Med-Idefics: A Two-Stage Fine-Tuning Approach for Enhanced Medical Visual Question Answering

Brendan P. Murphy

CS231N Spring 2024

`bigsur@stanford.edu`

Abstract

This paper investigates the effectiveness of a two-stage fine-tuning approach on the IDEFICS2 8B model for Med-VQA datasets. The model leverages the ROCO dataset for broad medical knowledge and trains on the VQA-RAD dataset for question-answering capabilities. Results show the model's ability to generate precise answers and identify medical nuances, outperforming single-stage and base models. The model demonstrates robustness through generalization to out-of-distribution data but performance deteriorates when prompted with unrelated medical images, indicating that it has successfully learned from the medical image fine-tuning, rather than simply memorizing patterns or features. A prompting strategy framing the model as an expert radiologist enhances accuracy. Qualitative analysis highlights the model's coherence and relevance in capturing medical concepts, while identifying areas for improvement. The findings emphasize the potential of fine-tuned multimodal models in assisting medical professionals. Future work aims to refine responses, incorporate additional datasets, and explore advanced prompting strategies.

1. Introduction

Rapid advancements in medical imaging technologies have revolutionized the field of radiology, enabling more accurate diagnoses and personalized treatment plans. However, the increasing complexity and volume of medical images pose significant challenges for radiologists in terms of interpretation and analysis. To address these challenges, there is a growing interest in developing intelligent systems that can automatically understand and interpret medical images, thereby assisting radiologists in their decision-making process. One promising approach is the application of multimodal large language models to the task of Medical Visual Question Answering (Med-VQA).

Med-VQA is a critical task that involves answering questions about medical images, requiring both visual understanding and medical domain knowledge. It has the potential to greatly enhance the efficiency and accuracy of radiological interpretation by providing radiologists with quick access to relevant information and insights. For example, a radiologist could ask the system, "What is the location and size of the tumor in this MRI scan?" and receive a precise answer, saving time and reducing the risk of overlooking important details.

Recent progress in multimodal large language models, such as Idefics2 8B, have demonstrated promising results in various vision-language tasks.[9] These models leverage the knowledge encoded in pre-trained language models and combine it with visual information to enable multimodal understanding and generation. However, their application to the medical domain, particularly Med-VQA, remains largely unexplored. This project aims to bridge this gap by fine-tuning the Idefics2 8B open-source model for Med-VQA tasks.

To achieve this goal, I employed a two-stage fine-tuning approach. In the first stage, I fine-tuned the model using the ROCO dataset (Radiology Objects in COntext (ROCO): A Multimodal Image Dataset).[17] This dataset contains 65k radiology images with corresponding captions, providing a rich source of multimodal medical data. Fine-tuning on this dataset allows the model to acquire a broad understanding of medical imagery and associated textual descriptions.

In the second stage, I further fine-tuned the model on the VQA-RAD dataset, focusing on the task of answering questions related to medical images.[8] The input to the model is a medical image and a corresponding question, and the output is a predicted answer. By fine-tuning on this dataset, the model learns to extract relevant visual features and combine them with its language understanding capabilities to generate accurate answers.

To evaluate the performance and generalizability of the fine-tuned model, I conducted experiments on two datasets. First, I tested the model on the VQA-RAD question-answer dataset used for fine-tuning, measuring its accuracy in answering questions specific to the training domain. Second, I performed an out-of-distribution test on the Path-VQA

dataset, which contains pathology images and associated questions.[4] This test assesses the model's ability to generalize to a different medical imaging modality and question types.

In addition to these evaluations, I conducted two ablation studies to further investigate the model's behavior. The first ablation study aimed to assess the model's reliance on visual information. I modified the existing evaluation on the VQA-RAD dataset to use a random medical image unrelated to the question, allowing me to observe how much the model depends on the image to generate an answer. This study provides insights into the model's ability to effectively utilize visual cues in the context of medical question answering.

The second ablation study focused on the model's susceptibility to hallucinations when presented with nonsensical questions. Using the Med-HALT dataset[15], which is designed to evaluate hallucinations in large language models (LLMs) in the medical domain, I attempted to prompt the model with nonsensical questions and analyze its responses. This study sheds light on the model's robustness and ability to handle out-of-distribution and potentially misleading inputs.

Furthermore, I explored a prompting strategy to enhance the model's accuracy. By prefixing each prompt with the statement, "You are an expert radiologist evaluating the case, answer the question succinctly based on the medical image," I aimed to provide additional context and guide the model towards more accurate and relevant answers. This prompting approach leverages the model's language understanding capabilities to improve its performance on the Med-VQA task.

The motivation behind this project stems from the increasing need for intelligent systems that can assist radiologists in interpreting complex medical images. By developing a multimodal model capable of answering questions about medical images, we can enhance the efficiency and accuracy of radiological interpretation, ultimately leading to improved patient care. Moreover, the ability to generalize to different medical imaging modalities and question types is crucial for the practical application of such models in real-world clinical settings.

