

# KGVL-BART: Knowledge Graph Augmented Visual Language BART for Radiology Report Generation

Kaveri Kale, Pushpak Bhattacharyya

IIT Bombay

{kaverikale,pb}@cse.iitb.ac.in

**Milind Gune**

Augnito India Pvt Ltd  
dgune@rediffmail.com

**Aditya Shetty**

Breach Candy Hospital, India  
adityashetty01@gmail.com

**Rustom Lawyer**

Augnito India Pvt Ltd  
rustom@augnito.ai

## Abstract

Timely generation of radiology reports and diagnoses is a challenge worldwide due to the enormous number of cases and shortage of radiology specialists. In this paper, we propose a **Knowledge Graph Augmented Vision Language BART (KGVL-BART)** model that takes as input two chest X-ray images- one frontal and the other lateral- along with *tags* which are diagnostic keywords, and outputs a report with the patient-specific findings. Our system development effort is divided into 3 stages: i) construction of the Chest X-ray KG (referred to as chestX-KG), ii) image feature extraction, and iii) training a KGVL-BART model using the visual, text, and KG data. The dataset we use is the well-known Indiana University Chest X-ray reports with the train, validation, and test split of 3025 instances, 300 instances, and 500 instances respectively. We construct a Chest X-Ray knowledge graph from these reports by extracting entity1-relation-entity2 triples; the triples get extracted by a rule-based tool of our own. Constructed KG is verified by two experienced radiologists (with experience of 30 years and 8 years, respectively). We demonstrate that our model- KGVL-BART- outperforms State-of-the-Art transformer-based models on standard NLG scoring metrics. We also include a qualitative evaluation of our system by experienced radiologist (with experience of 30 years) on the test data, which showed that 73% of the reports generated were fully correct, only 5.5% are completely wrong and 21.5% have important missing details though overall correct. To the best of our knowledge, ours is the first system to make use of multi-modality and domain knowledge to generate X-ray reports automatically.

## 1 Introduction

Medical imaging techniques are widely used in hospitals across the world. The detailed informa-

tion extracted from medical images is crucial for proper diagnosis and treatment. An experienced and skilled radiologist is required to prepare an accurate full text diagnostic report. However, due to a lack of experts, many reports contain indecisive findings, forcing patients to undergo further tests involving pathology or other advanced imaging methods. In addition, the time consuming process of full text radiology report generation is one of the biggest challenges. All over the world, the ratio of radiologists to patients is very low. The ratios in the US, China, and India are 1:10,000, 1:14,772, and 1:100,000 respectively (Arora, 2014). Given the huge number of cases and the shortage of radiology experts, timely report generation and diagnosis is a huge challenge worldwide.

In the case of a chest ailment, X-rays produce images of chest organs like, lungs, spinal bones, heart, airways, and blood vessels. These images help doctors ascertain an exact findings such as pneumonia, collapsed lung, emphysema, cancer, broken ribs, *etc.* Manual examination of X-ray images for a large number of patients can be time consuming, leading to delays. Human errors may further add to the challenges. This prompted the researchers to use deep learning models capable of automated report generation to address the above mentioned challenges. With deep learning based automatic report generation, the reports can be generated with minimal delay and free from any human errors. Large-scale pre-trained language models have recently expanded into multimodality learning, improving representations by combining visual and semantic features (Cho et al., 2021; Sollami and Jain, 2021; Mustafa et al., 2022). However, progress in adapting language models toward conditional Natural Language Generation (NLG) is limited to image and text modalities. The Natural Language Processing (NLP) community is moving towards transformer-based models. The major-

ity of NLP research today produces better results by tweaking an already-trained transformer model across a large corpus. This prompts us to research pre-trained transformers like BART that have generative capabilities while conditioning them on visual, textual, and KG data. This study introduces the KGVL-BART conditioned transformer-based model, which, produces a comprehensive report given an X-ray image, tags associated with X-ray image.), and chest X-ray KG. We train our model using the publicly accessible Indiana University chest X-ray dataset (referred to as IU-XRay) (Fischer et al., 2022). The encoder accepts input from tags, images, and KG in three different modalities. We create embeddings for each of these modalities before sending input to the multimodal encoder. We compare our quantitative findings to earlier transformer-based models. Furthermore, we provide qualitative analysis on test set by a radiologist.

**Problem Statement:** Design a system that generates a structured patient-specific report from radiology images, image tags and domain knowledge. The input to the system is

- two chest X-ray images- one frontal and the other lateral
- tags

The output of the system is

- radiology report with patient-specific findings

Domain knowledge comes from the Knowledge Graph.

Our **contributions** are:

1. A knowledge-enhanced BART-based Vision-Language Model which we call KGVL-BART and which generates chest X-ray reports with accuracy better than SoTA.
2. Chest X-ray knowledge graph created from IU Chest X-ray reports.
3. Demonstration of the fact that multi-modality helps in radiology report generation; we use both the image and its tags plus triples from the knowledge graph.

## 2 Fundamental Definitions

**Vision-Language Model:** Vision-Language models are the deep learning models that learn both vision and language modalities together.

