

# Memory-aligned Knowledge Graph for Clinically Accurate Radiology Image Report Generation

Sixing Yan

Department of Computer Science,  
Hong Kong Baptist University,  
Hong Kong SAR, China  
cssxyan@comp.hkbu.edu.hk

## Abstract

Automatic generating the clinically accurate radiology report from X-ray images is important but challenging. The identification of multi-grained abnormal regions in image and corresponding abnormalities is difficult for data-driven neural models. In this work, we introduce a Memory-aligned Knowledge Graph (MaKG) of clinical abnormalities to better learn the visual patterns of abnormalities and their relationships by integrating it into a deep model architecture for the report generation. We carry out extensive experiments and show that the proposed MaKG deep model can improve the clinical accuracy of the generated reports.

## 1 Introduction

Medical images are complex and hard to understand without specialized expertise. Given that the volume of radiology images is large, automatically generating the reports by the computer-aided system can alleviate the radiologists from the time-consuming reporting task. Recently, many deep learning models are studied in the automated radiology report generation (Han et al., 2018; Xie et al., 2019; Yang et al., 2021; Chen et al., 2020).

The deep encoder-decoder architecture has been commonly adopted in the report generation, where visual features were extracted from the input medical images using a convolutional neural network and fed to a recurrent neural network to generate the report. Different from image captioning which inputs one image and output one sentence, the report has much longer length while the correctness of medical entities generated in the report is the core requirement. More than the requirement of detecting abnormalities accurately like classification, the report is expected to provide the support details of present abnormalities. Thus, generating accurate report with readable and logical descriptions by natural language generation model is the key challenge in the report generation task.

Generating correct reports is impossible if the pathology of abnormal regions and corresponding abnormalities cannot be identified at first. Most existing studies (Liu et al., 2021a; Chen et al., 2020, 2021; You et al., 2021) proposed the attention and memory mechanism to enhance the identification of abnormal regions. However, different status of the same abnormality may have their specifics and the correlations of these visual patterns are ignored. In addition, identifying the actual abnormalities from abnormal regions is also challenging since the complex and rare abnormalities are hard to determined without professional knowledge.

To incorporate the prior medical knowledge, several research (Li et al., 2019; Zhang et al., 2020; Liu et al., 2021b) applied medical knowledge graph of certain abnormalities in the report generation aiming to learn the abnormality relationships. The corresponding representations, i.e., graph embedding, are computed by graph neural network given the input images. However, such representations are affected by the inner-connections of abnormalities for each input where the general characteristic of abnormalities are missing. For example, the representations of “*Effusion*”, computed as graph embedding, are different when “*Effusion*” appears with or without “*Atelectasis*”. But the general characteristic of “*Effusion*” over all relevant observations, e.g., density or shapes, are only determined by itself independently. This general but independent characteristic is still missed to model by existing approaches which limits the effectiveness the knowledge graph.

To alleviate the above challenges, in this work, we propose to learn the memory-aligned graph model, aiming to enhances the pathology identification and prior medical knowledge incorporation. The memory features of possible abnormal regions are first aligned by the input visual feature in an alternative manner, and concatenated with a universal memory embedding before feeding to

the graph attention network to compute the graph embedding. The graph embedding are later learned by the classification and fine-tuned in the report generation. We evaluate the proposed approach using two publicly accessible datasets. The evaluation results show the effectiveness of utilizing memory-aligned knowledge graph in generating the clinically accurate radiology report.

## 2 The Proposed Method

### 2.1 Problem Formulation

Given the radiology image with extracted visual features as  $V$ , the model aims to generate a radiology report  $R = \{y_1, y_2, \dots\}$ . We introduce a **Memory-aligned Knowledge Graph (MaKG)** to explore multi-grained features of the abnormalities and their relationships. The multi-grained memory features  $\hat{M}$  are first aligned from the memory slots  $M$  by  $V$ , and concatenated with a meshed memory embedding  $E$  to learn the abnormality graph embedding  $G$  for generating radiology report  $R$ . This process can be formulated as,

$$\{V, M\} \rightarrow \hat{M}; \{\hat{M}, E\} \rightarrow G; G \rightarrow R. \quad (1)$$

**Implementation.** Following (Chen et al., 2020, 2021; Liu et al., 2021b), we adopt a memory slots  $M \in \mathbb{R}^{M \times D}$  to record the information of abnormal regions which would indicate the potential abnormalities. The memory slots are initialized as plain learnable vectors and updated together with other modules. The  $M$  stands for the total number of the knowledge corresponding to the abnormality identification. We also adopt a  $E \in \mathbb{R}^{N \times D}$  embedding to model the universal features of each abnormality. The  $N$  is equal to the number of the abnormalities. We follow (Zhang et al., 2020) to construct and initialize the abnormality knowledge graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E}); |\mathcal{V}| = N$  which is a universal structure in the training. The nodes  $\mathcal{V}$  cover the common chest abnormalities and grouped by their organ or body part appearances as edges  $\mathcal{E}$ . The graph embedding  $G \in \mathbb{R}^{N \times D}$  is computed by the graph attention network. A overview of this framework is shown in Fig. 1.

### 2.2 Memory-aligned Graph Embedding

To learn the visual patterns of possible abnormal regions, we apply Multi-Head Attention (MHA) (Vaswani et al., 2017) to query the responding memory features from the memory slots  $M$ . The MHA computes the associated

weighted between different features which allows the abnormality-related memory features to be distilled from original  $M$ . To align different level of the alignment, we can perform the alignment attention alternatively as,

$$\begin{aligned} V'_{i+1} &= \text{MHA}(M_i, V_i); \\ M'_{i+1} &= \text{MHA}(V'_{i+1}, M_i), \end{aligned} \quad (2)$$

where  $V_0 = V$ ,  $M_0 = M$ ,  $V'_i$  and  $M'_i$  denote  $i$ -th step aligned visual and memory features, respectively. As observed, the patterns of abnormal regions should be learned in different fine-grained ways due to their variable shapes and sizes. Thus, we follow (You et al., 2021) to repeat the alignment  $K$  times and obtain multi-grained memory features  $\{M'_i\} = \{M'_1, M'_2, \dots, M'_K\}$ . We then aggregate the multi-grained memory features as  $\hat{M} = \text{MHA}(M'_*, M'_*)$ , where  $M'_* = \bigoplus_{i=1}^K M'_i$  and  $\hat{M} \in \mathbb{R}^{M \times D}$ .

To model the prior knowledge on the global characteristic of each abnormality which may not depend on the current input  $V$ , we add an meshed memory embedding  $E \in \mathbb{R}^{N \times D}$  of which each row represent one particular abnormality. We compute the graph embedding  $G \in \mathbb{R}^{N \times D}$  using graph attentional layer  $\text{GAT}(\cdot)$  (Veličković et al., 2017) as,

$$G = \text{GAT}(\text{FFN}(MW^G \oplus E)) \quad (3)$$

where  $\text{FFN}(x) = \text{ReLU}(xW_1^{\text{ff}} + b_1^{\text{ff}})W_2^{\text{ff}} + b_2^{\text{ff}}$ ,  $W_1^{\text{ff}}, W_2^{\text{ff}} \in \mathbb{R}^{D \times D}$  and  $W^G \in \mathbb{R}^{M \times N}$  are learnable parameters,  $b_1^{\text{ff}}, b_2^{\text{ff}}$  are learnable bias vectors. We learn  $G$  by adding a fully-connected layer with *Sigmoid* activation for each node and serving it as a binary classifier. Each node embedding is used to predict the existence probability of corresponding abnormality, and the classifier is trained using weighted binary cross entropy loss. The details can be found in (Zhang et al., 2020).

### 2.3 Report Generation by Graph embedding

For each decoding step  $t$ , the hidden stats  $h_t$  is encoded from the input word features  $x_t$  by the standard encoder from Transformer,

$$x_t = w_t + e_t; h_t = \text{MHA}(x_t, x_{1:t}), \quad (4)$$

where  $w_t$  and  $e_t$  are the word embedding and positional embedding, respectively. A  $L$  layers Transformer decoder is employed to generate the proper report by the attending MaKG embeddings  $G$  as,

$$\begin{aligned} h'_t &= \text{MHA}(h_t, G); \\ y'_t \sim p_t &= \text{Softmax}(h'_t W + b). \end{aligned} \quad (5)$$

