Smart Radiology: Context-Enhanced Report Generation Using BLIP2 T5

Chandana Pamidi, email: cpamidi@umass.edu

1. Abstract

In the realm of healthcare, radiology reports play a crucial role in diagnosing and understanding medical conditions. However, for commoners without a medical background, comprehending these reports can be challenging due to the prevalence of complex medical terminology. This project aims to bridge this gap by developing an AI-driven system that generates easy-to-understand radiology reports from X-ray images. Leveraging the power of the BLIP2 T5 model and in-context learning, the system ensures accuracy while providing contextually relevant explanations. This innovative approach integrates advanced natural language processing (NLP) techniques to avoid misinformation and enhance the accessibility of medical data for non-experts, empowering patients with clear and comprehensible information about their health.
KEYWORDS : BLIP, in-context learning, Advance NLP

2. Introduction

2.1. Background

Radiology reports are integral to medical diagnostics, offering detailed insights into patient health through the interpretation of X-ray images. These reports, however, often contain specialized medical jargon that can be perplexing for individuals without a medical background. Misunderstanding or misinterpreting these reports can lead to confusion and anxiety among patients. Therefore, there is a pressing need for a solution that translates complex medical information into easily digestible language without compromising accuracy.

2.2. Objective

This project proposes the development of an AI-driven system that generates simplified radiology reports from X-ray images. By integrating the BLIP2 T5 model with in-context learning, the system aims to produce reports that are not only accurate but also contextually relevant and easy for commoners to understand. The focus is on minimizing noisy data and enhancing the learning of contextual relationships between tokens to ensure the generation of reliable and comprehensible information.

3. Methodology

3.1. Baseline Approach

My baseline approach centers on leveraging the BLIP2-T5 model from the LAVIS (Language-and-Vision) library, renowned for its robust resources tailored to language and vision integration research. This choice is motivated by BLIP2-T5's exceptional ability to handle multimodal data, seamlessly blending textual and visual information. By harnessing pre-trained representations within BLIP2-T5, I tap into a vast reservoir of knowledge acquired from extensive pre-training on diverse datasets. This equips the model with nuanced features crucial for processing complex inputs like X-ray images and textual descriptions. BLIP2-T5's support for in-context learning empowers it to adapt dynamically to my task's specific context, ensuring the generation of accurate and contextually relevant reports.

3.2. Finetuning

In my project, I employed the BLIP2-T5 model and leveraged its support for in-context learning, enabling dynamic adaptation to the specific context of my task. This unique capability ensured the generation of accurate and contextually relevant reports tailored to the nuances of medical imaging data.

To fine-tune the BLIP2-T5 model for my specific task, I utilized the "mimic_chest_xray_v1" dataset, which contains a diverse collection of chest X-ray images. By inputting the model with context, consisting of both the chest X-ray images and their associated textual descriptions, I trained the model to generate informative reports based on the provided context.

During the fine-tuning process, I conducted hyperparameter tuning to optimize the model's performance. This involved experimenting with various hyperparameters such as learning rate, batch size, and number of training epochs to identify the configuration that yielded the best results.

By fine-tuning the BLIP2-T5 model on the "mimic_chest_xray_v1" dataset and performing hyperparameter tuning, I aimed to enhance the model's ability to generate accurate and clinically relevant reports from chest X-ray images.

3.3. Dataset

For my experiments, I selected the "mimic_chest_xray_v1" dataset curated by Hongrui. This dataset contains a diverse collection of chest X-ray images, making it suitable for training and evaluating models for various medical imaging tasks. To ensure robust performance assessment, I partitioned the dataset into three subsets: training, validation, and testing, with proportions of 60%, 20%, and 20%, respectively.

The training set, comprising 60% of the data, serves as the primary data source for model training. Here, the model learns to extract meaningful features from chest X-ray images and associated textual descriptions, which are crucial for generating accurate reports.

The validation set, constituting 20% of the dataset, plays a pivotal role in hyperparameter tuning and model selection. During training, the model's performance on this subset is monitored closely to prevent overfitting and ensure generalization to unseen data.

Finally, the testing set, also representing 20% of the dataset, serves as an independent benchmark for evaluating the model's performance. Here, the trained model's ability to generate informative and contextually relevant reports from unseen chest X-ray images is thoroughly assessed.

4. Pros and Cons

4.1. Pros

- The automated report generation streamlines the process of analyzing medical images and generating reports, saving time for both healthcare professionals and patients.
- The project serves as an educational tool, helping users learn about medical conditions and diagnostic interpretations through interactive reports and explanations.

4.2. Cons

- Lay users may misinterpret or misunderstand the generated reports, leading to incorrect assumptions or decisions about their health.
- Users may become overly reliant on the automated reports, potentially overlooking the importance of consulting healthcare professionals for accurate diagnoses and treatment recommendations.

5. Experiment Analysis

Through experimentation involving baseline and fine-tuned models, zero-shot and five-shot predictions, with and without rationale, I conducted an analysis based on ROUGE and METEOR scores. Despite observing improvements in the fine-tuned model, the results failed

to demonstrate a significant enhancement in prediction quality attributed to in-context learning. Surprisingly, there was no discernible difference in scores between experiments conducted with and without rationale. This outcome suggests that while fine-tuning the model yielded advancements, the efficacy of in-context learning in improving prediction quality remains inconclusive based on the evaluation metrics utilized. Further investigation and refinement may be necessary to elucidate the true impact of in-context learning in the context of medical image report generation.

Table 1. Experiment Results

	Rouge1	Rouge2	RougeL	Meteor
Baseline Zero-shot w/o Rationale	0.0592935	0.0029857	0.0484824	0.0181858
Baseline Zero-shot with Rationale	0.0592796	0.0029830	0.0484821	0.0181851
Baseline Zero-shot w/o Rationale	0.0593089	0.0029857	0.0485052	0.0181815
Baseline Five-shot with Rationale	0.0593057	0.0029874	0.0485187	0.0181813
Fine-tuned Zero-shot w/o Rationale	0.2526773	0.0889489	0.1933804	0.1415361
Fine-tuned Zero-shot with Rationale	0.2527088	0.0889440	0.1933961	0.1415300
Fine-tuned Five-shot w/o Rationale	0.25264551	0.0889276	0.1933679	0.1415228
Fine-tunes Five-shot with Rationale	0.2527015	0.0889791	0.1934306	0.1415332