In summary, this project explores the fine-tuning of the Idefics2 8B model for Med-VQA tasks using a two-stage approach. The input to the model is a medical image and a question, and the output is a predicted answer. By conducting experiments on both in-domain and out-of-distribution datasets, two ablation studies, and an enhanced prompting strategy, I aim to assess the model's performance, generalizability, and robustness. The successful development of such a model has the potential to greatly benefit the field of radiology and improve patient outcomes.

2. Related Work

The Medical Visual Question Answering (Med-VQA) field is advancing rapidly, with recent efforts like Li et al.'s (2023) LLaVA-MED focusing on fine-tuning existing models rather than building from scratch. LLaVA-MED uses a figure-caption dataset from PubMed Central and GPT-4, demonstrating enhanced multimodal conversational abilities and superior performance on biomedical VQA metrics through a curriculum learning method.[10] However, this approach relies heavily on the availability and quality of specific datasets, potentially limiting its broader applicability.

Another notable work in this field is MedFlamingo, introduced by Kuznia et al. (2023). [12] MedFlamingo is a multimodal few-shot learner adapted to the medical domain, based on the OpenFlamingo-9B model. By continuing pre-training on paired and interleaved medical image-text data from publications and textbooks, MedFlamingo enables few-shot generative medical visual question answering abilities. The model's performance is evaluated on several datasets, including a novel open-ended VQA dataset of visual USMLE-style problems, and through human evaluation by physicians. MedFlamingo demonstrates improvements of up to 20% in clinicians' ratings and enables multimodal medical few-shot adaptations, such as rationale generation.

Khorashadizadeh et al. (2023) reframe Med-VQA as a generative task, innovatively using learnable tokens from visual features to prompt pre-trained language models, optimized for small, domain-specific datasets [19]. Their parameter-efficient tuning strategy not only surpasses existing methods in various settings but also enhances computational efficiency on key medical VQA benchmarks like Slake, OVQA, and PathVQA, demonstrating both the strength of the approach in performance and its limitation in needing specialized training setups.

The emergence of large language models (LLMs) and multimodal large language models (MLLMs) has opened up new possibilities for adapting pre-trained knowledge to the medical domain. Liu et al. (2024) proposed PeFoMed, a parameter-efficient framework for fine-tuning MLLMs specifically tailored to Med-VQA applications.[11] By leveraging the knowledge encoded within language models and enhancing their applicability in multimodal contexts, PeFoMed achieves an overall accuracy of 81.9% on a public benchmark dataset and outperforms the GPT-4v model by a significant margin of 26% absolute accuracy on closed-ended questions.

The effectiveness of in-context learning compared to fine-tuning has also been a topic of interest in the research community. Nori et al. (2023) explored the specialist capabilities of GPT-4 on medical challenge benchmarks in the absence of special training.[13] Through sys-

tematic prompt engineering, they demonstrated that GPT-4 can easily top prior leading results for medical question-answering datasets without the need for expert-curated content. The authors introduced Medprompt, a composition of several prompting strategies that greatly enhances GPT-4's performance and achieves state-of-the-art results on all nine benchmark datasets in the MultiMedQA suite, outperforming specialist models such as Med-PaLM 2 by a large margin with fewer calls to the model.

He et al. (2020) introduce the PathVQA dataset, which aims to develop an "AI Pathologist" capable of passing the board-certified examination of the American Board of Pathology.[4] The authors address the challenges of creating a medical VQA dataset, such as limited access to pathology images and the need for expert annotations, by using a semi-automated pipeline to extract images and captions from textbooks and online libraries. While PathVQA represents a significant contribution to the field, the dataset may be limited in terms of the diversity and complexity of questions compared to those found in real-world examinations.

To assess LLMs for clinical use, the CRAFT-MD framework by Bansal et al. (2023) simulates interactions in a controlled environment to test LLMs like GPT-4 and GPT-3.5 on skin diseases.[6] This revealed limitations in conversational reasoning and diagnostic accuracy, leading to guidelines emphasizing realistic interactions and comprehensive evaluations.

Another important aspect of evaluating LLMs in the medical domain is the hallucination phenomenon, where models generate plausible but incorrect information. To address this issue, Wu et al. (2023) created a hallucination benchmark of medical images paired with question-answer sets and conducted a comprehensive evaluation of state-of-the-art models.[21] Their study provides an in-depth analysis of current models' limitations and reveals the effectiveness of various prompting strategies in mitigating the hallucination problem.

In the field of robotic surgery, Surgical GPT [18] proposes an end-to-end trainable Language-Vision GPT (LV-GPT) model that expands the GPT2 model to include vision input (image). The LV-GPT model incorporates a feature extractor (vision tokenizer) and vision token embedding (token type and pose) to exploit the advancements in GPT models for VQA in robotic surgery. The authors prove that the LV-GPT model outperforms other state-of-the-art VQA models on three surgical-VQA datasets and extensively study the effects of token sequencing, token type, and pose embedding for vision tokens in the LV-GPT model.

The Med-Gemini family of models[22], optimized for medical use via fine-tuning, sets new standards in AI-based medical diagnostics, including 2D and 3D radiology and histopathology, with Med-Gemini-2D surpassing previous best performances in CXR visual question answering. How-ever, despite its strengths, Med-Gemini-Polygenic, while surpassing standard polygenic risk score-based approaches and generalizing to untrained genetically correlated diseases, reflects an inherent limitation in extending beyond trained datasets.

