Attributed Abnormality Graph Embedding for Clinically Accurate X-Ray Report Generation

Sixing Yan, William K. Cheung, Keith Chiu, Terence M. Tong, Ka Chun Cheung, and Simon See

Abstract—Despite the recent success of deep learning models for text generation, generating clinically accurate reports remains challenging. More precisely modeling the relationships of the abnormalities revealed in an X-ray image has been found promising to enhance the clinical accuracy. In this paper, we first introduce a novel knowledge graph structure called an attributed abnormality graph (ATAG). It consists of interconnected abnormality nodes and attribute nodes for better capturing more fine-grained abnormality details. In contrast to the existing methods where the abnormality graph are constructed manually, we propose a methodology to automatically construct the fine-grained graph structure based on annotated X-ray reports and the RadLex radiology lexicon. We then learn the ATAG embeddings as part of a deep model with an encoder-decoder architecture for the report generation. In particular, graph attention networks are explored to encode the relationships among the abnormalities and their attributes. A hierarchical attention and a gating mechanism are specifically designed to further enhance the generation quality. We carry out extensive experiments based on the benchmark datasets, and show that the proposed ATAG-based deep model outperforms the SOTA methods by a large margin in ensuring the clinical accuracy of the generated reports.

Index Terms—Medical report generation, deep learning models, attributed abnormality graphs, radiology lexicon, clinical accuracy.

I. INTRODUCTION

AUTOMATIC generation of medical reports from X-ray images aims to assist radiologists to perform the reporting task which is important and yet time-consuming. An X-ray report prepared by the radiologist (as shown in Fig. 1) typically contains a paragraph with multiple sentences describing the presence/absence of abnormalities revealed in the images (called findings) together with a short conclusion (called impression). For a report generation system to be clinically accurate, findings of abnormalities revealed in the X-ray images are expected to be correctly reported. In the literature, the deep learning models with the encoder-decoder architecture have been found promising for the medical report generation, where visual features were extracted from the input images by a convolutional neural network and then fed to a sequential model like recurrent neural network [1], [2], [3], [4], and more recently the Transformer [5], [6], [7] to generate the report. As X-ray report generation expects not only sentence fluency but, more importantly, also clinical accuracy, various approaches have been proposed to achieve the goal. Semantic annotations [8] (see also Fig. 1) and concepts extracted from the medical reports [9] have been used to learn semantic features to assist the report generation. Also, the reinforcement learning approach has been used to increase the accuracy of the clinical keywords to be included in the generated reports [10], [11]. Recently, prior knowledge of abnormalities modeled as a graph has been integrated into the deep model architecture [12], [13], [14], and achieved the state-of-the-art performance on clinical accuracy.

Incorporation of the abnormality graph, while shown to be effective, requires the graph to be manually constructed, and thus could inevitably be sub-optimal. In addition, a medical report contains not only information about the observed abnormalities (e.g., “calcified granuloma”), but also their associated “attributes” (e.g., “left upper lobe” as its location). Therefore, it is important to properly consider both the abnormalities and their attributes in the representation to enable reports with the related details to be generated. Related work with attributed abnormalities explicitly considered is still rare.

To this end, we propose a novel fine-grained knowledge graph structure called ATributed Anormality Graph (ATAG) to represent the attributed abnormalities and integrate that
with a deep learning model for generating clinically accurate reports. We first propose a methodology to automatically construct ATAG using annotated X-ray reports and the RadiLex radiology lexicon. ATAG represents each attributed abnormality as an abnormality node and an associated set of attribute nodes to model the abnormality details. The inter-related abnormalities and attributes are linked up with edges. We then take ATAG as an overall medical knowledge graph and integrate it into a deep encoder-decoder model to learn the corresponding embeddings. The objective is to optimize the abnormality classification accuracy and the X-ray report generation quality. A hierarchical attention and a gating mechanism are specifically designed to allow the information encoded by ATAG to be better aligned with decoders implemented using either LSTM or Transformer to further enhance the clinical accuracy. We evaluate the proposed ATAG-based approach using the publicly available IU-XRay [15] and MIMIC-CXR [16] datasets and show that it can achieve higher accuracy on abnormality classification, and generate more clinically accurate reports compared to the SOTA methods according to the clinical accuracy metrics by a large margin.

To summarize, the main contributions of this paper include:

1) A methodology to automatically construct a fine-grained attributed abnormality graph (ATAG) from X-ray reports and a radiology ontology for representing abnormalities and their attributes;

2) An algorithm to learn the attributed abnormality embeddings using graph attention networks derived based on ATAG; and

3) A hierarchical attention and a gating mechanism to facilitate adaptive integration of the ATAG embeddings with various decoders to generate detailed X-ray reports with clinically accurate attributed abnormalities.

The rest of this paper is organized as follows. Section II reviews the related work and Section III presents an overview of the proposed framework. The algorithms for constructing the attributed abnormality graph and learning the ATAG embeddings are detailed in Sections IV and V respectively. Section VI shows how to utilize the ATAG embeddings to generate reports using different decoders. Details of the experiments can be found in Sections VIII, and IX concludes the paper.

II. RELATED WORK

The earlier effort on generating textual output based on visual input can be traced back to automatic image captioning [17], [18]. Radiology report generation is similar to some more recent image captioning works where the output involves multiple sentences to describe the objects in the image. For a radiology report, each sentence usually focuses on one topic, i.e., a clinical observation, with some fine-grained supporting details. In the literature, deep learning based methods using CNN encoder and RNN decoder for the report generation have been found promising [1], [2], [3], [4]. To achieve clinically accurate report generation, approaches incorporating semantic labels and knowledge graphs have been explored. Also, aligning the input visual features and the contents of the generated reports is vital to enhance clinical accuracy.

A. Use of Semantic Labels

Based on annotations in X-ray datasets, Yuan et al. [9] adopted Semrep (https://semrep.nlm.nih.gov/) to extract 69 concepts from the medical reports, and trained a CNN for concept classification and report generation. Miura et al. [6] learned an image-to-text model, and then fine-tuned it by increasing the number of the matched clinical entities in the generated report. Wang et al. [19] tried to optimize the generated reports by multi-label text classification. The aforementioned methods rely on either some taggers or manual effort to extract the medical concepts and abnormalities.

B. Use of Knowledge Graphs

Medical knowledge graphs representing the relationships of medical concepts and abnormalities, can be integrated with deep learning models to enhance the clinical accuracy of X-ray report generation. Li et al. [12] proposed an abnormality-and-attributes graph with nodes corresponding to 80 abnormality phrases (manually chosen) frequently appearing in reports and edges constructed based on occurrence frequencies. Zhang et al. [13] manually constructed a knowledge graph with 20 common abnormalities where the nodes are connected according to the body parts. In this paper, we propose to automatically construct an abnormality graph with fine-grained attributes using annotated X-ray report and the radiology lexicon and learn the corresponding embeddings to generate clinically accurate reports.