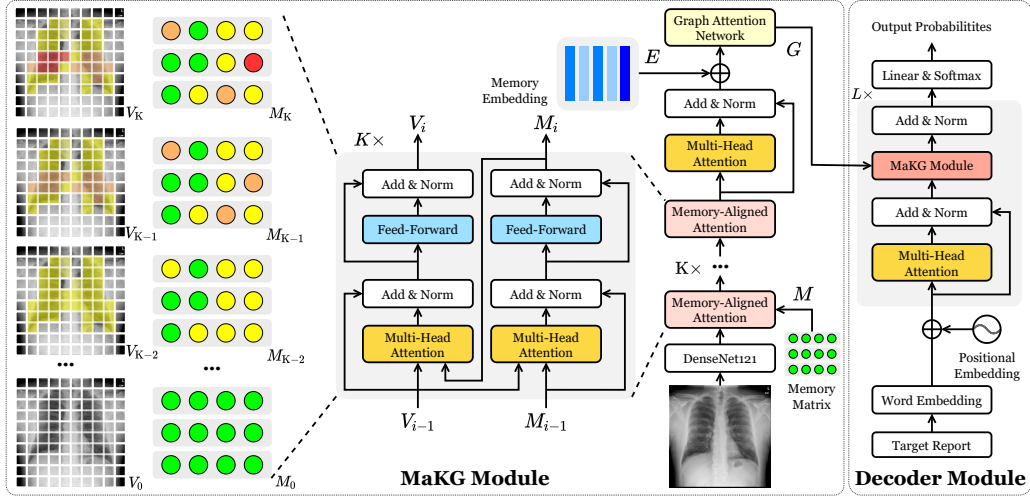


Figure 1: The MaKG-based deep model architecture.

Both encoder and decoder are trained by minimizing the cross-entropy loss  $L_{gen}(\theta) = -\sum_{t=1}^T \log(p_t | p_{1:t-1})$ .

### 3 Experiments

#### 3.1 Datasets, Metrics and Settings

We use two publicly available datasets IU X-Ray (Demner-Fushman et al., 2016) and MIMIC CXR (Johnson et al., 2019) for evaluating the model performances. For the IU X-Ray dataset, we collect 2,848 reports and 5,696 images containing both frontal and lateral chest X-rays. We partitioned the data into train/validate/test set by 7:1:2 for cross validation. For MIMIC CXR dataset, we follow original split set with train/validate/test size as 222,705 / 1,807 / 3,269 and report the average scores of three different runs.

For report quality, we adopt the language generation metrics including BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004) and CIDEr (Vedantam et al., 2015). To measure the clinical accuracy, we adopt the Clinical Efficacy (CE) (Chen et al., 2020) and Clinical Metrics (CM) (Miura et al., 2021) for common and critical observation accuracy, and MIRQI (Zhang et al., 2020) to evaluate accuracy of 14<sup>1</sup> observations and their associated attributes. The micro-avg F1-measure scores are reports.

To compare with the proposed model **TRANS.+MAKG**, we employ the basic vanilla

<sup>1</sup>14 clinical observations includes: *No finding, Enlarged Cardiomeastinum, Cardiomegaly, Lung lesion, Lung opacity, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural effusion, Pleural other, Fracture, Support devices*

**TRANS.** with three layers, 8 heads and 512 hidden state dimension, and an integration knowledge graph used in (Zhang et al., 2020) denoted as **TRANS.+KG**. We also compare **TRANS.+MAKG** with several report generation models, including **WORDSAT** (Xu et al., 2015), **ADAATTN** (Lu et al., 2017), **SENTSAT** (Krause et al., 2017), **COATTN** (Jing et al., 2018), **SENTKG** (Zhang et al., 2020), **M<sup>2</sup>TRANS** (Cornia et al., 2020), **R2GEN** (Chen et al., 2020) and **R2GEN-CMN** (Chen et al., 2021).

We adopt DenseNet121 (Huang et al., 2017) to extract the visual features. The dimensions of hidden state and number of heads in MHA are set as 512 and 8.  $K$  and  $M$  are set as 3 and 20. The model is trained with the learning rate  $5e-5$  in the end-to-end manner.

#### 3.2 Results on Multi-label Classification

For performance comparisons on image classification, we evaluate the proposed MAKG with the base DENSENET (Huang et al., 2017) and integrating with KG (Zhang et al., 2020) embedded with different graph neural network. Higher or equivalent scores are obtained for most of the classes as shown in Table 1. A possible explanation is that the alignment mechanism of MAKG enhances the learning of the abnormality patterns by distilling the irrelevant regions from the images.