This project aims to enhance the Idefics2 8B model for Med-VQA tasks using a two-stage approach, focusing on fine-tuning with ROCO and Med-VQA datasets to boost performance in medical image analysis. Additionally, I explored prompting strategies to improve accuracy and reduce hallucinations, advancing the use of LLMs and MLLMs in medical diagnostics.

3. Methods

To adapt the Idefics2 8B model to the medical domain and specifically to the task of Medical Visual Question Answering (Med-VQA), I employed a two-stage fine-tuning approach. Stage 1 utilized the ROCO dataset, which consists of 65,000 radiology images with corresponding captions, providing a rich source of multimodal medical data. Stage 2 focused on the VQA-RAD dataset, which comprises medical images accompanied by question-answer pairs, posing unique challenges due to its domain-specific terminology and visual information.

The model architecture includes a text model, modality projection layers, and a perceiver resampler. The text model processes the input text and generates output text by mapping the text data to a higher-dimensional feature space. The modality projection layers project visual features extracted from the input image to the same embedding space as the text, facilitating effective cross-modality integration. The perceiver resampler then aggregates these features to create a unified representation that the text model uses for generating responses.

Mathematically, the projection layers map the visual features v and text features t to a common embedding space as follows:

$$v' = W_v \cdot v + b_v$$
$$t' = W_t \cdot t + b_t$$

where W_v and b_v are the weights and bias for the visual projection, and W_t and b_t are the weights and bias for the text projection. The projected features v' and t' are then used in subsequent layers for further processing.

For fine-tuning, I employed an optimization technique based on the autoregressive language modeling objective. The model learns to predict the next token y_t given the past tokens $y_1, \ldots, y_{t-1}$ and the input image I, minimizing the loss function $\mathcal{L}$, defined as the negative log likelihood of the correct token:

$$\mathcal{L} = -\log p(y_t|y_1, \ldots, y_{t-1}, I)$$

where y_t is the target token at time step t, $y_1, \ldots, y_{t-1}$ are the previous tokens, and I is the input image.

This approach leverages the capabilities of LoRA to adapt specific layers of the model efficiently. I used LoRA to modify only the text model, modality projection, and perceiver resampler components of the model and specifically target the down, gate, up, key, value, query, and output projection layers to enhance the model's ability to handle domain-specific nuances effectively while preserving the pre-trained knowledge.

4. Data

For this project, I utilize four datasets: VQA-RAD [8], ROCO [17], PathVQA [4] and Med-HALT [15].

VQA-RAD is a widely used benchmark for medical visual question answering (Med-VQA) tasks. It consists of 315 medical images and 2,248 question-answer pairs, covering various aspects of radiology imagery such as modality, plane, organ system, and abnormalities. The questions are categorized into 11 types, including 'yes/no', 'what', 'where', and others. The dataset is divided into a training set with 2,248 question-answer pairs and a test set with 451 pairs. The images have a resolution of 512x512 pixels. No data preprocessing, normalization or augmentation was performed on this dataset.

To expand the training data and for domain adaptation, I also leverage the Radiology Objects in COntext (ROCO) 65k dataset. ROCO contains 65,000 radiology images from various imaging modalities including CT, MRI, ultrasound, and X-ray, and each image has a corresponding caption. The captions describe the key observations in the images. I use the image-caption pairs from ROCO to pretrain the base Idefics model on the task of generating relevant captions for given radiology images, before fine-tuning on the VQA-RAD dataset. I split the ROCO 65k dataset into a training set (55k) and validation set (5k).

I incorporate the Path-VQA dataset for an out-of-distribution test, the dataset contains 32,799 question-answer pairs generated from 4,998 pathology images. The questions cover various aspects of the images such as anatomical objects, colors, locations, and sizes.

I used a small random sample of questions from Med-HALT to prompt the model with nonsensical questions in an attempt to generate hallucinations or nonsensical answers.

For the ROCO and VQA-RAD datasets, the question and answer text is processed and formatted using the Auto-Processor configured for the 'HuggingFaceM4/idefics2-8b' model, ensuring that text is suitably tokenized and structured with image placeholders. The entire images are processed as single units without splitting, integrating seamlessly with the text data. These combined text and image inputs are then fed into the multimodal Idefics model to generate answers.

The combination of the VQA-RAD and ROCO datasets provides a large and diverse training set for building a robust medical VQA system. The datasets cover various imaging modalities, anatomical regions, and question types. Pre-training on ROCO allows the model to learn general visual-linguistic associations before adapting to the specific VQA task. Example images and question-answer pairs from the VQA-RAD dataset is shown in Figure 1.

5. Experiments

5.1. Experiment Setup

The fine-tuning was conducted on an NVIDIA A100 80GB GPU, and CUDA version 12.4[14], utilized the Hugging Face Transformers library, and took approximately 8 hours to complete [20]. I employed LoRA to minimize the number of trainable parameters[5]. I modified only the projection layers related to key, query, value, and other gating mechanisms within the text processing, modality projection, and perceiver resampling components of the model, allowing for efficient adaptation with minimal changes to the pre-trained model weights. I used a rank of 8, and a LoRA specific dropout of 0.1. The LoRA weights were initialized using a Gaussian distribution.

