Transforming Radiology Workflows: Pretraining for Automated Chest X-ray Report Generation

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

Automated chest X-ray report generation using machine learning has emerged as a promising technology for improving the accuracy and efficiency of chest X-ray interpretation. In this paper, we present a novel approach for automated report generation that combines the power of vision transformers for image information encoding and PubMedBERT for text decoding. Our model extracts image features using a vision transformer and text features using PubMedBERT (Gu et al., 2020). The encoded features are then fed into a text decoder to generate standardized reports. We trained our model on a dataset of chest X-rays and corresponding report findings (IU dataset) and evaluated its performance on a small subset of the MIMIC-CXR dataset.

Keywords: Chest X-ray, BLIP, PubMedBERT, ViT, Pre-Training.

1. Introduction

Chest X-rays are widely used for diagnosing chest-related conditions but require specialized expertise for interpretation, which can be time-consuming and subject to errors. Manual report writing is also costly, prone to variability, and may delay treatment. Healthcare professionals may interpret the same image differently, leading to inconsistent diagnoses and delays in treatment. Recent advancements in machine learning may improve the efficiency and accuracy of chest X-ray interpretation by automating the report-generating process. This could reduce wait times for patients, minimize errors, and make interpretation more accessible while also being cost-effective.

In this research paper, we present a novel machine-learning model for generating chest X-ray reports. Our model utilizes vision transformer (ViT) (Dosovitskiy et al., 2021) to extract features from the chest X-ray images, followed by a text decoder to generate standardized reports. The reports include key features of the X-ray image, such as lung function, the presence of any abnormalities, and a differential diagnosis based on the identified features. We train our model on the IU dataset of chest X-rays and corresponding reports and evaluated it on a small sample of the MIMIC-CXR dataset.

Similar work has been performed by researchers in the past. (Wu et al., 2022) presents DeltaNet for automatically generating medical reports which applies a conditional generation process. (Najdenkoska et al., 2021) proposes variational topic inference for automatic report generation by introducing a set of topics as latent variables to guide sentence generation by aligning image and language modalities in a latent space. (Liu et al., 2021) proposes a Contrastive Attention (CA) model for X-ray report generation.