C. Aligning Visual Features and Report Contents

As the report is generated based on the visual features attended by different abnormalities, correctly aligning the
visual features and the corresponding report contents is considered important to increase the reliability of the report generation system. To achieve that, You et al. [20] developed an alignment-enhanced Transformer to refine the visual features with the semantic features of disease labels. Orthogonal to this direction, a multi-aligned memory mechanism was proposed in [21] to refine the visual features in a fine-grained manner. In addition, the memory mechanism has been employed for the report generation by memorizing the visual patterns of observations via external memory construction. For instance, R2Gen [5] utilizes a memory matrix to memorize the projection between the visual and language patterns. A conditional layer normalization is adopted to update the memory matrix in the decoding process. Similarly, a cross-modal memory network was proposed in [22] to learn the latent features of abnormalities based on the visual and language features aiming to effectively transform the cross modalities. A memory distilling mechanism was proposed in [14] to query the related graph embedding for current decoding token. In this paper, we propose a specific hierarchical attention and a gating mechanism to facilitate the attended but not yet decoded contents in the form of an adaptive context vector computed based on the ATAG embeddings for the subsequent decoding steps, aiming to better utilize the fine-grained graph embedding for the report generation.

III. AN OVERVIEW OF THE PROPOSED FRAMEWORK

An X-ray report generation system extracts a set of visual features from one or multiple input X-ray images (e.g., frontal and lateral views), denoted as \( F(V) \), and generates a radiology report as a sequence of tokens \( (y_i) \), denoted as \( R = \{y_1, y_2, \ldots \} \), to describe findings on abnormalities. We adopt the widely used deep encoder-decoder architecture and propose to incorporate the knowledge about the relationships of abnormalities and their associated attributes as a fine-grained ATTributed Abnormality Graph (ATAG), denoted as \( G \).

In the following sections, we first present a methodology to automatically construct the ATAG \( G \) based on a set of annotated X-ray reports and a radiology lexicon. Given the ATAG structure, we then explain how to leverage Graph Attention Network [23] to learn the ATAG embeddings, denoted as \( Z \), based on the input visual feature \( F(V) \). After encoding the input as the ATAG embeddings \( Z \), we propose algorithms with a hierarchical attention and a gating mechanism for learning the decoder so that it can attend to the proper attributed abnormality node embeddings to depict different observations in the generated report \( \hat{R} \). We examine both LSTM-based and Transformer-based decoders.

IV. AUTOMATIC CONSTRUCTION OF ABNORMALITY GRAPH WITH FINE-GRAINED ATTRIBUTES

An X-ray medical report consists of concepts of abnormalities and their associated details such as the corresponding anatomical part, location, status, etc., and the concepts are inter-related. To represent the concepts and their relationships as a graph, we first define \( G^{(A)} = (V^{(A)}, E^{(A)}) \) where \( V^{(A)} \) represents a set of abnormality nodes and \( E^{(A)} \) represents a set of edges connecting them if the abnormalities are related. We also introduce an additional node \( v_0^{(A)} \) connecting to all the abnormality nodes to represent absence of all abnormalities.

In addition, for each abnormality node \( v_i^{(A)} \in V^{(A)} \) (except \( v_0^{(A)} \)), it is paired with an attribute graph \( G^{(B_i)} = (V^{(B_i)}, E^{(B_i)}) \) where \( V^{(B_i)} \) represents a set of associated attribute nodes and \( E^{(B_i)} \) represents a corresponding set of edges connecting the attribute nodes (indicating that they are inter-related for abnormality \( i \)). For each attribute graph \( G^{(B_i)} \), we also introduce an additional node connecting to all the other attribute nodes to indicate that no attributes to be specified. The attributed abnormality graph can therefore be denoted as \( G := (G^{(A)}, \{G^{(B_i)}\}) \) with each \( G^{(B_i)} \) corresponding to a distinct \( v_i^{(A)} \in V^{(A)} \).

To construct the ATAG automatically, we first extract the set of abnormalities and their associated attributes from i) the annotations in the X-ray datasets, and ii) the X-ray reports. Then, we connect them based on their occurrence statistics.

A. Extracting Abnormalities From X-Ray Annotations

If annotated X-ray datasets (e.g., IU XRays) are provided, the abnormality and attribute concepts can readily be extracted from the terms in the annotations. In particular, we adopt the RadLex lexicon (http://radlex.org/) used in the IU XRays. An annotation term is considered as an abnormality \( a_i \) if it can be found under RadLex’s “Clinical Finding” category, or an attribute \( b_j \) if it can be found under RadLex's different descriptor categories. E.g., “atelectasis” is a “clinical finding” and “right” is a “location descriptor” in RadLex. For the annotation where we cannot identify any clinical finding term, we use “other, [anatomical-part descriptor]” to denote the abnormality. We applied this methodology to the IU XRays dataset and tried to keep the terms of which the occurrence frequency is larger than different thresholds for extracting abnormalities and attributes, as reported in Table. I.

B. Extracting Abnormalities From X-Ray Reports

Alternatively, we can also extract abnormality and attribute concepts from X-ray datasets without annotations. For instance, MIMIC CXR [16]) is a big X-ray dataset with only paired images and reports where the ground-truth annotations are absent. We first extract clinical terms from the reports using the open annotation API provided by RadLex.org [24]. For the extracted terms which can be found under the “Clinical Finding” category, the corresponding abnormality nodes are created. Also for those under the “RadLex Descriptor” category, attribute nodes are created accordingly. The associations between the abnormalities and the attributes, and thus the corresponding edges, are established using the dependency parser RadGraph. In particular, we use the parsed relationships “located_at” and “modify” provided by RadGraph to determine the abnormality-to-attribute associations (as illustrated in Fig. 2). We applied this methodology to the MIMIC CXR dataset and managed to extract different numbers of terms depending on the threshold putting on their occurrence frequency as reported in Table. I.
| Dataset     | Freq | # Abn | # Atr | Max / Min / Avg (Atc. per Abn.) |
|-------------|------|-------|-------|----------------------------------|
| RUI XRay    | 10   | 41    | 106   | 47 / 1 / 11.5                   |
|             | 20   | 28    | 79    | 40 / 1 / 13.9                   |
|             | 30   | 23    | 64    | 34 / 1 / 14.0                   |
| MIMIC CXR   | 500  | 47    | 209   | 178 / 17 / 69.1                 |
|             | 1,000| 35    | 165   | 142 / 19 / 69.4                 |
|             | 2,000| 26    | 129   | 116 / 37 / 70.0                 |

V. LEARNING ATAG EMBEDDINGS

We integrate the proposed ATAG into an encoder-decoder architecture as shown in Fig. 3. DenseNet [26] is used to extract the visual feature for computing the ATAG embeddings. Specific graph attention layers are introduced to aggregate the representations from both the abnormality and attribute nodes. The ATAG embeddings are then learned using the multi-abnormality and multi-attribute classification as the learning objective.

