SIIM CMIMI24
Oct 21-22 | Boston, MA
Conference On Machine Intelligence In Medical Imaging

Generating Structured Radiology Reports of Chest Radiographs using Retrieval Augmented Generation

Yash S. Saboo, University of Texas at Austin; Aaron Fanous, MD, MS; Kal L. Clark, MD

Introduction/Background

Radiologist workload has increased over the past decade, increasing burnout and the risk of diagnostic inaccuracies. Artificial intelligence (AI) algorithms have been developed for tasks such as disease detection, image segmentation, and impression-generation. However, much work remains in using AI to generate comprehensive radiology reports. The purpose of this study is to develop a retrieval-augmented generative AI model that accurately generates radiology reports of chest radiographs (CXRs).

Methods/Intervention

We trained the DenseNet-121 model on 13964 CXRs from the VinDr-CXR dataset to classify the CXRs into seven classes: aortic enlargement, cardiomegaly, interstitial lung disease, lung opacity, pleural effusion, pneumothorax, no finding. We then used the trained DenseNet-121 model as an encoder to generate embeddings for 159,970 CXRs from the Medical Information Mart for Intensive Care Chest X-ray JPG (MIMIC-CXR-JPG) dataset. The embeddings were stored in a vector database. We generated reports for a separate test set of 59 CXRs from MIMIC-CXR-JPG using similarity search, where we compared the vector embedding of each of the 59 test CXRs with the embeddings in the vector database. The most similar embedding in the vector database for each of the 59 CXRs was identified using cosine similarity, and the most-similar embedding's associated report was retrieved and restructured into six distinct sections (cardiomediastinum, pleural space, lungs, bones, hardware, other) using Generative Pretrained Transformer (GPT) 4. These restructured reports were recommended as the reports for the 59 CXRs, respectively.

Results/Outcome

On the test set of 59 CXRs, the model achieved a median BLEU score of 0.0701, median BERT score of 0.216, median CheXbert score of 0.227, and median RadCliQ score of 1.713. Additionally, a board-certified radiologist assigned a RADPEER Score, ranging from 1 to 3, to each section (cardiomediastinum, pleural space, lungs, bones, hardware, other) of the 59 AI-generated reports, as shown in Figure 1. Averaging across all sections, the model achieved a RADPEER Score of 1.548.

Conclusion

This retrieval-augmented generative AI model has the potential to assist radiologists with generating structured radiology reports.

Statement of Impact

This approach of generating structured radiology reports on CXRs may increase workflow efficiency and reduce radiologist burnout.

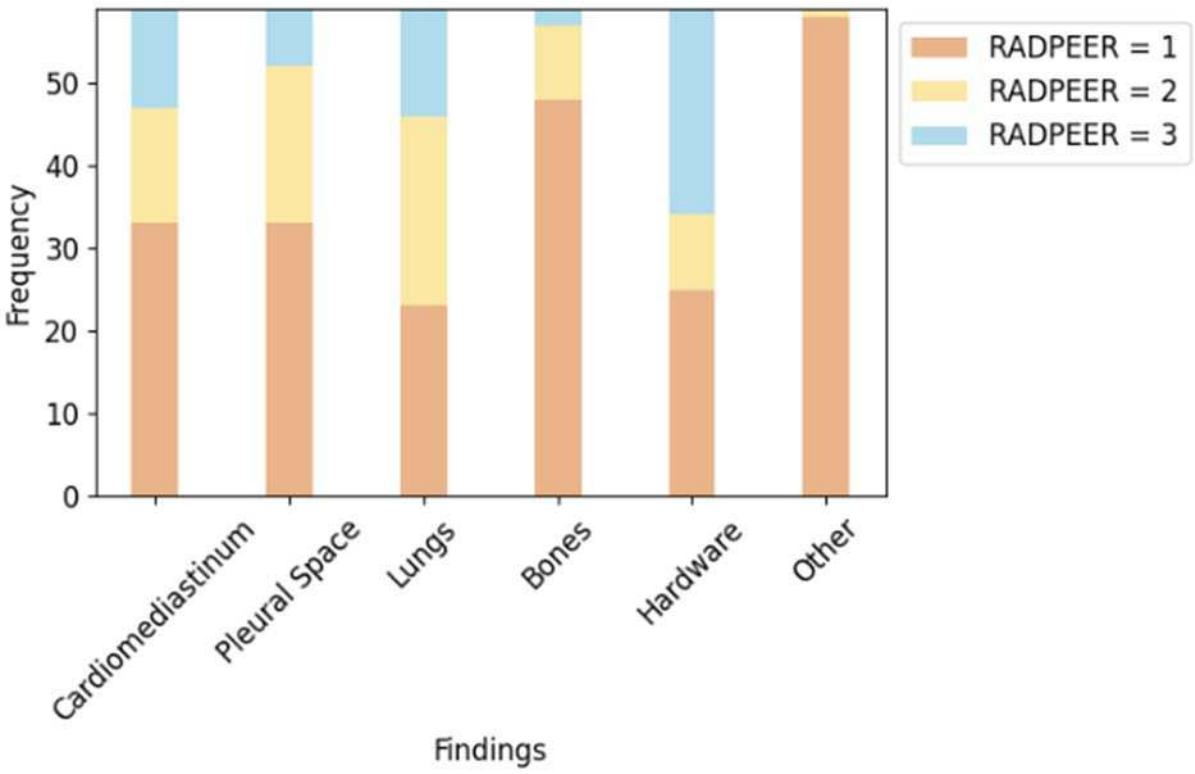

Fig. 1 shows the distribution of RADPEER scores per section of the chest radiograph. The model achieves a RADPEER Score of 1 56% of the time on the cardiomediastinum and pleural space region, 39% of the time on the lungs region, 81% of the time on the bones region, 42% of the time on the hardware region, and 98% of the time on the Other region

Keywords

Generative AI; Retrieval Augmented Generation; Natural Language Processing; Large Language Models; Chest Radiographs; Report Generation