Generating Reports for X-ray images

Ali Derakhshesh, Abolfazl Malekahmadi, MohammadTaha Teimury,
Mehrab Moradzadeh, Alireza Aghaei

CE Department @ Sharif University of Technology

Abstract:

Image captioning is a challenging task that combines computer vision and natural language processing to generate descriptive textual descriptions for visual content. This project presents a novel approach for image captioning specifically tailored for chest X-ray images. As chest X-rays play a crucial role in diagnosing respiratory diseases, the ability to automatically generate accurate and informative captions can aid radiologists and healthcare professionals in interpreting and understanding these images more efficiently. Our proposed method utilizes a fine-tuned transformer-based network for feature extraction and caption generation. Specifically, we employ a Blip Network, a state-of-the-art transformer-based model, to extract high-level visual features from chest X-ray images. The Blip and GIT Networks have been pre-trained on large-scale image datasets and fine-tuned on Indiana Chest-xray dataset to capture relevant visual information effectively. Additionally, we have implemented an LSTM-based model for comparison purposes.

Addressing the unique challenges posed by the X-ray images, such as the presence of anatomical structures and specific medical terminologies, our approach incorporates domain-specific knowledge and medical ontologies into the caption generation process. This ensures the accuracy and relevance of the generated captions, making them more clinically meaningful. We evaluate the performance of our approach on a comprehensive dataset of annotated Indiana Chest-xray dataset. The experimental results demonstrate the effectiveness of our fine-tuned Blip and GIT Networks in generating accurate and clinically relevant captions for chest X-rays. By assisting radiologists in the interpretation and analysis of chest X-ray images, our proposed image captioning system has the potential to improve diagnostic accuracy, workflow efficiency, and patient care.

Keywords: Image captioning, chest X-ray images, deep learning, transformer-based network, Blip and GIT Network, medical imaging, radiology, respiratory diseases.

Introduction:

1. Motivation:

Image captioning is a challenging task that combines the fields of computer vision and natural language processing to generate descriptive textual captions for visual content. The ability to automatically generate accurate and informative captions for images has gained significant attention in recent years, as it has the potential to enhance various applications, including image retrieval, accessibility for visually impaired individuals, and content understanding.

In the medical domain, image captioning plays a crucial role in assisting healthcare professionals, particularly radiologists, in the interpretation and analysis of medical images. Chest X-ray images, in particular, are widely used in diagnosing respiratory diseases and provide valuable insights into the condition of a patient's lungs and thoracic region. However, the interpretation of chest X-rays can be challenging and time-consuming, requiring expertise and domain knowledge.

The development of automated image captioning techniques tailored specifically for chest X-ray images can significantly aid radiologists in their diagnostic workflow. By generating accurate and informative captions, these techniques can assist in identifying abnormalities, highlighting specific anatomical structures, and providing insights into potential respiratory conditions. Ultimately, such automated systems have the potential to improve diagnostic accuracy, streamline workflow efficiency, and enhance patient care in the field of chest X-ray analysis.

The motivation behind this research is to address the existing challenges faced by radiologists and healthcare professionals in the interpretation of chest X-ray images. While previous studies have explored image captioning techniques in general medical imaging or non-medical image domains, the specific requirements and characteristics of chest X-ray images necessitate a tailored approach.

In this project, we propose an approach for image captioning specifically designed for chest X-ray images. Our methodology utilizes fine-tuned transformer-based networks, known as the Blip and GIT Network, for feature extraction and caption generation. By leveraging the power of transformers, which have demonstrated remarkable success in various natural language processing tasks, we aim to capture relevant visual information effectively and generate accurate captions that are clinically meaningful.

2. Related Works:

Several studies have explored image captioning techniques in the broader context of medical imaging and computer vision. However, the specific application of image captioning for chest X-ray images requires a focused investigation due to the unique characteristics and diagnostic importance of these images. In this section, we review the related work in medical image captioning, particularly in the domain of chest X-ray analysis.

One approach to medical image captioning involves the use of deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). For instance, researchers have applied CNNs for feature extraction from chest X-ray images, followed by RNN-based models for caption generation [1]. These studies have achieved promising results in generating captions for medical images, including chest X-rays. Furthermore, the chexNet by Rajpurkar et al. [2], a convolutional neural network trained on a dataset of 100,000 chest X-ray images. While primarily for diagnostic purposes, the underlying neural network structure presents useful insights for image captioning. However, these methods were limited by the training efficiency and expression ability, thus researchers began to explore the CNN-Transformer based models and achieved great success.