A. Abnormality Node Embedding Using GAT

Given the visual features extracted from both the frontal and lateral chest x-ray images $F^{(V)} \in \mathbb{R}^{H \times W \times 2D}$ where $H$ and $W$ refer to the height and width of the feature map and $D$ is the feature dimension per image, we define a corresponding set of abnormality node features $F^{(A)} \in \mathbb{R}^{|A| \times 2D}$ for the ATAG $G^{(A)}$ which are computed using a spatial attention mechanism which is implemented by a convolutional layer as in [13]. In particular, we set up $|A|$ channels with $1 \times 1$ kernel denoted as $\text{Conv}^{(A)}(F^{(V)}, |A|)$. For each channel, we compute the attention weight $\alpha_i^{(A)} \in \mathbb{R}^{1 \times H \times W}$ over $2D$ channels of $F^{(V)}$ to indicate the particular image region to be attended by abnormality node $v_i^{(A)}$. The overall attention-weighted visual features $F^{(A)}$ for $G^{(A)}$ is denoted by concatenating all the channels, defined as: $F^{(A)} = \bigoplus_{i=1}^{|A|} \alpha_i^{(A)} F^{(V)}$, where $\bigoplus$ denotes the concatenation operation. Regarding the visual feature $F_0^{(A)}$ for $v_0^{(A)}$, we use average pooling over the visual features of all the other nodes.

In addition, for all the abnormalities, we define a set of abnormality concept features $E^{(A)} \in \mathbb{R}^{|A| \times D}$ (to be learned) to encode the intrinsic information for the abnormalities and attributes which are fixed after training. Then, we aggregate $F^{(A)}$ and $E^{(A)}$ to define abnormality node embeddings, denoted as $Z^{(A)} \in \mathbb{R}^{|A| \times (1+1) \times D}$, which can be computed using the graph attention layer (GAT) [23]. Specifically, we first concatenate the 2D-dimensional attention-weighted visual features $F^{(A)}$ and $D$-dimensional concept feature $E^{(A)}$, and then project it back to a $D$-dimensional embedding via $W^{(A)} \in \mathbb{R}^{3D \times D}$, and then feed it to a GAT($\cdot$) defined on $G^{(A)}$, given as:

$$Z^{(A)} = \text{GAT}^{(A)}((F^{(A)} \oplus E^{(A)}) W^{(A)}, G^{(A)}).$$

B. Per-Abnormality Attribute Embedding

To compute the attribute-related visual features $F^{(B_i)}$ associated with the abnormality node $v_i^{(A)}$, we feed the attention-weighted visual features of each abnormality to a convolutional layer $\text{Conv}^{(B_i)}$ to attend values associated to compute the attribute nodes in $B_i$, given as:

$$\alpha^{(B_i)} = \text{Conv}^{(B_i)}(\alpha_i^{(A)} F^{(V)}, |B_i|).$$

The intuitive idea is to put focus on the local regions where the abnormality node is attending for computing the corresponding attribute-related visual features. We also define for all the possible attributes a corresponding set of intrinsic attribute concept features $E^{(B)} \in \mathbb{R}^{|B| \times D}$ where $B = \bigcup_{i=1}^{|A|} B_i$. Here we assume that they are abnormality-invariant. Then, again we concatenate the attribute-related visual features and their concept features, and feed it after a projection $W^{(B)} \in \mathbb{R}^{3D \times D}$ to the GAT($\cdot$) to compute the attribute embedding for $G^{(B_i)}$, given as $Z^{(B_i)} \in \mathbb{R}^{(|B_i|+1) \times D}$.

C. Overall Loss Function

To obtain the ATAG embeddings $Z^{(A)}$ and $Z^{(B_i)}$, we learn the deep learning model by feeding them to a fully-connected layer with the sigmoid function for multi-label classification in an end-to-end manner. We sum up the binary cross-entropy loss of all the abnormality classes $L^{(A)}_{\text{CLS}}$ and that of the attribute classes per abnormality $L^{(B_i)}_{\text{CLS}}$ to define the overall loss $L_{\text{CLS}}$ for training. As in [13], to alleviate the class imbalance in the dataset, the loss for each abnormality is weighted by $w_i^{(A)} = \text{Freq}_i / \sum_{i=1}^{|A|} \text{FREQ}_i$.
Fig. 3. An example of ATAG structure is presented in Part I, with the process of computing ATAG embedding shown in Part II.

\(| \mathcal{R} \setminus \mathcal{R}^{(a_i)} | \mathcal{R}^{(a_i)} | \) for each \(a_i\), where \(\mathcal{R}\) is the overall set of the reports and \(\mathcal{R}^{(a_i)} \subseteq \mathcal{R}\) is the set of reports with \(a_i\) mentioned in their annotations. Similarly, the loss for each node in the attribute graph \(\mathcal{G}(B)\) is weighted by \(w_j(B) = | \mathcal{R} \setminus \mathcal{R}^{(B)} | \mathcal{R}^{(B)} | \) where \(h_{i,j}\) refers to the attribute \(j\) associated with the abnormality \(i\). Given the abnormality and attribute ground-truth labels, the overall loss function is given as:

\[
\mathcal{L}_{\text{CLS}} = \beta^{(A)} \mathcal{L}_{\text{CLS}}^{(A)} + (1 - \beta^{(A)}) \sum_i \beta^{(a_i)} \mathcal{L}^{(B)}_i .
\]  

where \(\beta^{(A)}\) is a trade-off parameter between the abnormality and attribute losses, and \(\beta^{(a_i)} = w_j^{(A)} / \sum_{k=1}^{A} (w_k)^2\) balances the attribute losses among different abnormalities.

VI. REPORT GENERATION WITH ATAG EMBEDDINGS

With the input X-ray images encoded as the ATAG embeddings \((Z^{(A)}, \{Z^{(B)_i}\})\), we derive a context vector \(c^{(A+B)}\) to guide the report generation process using a decoder. In particular, to allow \(c^{(A+B)}\) to reveal the relevant information captured in \(Z^{(A)}\) and \(\{Z^{(B)_i}\}\) for attending to different sentences in the report are being generated, we first propose a hierarchical attention mechanism by making reference to the decoder’s evolving hidden state (denoted as \(h_t\) at time step \(t\)). We also introduce a gating mechanism for adapting the abnormality graph embedding \(Z^{(A)}\) and attribute graph embedding \(\{Z^{(B)_i}\}\), and show how the gating mechanism can be applied to LSTM- or Transformer-based decoders which are widely used for text generation.

