.

William Boag, Tzu-Ming Harry Hsu, Matthew Mcdermott, Gabriela Berner, Emily Alesentzer, and Peter Szolovits. 2020. Baselines for chest x-ray report generation. In *NeurIPS Workshop on Machine Learning for Health*, pages 126–140.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020.

Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR.

J. Cohen et al. 2020. Covid-19 image data collection. *ArXiv preprint*, abs/2003.11597.

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. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 2625–2634. IEEE Computer Society.

Mohammad Eslami, Solale Tabarestani, Shadi Albarqouni, Ehsan Adeli, Nassir Navab, and Malek Adjouadi. 2020. Image-to-images translation for multi-task organ segmentation and bone suppression in chest x-ray radiography. *IEEE Transactions on Medical Imaging*, 39(7):2553–2565.

Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. 2019. Unsupervised image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4125–4134. Computer Vision Foundation / IEEE.

Aydan Gasimova, Gavin Seegoolam, Liang Chen, Paul Bentley, and Daniel Rueckert. 2020. Spatial semantic-preserving latent space learning for accelerated dwi diagnostic report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 333–342. Springer.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269. IEEE Computer Society.

Jia-Hong Huang, C-H Huck Yang, Fangyu Liu, Meng Tian, Yi-Chieh Liu, Ting-Wei Wu, I Lin, Kang Wang, Hiromasa Morikawa, Hernghua Chang, et al. 2021. Deepopht: medical report generation for retinal images via deep models and visual explanation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2442–2452.

Lun Huang, Wenmin Wang, Jie Chen, and Xiaoyong Wei. 2019. Attention on attention for image captioning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4633–4642. IEEE.

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 *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 590–597. AAAI Press.

Baoyu Jing, Zeya Wang, and Eric Xing. 2019. Show, describe and conclude: On exploiting the structure information of chest X-ray reports. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6570–6580, Florence, Italy. Association for Computational Linguistics.

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, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chihying 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):1–8.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *2017 IEEE*

Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 3337–3345. IEEE Computer Society.

Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2019. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6666–6673. AAAI Press.

Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1537–1547.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR.

Justin Lovelace and Bobak Mortazavi. 2020. Learning to generate clinically coherent chest X-ray reports. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1235–1243, Online. Association for Computational Linguistics.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3242–3250. IEEE Computer Society.

Zhichao Lu, Kalyanmoy Deb, and Vishnu Naresh Boddeti. 2020a. Muxconv: Information multiplexing in convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 12041–12050. IEEE.

Zhichao Lu, Ian Whalen, Yashesh Dhebar, Kalyanmoy Deb, Erik Goodman, Wolfgang Banzhaf, and Vishnu Naresh Boddeti. 2020b. Multi-objective evolutionary design of deep convolutional neural networks for image classification. *IEEE Transactions on Evolutionary Computation*.

Luyang Luo, Lequan Yu, Hao Chen, Quande Liu, Xi Wang, Jiaqi Xu, and Pheng-Ann Heng. 2020. Deep mining external imperfect data for chest x-ray disease screening. *IEEE Transactions on Medical Imaging*, 39(11):3583–3594.

Toru Nishino, Ryota Ozaki, Yohei Momoki, Tomoki Taniguchi, Ryuji Kano, Norihisa Nakano, Yuki Tagawa, Motoki Taniguchi, Tomoko Ohkuma, and Keigo Nakamura. 2020. Reinforcement learning with imbalanced dataset for data-to-text medical report generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2223–2236, Online. Association for Computational Linguistics.

Yujin Oh, Sangjoon Park, and Jong Chul Ye. 2020. Deep learning covid-19 features on cxr using limited training data sets. *IEEE Transactions on Medical Imaging*, 39(8):2688–2700.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Wenjie Pei, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. 2019. Memory-attended recurrent network for video captioning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8347–8356. Computer Vision Foundation / IEEE.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *ArXiv preprint*, abs/1711.05225.

Ekagra Ranjan, Soumava Paul, Siddharth Kapoor, Aupendu Kar, Ramanathan Sethuraman, and Debdoot Sheet. 2018. Jointly learning convolutional representations to compress radiological images and classify thoracic diseases in the compressed domain. In *Proceedings of the 11th Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–8.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1179–1195. IEEE Computer Society.

I Satia, S Bashagha, A Bibi, R Ahmed, S Mellor, and F Zaman. 2013. Assessing the accuracy and certainty in interpreting chest x-rays in the medical division. *Clinical medicine*, 13(4):349.

Sonit Singh, Sarvnaz Karimi, Kevin Ho-Shon, and Len Hamey. 2019. From chest x-rays to radiology reports: a multimodal machine learning approach. In *2019 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE.

Preethi Srinivasan, Daksh Thapar, Arnav Bhavsar, and Aditya Nigam. 2020. Hierarchical x-ray report generation via pathology tags and multi head attention. In *Proceedings of the Asian Conference on Computer Vision*.

Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik G. Learned-Miller. 2015. Multi-view convolutional neural networks for 3d shape recognition. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 945–953. IEEE Computer Society.

Alasdair Tran, Alexander Patrick Mathews, and Lexing Xie. 2020. Transform and tell: Entity-aware news image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 13032–13042. IEEE.

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 *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3156–3164. IEEE Computer Society.

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 *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 9049–9058. IEEE Computer Society.

Yuxuan Xiong, Bo Du, and Pingkun Yan. 2019. Reinforced transformer for medical image captioning. In *International Workshop on Machine Learning in Medical Imaging*, pages 673–680. Springer.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov,

Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2048–2057. JMLR.org.

Yuan Xue, Tao Xu, L Rodney Long, Zhiyun Xue, Sameer Antani, George R Thoma, and Xiaolei Huang. 2018. Multimodal recurrent model with attention for automated radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 457–466. Springer.

C. Yin et al. 2019. Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In *ICDM*, pages 728–737.

Changchang Yin, Buyue Qian, Jishang Wei, Xiaoyu Li, Xianli Zhang, Yang Li, and Qinghua Zheng. 2019. Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 728–737. IEEE.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4651–4659. IEEE Computer Society.

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.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2242–2251. IEEE Computer Society.