Transformer-based models, originally introduced for natural language processing, have demonstrated remarkable performance in various tasks, including image captioning. Recent studies have leveraged pre-trained transformer models, such as the Blip Network, for image captioning in general domains [3]. These models excel at capturing complex visual information and generating coherent and contextually relevant captions. However, their application in the context of chest X-ray image captioning is relatively limited.

While existing literature provides valuable insights into image captioning in medical imaging, including chest X-ray analysis, the specific requirements and complexities of this domain necessitate further research. In this project, we aim to bridge this gap by proposing an approach that leverages fine-tuned transformer-based networks, the Blip and GIT Networks, for accurate caption generation in the context of chest X-ray images. Our approach incorporates domain-specific knowledge and medical ontologies, enhancing the clinical relevance and accuracy of the generated captions.

3. Methodology:

Our proposed methodology for chest X-ray image captioning combines the power of transformer-based models with domain-specific knowledge and medical ontologies, utilizing both the GIT Network and the BLIP Network. The overall pipeline consists of three main steps: Implementing an LSTM base model, fine-tuning the BLIP Network and GIT Network, and caption generation and evaluation with BLEU metric.

In the base model, we have trained an LSTM-CNN model which uses greedy and beam search algorithms to generate captions. It is observed that the beam search algorithm works considerably better than the greedy algorithm. We can see this in the results where greedy search achieves a maximum bleu score of 0.2 and beam search reaches a bleu score of 0.4.

In the second step, we have fine tuned two pre-trained BLIP and GIT networks on the Indiana Chest-xray dataset.

Finally, in the evaluation part, we have implemented the BLEU metric for all of our three models including the base model (LSTM based model), BLIP, and GIT model.

Models:

1. BASE MODEL (LSTM):

Our base model is based on encoder-decoder architecture as shown in Fig 1. In the image encoder part, we will use pre-trained CheXnet[2] which is a DenseNet with 121 layers shown in Fig 2 and has been trained on 112,000 chest X-ray images.