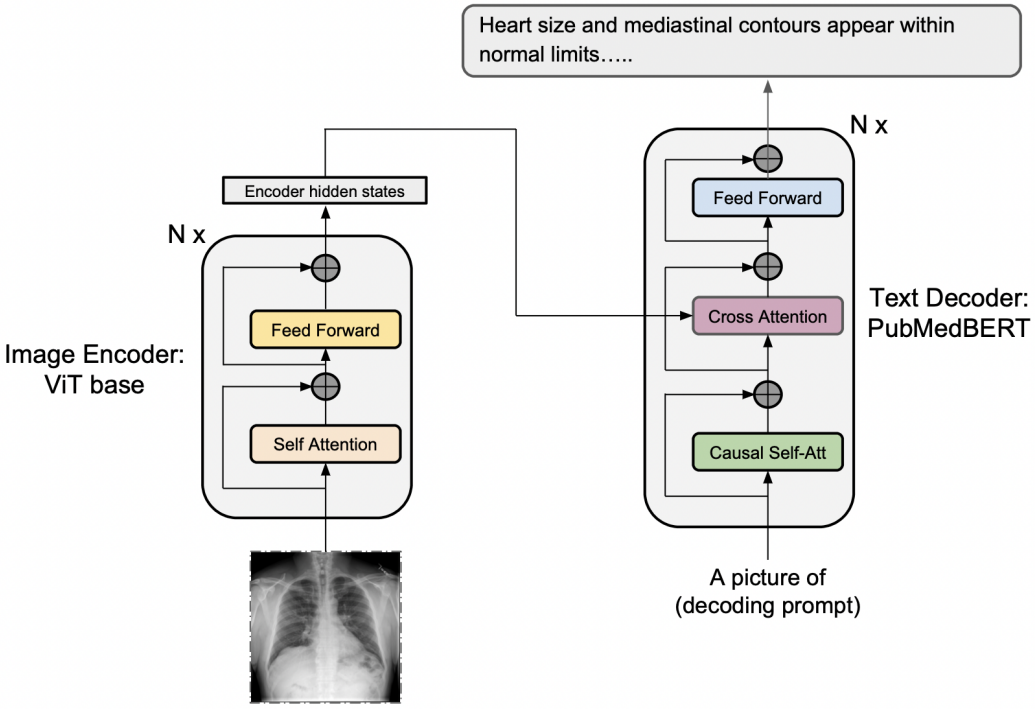

Figure 1: The encoder-decoder architecture of our image-generation framework

2. Methodology

We adopt the architecture from BLIP (Li et al., 2022), a bootstrapping language-image model pre-trained for both understanding-based and generation-based objectives. The model pre-training takes the input of pairs of images and the corresponding text and afterward generates radiology reports for given X-ray images. The framework uses a multimodal mixture of encoder-decoder (MED) model architecture that enables effective multi-task pre-training. MED can operate as a unimodal encoder, an image-grounded text encoder, or an image-grounded text decoder. Our model is jointly pre-trained with three vision-language losses: image-text contrastive learning, image-text matching, and image-conditioned language modeling. For image-text contrastive learning loss and image-text matching loss, we follow by (Li et al., 2021). We use the language modeling loss which is a cross-entropy loss for training the model to maximize the likelihood of the next token in the text.

In this work, we pre-train a BLIP model from the beginning with X-ray images and corresponding findings from reports. Our model architecture employs a visual transformer (ViT) as its image encoder, which divides an input image into patches and encodes them as a sequence of embeddings. Additionally, a [CLS] token is used to represent the global image feature. The text encoder is initialized with PubMedBERT(Gu et al., 2020) which is pre-trained on PubMed abstracts. To serve as a framework for text generation, our model replaces the bi-directional self-attention layers in PubMedBERT with causal self-attention layers that can operate as an image-grounded text decoder. At inference time, our model

Table 1: Scores calculated on 500 random samples from MIMIC-CXR Dataset

	Jaccard	ROUGE-2	ROUGE-l	METEOR
ViT + PubMedBERT	0.16	0.079	0.23	0.20

provides an encoder-decoder architecture for generating radiology reports with a given X-ray image, which is shown in Figure 1.

3. Experiments and Results

We utilize the IU dataset(OpenI) for chest X-ray to pre-train our model, which comprises frontal X-ray images of patients. The dataset was obtained from Kaggle and consisted of 3818 images with corresponding reports. To create captions for our images, we use the findings from the report, rather than the impression, as they provide a more objective description. We select frontal X-ray images for pre-training as they contain more informative features. Our model was trained using image-text contrastive loss, image-text matching loss, and language modeling loss, with the same objective as BLIP to improve language generation. We use ViT-base as our image encoder and PubMedBERT-base as our text decoder. We resize all the images to 224 by 224 and pre-train our model for 20 epochs with batch size 8. We use the initial learning rate of 3e-4 with 3000 warm-up steps.

To evaluate the performance of our pre-trained model, we randomly selected 500 samples from the MIMIC-CXR dataset(Johnson et al., 2019), which is a large, publicly available dataset of chest X-ray images with corresponding radiology reports. We then compared the system-generated findings with the original findings in the reports and calculated various metrics, including Jaccard similarity, ROUGE, and METEOR scores to measure the accuracy and quality of the generated reports. The scores are displayed in Table 1.

4. Conclusion

Our pre-trained model is aimed at generating X-ray reports, which can be helpful for reducing the workload of radiologists and facilitating quick diagnoses. To this end, we employed the BLIP architecture, which is known for its high accuracy and efficiency. The image encoder we used is a Vision Transformer, which has shown promising results in computer vision tasks, while the language encoder we used is PubMedBERT, a pre-trained language model specifically designed for biomedical applications.

While our current pre-trained model has shown some promise, its performance is limited due to the small size of the pre-training dataset. However, we believe that using the full MIMIC-CXR dataset for pre-training will greatly improve our model's performance and accuracy.

By utilizing the full MIMIC-CXR dataset, which will provide us with a much larger and more diverse set of training data, we hope to achieve higher accuracy and more robustness in our model, which will make it a more useful tool for radiologists and medical professionals. Our goal is to develop a reliable and efficient pre-trained model that can generate accurate X-ray reports quickly and easily, ultimately improving patient care and outcomes.

References

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing, 2020.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6 (1):317, 2019.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. Contrastive attention for automatic chest X-ray report generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 269–280, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.23. URL https://aclanthology.org/2021.findings-acl.23.

Ivona Najdenkoska, Xiantong Zhen, Marcel Worring, and Ling Shao. Variational topic inference for chest X-ray report generation, 2021.

OpenI. Indiana university - chest X-rays (png images). URL https://openi.nlm.nih.gov/faq.php.

Xian Wu, Shuxin Yang, Zhaopeng Qiu, Shen Ge, Yangtian Yan, Xingwang Wu, Yefeng Zheng, S. Kevin Zhou, and Li Xiao. DeltaNet: Conditional medical report generation for COVID-19 diagnosis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2952–2961, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.261.