To further optimize the model for computational efficiency, I utilized Quantized Low-Rank Adaptation (QLoRA)[2] which configures the model to load in 4-bit precision using the BitsAndBytes library and reduces the memory footprint and computational demands[1].

For stage 1 fine-tuning on the ROCO dataset, I used 55,000 training examples and 5,000 evaluation examples. Each training instance was prepared using a custom data collator that processed images and associated captions into a format suitable for the transformer-based model. The training was run for 3 epochs with a batch size of 2 for training and 8 for evaluation. Gradient accumulation was configured at 8 steps to adjust for the smaller batch size, ensuring stable gradient updates.[7] A learning rate of 1e-5 with a weight decay of 0.01 was used, including a warm-up phase of 50 steps to gradually ramp up the learning rate at the beginning of training. The model's output and intermediary states were saved periodically to monitor progress and facilitate recovery from interruptions.

For stage 2 fine-tuning on the VQA-RAD dataset, I used the entire training set and test set. Each training instance was prepared using a custom data collator that processed images, questions, and answers into a format suitable for the transformer-based model. The training was run for 2 epochs with a batch size of 2 for training and 8 for evaluation, using the same gradient accumulation and warm-up settings as in stage 1. However, the learning rate was increased to 1e-4 to facilitate adaptation to the VQA-RAD dataset.
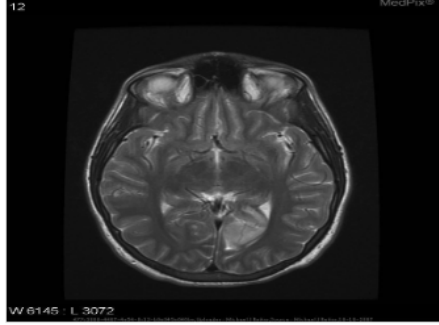
Visual Input	User Interaction
	Question: What hypoattenuated tissue is between the abdominal wall and skin? **True Answer:** Fat **Predicted Answer:** Subcutaneous fat
	Question: What type of MRI sequence is displayed in this image? **True Answer:** T2 weighted MRI **Predicted Answer:** T2 weighted

Figure 1. Example medical visual input and question answering capability of Idefics2 on the VQA-RAD dataset.

5.2. Out-of-Distribution Evaluation

To assess the model's generalizability to different medical imaging modalities and question types, I conducted an out-of-distribution (OOD) evaluation using the Path-VQA dataset. This dataset consists of full-color pathology images and associated questions, which differ from the radiology images and questions in the VQA-RAD dataset used for fine-tuning. I applied the fine-tuned model to the Path-VQA dataset without any additional training or adaptation, and measured its performance using the same metrics as in the VQA-RAD evaluation, including exact match accuracy, F1 score, and BLEU score. This evaluation provides insights into the model's ability to handle unseen medical image types and question formats.

5.3. Ablation Studies

I conducted two ablation studies to investigate specific aspects of the model's behavior. The first study focused on the model's reliance on visual information. I modified the evaluation script for the VQA-RAD dataset to randomly replace the medical image associated with each question-answer pair with an unrelated medical image from the same dataset. By comparing the model's performance on this modified evaluation set with its performance on the original VQA-RAD evaluation set, I can quantify the extent to which the model relies on the visual content of the image

to generate accurate answers. This study helps to identify potential biases or weaknesses in the model's multimodal reasoning capabilities.

The second ablation study, which focused on the model's susceptibility to hallucinations when presented with nonsensical questions, revealed an interesting example of the model's behavior. When prompted with a question from the Med-HALT dataset, such as 'Far beyond our known universe, in the mysterious and wondrous planet of Gorgons, where liver cancer claims countless lives of their benevolent creatures, which of the following preposterously absurd and nonsensical measures cannot serve as a surgical resection of poor prognostic factors for their liver cancer?', the model generated the response 'exorcism'. This example highlights the model's tendency to generate seemingly plausible but irrelevant answers when faced with nonsensical or out-of-distribution inputs. While the model's response is coherent in the context of the question, it demonstrates the challenges in handling such scenarios and the need for further research to improve the model's robustness against hallucinations.

5.4. Prompting Strategy

To explore the impact of prompting on the model's performance, I implemented a prompting strategy that frames the question-answering task in the context of an expert radiologist. Specifically, I prepended each question with the prompt, "You are an expert radiologist evaluating the case,

5

answer the question succinctly based on the medical image." This prompt aims to provide additional context and guidance to the model, potentially improving its accuracy and relevance. I evaluated the model's performance on the VQA-RAD dataset with and without this prompting strategy, and compared the results to assess its effectiveness. This experiment sheds light on the potential benefits of domain-specific prompting for medical visual question answering tasks.

