PEKING UNIVERSITY

Auto-Encoding Knowledge Graph for Unsupervised Medical Report Generation

Fenglin Liu[1], Chenyu You[3], Xian Wu[4], Shen Ge[4], Sheng Wang[2], Xu Sun[1]

[1] Peking University,

[2] Paul G. Allen School of Computer Science and Engineering, University of Washington,

[3] Department of Electrical Engineering, Yale University, [4] Tencent

Introduction

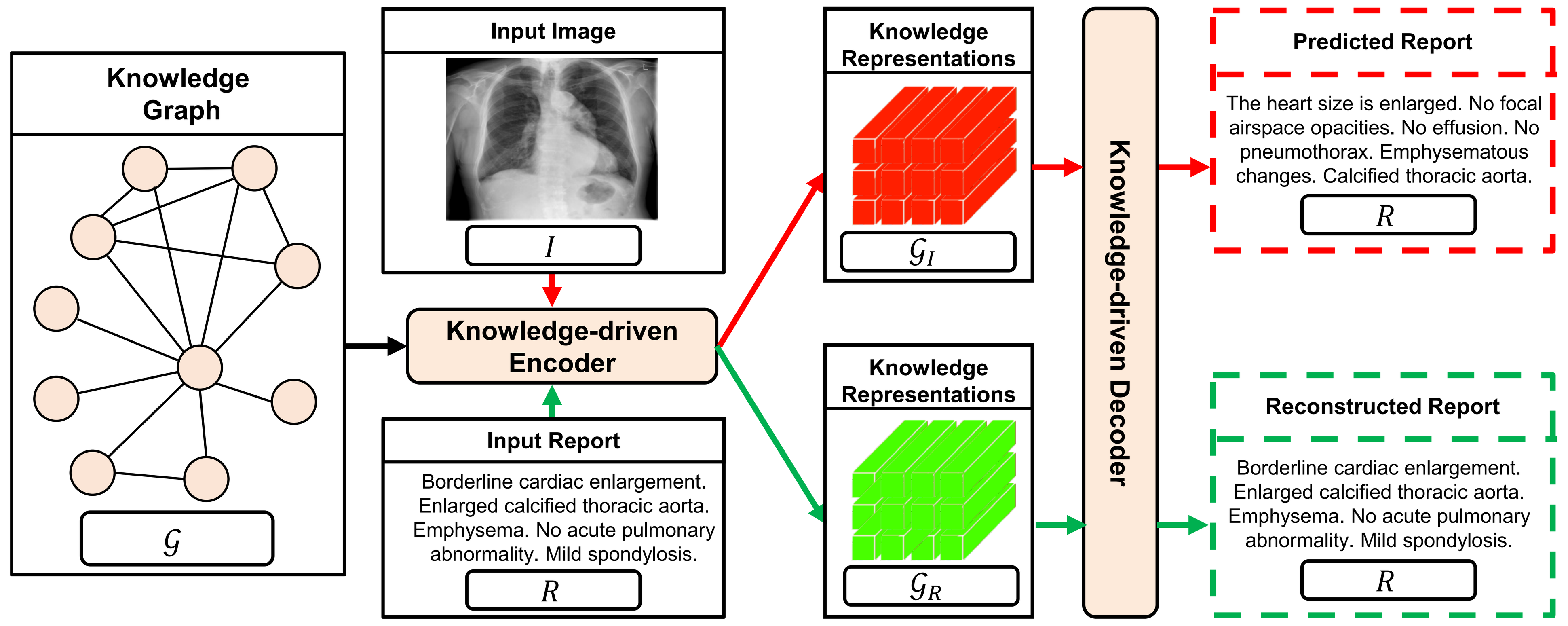

Figure 1. Illustration of our Knowledge Graph Auto-Encoder, which consists of a pre-constructed knowledge graph, a knowledge-driven encoder and a knowledge-driven decoder. The Green and Red lines denote the data flow in the training process and testing process of report generation, respectively.

Background:

- As shown in Figure 1, Medical Report Generation task aims to describe the clinical findings (R) in the input medical image (I), which can assist radiologists in clinical decision-making.
- Currently, the data-driven deep neural models, particularly those based on the encoder-decoder frameworks have achieved great success in advancing the state-of-the-art of medical report generation.

