Improving Performance of Vision Encoding Large Language Models with Contextual Prompts

Luke Taylor
lwtaylor@ucsd.edu

Raymond Song
rysong@ucsd.edu

Muchan Li
mul005@ucsd.edu

Albert Hsiao
a3hsiao@health.ucsd.edu

Abstract

Recent advancements in AI have led to the ability for large language models to understand data from different modalities. In this paper, we explain our methods to fine tune LLaVA, an open source Large Language and Visual Assistant similar to GPT-4, to read chest radiographs and generate a report in a radiologist's style.

Website: https://ai-xray.netlify.app/
Code: https://github.com/raymondsong00/Xray-Report-Generator

1 Introduction

Previously in our Quarter 1 Project, we explored and demonstrated how the power of deep learning models like convolutional neural networks (CNN) can contribute to the automation of radiograph analysis, specifically concentrated at detecting pulmonary edema. We also showed that there remain many challenges with CNN. For example, we were only able to achieve around 0.7 F1-score on the testing radiographs due to CNN's inability to capture more complex and often noisy patterns. Additionally, a great portion of the raw data, as radiographs are typically accompanied by reading reports from radiologists, are left out of the training and evaluation process. Consequently, with the advent of Large Language Models (LLM) and Vision Transformers (ViT), we turned our eyes onto these models, or more accurately AI systems, since their transformer-based architecture are capable of compressing and capturing more feature patterns. Additionally, as inspired by this Google Research blog post that explains and summarizes the feasibility of joining ViT and LLM to make multimodal (vision and language) models, which should now enable us to train models that can understand both plain-text reports and high-resolution radiographs, we decided to delve into this focus and find better solution.

2 Objectives

- Navigate how multi-modal models are used to leverage both X-Ray reports and images to generate reports.
- Assess whether the generated reports mimic the style of reports written by the expert readers.
- Assess the accuracy of the generated reports in identifying pathologies.
- Determine whether additional context in the text input of an LLM improves generated text outcomes.
- Explore the effect of prompt engineering on report generation and accuracy.
- Evaluate and quantify the improvement in the LLM for Chest X-rays given more context.

3 Methods

The general approach is to finetune a Large Language Model (LLM) with vision capabilities instead of training both a vision CNN tower and a LLM from scratch. We will present our design decisions, data wrangling, prompt design, and evaluation metrics in the following subsections.

3.1 Model Selection

Many new multimodal large model architectures have been proposed lately. For example, Li et al. (2023) famously proposed the BLIP-2 architecture through which they boostrapped vision-language pre-training by training a lightweight Query Transformer to efficiently connect the modality gap between pre-trained vision and language models. Similarly, Xu et al. (2023) built on top of the BLIP-2 framework and proposed a novel way of training the adapter between image encoder and language model using radiology-specific images and free-text. Most recently, Liu et al. (2023) presented a breakthrough in their Large Language and Vision Assistant (LLaVA) framework where they leveraged GPT to generate multi-turn visual instructions with regard to a single image-text pair and sequentially train the image encoder and language model on the generated visual instruction sequence. LLaVA's visual-instruction tuning mechanism offers the most state-of-the-art question-answer performance, hence we chose to use LLaVA as our base model and fine-tune LLaVA with our chest radiographs and paired radiologist reports dataset.

3.2 Data Wrangling and Fine-tuning

Initially, our dataset contained only large repositories of X-rays images consolidated in HDF5 files and raw text reports with patient information stored in spreadsheets. We developed pipelines to extract images and store alternatively as JPG files and extract patient-related information such as clinical history, age, gender, radiologist name, and reported findings and impressions. All patient-related information are then embedded in question prompts for fine-tuning LLaVA. We then fed the question prompt and the ground truth radiologist response as question-answer pairs to the model. The fine-tuning of the model was done using a Nvidia RTX A6000 with LORA adapters due to GPU memory constraints.

3.3 Prompt Design

We devised two prompts for model fine-tuning, one for benchmarking LLaVA's performance in reading X-rays and one for testing the factors within the question prompt that would influence LLaVA's performance. Our benchmark prompt, called the generic prompt, was designed similar in spirit to the prompt used by Sun et al. (2023), much like a general and straightforward question. Our improved prompt, called the context-embedded prompt, was designed similar in spirit to the few-shot prompts presented by Brown et al. (2020). We included patient-specific context as the examples and envisioned that the model would align these added contexts into report generation, mimicking the diagnostic procedures of a human radiologist.

3.4 Evaluation Metrics

We designed two metrics to quantify and evaluate LLaVA's performance. One, we employed model (as proposed by Remy, Demuynck and Demeester (2023)) to transform LLaVA-generated and radiologist-provided reports into word embedding vectors and compared the cosine similarity between the two to quantify how similar LLaVA's responses mimicked the styles of the ground truths. This became our first metric of measuring performance.