A. Hierarchical Attention on ATAG Embeddings

Decoders for text generation make use of the hidden state \(h_{t-1} \in \mathbb{R}^{D_h}\) to represent the historical information of the text generated up to time step \(t - 1\). We incorporate the ATAG embedding to update the hidden state for guiding the report generation process. To achieve that, we introduce a context vector at the time step \(t\) using a hierarchical attention mechanism with \(Z^{(A)}\) and \(\{Z^{(B)_i}\}\) both considered. We first make reference to the hidden state of the decoder the \(h_{t-1}\) and compute for each \(\mathcal{G}(B)\) at time step \(t\) an attention-weighted context vector \(c^{(B)}_t\) given as:

\[
c^{(B)}_t = \sum_{j=0}^{B} \xi^{(B)}_{t,j} \cdot Z^{(B)_j} , \quad \xi^{(B)}_{t,j} = \text{Attn}(h_{t-1}, Z^{(B)_j})
\]

where \(\xi^{(B)}_{t,j} \in \mathbb{R}^{B+1}\) is the attention weight vector, \(\zeta^{(B)}_t\) is the weighting of node \(j\)’s embedding \(Z^{(B)_j} \in \mathbb{R}^D\) in the aggregated context vector \(c^{(B)}_t\), and \(\text{Attn}(\cdot, \cdot)\) is defined as:

\[
\text{Attn}(h_{t-1}, Z^*_t) = \text{softmax}(h_{t-1}) W^D \text{tanh}(Z^*_t)
\]

The overall context vector is then derived by applying the attention mechanism again to the concatenation of the abnormality graph embedding \(Z^{(A)}\) and \(c^{(B)}_t\), given as:

\[
c^{(A+B)}_t = \sum_{i=0}^{A} c^{(A)}_{t,i} \cdot Z^*_i , \quad \zeta^{(A)}_t = \text{Attn}(h_{t-1}, Z^*_t)
\]

where \(Z^*_i \in \mathbb{R}^A\) and \(\zeta^{(A)}_t \in \mathbb{R}^{A+1}\) is the attention weight vector with each element corresponding to an abnormality node at time step \(t\).

B. Gating Mechanism for ATAG Embeddings

The ATAG embeddings \((Z^{(A)}, \{Z^{(B)_i}\})\) being fed to the decoder for report generation so far is assumed to be fixed during the generation process, as in [7], [12], [13], [14]. With the objective to avoid findings of repeated contents to be generated during the decoding, we propose a gating mechanism to adapt the ATAG embeddings for generating the next token or sentence. The intuition behind is to “forget” the information related to the findings which have been generated and “keep” the information content related to those which have not been generated. For clarity, we here use a generic notation \(Z_t\) to represent the embedding at time step \(t\) for presenting the proposed gating mechanism.

First, we allow \(Z_t\) to be element-wise multiplied by the corresponding attention value \(\zeta_t\) to give \(\hat{c}_t = \bigoplus_{i=1}^{A} \xi^{(A)}_{t,i} \cdot Z^{(B)_i}\). Then, to differentiate the information of the attended and unattended abnormalities in \(Z_t\) as indicated by \(\hat{c}_t\), we derive:

\[
\hat{Z}_t = \text{tanh}(\text{FCN}(Z_t + \hat{c}_t)) + Z_t + \hat{c}_t,
\]

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
where FCN(·) is a two-layers full connected NN with the Gaussian error linear unit as the activation function. To remove from \(Z_t\) the information of the unattended abnormalities and the attended but already generated ones, we compute:

\[
Z_t^{(F)} = O_t^{(F)} Z_t, \quad O_t^{(F)} = \text{gate}^{(F)}(H_t, Z_t) \quad (9)
\]

where \(H_t\) is the matrix with the hidden state \(h_t \in \mathbb{R}^D\) duplicated to have the same dimensions of \(Z_t\), and the gating network \(\text{gate}^{(\cdot)}(\cdot)\) is defined as:

\[
\text{gate}(H_t, Z_t) = \text{sigmoid}(H_t U_1 + \text{tanh}(Z_t U_2)) \quad (10)
\]

where \([U_1, U_2] \in \mathbb{R}^{D \times D}\) are the learnable parameters. The gating matrix \(O_t^{(F)}\) contains the weights indicating the information which have not been generated in \(h_t\).

To “keep” the information which has not been decoded so far, we compute \(Z_t^{(I)}\) based on \(Z_t\):

\[
Z_t^{(I)} = O_t^{(I)} Z_t, \quad O_t^{(I)} = \text{gate}^{(I)}(H_t, Z_t). \quad (11)
\]

The resulting embedding becomes \(Z_{t+1} = Z_t^{(F)} + Z_t^{(I)}\). For the clarity reason, we denote the overall gating mechanism as:

\[
Z_{t+1} = \text{GATE}(h_t, \zeta_t, Z_t). \quad (12)
\]

### C. Decoding With Two-Level LSTM

The two-level long short-term memory (LSTM) network [27] is commonly adopted as the decoder for the report generation in related works, where a top-level LSTM is used to predict the abnormalities (topics) for each sentence, and a bottom-level LSTM is to generate the description for the particular abnormalities. We propose to update of the hidden state for the top-level LSTM at time step \(t\) for generating the “topics” of the \(t\)th sentences with the consideration of also the ATAG context vector \(c_t^{(A+B)} \in \mathbb{R}^{2D}\) as defined in Section VI-A, given as:

\[
h_t^{(S)} = \text{LSTM}^{(S)}(c_t^{(A+B)}, h_{t-1}^{(S)}),
\]

\[
c_t^{(A+B)} = \text{HAT}(h_{t-1}^{(S)}, Z^{(A)}, \{Z^{(B)}\}). \quad (13)
\]

The initial hidden state \(h_0^{(S)}\) can be set to be the concatenation of the global average of abnormality and attribute node embeddings. For a particular \(h_t^{(S)}\) corresponding to the topic of the \(t\)th generated sentence, the bottom-level LSTM \(\text{LSTM}^{(W)}\) computes the hidden state \(h_t^{(W)}\) for generating the \(t\)th word token \(w_{t,\tau}\) of the sentence as:

\[
h_t^{(W)} = \text{LSTM}^{(W)}(h_t^{(S)} \oplus c_t^{(A+B)}, h_{t-1}^{(W)}), \quad (14)
\]

with \(h_{t,0}^{(W)}\) initialized as a zero vector. Then the \(t\)th word token of the \(t\)th sentence is predicted by \(\hat{y}_{t,\tau} \sim p_{t,\tau} = \text{softmax}(h_t^{(W)} W_y + b_y)\).