Limitation & Challenge:

- Existing models are trained in a supervised learning manner and heavily rely on labeled paired image-report datasets, which are not easy to acquire in the real world.
- The medical-related data can only be manually labeled by professional radiologists, and also involves privacy issues.
- The scales of widely-used datasets for medical report generation are relatively small compared to natural image related datasets.

To relax the reliance on the paired datasets, making use of all available data, like independent image or report sets, is important.

Approach

In this paper, we propose an unsupervised model Knowledge Graph Auto-Encoder (KGAE), which utilizes independent sets of images and reports in training (the image and report set are separate and have no overlap). As shown in Figure 1, our proposed KGAE consists of a pre-constructed **knowledge graph**, a **knowledge-driven encoder** and a **knowledge-driven decoder**.

Pre-constructed Knowledge Graph

In particular, we construct an off-the-shelf global medical knowledge graph $\mathcal{G} = (V, E)$ covering the common abnormalities and normalities, where $V = \{v_i\}_{i=1:N_{\mathrm{KG}}} \in \mathbb{R}^{N_{\mathrm{KG}} \times d}$ is a set of nodes and $E = \{e_{i,j}\}_{i,j=1:N_{\mathrm{KG}}}$ is a set of edges. In detail, based on the report corpus, i.e., MIMIC-CXR [1], we consider the N_{KG} frequent clinical abnormalities (e.g., "enlarged heart size") and normalities (e.g., "heart size is normal" and "lungs are clear") as nodes. The edge weights are calculated by the normalized co-occurrence of different nodes computed from report corpus.

Knowledge-driven Encoder

The knowledge-driven encoder, including a common mapping function $\mathcal{F}$, take either the image I or the report R as queries and project them to the same latent space, acquiring $\mathcal{G}_I$ and $\mathcal{G}_R$.

$$\mathcal{G}_I = \mathrm{KE}_I(I, \mathcal{G}) = \mathcal{F}(\mathrm{Attention}_I(I', V')); \quad \mathcal{G}_R = \mathrm{KE}_R(R, \mathcal{G}) = \mathcal{F}(\mathrm{Attention}_R(R', V'))$$

$$\mathrm{Attention}(x, y) = \mathrm{softmax}\left(\frac{xW_{\mathrm{q}}(yW_{\mathrm{k}})^{\top}}{\sqrt{d}}\right) yW_{\mathrm{v}}.$$

As a result, our encoder can extract the image and report knowledge representations $\mathcal{G}_I$ and $\mathcal{G}_R$, i.e., the knowledge related to the image and report, they (image, report knowledge) share the common latent space, which allows our model to **bridge the gap between vision and language domains without the training on the pairs of image and report**.

Knowledge-driven Decoder

The knowledge-driven decoder adopts the Transformer [2] to exploit $\mathcal{G}_I$ and $\mathcal{G}_R$ to generate report.

Unsupervised Training Details In the training stage, we estimate the parameters of the decoder by reconstructing the input report R based on $\mathcal{G}_R$, i.e., $R \rightarrow \mathcal{G}_R \rightarrow R$ auto-encoding pipeline; In the prediction stage, we directly input $\mathcal{G}_I$ into the trained decoder to generate the report, i.e., $I \rightarrow \mathcal{G}_I \rightarrow R$. In this way, our approach can produce desirable reports without any labeled image-report pairs.

Semi-Supervised and Supervised Training Details We fine-tune the unsupervised KGAE using partial and full image-report pairs to acquire the KGAE-Semi(-Supervised) and KGAE-Supervised, respectively. In the (semi-)supervised setting, given the image-report pairs, i.e., I-R, we train our approach by generating the ground truth report in the $I \rightarrow \mathcal{G}_I \rightarrow R$ pipeline.

Experiments

- We evaluate our approach under three settings on two public datasets MIMIC-CXR [1] and IU X-ray [3].