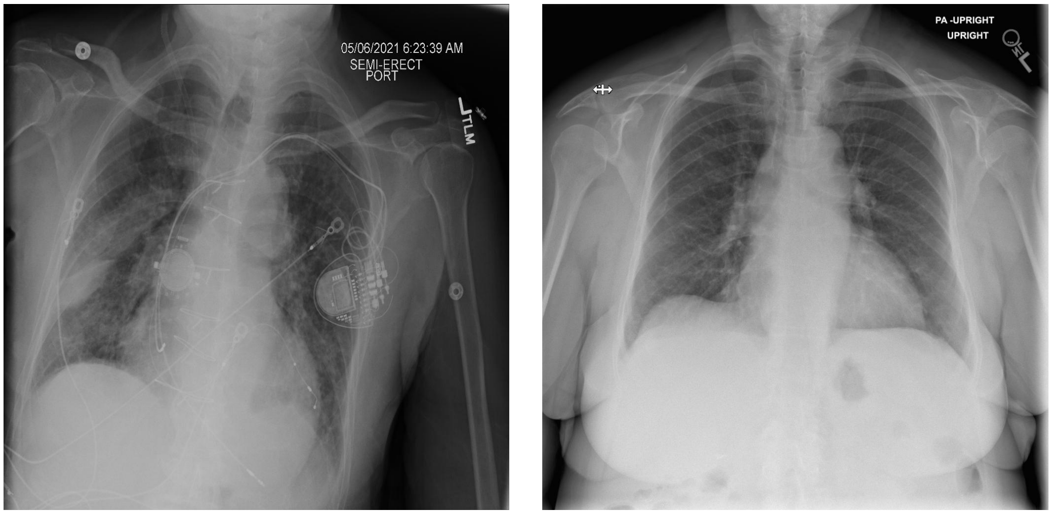
Two, we used PEGASUSScore-PET by Tie et al. (2023), a summarizer for PET scan reports, to summarize LLaVA generated reports and radiologist written reports into more concise language, given it's superior understanding of complex medical language compared to BART. With clear, summarized reports, we then fed these summaries into BART (Bidirectional Auto-Regressive Transformers) from Lewis et al. (2019) to perform zero shot classification. We gave BART "*pneumothorax*", "*pneumonia*", "*pleural effusion*", "*cardiomegaly*", "*edema*", "*rib fracture*" as candidate labels for every LLaVA and radiologist written report in our validation set. BART assigned probability values to each of these labels based on its confidence of disease's presence and severity in a report. We then used these probabilities to create binary labels for the radiologist reports with a threshold of 0.5 to create labels to compare against for an ROC curve between the LLaVA reports and corresponding radiologist reports. While these methods are crude and may not be as accurate as a trained classifier, we found them to be effective at evaluating our model's performance, similar to what others have done using models like GPT-4 to labels for disease labels and findings.

The third metric we used to evaluate the model was the number of distinct reports it produced for the testing set. This allows us to naively determine whether our model is just "memorizing" and "repeating" common and patterns answers or is indeed learning features specific to X-rays and pathological entities.

4 Results

As presented in Fig 2, when LLaVA is prompted with the generic prompt, it repeated the exact same response to two distinct X-Rays. This suggests that either the model has not learned any X-ray specific imaging feature and just memorized the report text (since the prompt are first fed to the model via fine-tuning), or that the prompt is not specific enough to elicit reasoning. However, when the model is prompted with the context-embedded prompt, the model can now articulate distinct findings with regard to each X-ray. Specifically, since we provided the corresponding radiologist reader, the model did better in memorizing and mimicking reader-specific reporting style. Moreover, the seemingly increased pathology detection accuracy–as an example, LLaVA correctly identified *interval removal of a right IJ (sheath)*–suggesting that model is now better at comprehending the task and retrieving the relevant vision and language features.

The ROC curves in Fig 3a roughly all fall around the (**0.5**) diagonal line, showing poor discriminative ability that only comparable to random chance. The ROC curves in Fig 3b shows greater area under the curve, showing model's enhanced dicriminative ability, with



(a) X-Ray 1 (b) X-Ray 2

Figure 1: One pathology-present X-ray (left) and one pathology-free X-ray (right) for demonstration

the best performance in detecting pneumothorax at **0.73**. Answers prompted using the generic prompt have similarity scores distributed normally, centered around **0.52**, as shown in Fig 4a; whereas the distribution of scores in Fig 4b suggest that the model has indeed comprehended the task better with the majority of the answers having similarity scores above **0.63** and only a diminishing amount of results in the low scoring tail.