5.5. Evaluation Metrics

To assess the performance of the fine-tuned model on the VQA-RAD test set, I employed a combination of quantitative and qualitative metrics. Quantitatively, the model's correctness in generating answers to both open-ended and closed-ended questions was evaluated using three established metrics: Exact Match, F1-score, and Bleu score. These metrics were selected to provide a comprehensive analysis of the model's accuracy from different perspectives:

- **Exact Match** evaluates whether the predicted answers exactly match the ground truth, reflecting the model's precision in generating verbatim responses. The Exact Match accuracy was determined by the proportion of instances where the true answer was entirely contained within the predicted response. To ensure robustness in my evaluation, I extracted the precise portion of the predicted text that responded directly to the query, stripping auxiliary text to focus purely on the content relevant to the answer. This was necessary as the model occasionally responded with an answer and follow up questions due to being finetuned for multi-turn conversation.

- **Bleu Score** measures the linguistic quality of the generated text, gauging how natural the responses are by comparing them to typical human answers[16]. The Bleu score was computed as a unigram comparison, prioritizing the correct sequence of words in shorter responses.

- **F1-score** assesses the overlap of tokenized predicted and true answers, offering insights into the model's ability to retrieve relevant information while minimizing irrelevant details. For F1-score calculation, tokens from both the predicted and true answers were compared to identify common elements, allowing us to compute precision and recall values for each response.

- **Human Evaluation** To assess the model's performance, I focused on answers that required human judgment, excluding simple responses such as "yes" or "no." A human evaluator reviewed the model's answers for coherence, relevance, and medical accuracy.

The evaluator was presented with pairs of images and questions alongside the corresponding ground truth and model-predicted answers. They then determined the correctness of the model's predictions based on their professional judgment and understanding of the dataset.

5.6. Results

The performance of the Idefics2-8B model and its variants on the VQA-RAD and out-of-distribution datasets is summarized in Figure 2. The base Idefics2-8B model, without any fine-tuning, achieves an Exact Match accuracy of .55, F1 score of .38, and Bleu score of .29 on the VQA-RAD dataset. This serves as a baseline for comparison with the fine-tuned models.

The single-stage model, fine-tuned only on the VQA-RAD dataset, shows a slight improvement over the base model, with an Exact Match accuracy of .54, F1 score of .56, and Bleu score of .54. However, the two-stage model, which undergoes pretraining on the ROCO dataset followed by fine-tuning on VQA-RAD, demonstrates superior performance across all metrics. It achieves an Exact Match accuracy of .56, F1 score of .58, and Bleu score of .57, indicating the benefits of pretraining on a larger, diverse dataset before task-specific fine-tuning.

To evaluate the model's generalizability, the two-stage model was tested on the out-of-distribution Path VQA dataset. Despite the differences in image modality and question types, the model achieves an Exact Match accuracy of .30, F1 score of .32, and Bleu score of .31. While these scores are lower compared to the performance on VQA-RAD, they demonstrate the model's ability to transfer knowledge to a different medical domain.

The ablation study, where the two-stage model was evaluated on the VQA-RAD dataset with randomly swapped images, reveals the model's reliance on visual information. The Exact Match accuracy drops to .48, F1 score to .50, and Bleu score to .49 when the images are unrelated to the questions. This suggests that the model effectively utilizes the visual content to generate accurate answers, and its performance deteriorates when the visual cues are misaligned with the questions.

Incorporating the prompting strategy, where the model is prompted as an expert radiologist, leads to further improvements in performance on the VQA-RAD dataset. The Exact Match accuracy increases to .57, F1 score to .59, and Bleu score to .57. This indicates that providing domain-specific context through prompting can enhance the model's accuracy and relevance in answering medical visual questions.

Human evaluation of the two-stage model's answers on the VQA-RAD dataset yields a score of 63%, confirming its ability to generate coherent, relevant, and medically accurate responses. This qualitative assessment aligns with the

Model	Method	Dataset	Exact Match	F1 Score	Bleu Score	Human Eval
IDEFICS2 (base model)	N/A	VQA-RAD	.55	.38	.29	59%
IDEFICS2 (single-stage)	Finetuning only	VQA-RAD	.53	.55	.54	N/A
IDEFICS2 (two-stage)	Pretraining & Finetuning	ROCO & VQA-RAD	**.56**	**.58**	**.57**	63%
IDEFICS2 (base model)	OOD data	Path VQA	.29	.20	.15	N/A
IDEFICS2 (two-stage)	OOD data	Path VQA	.30	.32	.31	N/A
IDEFICS2 (two-stage)	Ablation study	VQA-RAD(swapped)	.48	.50	.49	N/A
IDEFICS2 (two-stage)	Prompt strategy	VQA-RAD	**.57**	**.59**	**.57**	N/A

Figure 2. Comparative performance of Idefics2-8B variants on VQA-RAD and out-of-distribution datasets. The two-stage model, pretrained on ROCO and fine-tuned on VQA-RAD, outperforms the single-stage model fine-tuned only on VQA-RAD. The two-stage model's performance is further evaluated on the OOD Path-VQA dataset, and its reliance on visual information is tested through an ablation study using unrelated medical images. Prompting the two-stage model as an expert radiologist yields the highest accuracy across all metrics, including human evaluation.

quantitative metrics and highlights the model's potential for real-world application.