**Tags:** Tags are the diagnostic keywords (e.g., left

lung, pulmonary atelectasis, hernia, pneumonia, etc.) associated with X-ray image.

**Findings:** The findings section in radiology reports is the clinical description of abnormalities and normalities observed on radiology image.

**Impression:** The impression section in radiology reports is the summary of findings section.

**Knowledge Graph:** Paulheim (2017) defines Knowledge Graph (KG) as "A knowledge graph (i) mainly describes real-world entities and their interrelations, organized in a graph, (ii) defines possible classes and relations of entities in a schema, (iii) allows for potentially interrelating arbitrary entities with each other and (iv) covers various topical domains."

**Knowledge Graph Embeddings (KGEs):** Knowledge graph embeddings represent the entities and relations in lower-dimensional vectors.

**KG-Grounding:** The KG grounding is the process of extracting the subgraph (referred to as grounded KG) from domain-specific KG (in our case chestX-KG). A grounded KG is a subgraph from the chestX-KG whose nodes represents tags present in the input tag set plus additional relevant nodes.

## 3 Related Work

Researchers (Jing et al., 2017; Zhang et al., 2017; Yuan et al., 2019) studied the issue of automatic report generation. Their research looked into the visual attention given to recurrent decoders and convolution-recurrent architectures (CNN-RNN) that were first introduced by Vinyals et al. (2015) on image captioning. Transformers, attention-only based models that have replaced recurrent models in the NLP community (Vaswani et al., 2017; Devlin et al., 2018). Several attempts have been made in the medical field to create medical reports from the corresponding images. Most authors use multilabel image captioning to produce X-ray reports, and they subsequently use those captions as textual features. The IU-Xray dataset's chest X-ray images were used to generate the first structured report using tags predicted by a CNN-RNN model (Shin et al., 2016). In (Wang et al., 2017b), a system for generating natural reports for the Chest-Xray14 dataset, employing private reports, was presented. This framework used a non-hierarchical CNN-LSTM architecture and focused on semantic and visual aspects. The IU-Xray dataset was created by Jing et al. (2017) to generate radiology

reports automatically.

There is a lot of research performed in NLG on multimodal constraints. NLG models based on transformers (Sollami and Jain, 2021) propose a model called MAnTiS, Multimodal Adaptation for Text Synthesis, as a general method for multimodal conditionality. In this method, separator tokens are used to separate each modality type, and modality-specific encoders are used to encode each modality type. Liu et al. (2021) contend that pre-trained language models and textual concepts by themselves are insufficient to give enough data for generative commonsense reasoning. They supply the ConceptNet KG as input to the transformer model for generative commonsense reasoning in addition to text input. Liu et al. (2020) proposes a Knowledge-enabled Bidirectional Encoder Representation from Transformers (K-BERT). Since the parameters of all pre-trained BERT models are the same, K-BERT can load any of them. Additionally, K-BERT can easily integrate domain knowledge into the models by giving them access to a KG without prior training. Xing et al. (2021) propose KM-BART to conduct the task of Visual Commonsense Generation (VCG) by integrating visual features in pretrained BART model.

## 4 Methodology

As shown in the figure 1, the model architecture consists of six major components, namely, KG grounding, KG embedding, image embedding, text embedding, encoder, and decoder. The KG grounding module extracts the grounded KG from the chestX-KG. Grounded KG includes all of the nodes in the input tag set and their significant neighbors. The KG embedding module converts the grounded KG into vector form using the KGE technique. The image feature extractor module generates the feature vectors for input chest X-ray images. To compute the text features, we use the BART text encoder method. The encoder and decoder are multilayer transformers. This section explains all components in detail.

### 4.1 Knowledge Graph Grounding

The KG grounding module extracts the small sub-graph from chestX-KG for each report in the dataset, given a tag set as input. First, it adds all entities from chestX-KG that are present in the input tag set, and then it adds their significant neighbors.

The algorithm focuses on first finding the most

appropriate path in chestX-KG from tag entities to the root of chestX-KG and then adding neighbor nodes that are connected with *DefaultPropertyOf* (default property of entities) relation. The following are the steps to extract grounded KG:

- Find all possible candidate paths from a matched entity to the root node.
- Find the most appropriate top five paths by ranking based on the precision and recall of entities in the input tag set and entities in all possible candidate paths.
- We consider all paths, including those with matched entities that are absent from the selected top-ranking path.
- Instead of adding all neighbors of input tag entities, we add only significant entities that are default properties or default descriptors of input entities.

Our proposed method reduces noise by adding only significant nodes to grounded KG. We propose a context-aware KG grounding algorithm to select M triples from the chestX-KG for an entity candidate. The pseudocode of this algorithm is shown as follows.

---

#### Algorithm 1: KG Grounding

---

```

Input : K: Tags
          G(V, E) : chestX-KG
Output : Grounded KG
1 Find all candidate paths in G(V, E) that includes the
  node with input concept
2 path-dict -> initialize
3 for each path in possible candidate-paths do
4    $Precision = \frac{K \cap AllEntitiesinpath}{No.ofconceptsinK}$ 
5    $Recall = \frac{K \cap AllEntitiesinpath}{No.ofnodesinpath}$ 
6    $F-score = \frac{2 * Precision * Recall}{Precision + Recall}$ 
7   add path-dict -> (path:F-score)
8 end
9 Sort path-dict in descending order of F-score
10 Get top 5 paths
11 for each path in top-5-paths do
12   if  $len(set(K) - set(path)) > 0$  then
13     Add all triplets from that path in grounded
      KG triple set
14     for each node in path do
15       Find all neighbors of node with
        default-property relation
16       Add all triples of form (neighbor,
        DefaultPropertyOf, node) in
        grounded KG triple set
17     end
18   end
19 end
20 Return grounded KG triplets set.

```

---

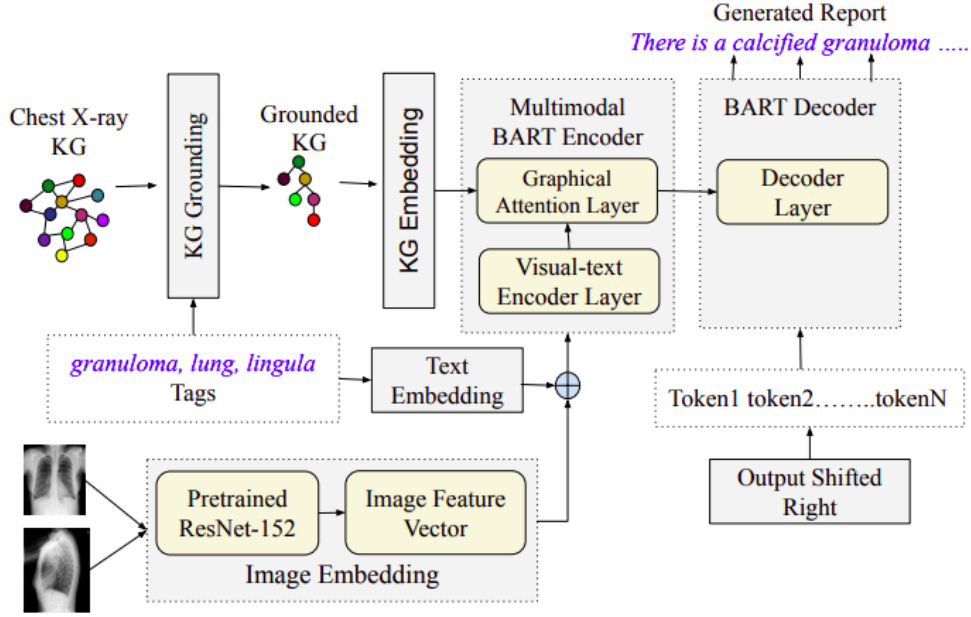


Figure 1: The architecture of our proposed KGVL-BART model. KGVL-BART has six important components: KG grounding, KG Embedding, image embedding, text embeddings, encoder, and decoder.

## 4.2 Knowledge Graph Embeddings

Knowledge Graph Embedding methods embed the components of a KG, including entities and relations, into continuous vector spaces. The chestX-ray KG is represented in low dimensional vector space using KGE. There are different techniques used for KG embeddings like TransR, TransH, TransE, TransD, *etc.* (Wang et al., 2017a). For simplicity and concreteness, in this work, we primarily consider the TransE (Bordes et al., 2013) model due to its state-of-the-art performance.

## 4.3 Text Embedding

The input embeddings in KGVL-BART are made of two separate embeddings, token embeddings and position embeddings. To get the final text embedding, we add the vectors of token embeddings and position embeddings.

### 4.3.1 Token Embeddings

Tokens are nothing but a word or part of a word. The textual encoder uses the vocabulary offered by large-cnn BART, and the token embedding is consistent with BART. Using a trainable lookup table, we transform each token in the input tag set into an embedding vector of dimension  $d$ .

In order to create these token embeddings, a method called BART tokenizer is used to tokenize the text. Input tag set  $T$  is tokenized as  $\{t_1, t_2, \dots, t_{|t|}\}$  and encoded as learned embedding

$e_t = \{e_{t_1}, e_{t_2}, \dots, e_{t_{|t|}}\}$ . For a token  $t_n$ , its embedding is  $e_{t_n} \in R^d$ , where  $d$  is the dimension of the token embeddings. The encoder, decoder, and language modeling head (Press and Wolf, 2016) all share the embedding parameters. Due to the permutation-invariance of the attention layers, BART learns positional embeddings for absolute token positions and adds them to the token embeddings (Vaswani et al., 2017; Devlin et al., 2018).

### 4.3.2 Positional Embeddings

Position embeddings represent the position of the word within that sentence that is encoded into a vector. We must introduce some information about the relative or absolute location of the tokens in the sequence because our model lacks recurrence and convolution and hence cannot use the sequence’s order. To do this, we augment the token embeddings at the base of the encoder and decoder stacks with positional embeddings. The positional encodings and token embeddings have the same dimension  $d$  allowing the token and positional embeddings to be added together. The text embeddings are the sum of the token embeddings and the positional embeddings, i.e.,  $e_{tp} = e_t + e_p$ , where  $e_p$  is the positional embeddings.

## 4.4 Image Embedding

For image feature extraction, we use three different methods: i) Pretrained CheXNet (Rajpurkar

et al., 2017) model (referred to as CheXNet), ii) Pre-trained ResNet-152 model (referred to as ResNet-152), and iii) Fine-tuned ResNet-50 for multilabel image classification on NIH chest-xray dataset (referred to as NIH-ResNet). We extract the image embeddings from each of these models, and we train KGVL-BART separately for these methods. In this section, we give details about the pretrained ResNet-152 feature extractor.

We extract the embedding form of the final fully connected layer of the pre-trained ResNet-152 model (He et al., 2016). We transform images using the same parameters as during pretraining, which include resizing, center cropping, and normalizing. We project the image feature vector through a linear layer with a learnable weight matrix  $W \in R^{N*d}$  onto the language model embedding space  $d$ .

To get the final input embeddings, we sum up the text and image embeddings. In addition to the original vocabulary of BART, for images, we use  $\langle img \rangle$  and  $\langle /img \rangle$  to indicate the start and the end of visual embeddings, respectively. Multimodal Feature Augmentation is done by adding image feature vector with the text feature vector to generate a single feature vector, i.e.,  $e_{tv} = e_{tp} + e_v$ .

#### 4.5 Encoder

The encoder uses two modalities—image and text, and text generation is conditioned on grounded KG. According to the figure 1, the KG enhanced encoder layer sits above the visual-textual encoder layer and is intended to enhance the visual-text representation  $\{e_{tv_1}, e_{tv_2}, \dots, e_{tv_n}\}$  by taking the KG structure into account. We use a graph attention layer to incorporate graph representations into the input encoding process. It uses explicit relations to help the model learn intra-concept relations more effectively. Formally, the grounded KG embedding, as well as the output visual-textual embeddings from the visual-textual encoder, are combined by the KG-augmented encoder to update the visual-textual token representation. Our self-attention layer and fully-connected layer with residuals make up the stack of  $m$  transformer blocks that make up our bidirectional multimodal encoder.

#### 4.6 Decoder

The decoder uses the text embedding module at the bottom layer to encode the text. The decoder in our model is also a multi-layer transformer. Our decoder is auto-regressive and unidirectional. We skip over a detailed explanation of these modules

because our textual transformers are the same as those used in BART (Lewis et al., 2019; Vaswani et al., 2017).

## 5 Experiments

The datasets, evaluation metrics, and baselines used for the training and evaluation of the KGVL-BART model are covered in detail in this section.

### 5.1 Dataset

We use the IU Chest X-ray dataset to train our model (Demner-Fushman et al., 2016). There are 3825 patient reports in this dataset. 7430 chest X-ray images from the front and sides contribute to this dataset. Each patient report contains two types of tags: MTI tags and manual tags from MESH<sup>1</sup> and RadLex<sup>2</sup>. Each report has three parts: an impression, which is a title or summary of the report; findings, which contain the report in detail; and manual tags. We use MESH tags (text) as one of the input to train our model. We concatenate impressions and findings and use it as target to train our model. IU Chest X-ray dataset includes normal and abnormal study reports. There are total 3825 reports out of which 1379 are normal and 1646 are abnormal. The dataset is balanced with respect to normal and abnormal reports. Additionally, we chose 500 samples at random to serve as the test set. Our split is 3025 for training, 300 for validation, and 500 for testing. Table 1 shows the samples from the IU-Chest X-ray dataset.

Table 1 shows an example from the dataset that we are using to train KGVL-BART model.

#### 5.1.1 Chest X-ray Knowledge Graph (chestX-KG)

Nodes in our knowledge graph represent all necessary information, like findings, observations, anatomy, properties, and modifiers related to the organ. We define eight logical relations to construct chestX-KG. i) *PartOf*: It represents the relation between anatomy and sub-anatomy, ii) *TypeOf*: It represents the relation between similar type of entities, iii) *ModifierOf*: It denotes the descriptors of findings, anatomical locations, properties, etc., iv) *ObservationOf*: It denotes the clinical observations observed for a particular finding, v) *DefaultObservationOf*: It denotes the observation that is associated by default with a particular anatomical

<sup>1</sup>[https://www.nlm.nih.gov/mesh/qualifiers\\_scopenotes.html](https://www.nlm.nih.gov/mesh/qualifiers_scopenotes.html)

<sup>2</sup><https://radlex.org/>


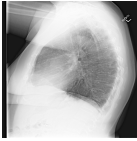



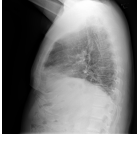


Frontal Image	Lateral Image	Tags	Target
		Osteophyte, thoracic vertebrae, multiple, small, Thickening, pleura, apex, bilateral, Lung, hyperdistention, mild	<b>Impression:</b> No acute cardiopulmonary abnormality. <b>Findings:</b> The cardiomeastinal silhouette and pulmonary vasculature are within normal limits. There is no pneumothorax or pleural effusion. There are no focal areas of consolidation. Cholecystectomy clips are present. Small T-spine osteophytes. There is biapical pleural thickening, unchanged from prior. Mildly hyperexpanded lungs.
		normal	<b>Impression:</b> No acute cardiopulmonary findings. <b>Findings:</b> Heart size and mediastinal contour are within normal limits. There is no focal airspace consolidation or suspicious pulmonary opacity. No pneumothorax or large pleural effusion. Mild degenerative change of the thoracic spine.
		Pulmonary Atelectasis, base, Spondylosis, thoracic vertebrae, Arthritis, cervical vertebrae	<b>Impression:</b> Basilar atelectasis. No confluent lobar consolidation or pleural effusion. <b>Findings:</b> The cardiac contours are normal. XXXX basilar atelectasis. The lungs are clear. Thoracic spondylosis. Lower cervical XXXX arthritis.
		Calcified Granuloma, lung, upper lobe, right	<b>Impression:</b> No acute cardiopulmonary process. <b>Findings:</b> The cardiomeastinal silhouette is within normal limits for size and contour. The lungs are normally inflated without evidence of focal airspace disease, pleural effusion, or pneumothorax. Stable calcified granuloma within the right upper lung. No acute bone abnormality.

Table 1: Samples from the IU Chest X-ray dataset. We use frontal images, lateral images, and tags as input to our model and target text as a concatenation of impression and findings columns from the IU Chest X-ray dataset. In this table, we show only the columns we use to train the KGVL-BART model.

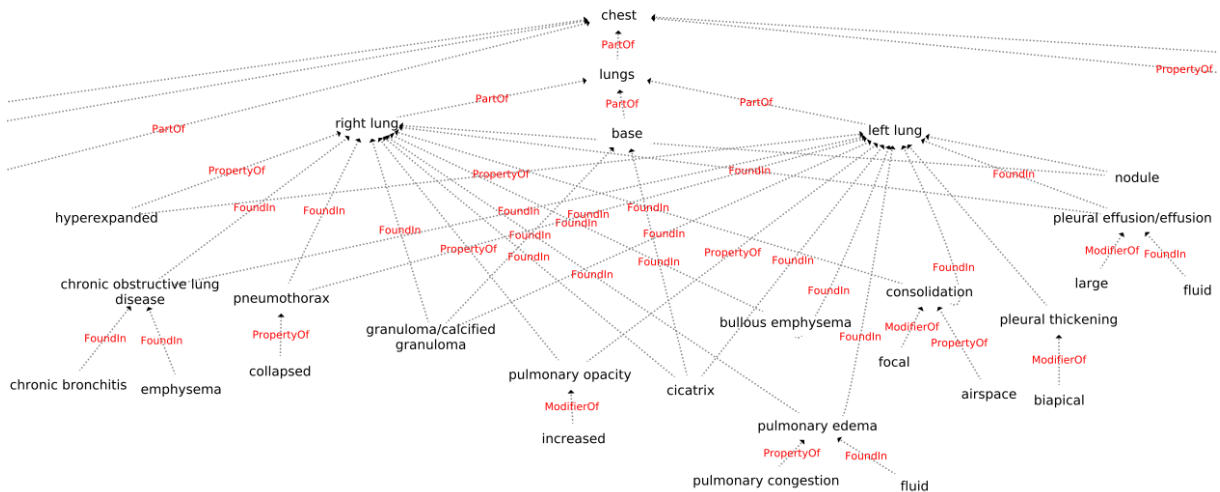


Figure 2: Chest X-ray KG constructed by extracting triplets from the free-text chest X-ray reports. Constructed KG is verified by radiologists. Only a portion of the entire KG is displayed due to space limitations.

location or particular finding. vi) *PropertyOf*: It denotes the relation between entities (anatomical entities, finding entities, observation entities, etc.) and their properties., vii) *DefaultPropertyOf*: It denotes the property that exist by default with particular anatomical location or particular finding., and viii) *FoundIn*: It denotes the relation between findings and corresponding anatomical location.

We use a rule-based and pattern-based approach

to extract the triples from the IU X-ray text report corpus. After extracting triples from text-report corpus we construct hierarchical KG with chest as root node and children represents its parts or associated findings. More details about radiology KG construction is given in paper (Kale et al., 2022). The constructed KG is verified by two radiologists who are involved in this research work. Figure 2 shows the part of chestX-KG that we have constructed.

## 5.2 Training

Three different techniques are used to extract the features from the frontal and lateral X-ray images: i) Pretrained CheXNet model, ii) Pretrained Resnet-152 model, and iii) Fine tuned Resnet-50 for multi-label image classification on the NIH chest X-ray dataset<sup>3</sup> (Wang et al., 2017b).

We have trained the KGVL-BART model for each of these image embeddings separately and also provided quantitative results. To implement the TransE model for KG embeddings, we use the open-source OpenKE<sup>4</sup> tool. The chestX-KG contains 106 nodes and 126 triples.

## 5.3 Pretraining Setup

For model pretraining we use ConceptNet KG (Speer et al., 2017) and common sense generation dataset (Lin et al., 2020). For pretraining, we do not use images. At the time of pretraining, we are not adding image features to text features. We use byte-pair encoding for tokenization, with a maximum length of 32 for the encoder and 64 for the decoder. We set the learning rate to 0.00001 and used AdamW with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.98$  for optimization. We set the batch size to 60. We trained the KGVL-BART for 10 epochs, and the gradients are accumulated every 6 steps. We apply dropout with a probability of 0.1 to avoid over-fitting. While inferencing, we use beam search with beam size 5 and a length penalty of factor 0.6.

## 5.4 Training Setup

We propose our own algorithm for KG-grounding tasks. We construct grounded KGs for each report tag set. We use pretrained weights to initialize the KGVL-BART model. We train our model on the IU Chest X-ray dataset.

We use byte-pair encoding for tokenization with a maximum length of 300 for the encoder and 500 for the decoder. We set the learning rate to 0.00001 and used AdamW with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.98$  for optimization. We set the batch size to 18. We trained the KGVL-BART for 15 epochs, and the gradients are accumulated every 6 steps. We apply dropout with a probability 0.1 to avoid over-fitting. We use beam search with beam size 5 and length penalty with factor 0.6 while inferencing. DGX A100-SXM-80GB GPU server takes approximately

15 minutes for a single epoch.

## 6 Evaluation

We evaluate our model by automatic metrics and human evaluation as well. For automatic evaluation we use word-overlap metrics. Manual evaluation is performed by two radiologists.

### 6.1 Quantitative Evaluation

We compare the performance of KGVL-BART model with previous state-of-the-art conditional transformer text generation models, i) the CNN-RNN model (Vinyals et al., 2015), ii) CDGPT2 for visual input only, text input only, and for both visual and text inputs (Alfarghaly et al., 2021). Following other conventional generation tasks, we use several widely-used automatic metrics to automatically assess the performance, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005), which mainly focus on measuring n-gram similarities. We have evaluated other NLG metrics like BLEURT (Selam et al., 2020) and BERTScore (Zhang et al., 2019). Table 2 shows the NLG metrics score of generated X-ray reports by KGVL-BART and baseline models vs. gold standard X-ray reports. Evaluation is done on a test dataset. We have added ablation study, which shows that knowledge infusion improves the report generation. Last row in table 2 shows the results of our model by removing KG augmented layer from it.

Even though the dataset is balanced with respect to normal and abnormal reports, However, multiple organ findings are included in a single report; if at least one organ is found to be abnormal, the report is classified as abnormal. Findings for the other organs are normal, despite the fact that this is classified as an abnormal report. As a result, our model may be overly focused on widely reported normal findings. Hence, we have added evaluation metrics only to abnormal studies as well. Table 3 shows the NLG metrics for abnormal studies. For evaluation, we consider abnormal studies from the test set. Our results show that our model is capable of generating better reports than SoTA models for abnormal findings as well.

### 6.2 Qualitative Evaluation

This section provides qualitative analysis by a radiologist having experience of thirty years. Access to images and the ground-truth reports was given

<sup>3</sup><https://www.kaggle.com/datasets/nih-chest-xrays/data>

<sup>4</sup><https://github.com/thunlp/OpenKE>

Method	Automatic Evaluation Metrics								
	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Rouge-L	Meteor	chrF++	BLUERT	BERTScore
CNN-RNN (Vinyals et al., 2015)	0.316	0.211	0.140	0.095	0.267	0.159	-	-	-
CDGPT2-vis-only (Alfarghaly et al., 2021)	0.340	0.209	0.138	0.091	0.281	0.153	-	-	-
CDGPT2-sem-only (Alfarghaly et al., 2021)	0.357	0.224	0.151	0.103	0.275	0.149	-	-	-
CDGPT2 (Alfarghaly et al., 2021)	<u>0.387</u>	<u>0.245</u>	<u>0.166</u>	<u>0.111</u>	0.289	0.164	0.370	0.460	0.888
CNN-TRG (Pino et al., 2021)	0.273	-	-	-	0.352	-	-	-	-
KGVL-BART (CheXNet)	0.326	0.139	0.080	0.050	0.340	0.387	0.453	0.473	<u>0.892</u>
KGVL-BART (NIH-ResNet)	0.324	0.144	0.090	0.060	<u>0.355</u>	<u>0.390</u>	<u>0.467</u>	0.468	0.889
KGVL-BART (ResNet-152)	<b>0.423</b>	<b>0.256</b>	<b>0.194</b>	<b>0.165</b>	<b>0.444</b>	<b>0.500</b>	<b>0.543</b>	<b>0.526</b>	<b>0.909</b>
Our model (without KG layer/ResNet-152)	0.341	0.188	0.142	0.119	0.351	0.376	0.424	<u>0.478</u>	0.892

Table 2: Results on whole test set (abnormal + normal studies): BLEU, ROUGE and METEOR score of generated X-ray reports by previous transformer-based models and our KGVL-BART models vs. gold standard X-ray reports. The best results are in bold font, and the second best is underlined.

Method	Automatic Evaluation Metrics					
	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Rouge-L	Meteor
CDGPT2 (Alfarghaly et al., 2021)	0.273	0.090	0.034	0.013	0.267	0.316
KGVL-BART (CheXNet)	<u>0.308</u>	<u>0.123</u>	<u>0.073</u>	<u>0.049</u>	<u>0.314</u>	<u>0.359</u>
KGVL-BART (NIH-ResNet)	0.299	0.119	0.067	0.046	0.309	0.352
KGVL-BART (ResNet-152)	<b>0.388</b>	<b>0.207</b>	<b>0.139</b>	<b>0.108</b>	<b>0.397</b>	<b>0.458</b>

Table 3: Results on abnormal studies from test set: BLEU, ROUGE and METEOR score of generated X-ray reports by previous transformer-based models and our KGVL-BART models vs. gold standard X-ray reports. The best results are in bold font, and the second best is underlined.

to the radiologist to evaluate the automatically-generated reports. The radiologist classified the generated reports into **accurate** (report with most of the vital information), **missing details** (reports with no false information but missing some vital details), and **false** predictions (report with false information and overall incorrect diagnosis). We have provided 200 random samples from the test dataset and their corresponding generated reports for qualitative analysis. Qualitative evaluation by the radiologist shows that 73% of the reports generated were fully correct, only 5.5% are completely wrong and 21.5% have important missing details though overall correct. Further, these random samples were classified into normal and abnormal reports. For normal studies 95.6% of the reports generated were fully correct, 4.40% have important missing details though overall correct and 0% false reports. For the abnormal cases, the model could generate 54.13% of the reports correctly, 35.78% missing details, and 10% had false reports. Table 4 contains the results of the qualitative analysis. The reports classified as missing details by the radiologist also contain useful information that can be used to speed up the report writing process. Overall results are promising however, there is still a scope for improvement. The results show that the model is a bit biased towards normal cases, which is often

the case with medical datasets as it lacks necessary details to describe the different irregularities present in the image.

Method	Samples	Accurate	Missing Details	False
Ours	All(500)	<b>73.00%</b>	<b>21.50%</b>	<b>05.50%</b>
	Normal(183)	95.60%	04.40%	<b>00.00%</b>
	Abnormal(317)	<b>54.13%</b>	<b>35.78%</b>	<b>10.00%</b>
CDGPT2	All(500)	61.60%	28.20%	10.20%
	Normal(201)	<b>99.00%</b>	<b>00.00%</b>	01.00%
	Abnormal(299)	36.50%	47.10%	16.40%

Table 4: Results of generated reports, manually evaluated by radiologist. Manual evaluation is done on the IU X-RAY dataset. Best results are shown in a bold face.

## 7 Summary, Conclusion and Future work

We introduce KGVL-BART a knowledge-enhanced Vision Language model to generate X-ray reports from chest X-ray images and tags. We develop a knowledge graph called chestX-KG, which is verified by two experienced radiologists. Overall the reports generated are accurate. Our approach to constructing grounded KGs is relatively noise-free since it considers only entities in the hierarchical path from tag set entities to the root node of the knowledge graph and only adds neighbors with "default" properties. Experimental results show the efficacy of our method and its superiority over SoTA.








	Normal	Ground Truth	Generated by KGVL-BART Model
<b>Accurate</b>		The lungs are clear. Heart size is normal. No pneumothorax, pleural effusion, or focal airspace disease. Bony structures appear intact.	The cardiomeastinal silhouette is within normal limits. No pneumothorax or pleural effusion. No acute bony abnormalities.
		Heart size is normal. <b>There are densely calcified mediastinal and right hilar lymph xxxx which suggest prior histoplasmosis exposure.</b> No consolidating airspace disease is seen within the lungs. No pleural effusion or pneumothorax. No convincing acute bony findings.	Lungs are clear bilaterally. Specifically, no evidence of focal consolidation, pneumothorax, or pleural effusion.. <b>Calcified right hilar lymph xxxx noted.</b> Cardio mediastinal silhouette is unremarkable. Visualized osseous structures of the thorax are without acute abnormality. <b>Stable enlargement of the cardiac silhouette, consistent with calcification. Aortic calcifications noted.</b> No acute bony abnormality identified.
<b>Missing Details</b>		<b>Lungs are clear bilaterally.</b> There is no focal consolidation, Pleural effusion, or pneumothoraces. Cardiomeastinal silhouette is within normal limits. xxxx are unremarkable.	Normal heart size and mediastinal contours. No focal airspace consolidation. no pneumothorax or pleural effusion. No acute bony abnormalities.
		Heart size is normal. No pneumothorax or pleural effusions. There is an <b>8 mm</b> calcified nodule in the left midlung. There is also a <b>7 mm</b> calcified nodule near the left hilum. <b>Hyperexpanded lungs consistent with chronic obstructive pulmonary disease.</b>	There are low lung volumes. the lungs are otherwise clear. There is a calcified left hilar lymph node and left midlung granuloma. No focal airspace consolidation to suggest pneumonia. No pleural effusion or pneumothorax. Normal heart size. No acute bony abnormality.
<b>False</b>		Heart size is moderately enlarged. The mediastinum are within normal limits. there is no pleural effusion or pneumothorax. There is suspected right lower lobe airspace opacity demonstrated on the lateral study. There is a fracture of superior sternotomy unchanged.	There is stable eventration of the right hemidiaphragm. There is no focal lung consolidation. Heart size is within normal limits. No pneumothorax or pleural effusion. No acute bony abnormalities.