For our best performing model, 52% of LLaVA's reports were distinct, compared with 97% of radiologists' reports being distinct. The baseline LLaVA model trained on the generic prompt only had 11% of its responses being distinct, providing evidence that LLaVA repeated reports it found most common in its training data. Its most common output was *FINDINGS:Devices, lines and Tubes: None Mediastinum: Unchanged cardiac silhouette and mediastinal contours. No lymphadenopathy is appreciated.Lungs: The lungs are clear.Pleura: No pneumothorax or effusion.Bones and soft tissues: Unchanged IMPRESSION: No acute cardiopulmonary findings.*

5 Discussion

The output of the generic report from LLaVA suggests that the model is not using the image to inform its text generation but rather takes an average report as its generated output based on the varying context in each prompt. As we include more context in the training of the model, the model produces more varied reports that more accurately resemble the models structurally and semantically seen in Figure 4b and Figure 3b. The majority of the improvement is likely in the text embedding that the language model LLaVA uses because the image is the identical. Perhaps the vision encoder that LLaVA uses does not fully capture the details of a chest radiograph that has many minuscule features that are not weighted

Radiologist Ground Truth Reports

Example X-Ray 1

FINDINGS: See impression.

IMPRESSION: Interval removal of a right IJ Swan-Ganz catheter. Sheath remains present. Increased interstitial and alveolar pulmonary edema. Stable small bilateral pleural effusions with slight loculation in the right major fissure.

Example X-Ray 2

FINDINGS: Unremarkable cardiomediastinal silhouette. No consolidation, pleural effusion or pneumothorax. Basilar reticular fibrotic changes are noted, unchanged. No acute osseous abnormality identified.

IMPRESSION: Stable chest x-ray. No acute cardiopulmonary findings.

LLM Generated Reports

Generic Prompt

Write a report for the given chest x-ray. It should contain a clear findings and impression section. Be explicit for any abnormalities or normal findings.

Generic Prompt

Write a report for the given chest x-ray. It should contain a clear findings and impression section. Be explicit for any abnormalities or normal findings.

LLaVA Output to Generic Prompt:

FINDINGS: Lines and Tubes: None. Mediastinum: The cardiomediastinal silhouette is unchanged. No lymphadenopathy is appreciated. Lungs: The lungs are clear. Pleura: No pneumothorax or effusion. Bones and soft tissues: Unchanged

IMPRESSION: No acute cardiopulmonary abnormality.

LLaVA Output to Generic Prompt:

FINDINGS: Lines and Tubes: None. Mediastinum: The cardiomediastinal silhouette is unchanged. No lymphadenopathy is appreciated. Lungs: The lungs are clear. Pleura: No pneumothorax or effusion. Bones and soft tissues: Unchanged.

IMPRESSION: No acute cardiopulmonary findings.

Context Embedded Prompt

AUTHOR: Hsiao, Albert, CLINICAL HISTORY: Swan placement, CONTEXT: Inpatient, AGE: 75, SEX: Male, Based on AUTHOR and CLINICAL HISTORY, suppose you were a radiologist on X-RAY, could you provide a detailed report from this chest X-ray?

Context Embedded Prompt

AUTHOR: Hsiao, Albert, CLINICAL HISTORY: Cough for 1 month, CONTEXT: Outpatient, AGE: 73, SEX: Female, Based on AUTHOR and CLINICAL HISTORY, suppose you were a radiologist on X-RAY, could you provide a detailed report from this chest X-ray?

LLaVA Output to Context Embedded Prompt:

FINDINGS: See impression.

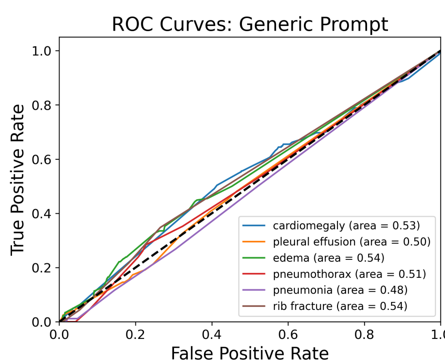
IMPRESSION: Interval removal of a right IJ sheath. Right chest tubes remain present. Stable small right pneumothorax. Stable interstitial edema.

LLaVA Output to Context Embedded Prompt:

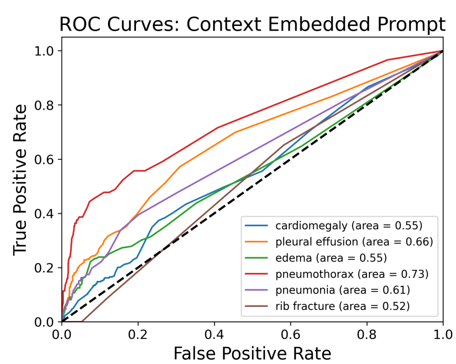
FINDINGS: Unremarkable cardiomediastinal silhouette. No consolidation, pleural effusion or pneumothorax. No acute osseous abnormality identified.