#### 3.3 Results on Report Generation

The main focus of this experiment is to evaluate the effectiveness of applying memory alignment knowledge graph (MaKG) in enhancing the clinical accuracy of the report generation.

Class	Integration Module			
	-	KG*	KG	MaKG
Normal/No Finding	0.795	<u>0.807</u>	0.806	<b>0.821</b>
Cardiomegaly	0.866	0.913	<u>0.922</u>	<b>0.930</b>
Scoliosis	0.664	0.663	<u>0.671</u>	<b>0.687</b>
F.B.	<u>0.695</u>	0.671	0.686	<b>0.727</b>
Effusion	0.921	0.942	<u>0.950</u>	<b>0.962</b>
Thickening	0.733	0.728	<u>0.753</u>	<b>0.785</b>
Pneumothorax	0.824	<u>0.843</u>	<u>0.843</u>	<b>0.889</b>
H.H	0.860	<b>0.884</b>	0.857	<u>0.870</u>
Calcinosis	0.676	<u>0.669</u>	<u>0.669</u>	<b>0.690</b>
Emphysema	0.892	0.890	<u>0.902</u>	<b>0.919</b>
Pneumonia	0.844	<b>0.863</b>	0.835	<u>0.861</u>
Edema	0.897	<u>0.931</u>	0.912	<b>0.949</b>
Atelectasis	0.788	<u>0.833</u>	0.823	<b>0.838</b>
Cicatrix	0.742	0.734	<u>0.745</u>	<b>0.774</b>
Opacity	0.796	0.803	<u>0.806</u>	<b>0.829</b>
Lesion	0.597	0.643	0.630	<u>0.647</u>
Airspace Disease	0.830	<b>0.857</b>	0.823	<u>0.846</u>
Hypoinflation	0.768	<u>0.775</u>	0.767	<b>0.791</b>
Medical Device	0.775	<u>0.805</u>	0.798	<b>0.825</b>
Other	0.595	<u>0.596</u>	<u>0.607</u>	<b>0.653</b>
Average	0.778	<u>0.792</u>	0.867	<b>0.879</b>

Table 1: Performance on multi-label classification (AUC) on IU XRay dataset. The best scores are in bold face and the second best are underlined.

**Clinical Accuracy Metric** As shown in Table 3, TRANS.+MAKG achieves the first and second best performances over all clinical accuracy related metrics, and outperforms TRANS+KG with significantly improvement in MIRQI score which evaluates the accuracy of both abnormalities and their associated attributes. It indicates integrating MaKG is able to enhance the generation of clinically accurate report by providing correct attribute descriptions in the fine-grained level. This observation is important because the correctness of the associated attributes is necessary for the correctness of the abnormality descriptions. The incomplete or incorrect attributes of the same abnormalities would result different or even incorrect follow-up treatments. Noted that TRANS.+MAKG does not obtain the first best score in CE which measures the accuracy of 13 clinical observations and normality observation. However, the best scores of CM and Hits are observed shows that TRANS.+MAKG is able to identify the most critical abnormalities and cover most of the abnormalities that are frequently mentioned in the report repositories.

As observed from Table. 3, no model could detect all evaluated abnormalities for IU XRay dataset. Thus, we further study the detailed results as shown in Table. 2. As observed, there are some abnormalities of which appearance ratio is around 5% in the whole training set which is relatively rare.

The failed detection could be caused by different reasons, such too few training data (e.g., “Fracture”) or too hard to learning (e.g., “Pneumothorax”) which is also very hard for clinicians to determine).

Class (%)	Integration Module		
	-	KG	MaKG
No Finding (31.72%)	<b>0.603</b>	0.500	0.456
Enlarged Cardio. (13.3%)	0.000	0.000	<b>0.034</b>
Cardiomegaly (15.6%)	0.265	<b>0.392</b>	0.341
Lung Lesion (5.2%)	0.000	0.000	<b>0.054</b>
Lung Opacity (21.3%)	0.181	0.209	<b>0.278</b>
Edema (4.7%)	0.000	0.000	<b>0.160</b>
Consolidation (5.2%)	0.000	0.038	<b>0.073</b>
Pneumonia (3.0%)	0.000	0.000	0.000
Atelectasis (8.1%)	0.000	0.087	<b>0.227</b>
Pneumothorax (6.6%)	0.000	0.000	0.000
Pleural Effusion (10.2%)	0.089	0.172	<b>0.278</b>
Pleural Other (1.6%)	0.000	0.000	0.000
Fracture (2.9%)	0.000	0.000	0.000
Support Devices (3.9%)	0.091	0.114	<b>0.242</b>