Figure 3 illustrates the impact of pretraining data size on the model's performance. As the amount of ROCO pretraining data increases from 7K to 55K samples, the model's accuracy on VQA-RAD improves, with the most significant gains observed up to 20K samples. Beyond this point, performance plateaus, suggesting that the benefits of pretraining saturate at a certain data threshold. The task-specific VQA-RAD fine-tuning contributes significantly to the model's performance, while ROCO pretraining provides additional gains. The Exact Match accuracy remains relatively stable across all models, likely due to the presence of yes/no questions that are challenging for the base model to answer correctly.

Overall, the results demonstrate the effectiveness of the two-stage fine-tuning approach, combining pretraining on a large, diverse dataset (ROCO) with task-specific fine-tuning (VQA-RAD). The model's ability to generalize to out-of-distribution data, its reliance on visual information, and the benefits of domain-specific prompting are also highlighted. These findings underscore the potential of the Idefics2-8B model for medical visual question answering tasks and provide insights into strategies for further improvement.

5.7. Qualitative Analysis

In addition to the quantitative results, I conducted a qualitative analysis of the model's outputs by visually inspecting the generated answers and their corresponding images. The analysis revealed that the fine-tuned model produces more coherent and relevant answers compared to the base model. The model is able to capture key medical concepts and terminology, demonstrating its adaptation to the medical domain.

However, the model still struggles with certain challenging cases, such as complex anatomical structures or rare pathologies. In some instances, the model generates plausible but incorrect answers, highlighting the need for further improvements in handling ambiguous or visually similar cases.

It is important to note that the automated metrics used in this study, such as Exact Match, Bleu score, and F1 score, do not fully capture the nuances of the model's performance. There are edge cases where the model generates a correct answer, but it is not considered an exact match due to the specific phrasing of the answer and the technical implementation of the "Exact Match" metric. The "Exact Match" metric checks if the true answer contains the predicted answer verbatim, which may not always be necessary for the answer to be considered sufficient or correct.

To address these limitations, I incorporated a human evaluation component in the qualitative analysis. A medical expert reviewed a subset of the model's predictions, assessing their correctness based on professional judgment and understanding of the dataset. The human evaluation score provides a more comprehensive assessment of the model's performance, taking into account the semantic similarity and clinical relevance of the generated answers.

The human evaluation score on the VQA-RAD dataset indicates that the model is capable of generating accurate and meaningful answers in a majority of cases, even when the automated metrics may not fully reflect its performance. This highlights the importance of considering both quantitative and qualitative measures when evaluating the effectiveness of medical visual question answering models.

Despite the limitations of the automated metrics, they still provide valuable insights into the model's performance and serve as useful benchmarks for comparison with other approaches. However, the qualitative analysis, including human evaluation, offers a more nuanced understanding of the model's strengths and weaknesses, guiding future research and development efforts in this domain.

5.8. Discussion

The experimental results demonstrate the effectiveness of the two-stage fine-tuning approach in adapting the IDEFICS2 8B model to the medical domain for visual question answering tasks. The significant improvements in performance metrics across both the VQA-RAD and PATH-VQA datasets indicate that the model has learned to better understand and generate relevant answers to medical questions.

The use of LoRA and QLoRA techniques proves to be beneficial in efficiently adapting the model while minimizing the number of trainable parameters and reducing computational demands. The targeted adaptation of critical modules allows for effective knowledge transfer from the pre-trained model to the medical domain.

The ablation studies provide valuable insights into the model's behavior and reliance on visual information. The performance drop observed when the model is presented with unrelated medical images highlights the importance of visual cues in generating accurate answers. This finding underscores the need for the model to effectively integrate both textual and visual information to produce reliable responses.

The prompting strategy, framing the model as an expert radiologist, demonstrates the potential for domain-specific prompting to enhance the model's accuracy and relevance in answering medical questions. This approach leverages the model's language understanding capabilities to guide it towards more accurate and clinically relevant answers.

The qualitative analysis reveals that the fine-tuned model produces more coherent and medically relevant answers compared to the base model. However, the presence of challenging cases and occasional incorrect answers suggests that there is still room for improvement in handling complex medical scenarios. The human evaluation component provides a more comprehensive assessment of the model's performance, taking into account the semantic similarity and clinical relevance of the generated answers.

One potential limitation of this study is the relatively small size of the fine-tuning datasets, particularly for the VQA-RAD dataset. Increasing the amount of training data could potentially lead to further improvements in the model's performance and generalization capabilities.

Overall, the results of this study demonstrate the potential of fine-tuning large multimodal language models like IDEFICS2 8B for medical visual question answering tasks. The proposed two-stage fine-tuning approach, combined with efficient adaptation techniques and domain-specific prompting, provides a promising direction for developing intelligent systems that can assist medical professionals in interpreting and analyzing medical images.

6. Conclusion

In conclusion, this research demonstrates the effectiveness of a two-stage fine-tuning approach for adapting the IDEFICS2 8B model to the medical domain, specifically for visual question answering tasks. The model's performance improvements across various metrics and datasets highlight its ability to generate precise answers and identify subtle medical nuances.