To apply the gating mechanism, we first aggregate the hidden state of the top-level LSTM \(h_t^{(S)}\) and the self-attended hidden state of the bottom-level LSTM \(h_t^{(W)}\), given as:

\[
h_t^{(S+W)} = (h_t^{(S)} \oplus c_t^{(W)}) W
\]

\[
c_t^{(W)} = \sum_{\tau=1}^{T} s_{t,\tau} h_{t,\tau} s_t^{(W)} = \text{Attn}(h_t^{(W)}, h_t^{(W)}) \quad (15)
\]

where \(H_t^{(W)} \in \mathbb{R}^{T \times D}\) is the hidden state matrix by stacking \(h_{t,1}^{(W)} \) over \(\tau, W \in \mathbb{R}^{D \times D}\) is a projection matrix which is to be learned, and \(T\) is the sentence length. Then, we adapt the abnormality embeddings \(Z_t^{(A)}\) and the attribute embeddings \(Z_t^{(B)}\) over the generation process as (Fig. 4):

\[
Z_{t+1}^{(A)} = \text{GATE}(h_{t+1}^{(S+W)}, \zeta_t^{(A)}, Z_t^{(A)}),
\]

\[
Z_{t+1}^{(B)} = \text{GATE}(h_{t+1}^{(S+W)}, \zeta_t^{(B)}, Z_t^{(B)}). \quad (16)
\]

### D. Decoding With Transformer

The Transformer is another powerful model for the decoder which adopts multiple layers of multi-head attention (MHA) for sequence modeling and generation [28]. Given the embedding of the \(t\)th word token \(w_t\) and the corresponding positional embedding \(e_t\), a Transformer feeds the input \(x_{t,0} = w_t + e_t\) to go through \(L\) layers of MHA to generate the next word \(y_t\). For the \(l\)th layer, we define the hidden state of the \(t\)th token as \(h_{t,l} = \text{MHA}(x_{t,l-1}, x_{t+1,l-1})\), and then the ATAG context vector for the \(l\)th layer as:

\[
c_t^{(A+B)} = \text{HAT}(h_{t,l}, Z^{(A)}, \{Z^{(B)}\}). \quad (17)
\]

We then compute the output of the \(l\)th layer \(x_{t,l}\) by:

\[
x_{t,l} = \text{FCN}(h_{t,l}),
\]

\[
\hat{y}_{t,l} = \text{FCN}(\text{MHA}(h_{t,l}, c_t^{(A+B)})), \quad (18)
\]

and the last layer’s output \(x_{t,L}\) is used to predict the token \(y_t\).

We also apply the gating mechanism to each layer of MHA as shown in Fig. 5 to allow the ATAG embeddings to be adaptive. In particular, we update the abnormality embedding \(Z_t^{(A)}\) and attribute embeddings \(Z_t^{(B)}\) based on \(h_{t,l}\) as:

\[
Z_{t,l}^{(A)} = \text{GATE}(h_{t,l}, \zeta_{t,l}^{(A)}, Z_{t,l}^{(A)}),
\]

\[
Z_{t,l}^{(B)} = \text{GATE}(h_{t,l}, \zeta_{t,l}^{(B)}, Z_{t,l}^{(B)}). \quad (19)
\]

In our experiment, we repeat the gating mechanism multiple times for each layer and can obtain better performance.
VII. CLINICAL ACCURACY EVALUATION

Clinical accuracy, in addition to language quality, is one of the main concerns for the quality of the generated reports. So we need not only language quality performance metrics but also those for measuring clinical accuracy. We make reference to MIRQI-F1 [13] and propose a new metric named \textbf{Rad}iology \textbf{R}eport \textbf{Q}uality \textbf{I}ndex \textbf{F}1 score (RadRQI-F1) with the objective to achieve more reliable and fine-grained evaluation, where we make use of the RadLex lexicon [24] and RadGraph [29].

A. RadRQI-F1

To define the proposed RadRQI-F1 score, the general idea is to first identify the positive and negative mentions of abnormalities and their attributes in the generated reports and the ground truths, and compute the F-1 measure. We first extract the tuples of “(Abnormality, Negation, Attribute1, Attribute2, …)" from both the generated and ground-truth reports using the RadLex lexicon [24]. The keywords if found under RadLex’s “Clinical Finding" category will be taken as abnormalities, and those under “Clinical Descriptor" as their attributes. Next, whether an abnormality or attribute is present or not, and their associations are determined using RadGraph [29]. For instance, the terms will be taken as “positive mentions" if they are labeled as “present" by RadGraph, and as “negative mentions" if labeled as “absent”. Then, the association between abnormality and attribute is established by detecting the “modify" and “located At" relations.

The RadRQI-F1 is defined as the F-1 measure of the positive mentioned abnormalities in the generated reports:

\[
\text{RadRQI-F1} = \frac{2TP}{2TP + FP + TF} \tag{20}
\]

To evaluate “positive mentions” of abnormalities and their attributes [13], we count the True Positives of abnormalities \(TP^{(A)}\) and their corresponding attributes \(TP^{(B)}\) to give:

\[
TP = (1 - \alpha^{(B)})TP^{(A)} + \alpha^{(B)}TP^{(B)} \tag{21}
\]

where \(\alpha^{(B)}\) is the trade-off weight between the abnormality and attribute term accuracy. For False Positive FP, we count negative- and not-mentioned abnormalities which are positive-mentioned in the generated report. For False Negative FN, we count positive-mentioned abnormalities which are negative- or not-mentioned in the generated report. We also count the number of abnormalities with non-zero RadRQI-F1 score, denoted as RadRQI-Hits, to indicate the coverage of distinct abnormality categories in the generated reports.

RadRQI-F1, compared to MIRQI-F1, is considered more reliable as RadRQI-F1 covers more abnormalities and attributes while MIRQI-F1 evaluates only a small set of abnormalities under 12 disease categories and may take also irrelevant words (e.g., stop words) as abnormalities’ attributes. Also, MIRQI-F1 does not consider terms which are found in the ground truth but not mentioned in the generated report, nor those found in the generated reports but not mentioned in the ground truth. By ignoring those not-mentioned terms, it will favor methods which generate only a few correct abnormalities but missing many others. This results in misleading evaluation results. RadRQI-F1 considers both positive and negative mentions in calculating the F1 measure.

VIII. EXPERIMENTS

A. Datasets and Evaluation Metrics

We use two publicly available datasets IU XRay [15] and MIMIC CXR [16] for performance evaluation. For IU XRay, we extract only the reports with both frontal and lateral view images, complete finding/impression sections and annotations available, resulting 2,848 cases and 5,696 images. We tokenize reports and filter out tokens with frequency less than three, resulting in 1,028 unique tokens. We partition the data into training/validation/test set by 7:1:2 for five-fold cross validation. For MIMIC CXR, we extract findings/impression sections as the target report and filter out tokens with frequency less than 10, resulting 4,936 distinct tokens and following the sections as the target report and filter out tokens with frequency less than three, resulting 2,848 cases and 5,696 images. We use two publicly available datasets IU XRay [15] and MIMIC CXR [16] for performance evaluation.