Figure 3: Examples of ground truth and reports generated by KGVL-BART (ResNet-152). The first slot shows the instance of accurate prediction by the KGVL-BART model for normal and abnormal cases. Abnormal findings are highlighted in blue. The second slot shows the example of partially correct predictions by the KGVL-BART model for normal and abnormal cases. Details that are present in ground truth but missing in the predicted report is highlighted in red. The third slot shows the example of false predictions by the KGVL-BART model for abnormal case. For normal case model does not generate false report.

Our future work consists in training our model on the much larger MIMIC-CXR (Johnson et al., 2019) dataset. We would also like to expand the scope of our work to CT, MRI, etc.

### Limitations

The IU Chest X-ray and the MIMIC-CXR datasets are publicly available that links chest X-ray images with text radiology reports. The IU Chest X-ray dataset available for general use and the MIMIC-CXR dataset has restricted access. Annotating medical reports requires domain experts' knowledge, and it is costly. Medical data is likewise subject to strict privacy regulations and is governed, for example, by the Health Insurance Portability and Accountability Act (HIPAA). Therefore, only a small amount of data is accessible to the general (research/corporate/industry) use.

There are many more sentences in this dataset

describing normalities than abnormalities. As a result, most machine learning models are biased to produce normal reports more frequently than abnormal ones. Given the scarcity of examples, abnormalities are more challenging to find.

### Ethics Statement

IU Chest X-ray dataset's authors used appropriate techniques to de-identify the text reports. Data is anonymized; hence our model will not disclose information about the patient's identity.

### References

- Omar Alfarghaly, Rana Khaled, Abeer Elkorany, Maha Helal, and Aly Fahmy. 2021. Automated radiology report generation using conditioned transformers. *In-formatics in Medicine Unlocked*, 24:100557.
- Richa Arora. 2014. The training and practice of radiol-

- ogy in india: current trends. *Quantitative imaging in medicine and surgery*, 4(6):449.
- Satanjeev Banerjee and Alon Lavie. 2005. **METEOR: An automatic metric for MT evaluation with improved correlation with human judgments**. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.
- Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR.
- Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Tim Fischer, Steffen Remus, and Chris Biemann. 2022. **Measuring faithfulness of abstractive summaries**. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 63–73, Potsdam, Germany. KONVENS 2022 Organizers.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Baoyu Jing, Pengtao Xie, and Eric Xing. 2017. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*.
- 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*.
- Kaveri Kale, Pushpak Bhattacharyya, Aditya Shetty, Miling Gune, Kush Shrivastava, Rustom Lawyer, and Spriha Biswas. 2022. Knowledge graph construction and its application in automatic radiology report generation from radiologist’s dictation. *arXiv preprint arXiv:2206.06308*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. **CommonGen: A constrained text generation challenge for generative commonsense reasoning**. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908.
- Ye Liu, Yao Wan, Lifang He, Hao Peng, and S Yu Philip. 2021. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6418–6425.
- Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. 2022. Multimodal contrastive learning with limoe: the language-image mixture of experts. *arXiv preprint arXiv:2206.02770*.
- 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.
- Heiko Paulheim. 2017. **Knowledge graph refinement: A survey of approaches and evaluation methods**. *Semantic web*, 8(3):489–508.
- Pablo Pino, Denis Parra, Cecilia Besa, and Claudio Lagos. 2021. Clinically correct report generation from chest x-rays using templates. In *International Workshop on Machine Learning in Medical Imaging*, pages 654–663. Springer.
- Ofir Press and Lior Wolf. 2016. Using the output embedding to improve language models. *arXiv preprint arXiv:1608.05859*.
- Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.
- Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

- Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M Summers. 2016. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2497–2506.
- Michael Sollami and Aashish Jain. 2021. Multimodal conditionality for natural language generation. *arXiv preprint arXiv:2109.01229*.
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.
- Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017a. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743.
- Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017b. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106.
- Yiran Xing, Zai Shi, Zhao Meng, Gerhard Lakemeyer, Yunpu Ma, and Roger Wattenhofer. 2021. Kmbart: Knowledge enhanced multimodal bart for visual commonsense generation. *arXiv preprint arXiv:2101.00419*.
- Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. 2019. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 721–729. Springer.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. 2017. Mdnnet: A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6428–6436.