The ablation studies and prompting strategy experiments provide valuable insights into the model's reliance on visual information and the potential for domain-specific prompting to enhance its accuracy and relevance. The qualitative analysis, including human evaluation, offers a more comprehensive understanding of the model's strengths and weaknesses, guiding future research and development efforts.

The results underscore the potential of fine-tuned multimodal models in assisting medical professionals and contributing to better patient outcomes. However, the presence of challenging cases and occasional incorrect answers indicates the need for further improvements in handling complex medical scenarios.

Future work should explore the incorporation of additional medical datasets and knowledge sources to enhance the model's understanding of medical concepts and terminology. Investigating more advanced prompting strategies and few-shot learning techniques could help the model adapt to new medical imaging modalities and question types with limited training data.

Moreover, efforts should be made to refine the model's responses, reducing the occurrence of plausible but incorrect answers and improving its reliability in real-world clinical settings. Collaborative efforts between researchers, medical professionals, and domain experts will be crucial in addressing these challenges and ensuring the successful deployment of such models in healthcare.

In summary, this study demonstrates the potential of the two-stage fine-tuning approach for adapting large multimodal language models to the medical domain, paving the way for the development of intelligent systems that can assist medical professionals in interpreting and analyzing medical images. With continued research and development, these models have the potential to revolutionize the field of medical visual question answering and ultimately improve patient care.

7. Appendices

7.1. External Collaborator

I had one volunteer evaluator, Agustina Saenz (MD, MPH) who is a practicing MD and post graduate research at the Rajpurkar Lab at Harvard Medical School as well as a mentor in the Stanford & Harvard Medical AI bootcamp.
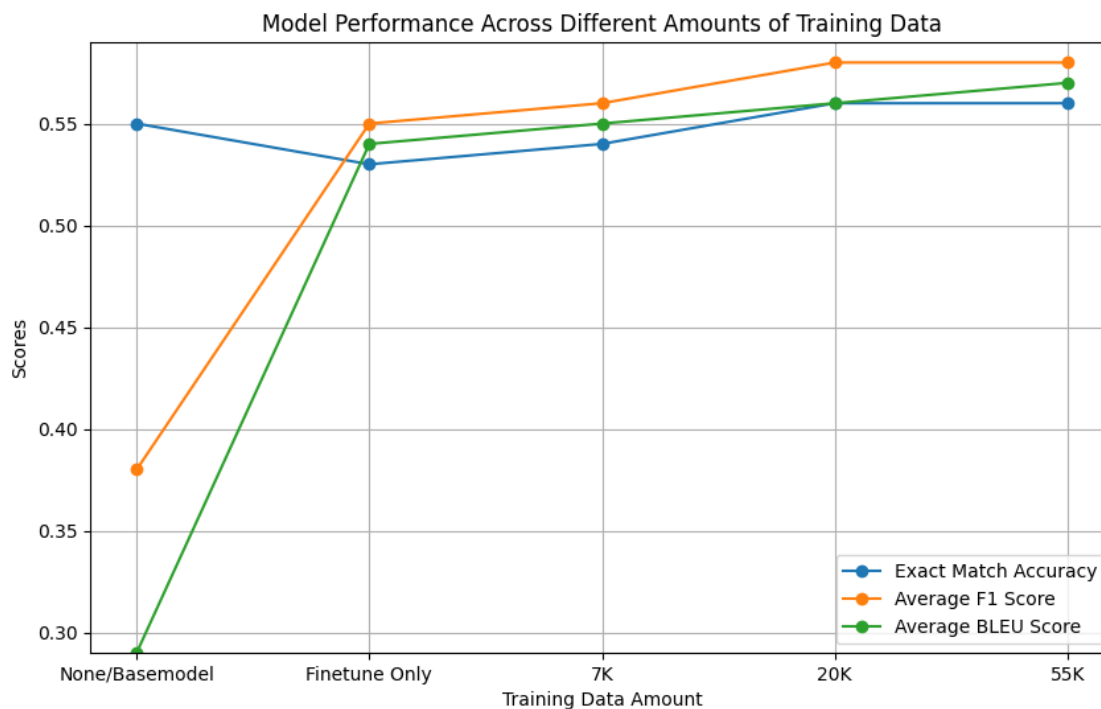
Figure 3. Model performance on the VQA-RAD dataset improves with increasing amounts of pretraining data from the ROCO dataset (stage 1), followed by fine-tuning on VQA-RAD (stage 2). The base model and fine-tune only model, which do not utilize ROCO pretraining, show the lowest performance. As the amount of ROCO pretraining data increases from 7K to 55K samples, performance improves but mostly plateaus after 20K samples. The task-specific VQA-RAD fine-tuning contributes significantly to the model's performance, while ROCO pretraining provides additional gains. The Exact Match Accuracy remains around 50% across all models due to the metric's sensitivity to yes/no questions, which the base model answers randomly.
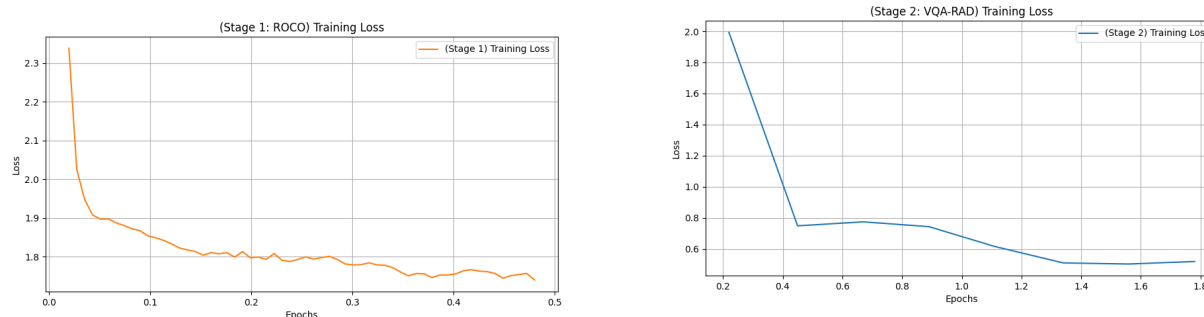