Table 2: Detailed CE evaluation results (F1-measure) of TRANSFORMER and integrating with KG and MAKG in IU XRay dataset, respectively. The best scores are in bold face

**Natural Language Generation Metrics** As the experimental results show, the higher NLG scores do not always indicate the clinically accurate reports are generated. While the clinical accuracy is a mission-critical requirement for radiology report generation, the generated report is expected to be clinically accurate using relatively readable sentences. The TRANS.+MAKG achieves similar NLG scores which indicates that the integration of MaKG is able to generate more reasonable descriptions of the abnormalities without decreasing the informativeness from TRANS. much. More powerful decoders (e.g., MemroyTrans. (Chen et al., 2020) or AlignTrans. (You et al., 2021)) should be able to enhance the overall performances.

**Qualitative Results** As shown in Fig. 2, two cases of ground truth and generated reports are visualized. The extracted clinical findings and the associated modifications are also attached. As observed, TRANS.+MAKG is able to detect more correct abnormalities in such cases than TRANS.+KG. It is believed to assistant clinicians to detect the abnormalities which are easy to ignored, thus increases the usability of applying the MaKG in improving the clinical accuracy in the report generation task.

Dataset	Model	NLG Metrics				Clinical Accuracy Metrics			
		B.	M.	R.	C.	CM	CE	MIRQI	Hits (14)
IU XRay	WORDSAT (Xu et al., 2015)	0.262	0.383	<u>0.369</u>	0.317	0.094	0.215	0.463	5.6
	ADAATTN (Lu et al., 2017)	0.269	0.379	0.367	0.358	<u>0.240</u>	0.338	0.474	6.6
	SENTSAT (Krause et al., 2017)	<u>0.274</u>	0.372	0.365	0.318	0.106	0.241	0.451	4.8
	CoATTN (Jing et al., 2018)	0.256	0.367	0.357	0.307	0.061	0.245	0.438	5.2
	SENTKG (Zhang et al., 2020)	0.271	<u>0.391</u>	0.367	0.304	0.067	0.242	0.490	4.8
	M <sup>2</sup> TRANS. (Cornia et al., 2020)	0.269	0.299	0.363	0.367	0.104	0.253	0.481	5.6
	R2GEN (Chen et al., 2020)	0.251	0.367	0.342	0.461	0.100	0.322	0.389	9.0
	R2GEN-CMN (Chen et al., 2021)	<b>0.294</b>	<b>0.392</b>	<b>0.370</b>	<b>0.681</b>	0.104	0.330	0.462	8.0
	TRANS. (Vaswani et al., 2017)	0.264	0.390	0.357	0.587	0.147	<b>0.394</b>	0.486	5.0
TRANS.+KG	0.265	0.380	0.353	<u>0.593</u>	0.205	0.320	<u>0.504</u>	<u>9.2</u>	
TRANS.+MAKG (ours)	0.265	0.378	0.353	0.523	<b>0.262</b>	<u>0.362</u>	<b>0.515</b>	<b>10.8</b>	
MIMIC CXR	WORDSAT (Xu et al., 2015)	0.160	0.284	0.249	0.082	0.354	0.324	0.391	10.0
	ADAATTN (Lu et al., 2017)	0.151	<b>0.301</b>	0.248	0.096	0.384	0.366	0.438	12.0
	SENTSAT (Krause et al., 2017)	<b>0.182</b>	0.236	<u>0.252</u>	0.073	0.412	0.364	0.411	11.3
	CoATTN (Jing et al., 2018)	<u>0.181</u>	0.235	<b>0.253</b>	0.070	0.423	0.364	0.418	9.7
	M <sup>2</sup> TRANS. (Cornia et al., 2020)	0.165	<u>0.299</u>	0.249	0.102	<b>0.458</b>	<b>0.469</b>	0.518	<u>13.7</u>
	R2GEN (Chen et al., 2020)	0.124	0.158	0.160	<b>0.170</b>	0.262	0.296	0.383	13.0
	R2GEN-CMN (Chen et al., 2021)	0.123	0.162	0.163	0.128	0.329	0.356	0.485	10.0
	TRANS. (Vaswani et al., 2017)	0.126	0.160	0.164	<u>0.167</u>	0.286	0.288	0.368	13.0
	TRANS.+KG	0.109	0.280	0.214	<u>0.119</u>	0.406	<u>0.398</u>	<u>0.535</u>	12.0
TRANS.+MAKG (ours)	0.137	0.284	0.228	0.120	<u>0.455</u>	<b>0.469</b>	<b>0.572</b>	<b>14.0</b>	

Table 3: Performance comparison of report generation models. The best scores are in bold face and the second best are underlined. “B.”, “M.”, “R.” and “C.” stand for BLEU, METEOR, ROUGE and CIDEr scores, respectively. The maximum number of “Hits” is 14 which is defined by CheXpert labeling toolkit.

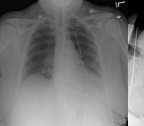

Ground Truth		The heart size is moderately enlarged. There is evidence of previous aortic valve replacement. XXXX sternotomy XXXX are grossly intact. The pulmonary XXXX and mediastinum are within normal limits. There is no pleural effusion or pneumothorax. There are chronically increased interstitial lung markings without superimposed focal airspace disease identified. There are degenerative changes of the spine. Cardiomegaly without superimposed acute disease noted.	
MIRQI Results	['calcifi', 'Calcinosis', 'POSITIVE', 'aorta'] ['airspace disease', 'Airspace Disease', 'POSITIVE', 'left/lobe'] ['contour', 'Enlarged Cardiomeastinum', 'POSITIVE', 'heart'] ['atherosclero', 'Other Finding', 'POSITIVE', 'disease'] ['infiltrate', 'Airspace Opacity', 'UNCERTAIN', 'atelectasis'] ['atelecta', 'Atelectasis', 'UNCERTAIN', 'infiltrate'] ['the heart', 'Cardiomegaly', 'UNCERTAIN', 'contours/stable']	['degenera', 'Other Finding', 'POSITIVE', 'changes'] ['cardiomegaly', 'Cardiomegaly', 'POSITIVE', 'disease'] ['interstitial lung', 'Airspace Opacity', 'POSITIVE', 'markings'] ['the heart', 'Cardiomegaly', 'POSITIVE', 'size'] ['valve', 'Support Devices', 'POSITIVE', 'replacement'] ['mediastinum', 'Enlarged Cardiomeastinum', 'NEGATIVE', 'limits'] ['effusion', 'Pleural Effusion', 'NEGATIVE', 'is/no/pleural/pneumothorax'] ['airspace disease', 'Airspace Disease', 'NEGATIVE', 'superimposed/focal'] ['pneumothorax', 'Pneumothorax', 'NEGATIVE', 'is/effusion']	
TRANS+KG	1. probable small bilateral pleural effusions . 2. possible lower thoracic <unk> deformity not well <unk> on todays study, the lungs are clear, there appear to be small bilateral pleural effusions . the heart is not grossly enlarged, there are atherosclerotic changes of the aorta, increased kyphosis is seen in the may be a thoracic <unk> deformity that is not <unk> . arthritic changes are seen .	stable cardiomegaly . no acute infiltrate or effusion . mildly enlarged, the cardiac silhouette and mediastinal contours are within normal limits . no pneumothorax or pleural effusion . clear .	
MIRQI Results	['deformity', 'Other Finding', 'UNCERTAIN', 'possible/thoracic'] ['kyphosis', 'Other Finding', 'POSITIVE', 'increase'] ['atherosclero', 'Other Finding', 'POSITIVE', 'changes'] ['the heart', 'Cardiomegaly', 'NEGATIVE', ''] ['effusion', 'Pleural Effusion', 'UNCERTAIN', 'probable/small/bilateral/pleural']	['cardiomegaly', 'Cardiomegaly', 'POSITIVE', 'stable'] ['infiltrate', 'Airspace Opacity', 'NEGATIVE', 'acute/effusion'] ['effusion', 'Pleural Effusion', 'NEGATIVE', 'acute/infiltrate'] ['cardiac silhouette', 'Cardiomegaly', 'NEGATIVE', 'contours'] ['pneumothorax', 'Pneumothorax', 'NEGATIVE', 'no/effusion'] ['contour', 'Enlarged Cardiomeastinum', 'NEGATIVE', 'cardiac/silhouette/mediastinal/limits']	
TRANS+MAKG	1. increased elevation right hemidiaphragm with right basilar atelectasis . left basilar airspace disease and pleural effusion unchanged . 2. interval removal of rightsided chest tube no pneumothorax . stable cardiomeastinal silhouette, there has been interval removal of the chest tube with increased elevation of the right hemidiaphragm and unchanged in the left basilar atelectasis .	1. no acute cardiopulmonary disease . 2. stable mild cardiomegaly . 3. prominent central vasculature . pa and lateral views of the chest were obtained . tracheostomy tube . probable mild cardiomegaly . prominence of the central vasculature unchanged . no pneumothorax pleural effusion or focal air space consolidation .	
MIRQI Results	['airspace disease', 'Airspace Opacity', 'POSITIVE', 'left/basilar/unchanged'] ['effusion', 'Pleural Effusion', 'POSITIVE', 'pleural/unchanged'] ['tube', 'Support Devices', 'NEGATIVE', 'chest'] ['atelecta', 'Atelectasis', 'POSITIVE', 'right/left/basilar'] ['elevation', 'Other Finding', 'POSITIVE', 'increased/hemidiaphragm'] ['mediastinal silhouette', 'Enlarged Cardiomeastinum', 'UNCERTAIN', 'cardiomeastinal']	['cardiomegaly', 'Cardiomegaly', 'POSITIVE', 'mild'] ['prominent', 'Other Finding', 'POSITIVE', 'vasculature'] ['tracheostomy', 'Other Finding', 'POSITIVE', 'tube'] ['tube', 'Support Devices', 'POSITIVE', 'tracheostomy'] ['consolidat', 'Consolidation', 'NEGATIVE', 'effusion/focal/air/space'] ['pneumothorax', 'Pneumothorax', 'NEGATIVE', 'effusion'] ['effusion', 'Pleural Effusion', 'NEGATIVE', 'no/pneumothorax/pleural/consolidation']	