Methods	Year	Ratio of Pairs	IU X-ray [9] B-1	B-2	B-3	B-4	M	R-L	MIMIC-CXR [17] B-1	B-2	B-3	B-4	M	R-L
NIC [39]	2015	100%	0.216	0.124	0.087	0.066	-	0.306	0.299	0.184	0.121	0.084	0.124	0.263
AdaAtt [31]	2017	100%	0.220	0.127	0.089	0.068	-	0.308	0.299	0.185	0.124	0.088	0.118	0.266
Att2in [35]	2017	100%	0.224	0.129	0.089	0.068	-	0.308	0.325	0.203	0.136	0.096	0.134	0.276
Transformer [6]	2020	100%	0.396	0.254	0.179	0.135	0.164	0.342	0.314	0.192	0.127	0.090	0.125	0.265
$\mathcal{M}^2$ Trans. [7]	2020	100%	0.437	0.290	0.205	0.152	0.176	0.353	0.238	0.151	0.102	0.067	0.110	0.249
R2Gen [6]	2020	100%	0.470	0.304	0.219	0.165	0.187	0.371	0.353	0.218	0.145	0.103	0.142	0.277
KGAE		**0%**	0.417	0.263	0.181	0.126	0.149	0.318	0.221	0.144	0.096	0.062	0.097	0.208
KGAE-Semi	Ours	60%	0.497	0.320	0.232	0.171	0.189	0.379	0.352	0.219	0.149	0.108	0.147	0.290
KGAE-Supervised		100%	**0.512**	**0.327**	**0.240**	**0.179**	**0.195**	**0.383**	**0.369**	**0.231**	**0.156**	**0.118**	**0.153**	**0.295**

Table 1. Performance in terms of natural language generation metrics. B-n, M and R-L are short for BLEU-n, METEOR and ROUGE-L.

Methods	Year	Ratio of Pairs	MIMIC-CXR [17] Precision	Recall	F1
NIC [39]	2015	100%	0.249	0.203	0.204
AdaAtt [31]	2017	100%	0.268	0.186	0.181
Att2in [35]	2017	100%	0.322	0.239	0.249
Up-Down [1]	2018	100%	0.320	0.231	0.238
$\mathcal{M}^2$ Trans. [7]	2020	100%	0.197	0.145	0.133
Transformer [6]	2020	100%	0.331	0.224	0.228
R2Gen [6]	2020	100%	0.333	0.273	0.276
KGAE		**0%**	0.214	0.158	0.156
KGAE-Semi	Ours	60%	0.360	0.302	0.307
KGAE-Supervised		100%	**0.389**	**0.362**	**0.355**

Table 2. Results in terms of clinical efficacy metrics, which measure the accuracy of descriptions for clinical abnormalities.

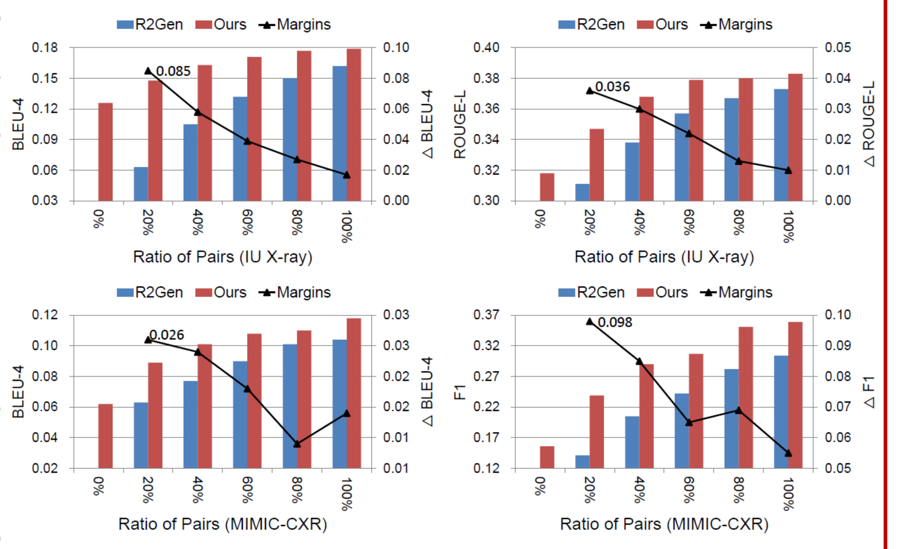

Table 3. Results of R2Gen [4] and ours with respect to various amount of I-R pairs for training. The margins in different ratios are shown with polyline and right y-axis. The fewer the pairs, the larger the margins.

- The unsupervised KGAE can even outperform several supervised models. By using only 60% of paired dataset, KGAE is able to achieve competitive results with current state-of-art models; By training on fully paired datasets as in existing works, KGAE can set new state-of-the-arts.

References

[1] MIMIC-CXR: A large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042, 2019.*

[2] Attention is all you need. *In NIPS, 2017.*

[3] Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Medical Informatics Assoc., 23(2):304–310, 2016.*

[4] Generating radiology reports via memory-driven transformer. *In EMNLP, 2020.*