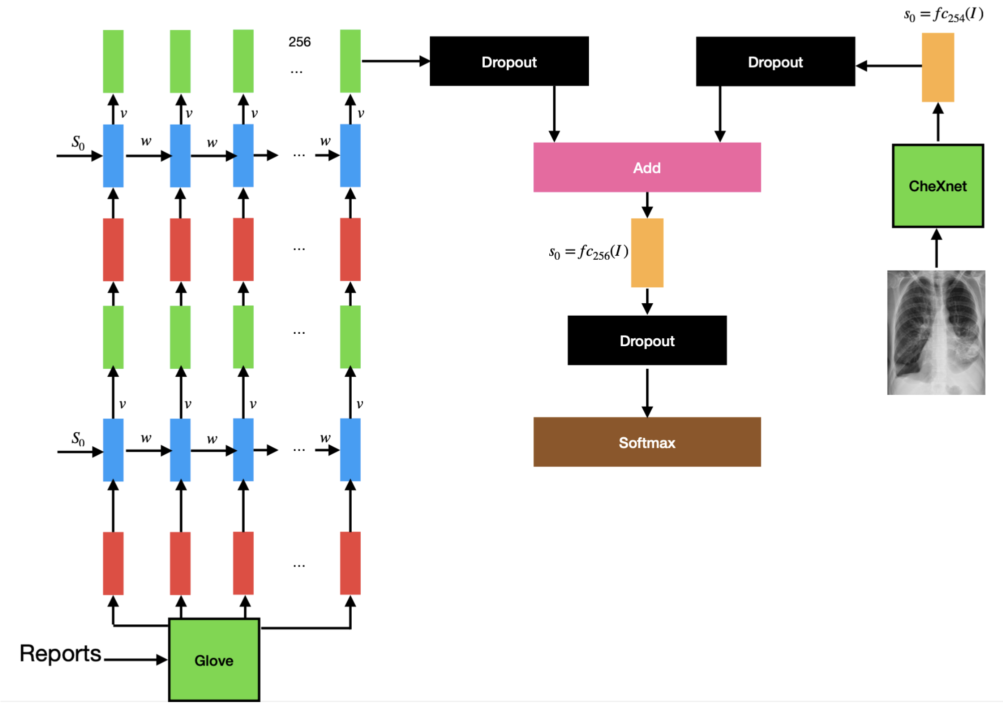


Fig 1: Base Model Architecture

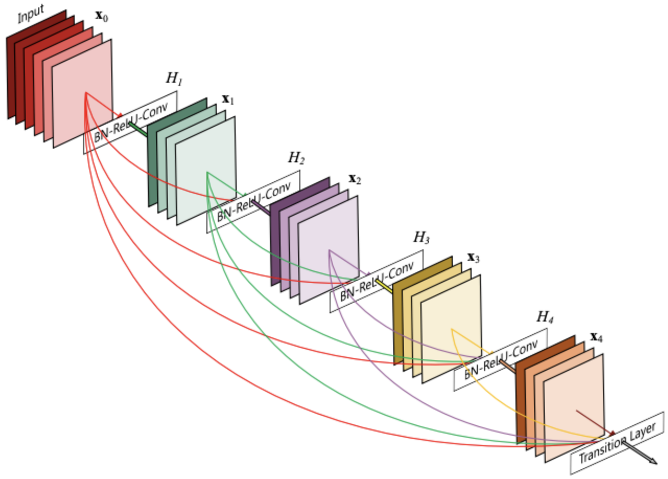


Fig 2: DenseNet Architecture

To obtain text embeddings, we have used the Glove model. The decoder part contains two LSTMs with 256 hidden units each. In the test phase, a caption is decoded and generated for each photo given to the model using beam search and greedy decoding methods.

2. **BLIP**:

The Blip paper accomplishes different tasks by using 3 different loss functions (ITC, ITM, and LM). Using Language Model (LM) loss function, we apply the image captioning task on X-ray images. In the Blip paper, which was state of the art at the time, the output is generated by the architecture shown in Fig 3. This model has used ViT-L/16 for the image encoder part and Bert model for the text encoder part.

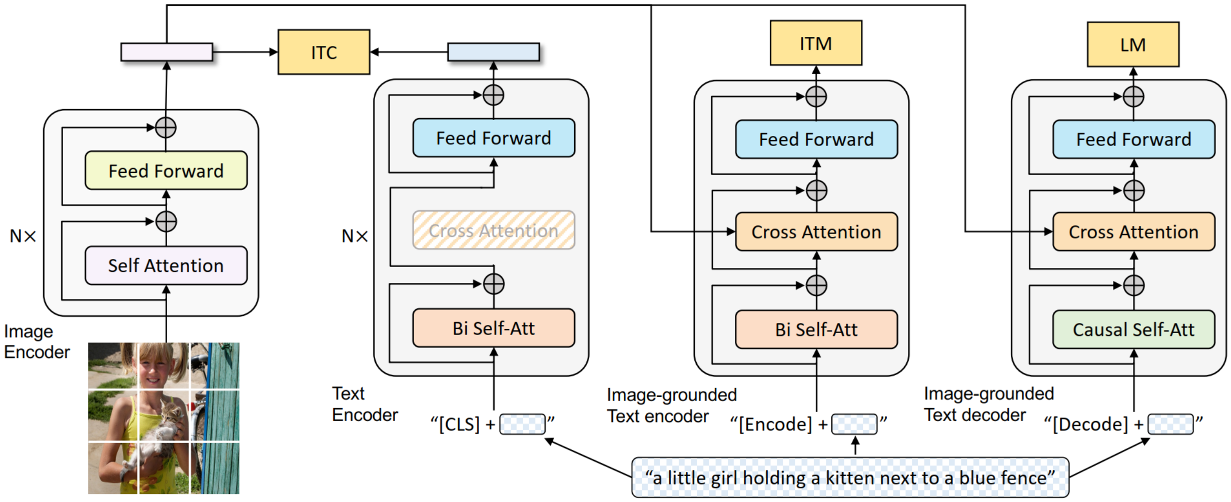
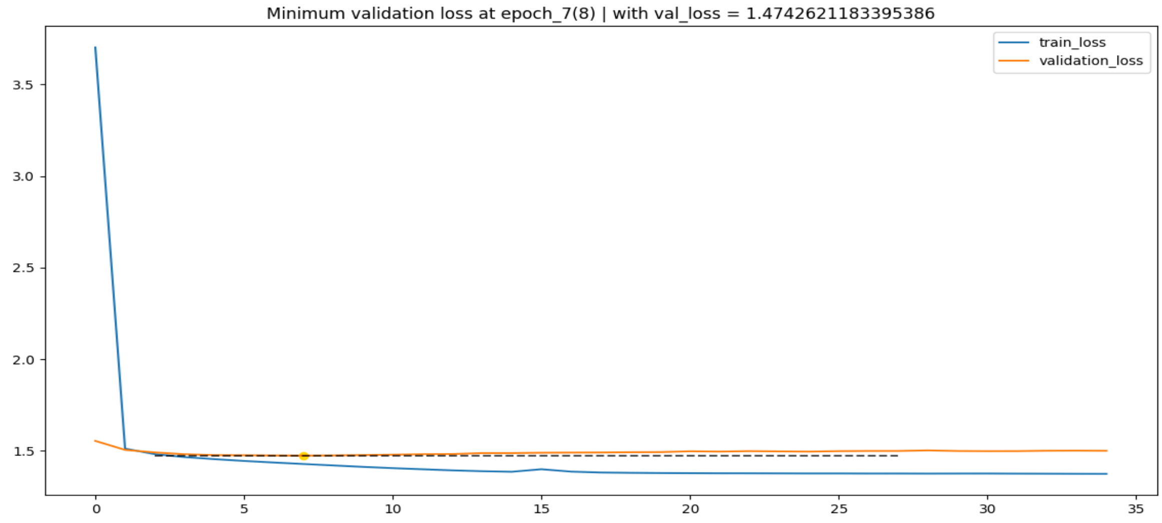