Figure 2: Illustration of reports generated by TRANS.+KG and TRANS.+MAKG. The extracted medical entities by MIRQI evaluation toolkit are attached as [“keyphrase”, “category”, “negation”, “attributes”].

## 4 Conclusions

In this work, we propose a memory-aligned knowledge graph (MaKG) to enhance the clinically accurate report generation by modeling the relationship between abnormal regions and particular abnormalities. The experiments prove the effectiveness of integrating MaKG with the generation model is able to generate descriptive report with both correct abnormalities and associated attributes. In ad-

dition, the proposed MaKG is not limited to the specific knowledge graph structure which give the opportunities on incorporating different professional knowledge for specific medical applications.

## 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 Ex-*

- trinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5904–5914.
- Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing 2020*, pages 1439–1449.
- Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587.
- 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.
- Zhongyi Han, Benzhen Wei, Stephanie Leung, Jonathan Chung, and Shuo Li. 2018. Towards automatic report generation in spine radiology using weakly supervised framework. In *Proceedings of the 21th International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 185–193.
- Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708.
- Baoyu Jing, Pengtao Xie, and Eric P. Xing. 2018. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2577–2586.
- Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.
- Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–325.
- Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2019. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 6666–6673.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021a. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13753–13762.
- Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Xu Sun, et al. 2021b. Auto-encoding knowledge graph for unsupervised medical report generation. *Advances in Neural Information Processing Systems*, 34.
- Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383.
- Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304.
- 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.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.
- Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.
- Xiancheng Xie, Yun Xiong, Philip S. Yu, Kangan Li, Suhua Zhang, and Yangyong Zhu. 2019. Attention-based abnormal-aware fusion network for radiology report generation. In *Proceedings of the 24th International Conference on Database Systems for Advanced Applications*, pages 448–452.

- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057.
- Xingyi Yang, Muchao Ye, Quanzeng You, and Fenglong Ma. 2021. Writing by memorizing: Hierarchical retrieval-based medical report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5000–5009. Association for Computational Linguistics.
- Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. 2021. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 72–82. 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 34th AAAI Conference on Artificial Intelligence*, pages 12910–12917.