For IU XRay, we only extract the reports with both frontal and lateral view images, complete finding/impression sections and annotations available, resulting 2,848 cases and 5,696 images. We use two publicly available datasets IU XRay [15] and MIMIC CXR [16] for performance evaluation.

For MIMIC CXR, we only extract the reports with both frontal and lateral view images, complete finding/impression sections and annotations available, resulting 2,848 cases and 5,696 images. We use two publicly available datasets IU XRay [15] and MIMIC CXR [16] for performance evaluation.
with the suffix (\([\#\text{Labels}]\)) indicating the number of abnormalities being evaluated. We use the CheXpert labeling toolkit to label 12 different thoracic diseases together with “support devices” and “normality”. We also report CE(Hits) as the average number of classes which have non-zero F1 scores. ii) For RadRQI-F1, we measure the scores of top-25 and top-50 abnormalities for IU XRay and MIMIC CXR, respectively. The $\alpha$ is set to 0.5 to put equal emphasis on abnormalities and attributes for the evaluation.

We also adopt as the common NLP metrics like BLEU [32], ROUGE [33] and CIDEr [34] to evaluate language quality.

### B. Experiment Settings and Baselines

To evaluate the performance of ATAG on multi-label classification, we extract the input visual features using DenseNet-121 [26], train the encoders of the ATAG-based models and compare them with some baselines for classifying abnormalities and their attributes. The encoders we tested include feed-forward network (FFN) [31], knowledge graph (KG) [13], the ATAG without attributes (AG), and the full version of ATAG (ATAG). We consider cases with different numbers of labels for the classification task. The corresponding models are denoted as \text{Encoder(\([\#\text{Labels}]\))} as shown in Table III. For example, \text{FFN(20)} stands for the setting where FFN is used as the encoder and 20 labels are used for the classification. Here, we adopt the 20 specific labels used in [13]. For the models trained with IU XRay, \text{FFN(41)} stands for the setting with the 41 abnormalities identified via the ATAG construction, and \text{FFN(41+106)} stands for the setting with further 106 distinct attributes identified and included. For the models trained with MIMIC CXR, we consider 47 abnormalities and 209 distinct attributes identified via the ATAG construction. In our experiments, all input images are resized to 512 × 512 before feeding into the DenseNet.

To evaluate the performance on report generation, we first train the encoders of the ATAG-based models and then compare them with some baselines for classifying abnormalities and attributes for the evaluation. The dimension of the hidden states in all the LSTM decoders is 512, and the number of hidden state dimensions, heads, layers and looping in all the Transformer-based decoders are set to be 512, 8, 2 and 3 respectively. For fair comparison, we follow the practice of related works [7], [12], [13], [14] and adopt two-phase training where the encoder is trained first and then fixed when the decoder is trained [13].

For training models based on IU XRay, the encoder is trained with the learning rate 1e-6 for 150 epochs, followed by the decoder with the learning rate 1e-4 for 100 epochs. The mini-batch size is 8. For those trained on MIMIC CXR, the encoder and decoder are trained for 32 epochs using the mini-batch size of 16 and the same learning rates as IU XRay. The proposed model is implemented by PyTorch with a Python Package DGL (https://www.dgl.ai/) adopted for the graph attention network. The experiment is conducted with Intel(R) Xeon Gold CPU (2.70GHz) and four sets of NVIDIA Tesla V100S GPU. For the LSTM-based model, the maximum number and length of sentences in a report is set to be (10, 30) and (15, 50) for IU XRay and MIMIC CXR, respectively. For the transformer-based model, the maximum length of a report is set to be 60 and 100 for IU XRay and MIMIC CXR, respectively. The beam size is 3 by default. The training / inference time is reported in Table II.

| Dataset | Model | Training (s) | Inference (s) |
|---------|-------|--------------|---------------|
| IU XRay | SentSAT+ATAG | 0.3 | 0.3 |
|         | SentSAT+ATAG+GATE | 0.4 | 0.8 |
|         | Trans+ATAG      | 0.6 | 5.1 |
|         | Trans+ATAG+GATE | 0.9 | 10.4 |
| MIMIC CXR | SentSAT+ATAG | 0.6 | 0.6 |
|         | SentSAT+ATAG+GATE | 0.8 | 1.2 |
|         | Trans+ATAG      | 1.5 | 18.9 |
|         | Trans+ATAG+GATE | 1.6 | 26.0 |

### Table II

**The Estimated Time of Training / Inference per Image**

| Dataset | Model | ROC-AUC (\([\text{std.}]\)) | Abn. | Attr. |
|---------|-------|--------------------------|------|-------|
| IU XRay | FFN(20) [31] | 0.740±0.019 | - | - |
|         | KG (20) [13] | 0.728±0.002 | - | - |
|         | FFN(41) | 0.890±0.009 | - | - |
|         | AG (41) | 0.888±0.003 | - | - |
|         | FFN(41+106) | 0.884±0.012 | 0.560±0.054 | - |
|         | ATAG (41+106) | 0.892±0.006 | 0.686±0.069 | - |
| MIMIC CXR | FFN(20) [31] | 0.099±0.024 | - | - |
|         | KG (20) [13] | 0.595±0.103 | - | - |
|         | FFN(41) | 0.793±0.099 | - | - |
|         | AG (41) | 0.795±0.102 | - | - |
|         | FFN(41+106) | 0.801±0.109 | 0.330±0.104 | - |
|         | ATAG (41+106) | 0.810±0.110 | 0.799±0.132 | - |

### Table III

**Performance on Multi-Label Classification (AUC) Over All the Categories Being Trained. “Abn.” and “Attr.” Stand for Abnormality and Attribute Classification, Respectively**

4The codes of WordSAT [17], SentSAT [9], CoAtt [8], and SentSAT+KG [13] are provided by [13]; Transformer [28] and AdaAttn [35]: https://github.com/ysmiura/ifcc/tree/master/clinicgen/models. M2 Trans. [36]: https://github.com/cuhksz-nlp/meshed-memory-transformer. R2Gen [5]: https://github.com/cuhksz-nlp/R2Gen. R2Gen-CMN [22]: https://github.com/cuhksz-nlp/R2GenCMN
**C. Performance on Multi-Label Classification**

We report the ROC-AUC and PR-AUC scores for all the evaluated models in Table III. For IU XRay, we compared FFN(20), KG(20), FFN(41), AG(41) and ATAG(41) with the settings as explained in the previous section. We observed that as the number of abnormalities considered in the model increases from 20 to 41, the performance improvement is significant in terms of both ROC-AUC ($P_{t-test} < 0.001$) and PR-AUC ($P_{t-test} < 0.001$) regardless AG/ATAG used or not. With the use of ATAG, the improvement on attributes classification is significant based on ROC-AUC ($P_{t-test} < 0.05$) and PR-AUC ($P_{t-test} < 0.01$). For PR-AUC scores, we notice that the variance is large. This is because the attributes of different abnormalities are imbalanced as reported in Table. I.