Fig 3: Blip Network Architecture

For this model we have used the Indiana Chest-xray dataset.
2559 frontal chest x-ray images for train also 320 images for validation and 320 images for the final test.
We will feed the images to the ViT image encoder and the corresponding doctor caption to the Bert encoder. The pretrained model validation loss was 13.051, after training the model for 8 epochs, the validation loss reached 1.474. In the test phase, we generate a text report for each input image. Moreover, our Bleu score is equal to 0.205.



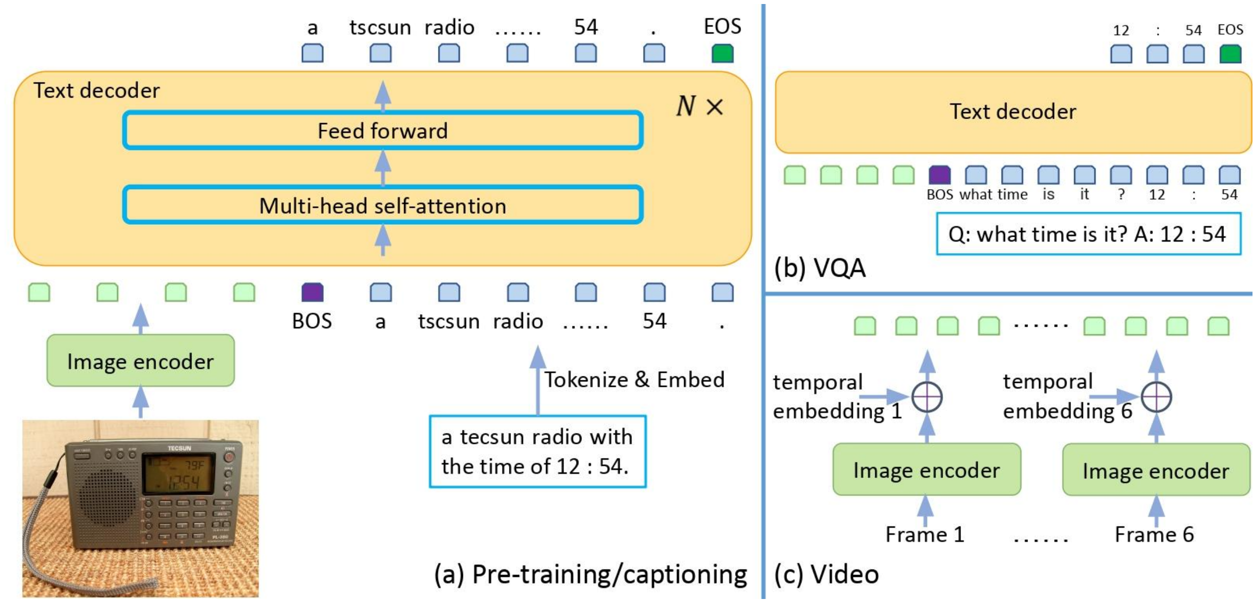
3. GIT

The only components for GIT are one image encoder and one text decoder. The image encoder is based on the contrastive pre-trained model. The input is the raw image and the output is a compact 2D feature map, which is flattened into a list of features. With an extra linear layer and a layernorm layer, the image features are projected into D dimensions, which are the input to the text decoder.