IMPRESSION: No acute cardiopulmonary findings.

Figure 2: LLaVA outputs in response to different prompts compared to the ground truths.

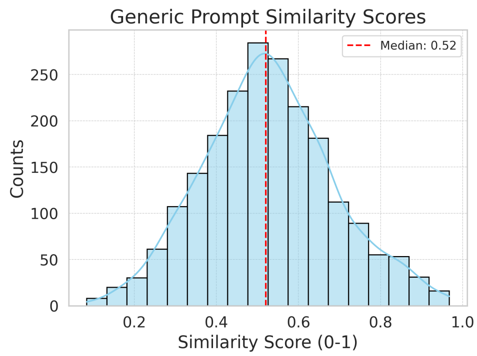


(a) Figure 3a

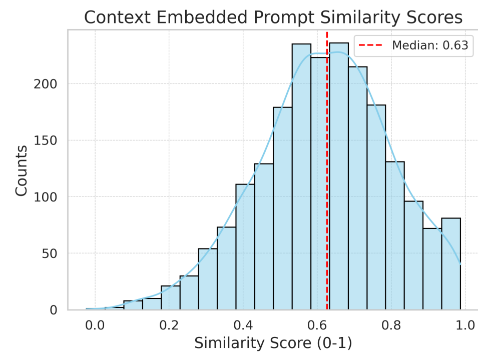


(b) Figure 3b

Figure 3: ROC curves for different pathologies and their corresponding AUC values given the BART scoring



(a) Figure 4a



(b) Figure 4b

Figure 4: Distribution of cosine similarity scores between text embedding of LLaVA generated prompt and actual radiologist report for both generic and context embedded prompts

enough in the model.

As a result, further experiments can look into doing an ablation study and remove a part of the context and see the contribution that each part of the context affects the outputs of the model. Because we added all of the context at once, it is difficult to do analysis on which factors the model is relying on. We believe that LLaVA mainly uses from the clinical history and the author name to generate the report. Additionally, to see if how much LLaVA's report generation is influenced by the image features on top of text features from reports, a good experiment would be to remove the vision encoder and train just the Vicuna-13b that LLaVA uses. Changing the vision encoder tower would also be an interesting experiment because the pre-trained CLIP ViT-L/14 is trained on general-purpose images and text labels and likely does not embed the X-ray's details well enough for the language model to use in its response.

6 Conclusion

Adding context to the prompts greatly improves the performance of LLaVA on generating radiologist-like reports on Chest X-rays. The fine-tuned models are also able to mimic the style that the expert readers use. The multi-modal models are able to use both the chest radiographs and reports to generate reports. Although the highest AUC for any pathology is **0.73** for pneumothorax, the performance of the model with additional information improves from random guessing. Further exploration into prompt engineering combined with more domain-specific training architecture would likely further improve the performance.

7 Acknowledgement

We would like to first and foremost thank our mentor Dr. Albert Hsiao in supporting us with ideas and directions to navigate throughout this project. His insights are invaluable. We would like to also thank Eric, a member of Dr. Hsiao's lab who has always inspired us with awesome tricks and insights on training, testing, and navigating new models. Lastly, we would like to thank the lab's machine Ashe and its powerful A6000 GPU, without it none of this project would be possible. This project was a true teamwork and we are all very grateful for this opportunity and journey.

References

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell et al. 2020. "Language models are few-shot learners." *Advances in neural information processing systems* 33: 1877–1901

Lewis, Mike, 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*

Li, Junnan, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." In *International conference on machine learning*. PMLR

Liu, Haotian, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. "Visual Instruction Tuning."

Remy, François, Kris Demuynck, and Thomas Demeester. 2023. "BioLORD-2023: Semantic Textual Representations Fusing LLM and Clinical Knowledge Graph Insights."

Sun, Yuxuan, Chenglu Zhu, Sunyi Zheng, Kai Zhang, Zhongyi Shui, Xiaoxuan Yu, Yizhi Zhao, Honglin Li, Yunlong Zhang, Ruojia Zhao et al. 2023. "Pathasst: Redefining pathology through generative foundation ai assistant for pathology." *arXiv preprint arXiv:2305.15072*

Tie, Xin, Muheon Shin, Ali Pirasteh, Nevein Ibrahim, Zachary Huemann, Sharon M Castellino, Kara M Kelly, John Garrett, Junjie Hu, Steve Y Cho et al. 2023. "Automatic personalized impression generation for pet reports using large language models." *ArXiv*

Xu, Shawn, Lin Yang, Christopher Kelly et al. 2023. "ELIXR: Towards a general purpose X-ray artificial intelligence system through alignment of large language models and radiology vision encoders." *Title of the Journal* Volume Number (Issue Number), p. Page Numbers