**D. Performance on Report Generation**

We evaluate the clinical accuracy of the reports generated by the baselines and ATAG-based methods based on IU XRay and MIMIC CXR.

1) **Results on IU XRay:** According to Tables IV and V, we observe that models with more abnormalities included again can lead to significant improvement in terms of both classification accuracy and NLG metrics. "Top-K" is set to Top-25 for IU XRAY and Top-50 for MIMIC CXR. The best scores are in bold face and the second best are underlined. "B.", "R." and "C." stand for BLEU, ROUGE-L and CIDEr scores. For BLEU, we take average of "BLEU-1", "BLEU-2", "BLEU-3" and "BLEU-4".
CE and RadRQI-F1. For example, a 103% improvement on RadRQI-F1 (Top-K) is achieved by SentSAT(41) compared to SentSAT(20), and a 30% more improvement is achieved by SentSAT+ATAG+GATE(41+106). In addition, we notice that among the models adopting LSTM-based decoders, SentSAT integrated with ATAG+GATE can give either superior or comparable performance on CE and RadRQI-F1 when compared with all the baselines. Specifically, incorporating ATAG+GATE can outperform the SOTA method based on CE(13), RadRQI-F1(13) and RadRQI-F1 (Top-K), and give comparable performance based on CE(5) and RadRQI-F1(5) with less abnormalities covered.

We also evaluate the abnormality coverage by Hits metric as the generated report is expected not to miss out true abnormalities as far as possible. The incorporation of ATAG and GATE is particularly effective in enhancing the coverage.

In addition, we notice that vanilla Transformer can achieve a high CE(13+NL) score but not for CE(13) and CE(Hits). Closer examination shows Transformer tends to generate common sentences like “No finding”, making the accuracy of “normality” high. Without taking into account “normality”, its CE(13) and CE(Hits) scores drop sharply, indicating that many abnormalities are in fact missed.

Regarding the language quality of the generated reports, the results are also reported in Table IV. For the decoders integrated with ATAG and GATE, although the best scores are not achieved, their performance is still comparable. For the various models we tested, the inconsistency between “BLEU” and “CIDEr” metrics is observed. A high “BLEU” score can be achieved if a report is “dominated” by the common normality descriptions even if no specific description is generated.

2) Results on MIMIC CXR: Tables IV and V also show the advantage of introducing ATAG and GATE for MIMIC CXR. In addition, by comparing the performance between the models with the Transformer as the decoder and those with the LSTM, the former being more powerful can outperform the latter significantly for MIMIC CXR. This however is not the case for IU XRay. This hints that the Transformer needs a large amount of data for training.

Also, among those using the Transformer-based decoder, Trans.+ATAG mostly gives the best performance in terms of CE while Trans.+ATAG+GATE generally underperforms in terms of RadRQI-F1. To better understand the cases where Trans.+ATAG+GATE underperforms Trans.+ATAG, we studied the failure cases found in Trans.+ATAG+GATE but not in Trans.+ATAG. We are mostly corresponding to Cardiomegaly which covers 35% of the testing cases in MIMIC CXR. Cardiomegaly is an abnormality related to heart enlargement, while Enlarged Cardiomegaly is another. These two abnormalities are separately evaluated in CE. Trans.+ATAG can detect heart abnormal enlargement, and generate descriptions of both Enlarged Cardiomegaly and Cardiomegaly when needed. On the contrary, Trans.+ATAG+GATE generates only either one, which is probably due to the gating mechanism which discourages generating repeated contents of the abnormality. This accounts for the lower accuracy of Trans.+ATAG+GATE for Cardiomegaly, and thus the lower overall micro-avg CE score.

How to better control the gating mechanism to alleviate such situations remains open.

E. Sensitivity to the Size of ATAG

To understand how the clinical accuracy is affected by the size of the ATAG, we constructed ATAGs with different numbers of abnormalities and attributes by putting thresholds on their occurrence frequencies (see Table I). The performance comparison results are shown in Table VI. According to the RadRQI-F1(13) and RadRQI-F1 (Top-K) columns, we observe the ATAG of larger size in general can achieve higher scores, indicating better clinical accuracy. This is consistent to our conjecture that the graph structure which can better represent more abnormalities in general can enable reports with more detailed abnormalities and attributes to be generated. Yet, we observe some exceptional cases and how to obtain the graph structure to achieve the optimal performance remains open.

We also notice that the improvement gained by larger models is more prominent for MIMIC CXR which is much larger than IU XRay. In particular, for IU XRay, the improvement from ATAG(28+79) to ATAG(41+106) is marginal, which is probably due to the fact that the additional abnormalities and attributes are rare and capturing them as well does not contribute much to the overall accuracy. For MIMIC CXR, it is much larger and the improvement gain due to the increase in the ATAG size is more obvious. E.g., Transformer-based decoder with ATAG(47+209) can achieve an improvement of 20.4% over ATAG(26+106) on RadRQI-F1 (Top-K).

F. Case Study

Fig. 6 shows two cases from IU XRay, each with the ground truth report and the reports generated by the baselines and the proposed models. We apply a post-processing step of removing duplicated or short sentences from the generated reports, and show the disease/abnormality keywords extracted by CheXpert labeler for CE metrics and by RadLex for...
Fig. 6. Illustration of reports generated by the baselines and models integrated with ATAG (yellow background color) and GATE (orange background color) on the IU XRay dataset. The correct abnormality and attribute terms are highlighted with green and blue colors. The expert-labeled annotations provided by [15] are also presented in dot-line boxes for reference.

RadRQI-F1 from the ground truth and the generated reports accordingly. We observe that integrating ATAG can generate more accurate abnormalities, and the gating mechanism can further increase the accuracy of the associated attributes.

Yet, some abnormalities cannot be well distinguished due to several reasons. For example, in Fig. 6 (left), Trans+ATAG and Trans+ATAG+GATE report to detect the eventration of hemidiaphragm where half of “base chest cavity” is denominated by the whiter density. However, it could be the whole chest cavity is elevated (moved up) instead of abnormal enlargement of the diaphragm in the base as predicted by the model. This false positive error may be caused by the non-standard radiographic positioning or anatomy variability.

Also, when the visual pattern hinting “atelectasis” is detected, the Trans+ATAG+GATE generates “less severe consolidation in the right lower lobe is either pneumonia or atelectasis” with three possible abnormalities of similar patterns. It suggests that the model tries to avoid missing potential abnormalities as far as possible. In addition, all models fail to detect “deformity of thoracic vertebrae” as indicated in Fig. 6 (left), even though some can predict certain bone-related abnormalities such as “calcification” and “degenerative changes of thoracic spine”.

In general, distinguishing similar abnormalities of the same anatomical part is challenging.

IX. CONCLUSION

In this paper, we propose to automatically construct a fine-grained attributed abnormality graph (ATAG) and the corresponding embeddings for representing abnormalities in X-ray images. To the best of our knowledge, this is the first attempt to construct the detailed attributed graph structure automatically from annotated reports and then learn the embeddings for the report generation. A hierarchical attention mechanism is proposed to aggregate the abnormality and attribute embeddings, and a gating mechanism is employed to allow more adaptive integration of ATAG embedding with...
either the LSTM-based or Transformer-based decoder. We performed comprehensive empirical evaluation on two benchmark datasets. The proposed ATAG-based models can outperform the SOTA methods in terms of abnormality classification accuracy and clinical accuracy of the generated report. The limitations of the methodologies are also discussed. Future research directions include consideration of ambiguous and potentially incorrect annotations, and integration of EHR data of different modalities to further enhance the clinical accuracy.

REFERENCES

[1] Y. Xue et al., “Multimodal recurrent model with attention for automated radiology report generation,” in Proc. 21st Int. Conf. Med. Image Comput. Comput.-Assist. Intervent., 2018, pp. 457–466.
[2] Z. Han, B. Wei, S. Leung, J. Chung, and S. Li, “Towards automatic report generation in spine radiology using weakly supervised framework,” in Proc. 21st Int. Conf. Med. Image Comput. Comput.-Assist. Intervent., 2018, pp. 185–193.
[3] X. Xie, Y. Xiong, P. S. Yu, K. Li, S. Zhang, and Y. Zhu, “Attention-based abnormal-aware fusion network for radiology report generation,” in Proc. 24th Int. Conf. Database Syst. Adv. Appl., 2019, pp. 448–452.
[4] X. Yang, M. Ye, Q. You, and F. Ma, “Writing by memorizing: Hierarchical retrieval-based medical report generation,” in Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process., 2021, pp. 5000–5009.
[5] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, “Generating radiology reports via memory-driven transformer,” in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2020, pp. 1439–1449.
[6] Y. Miura, Y. Zhang, E. Tsai, C. Langlotz, and D. Jurafsky, “Improving factual completeness and consistency of image-to-text radiology report generation,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol., 2021, pp. 5288–5304.
[7] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, “Exploring and distilling posterior and prior knowledge for radiology report generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 13753–13762.
[8] B. Jing, P. Xie, and E. Xing, “On the automatic generation of medical imaging reports,” in Proc. 56th Annu. Meeting Assoc. Comput. Linguistics, 2018, pp. 2577–2586.
[9] J. Yuan, H. Liao, R. Luo, and J. Luo, “Automatic radiology report generation based on multi-view image fusion and medical concept enrichment,” in Proc. 22nd Int. Conf. Med. Image Comput. Comput.-Assist. Intervent., 2019, pp. 721–729.
[10] G. Liu et al., “Clinically accurate chest X-ray report generation,” in Proc. Mach. Learn. Healthcare Conf., 2019, pp. 249–269.
[11] B. Jing, Z. Wang, and E. Xing, “Show, describe and conclude: On exploiting the structure information of chest X-ray reports,” in Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, 2019, pp. 6570–6580.
[12] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, “Knowledge-driven encode, retrieve, paraphrase for medical image report generation,” in Proc. AAAI, vol. 33, 2019, pp. 6666–6673.
[13] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, “When radiology report generation meets knowledge graph,” in Proc. 34th AAAI Conf. Artif. Intell., 2020, pp. 12910–12917.
[14] F. Liu et al., “Auto-encoding knowledge graph for unsupervised medical report generation,” in Proc. Adv. Neural Inf. Process. Syst., vol. 34, 2021, pp. 16266–16279.
[15] D. Demner-Fushman et al., “Preparing a collection of radiology examinations for distribution and retrieval,” J. Amer. Med. Inform. Assoc., vol. 23, no. 2, pp. 304–310, 2015.
[16] A. E. W. Johnson et al., “MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports,” Sci. Data, vol. 6, no. 1, p. 317, Dec. 2019.
[17] K. Xu et al., “Show, attend and tell: Neural image caption generation with visual attention,” in Proc. 32nd Int. Conf. Mach. Learn. (ICML), vol. 37, Feb. 2015, pp. 2048–2057.
[18] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, “A hierarchical approach for generating descriptive image paragraphs,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 317–325.
[19] Z. Wang, H. Han, L. Wang, X. Li, and L. Zhou, “Automated radiographic report generation purely on transformer: A multi-criteria supervised approach,” IEEE Trans. Med. Imag., vol. 41, no. 10, pp. 2803–2813, Oct. 2022.
[20] D. You, F. Liu, S. Ge, X. Xie, J. Zhang, and X. Wu, “AlignTransformer: Hierarchical alignment of visual regions and disease tags for medical report generation,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent., 2021, pp. 72–82.
[21] S. Yan, “Memory-aligned knowledge graph for clinically accurate radiology image report generation,” in Proc. 21st Workshop Biomed. Lang. Process., 2022, pp. 1–17.
[22] Z. Chen, Y. Shen, Y. Song, and X. Wan, “Cross-modal memory networks for radiology report generation,” in Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process., vol. 1, 2021, pp. 5904–5914.
[23] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph attention networks,” 2017, arXiv:1710.10903.
[24] M. Martinez-Romero, C. Jonquet, M. J. O’Connor, J. Graybeal, A. Pazos, and M. A. Musen, “NCBO ontology recommender 2.0: An enhanced approach for biomedical ontology recommendation,” J. Biomed. Semantics, vol. 8, no. 1, pp. 1–22, Dec. 2017.
[25] A. Hogan, “Knowledge graphs,” ACM Comput. Surv., vol. 54, no. 4, pp. 1–37, 2021.
[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4700–4708.
[27] A. Graves, “Long short-term memory,” in Supervised Sequence Labelling With Recurrent Neural Networks. Berlin, Germany: Springer, 2012, pp. 37–45.
[28] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5998–6008.
[29] J. Jain et al., “RadGraph: Extracting clinical entities and relations from radiology reports,” 2021, arXiv:2106.14463.
[30] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, “Hybrid retrieval-generation reinforced agent for medical image report generation,” in Proc. 32nd Int. Conf. Neural Inf. Process. Syst., 2018, pp. 1530–1540.
[31] J. Irvin et al., “CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison,” in Proc. 33rd AAAI Conf. Artif. Intell., vol. 33, Jul. 2019, pp. 590–597.
[32] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics, 2002, pp. 311–318.
[33] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81.
[34] R. Vedantam, C. L. Zitnick, and D. Parikh, “CIDEr: Consensus-based image description evaluation,” in Proc. 28th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 4566–4575.
[35] J. Lu, C. Xiong, D. Parikh, and R. Socher, “Knowing when to look: Adaptive attention via a visual sentinel for image captioning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 375–383.
[36] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, “Meshed-memory transformer for image captioning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 10578–10587.