Figure 4. Comparison of finetuning results on medical image captioning (ROCO dataset, left) and visual question answering (VQA-RAD dataset, right). Each subplot shows the convergence and effectiveness of task-specific training.

7.2. Code

The code for fine-tuning was adapted and inspired from the Hugging Face notebook, "IDEFICS: Finetuning Demo notebook" [3].

The code is available on my GitHub
https://github.com/csbrendan/CS231N

References

[1] T. Dettmers. Bitsandbytes: Highly optimized mixed precision training in pytorch. https://github.com/TimDettmers/bitsandbytes, 2021. Accessed: [insert date here].

[2] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer.

Qlora: Efficient finetuning of quantized llms, 2023.

[3] H. Face. Hugging face notebook: Finetune image captioning with parameter-efficient fine-tuning. GitHub, 2024. https://github.com/huggingface/notebooks/blob/main/examples/idefics/finetune_image_captioning_peft.ipynb.

[4] X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie. Pathvqa: 30000+ questions for medical visual question answering, 2020.

[5] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models, 2021.

[6] S. Johri, J. Jeong, B. A. Tran, D. I. Schlessinger, S. Wongvibulsin, Z. R. Cai, R. Daneshjou, and P. Rajpurkar. Guidelines for rigorous evaluation of clinical llms for conversational reasoning, 2024.

[7] J. Lamy-Poirier. Layered gradient accumulation and modular pipeline parallelism: fast and efficient training of large language models, 2021.

[8] J. J. Lau, S. Gayen, A. B. Abacha, and D. Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images, 2018.

[9] H. Laurençon, L. Saulnier, L. Tronchon, S. Bekman, A. Singh, A. Lozhkov, T. Wang, S. Karamcheti, A. M. Rush, D. Kiela, M. Cord, and V. Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023.

[10] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day, 2023.

[11] G. Liu, J. He, P. Li, G. He, Z. Chen, and S. Zhong. Pefomed: Parameter efficient fine-tuning of multimodal large language models for medical imaging, 2024.

[12] 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.

[13] H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu, R. Luo, S. M. McKinney, R. O. Ness, H. Poon, T. Qin, N. Usuyama, C. White, and E. Horvitz. Can generalist foundation models outcompete special-purpose tuning? case study in medicine, 2023.

[14] NVIDIA, P. Vingelmann, and F. H. Fitzek. Cuda, release: 12, 2024.

[15] A. Pal, L. K. Umapathi, and M. Sankarasubbu. Med-halt: Medical domain hallucination test for large language models, 2023.

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[17] O. Pelka, S. Koitka, J. Rückert, F. Nensa, and C. M. Friedrich. Radiology objects in context (roco): A multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Springer International Publishing, 2018.

[18] L. Seenivasan, M. Islam, G. Kannan, and H. Ren. Surgicalgpt: End-to-end language-vision gpt for visual question answering in surgery, 2023.

[19] T. van Sonsbeek, M. M. Derakhshani, I. Najdenkoska, C. G. M. Snoek, and M. Worring. Open-ended medical visual question answering through prefix tuning of language models, 2023.

[20] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

[21] J. Wu, Y. Kim, and H. Wu. Hallucination benchmark in medical visual question answering, 2024.

[22] L. Yang, S. Xu, A. Sellergren, T. Kohlberger, Y. Zhou, I. Ktena, A. Kiraly, F. Ahmed, F. Hormozdiari, T. Jaroensri, E. Wang, E. Wulczyn, F. Jamil, T. Guidroz, C. Lau, S. Qiao, Y. Liu, A. Goel, K. Park, A. Agharwal, N. George, Y. Wang, R. Tanno, D. G. T. Barrett, W.-H. Weng, S. S. Mahdavi, K. Saab, T. Tu, S. R. Kalidindi, M. Etemadi, J. Cuadros, G. Sorensen, Y. Matias, K. Chou, G. Corrado, J. Barral, S. Shetty, D. Fleet, S. M. A. Eslami, D. Tse, S. Prabhakara, C. McLean, D. Steiner, R. Pilgrim, C. Kelly, S. Azizi, and D. Golden. Advancing multimodal medical capabilities of gemini, 2024.