Enhancing Vision-Language Models for Specialized Medical Tasks: A Focus on Breast Cancer Diagnosis Detection and Analysis

Sid Potti
Stanford University
sidpotti@stanford.edu

Jayson Meribe
Stanford University
jmeribe@stanford.edu

Kyle Hu
Stanford University
kylehu@stanford.edu

The literature review of sources relating to our project is discussed within the appropriate parts of this milestone.

1. Introduction

The improved specialization of medical AI models will lead to a unrivaled emergence of improved healthcare treatment and positive patient outcomes. Breast cancer, one of such specialized medical problems, remains one of the most prevalent and deadly cancers affecting women globally. Early and accurate diagnosis is critical for improving patient outcomes, making it imperative to develop advanced diagnostic tools. Vision-language models have shown great promise in various medical imaging tasks by leveraging both visual and textual data. Having such data fusion is critical for accurate diagnosis and prediction. However, there is more that can be done. This paper explores the enhancement of vision-language models specifically tailored for breast cancer diagnosis and analysis. Our larger aim is for our findings to lead to a more generalized solution for improving vision-language models for specialized medical problems.

2. Problem Statement

In "Vision Language Models for Medical Report Generation and Visual Question Answering", it is explained that current vision language models fall within two categories: contrastive and generative. In contrastive learning, the model is trained to embed both positive and negative pairs within a shared space. Positive pairs are composed of related visual and textual content, such as an image and its corresponding description. Negative pairs, on the other hand, consist of unrelated content, like an image paired with a randomly selected, different description. The aim is to bring the embeddings of positive pairs closer together while pushing those of negative pairs further apart within this shared space.[3]. In "Frozen Large-scale Pretrained Vision-Language Models are an Effective Foundational Backbone for Enhancing Multimodal Breast Cancer Risk Assessment", the usage of such a contrastive model is discussed. The cropped lesion image input is passed through a CLIP vision encoder , while the corresponding EHR data is fit into a template and passed through a CLIP language Encoder. The embeddings are concatenated and passed through a fusion model with 2 Fully Connected Layers and a Dropout layer with a rate of 70%. Lastly it is passed through a classifier[4]. In contrast to contrastive models, all trained generative models use encoded representations of the images and text as context for their following autoregressive generation. However even with the success of both approaches, VLM's particularly in the medical context have a variety of problems - both in their generalizing and specializing abilities. Some issues include hallucination, patient data privacy, catastrophic forgetting (when a model learning new information inadvertently erases previously acquired knowledge), and the classification of generative pretrained models lagging behind contrastive models.

3. Method and Dataset

To restate, our goal is to improve the ability of medical VLM's to specialize in certain problems. In this paper we will be primarily focusing on breast cancer diagnosis detection and analysis. We hope to see outcomes of this improved ability in the form of improved classification, note generation, and question answering. For our first benchmark, we will finetune Med-Flamingo, an finetuned version of Google's Flamingo Vision Language Model on medical tasks. More specifically, Med-Flamingo is a multimodal few-shot learner based on the Flamingo Architecture [6]. The Flamingo model utilizes a pre-trained frozen CLIP vision encoder for visual feature generation and then converts these visual features into tokens via a trained perceiver resampler. The tokens undergo cross attention with tokenized text inputs inside of a pre-trained frozen LLM LLaMa-7B [3]. We will measure classification accuracy, QA accuracy, and text generation accuracy. The datasets we will be using for this are VinDr-mammo, CBIS-DDSM and EMBED. VinDr-mammo consists of 5,000 four-view exams with breast-level assessment and finding annotations.

CBIS-DDSM consists of ROIs annotations on 1939 images from 1076 women, and EMBED consists of 60000 annotated lesions linked with imaging descriptors and pathological outcomes. We build our own database of breast-cancer related questions and answers from these datasets to use for our model question-answering. We then finetune OpenFlamingo, an open source implementation of Google's Flamingo, on the same datasets as our second benchmark, quantifying the same measures. After finetuning Open-Flamingo, we will make our improvements to the open-source architecture. [2]

For our first improvement, we will be targeting the issue of having the classification accuracy of a generative architecture like Flamingo lagging behind contrastive models. To solve this, we will be implementing a instance matching technique outlined in "Enhancing the medical foundation model with multi-scale and cross-modality feature learning". To incorporate instance matching into the Med-Flamingo model, we begin by extracting features using the existing components of the model. The CLIP vision encoder generates visual features, which are then converted into tokens via the perceiver resampler. Simultaneously, the LLaMA-7B model processes textual inputs through its cross-attention layers. These visual and textual features are subsequently fused, typically through concatenation, to form a comprehensive representation of the input pair. To enable instance matching, a binary classification head is introduced, consisting of two linear layers designed to predict whether the input pairs (image and text) are matched or not. The training process is enhanced by incorporating a binary cross-entropy loss function, specifically tailored to train this classifier. This classifier and the associated loss are then integrated into the existing training loop of the Med-Flamingo model. By ensuring that gradients flow correctly and optimizing both instance matching and generative tasks, the model leverages the strengths of contrastive learning. This enhancement improves the Med-Flamingo model's performance in complex medical tasks, such as breast cancer diagnosis and analysis, by creating a more robust and reliable feature distribution combining the power of contrastive learning with generative modeling.[1]

Secondly, we would like to address the issue of hallucinations and limited specialized medical information. We propose a RAG approach that allows the model to incorporate relevant information that could be outside of what it was trained on. To integrate the Retrieval-Augmented Generation (RAG) framework into the Med-Flamingo model for breast cancer diagnosis, we enhance the model's text generation process by incorporating relevant external medical information. First, we index specialized medical texts related to breast cancer in a vector database using FAISS. When the model receives an input image, the CLIP vision encoder extracts visual features, which are then projected through a projection layer specific to the task. These projected visual features are used to query the FAISS index, retrieving the most relevant medical texts based on similarity. The retrieved texts are then combined to form a contextual input, which is tokenized and fed into the LLaMA-7B model. The model uses a gated cross-attention mechanism to incorporate the visual tokens as context while processing the retrieved text. This setup ensures that the generated diagnostic report is informed by both the visual features of the input image and the retrieved relevant medical texts, reducing hallucinations and improving the accuracy of the specialized medical information in the output. This approach is inspired by "Contrastive X-ray-Report Pair Retrieval based Generation (CXR-RePaiR-Gen)". [5]

Lastly, if time permits, we will be addressing the issue of catastrophic forgetting. Vision-language models like Med-Flamingo often struggle with catastrophic forgetting when fine-tuned on new tasks, resulting in the loss of previously acquired knowledge. As described in "Learning without Forgetting for Vision-Language Models", while LoRA (Low-Rank Adaptation) introduces low-rank updates to model weights and freezes most parameters to address this issue, it doesn't fully exploit cross-modal interactions or effectively retain task-specific knowledge. PROOF (Projection Fusion for VLM) addresses these limitations by implementing projection layers that map pre-trained visual and textual features into a new space, creating expandable projections for new tasks while freezing old ones to preserve earlier learnings. Additionally, PROOF employs self-attention mechanisms to fuse visual and textual information, enhancing the model's context-aware predictions. [2] To implement PROOF in Med-Flamingo for the task of breast cancer diagnosis, we plan to freeze 70% of the model's layers, retaining the stability of the pre-trained knowledge while allowing 30% of the layers to remain trainable for adaptation to the new task. Specifically, we will apply projection layers to the visual tokens from the perceiver resampler and the textual embeddings from the LLaMA-7B model. For this single task, we introduce new projection layers for both visual and text features, ensuring that the previous layers remain frozen to retain learned knowledge. Finally, we use the projected visual tokens as context within the gated cross-attention layers of the LLaMA-7B model, enabling the model to integrate visual and textual information effectively for accurate breast cancer diagnosis. This approach allows Med-Flamingo to adapt to the specific task of breast cancer diagnosis without forgetting previously learned information, leveraging both stability and adaptability for enhanced performance.

Conclusively, we will not only be applying state of the art Vision Language models to breast cancer detection and diagnosis but also improving upon the general ability of vision language models to specialize on medical problems such as breast cancer.

4. Work So Far and Preliminary Results

4.1. CBIS-DDSM Pre-Processing

The CBIS-DDSM dataset contains two groups of information for patients recorded. This being patient calcification data and mass data. The calcification data contains the calcification type and distribution along with intersecting information from the mass data. The mass data contains laterality, breast density, image view, pathology, mass shape, mass margins, physician subtlety score, and a BI-RADS assessment score. This information was used to form multiple choice question, as shown in 4.1.

Table 1. Mammography Assessment Questions

Question
Which breast is this left or right?
What is breast density shown here?
What view was used when taking this mammogram?
What is the calcification type show here?
What is the calcification distribution shown here?
On a scale of 1 to 5, describe the subtlety of the abnormalities?
What is your BI-RADS assessment rating?

4.2. Evaluation

To create a baseline to compare with the methods we introduce in this paper, we use Med-flamingo and few-shot prompting. From the dataset of compiled images and corresponding questions, we picked 3 question-image pairs. We interleave two of the question-image pairs with their corresponding question, choices, and answers. For the final question we append the image, the question and the choices, but leave out the answers for Med-flamingo to complete.

4.3. Results

Overall, performance of Med-flamingo without fine tuning on the MCQ dataset was very poor. This is likely because of the lack of similar examples in the training set used to finetune Med-flamingo. We believe that by finetuning Med-flamingo on our own dataset with specific mammography question we can dramatically increase this score.

Table 2. Results for Different Models

Model	BERT Score	MCQ
Med-flamingo	-	0.267
Med-flamingo (finetuned)	-	-
Openflamingo (finetuned)	-	-
EFS-VLM (Ours)	-	-

References

[1] Enhancing the medical foundation model with multi-scale and cross-modality feature learning. 2024.

[2] Learning without forgetting for vision-language models. 2024.

[3] I. Hartsock and G. Rasool. Vision-language models for medical report generation and visual question answering: A review. 2024.

[4] K. K. W. Hung Q. Vo, Lin Wang. Frozen large-scale pretrained vision-language models are an effective foundational backbone for enhancing multimodal breast cancer risk assessment. 2024.

[5] V. K. A. Y. N. P. R. Mark Endo, Rayan Krishnan. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. 2024.

[6] M. Moor, Q. Huang, S. Wu, M. Yasunaga, C. Zakka, Y. Dalmia, E. P. Reis, P. Rajpurkar, and J. Leskovec. Med-flamingo: a multimodal medical few-shot learner. 2023.