The approach is equivalent to separating the two tasks sequentially:

(i) using the contrastive task to pre-train the image encoder

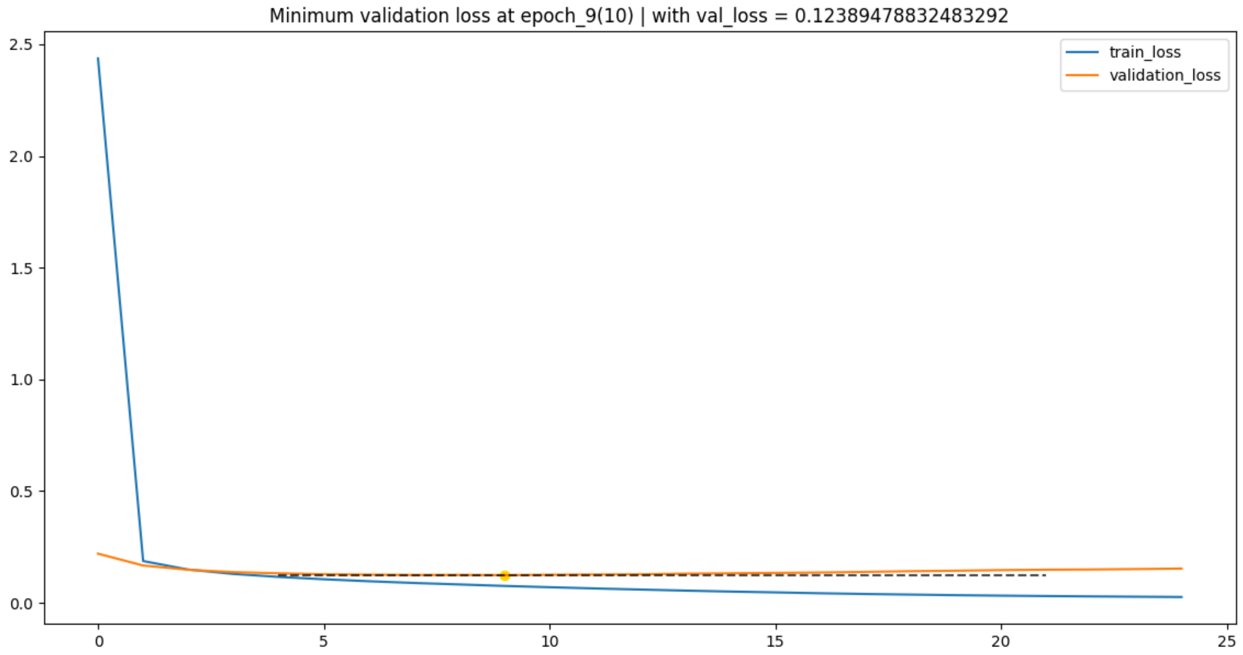
(ii) using the generation task to pre-train both the image encoder and text decoder



The text decoder is a transformer module to predict the text description. The transformer module consists of multiple transformer blocks, each of which is composed of one self-attention layer and one feed-forward layer. The text is tokenized and embedded into D dimensions, followed by an addition of the positional encoding and a layernorm layer.

The image features are concatenated with the text embeddings as the input to the transformer module. The text begins with the [BOS] token, and is decoded in an auto-regressive way until the [EOS] token or reaching the maximum steps. The seq2seq attention mask as is applied such that the text token only depends on the preceding tokens and all image tokens, and image tokens can attend to each other. This is different from a unidirectional attention mask, where not every image token can rely on all other image tokens.

For this part we have used the same data as for Blip section.
We have used GIT large-size which is pretrained on COCO. The pretrained validation loss was 11.530, after training the model for 10 epochs the validation loss reached 0.124.
In the test phase we generate a text report for each input image. Moreover, our Bleu score is equal to 0.212.



Results:

Conclusion:

Acknowledgement:

To achieve our project goals, a VPS has been bought from HPC of Sharif University of Technology with the following specifications:

GPU: A100 - CPU: 12 core Intel Xeon - RAM: 152 GB

References:

1. Efficient CNN-LSTM based Image Captioning using Neural Network Compression.
2. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning
3. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation