



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

Brendan P. Murphy
bigsur@stanford.edu

Introduction

Introduction: This research explores the application of multimodal large language models to the task of Medical Visual Question Answering (Med-VQA). By fine-tuning the IDEFICS2 8B model using a two-stage approach, leveraging the ROCO and VQA-RAD datasets, I aim to develop an intelligent system capable of accurately interpreting medical images and answering related questions [11] [10] [19]. The study investigates the model's performance, generalizability, and robustness through various evaluations and ablation studies, while also exploring prompting strategies to enhance accuracy. The findings highlight the potential of fine-tuned multimodal models in assisting radiologists and improving the efficiency of medical decision-making processes.

Datasets

ROCO

- Utilized for domain adaptation and pretraining the base Idefics model on generating relevant captions
- Contains 65,000 radiology images from various modalities
- Each image has a corresponding caption describing key observations

VQA-RAD

- A benchmark dataset for Med-VQA tasks
- Contains 315 medical images and 2,248 question-answer pairs
- Divided into a training set (2,248 pairs) and a test set (451 pairs)

Path-VQA

- Incorporated for an out-of-distribution test [6]
- Questions cover various aspects of the images, such as anatomical objects, colors, locations, and sizes

Med-HALT

- Aimed at generating hallucinations to assess the model's robustness [17]
- A small random sample of questions used to prompt the model with nonsensical questions


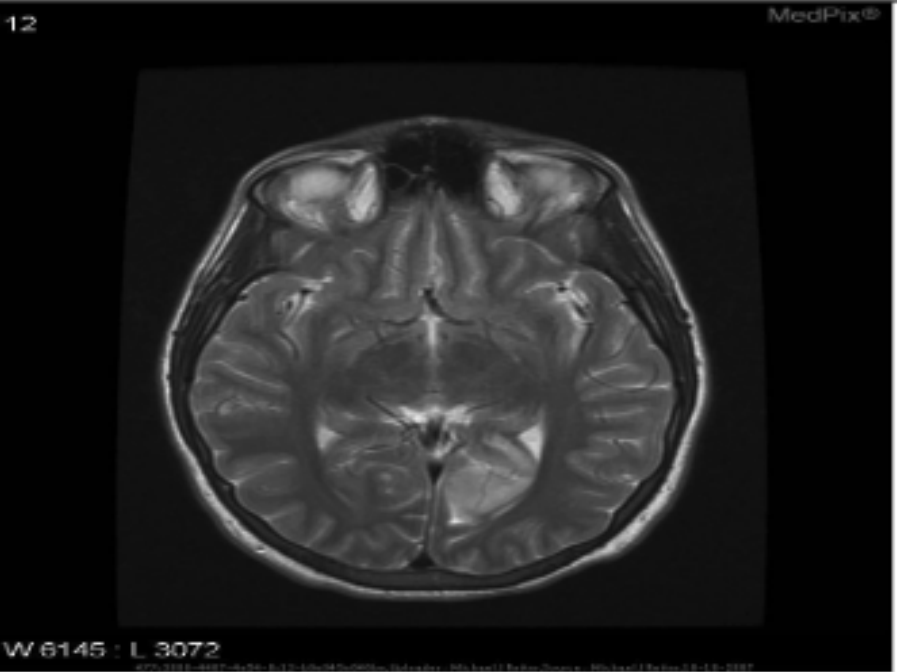
| Visual Input | User Interaction |
|---|--|
|  | <p>Question: What hypoattenuated tissue is between the abdominal wall and skin?</p> <p>True Answer: Fat</p> <p>Predicted Answer: Subcutaneous fat</p> |
|  | <p>Question: What type of MRI sequence is displayed in this image?</p> <p>True Answer: T2 weighted MRI</p> <p>Predicted Answer: T2 weighted</p> |

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

Methods

Model architecture

- Text model:** Processes input text and generates output text by mapping data to a higher-dimensional feature space.
- Modality projection layers:** Project visual features to the same embedding space as text for cross-modality integration.
- Perceiver resampler:** Aggregates features to create a unified representation for generating responses.

Two-stage fine-tuning approach

- Utilizes the ROCO dataset (65,000 radiology images with captions) for domain adaptation..
- Focuses on the VQA-RAD dataset (medical images with question-answer pairs) for task-specific training.

Fine-tuning optimization

- Employs the autoregressive language modeling objective to predict the next token based on past tokens and the input image.
- Utilizes LoRA to efficiently adapt specific layers (text model, modality projection, and perceiver resampler components) while preserving pre-trained knowledge.

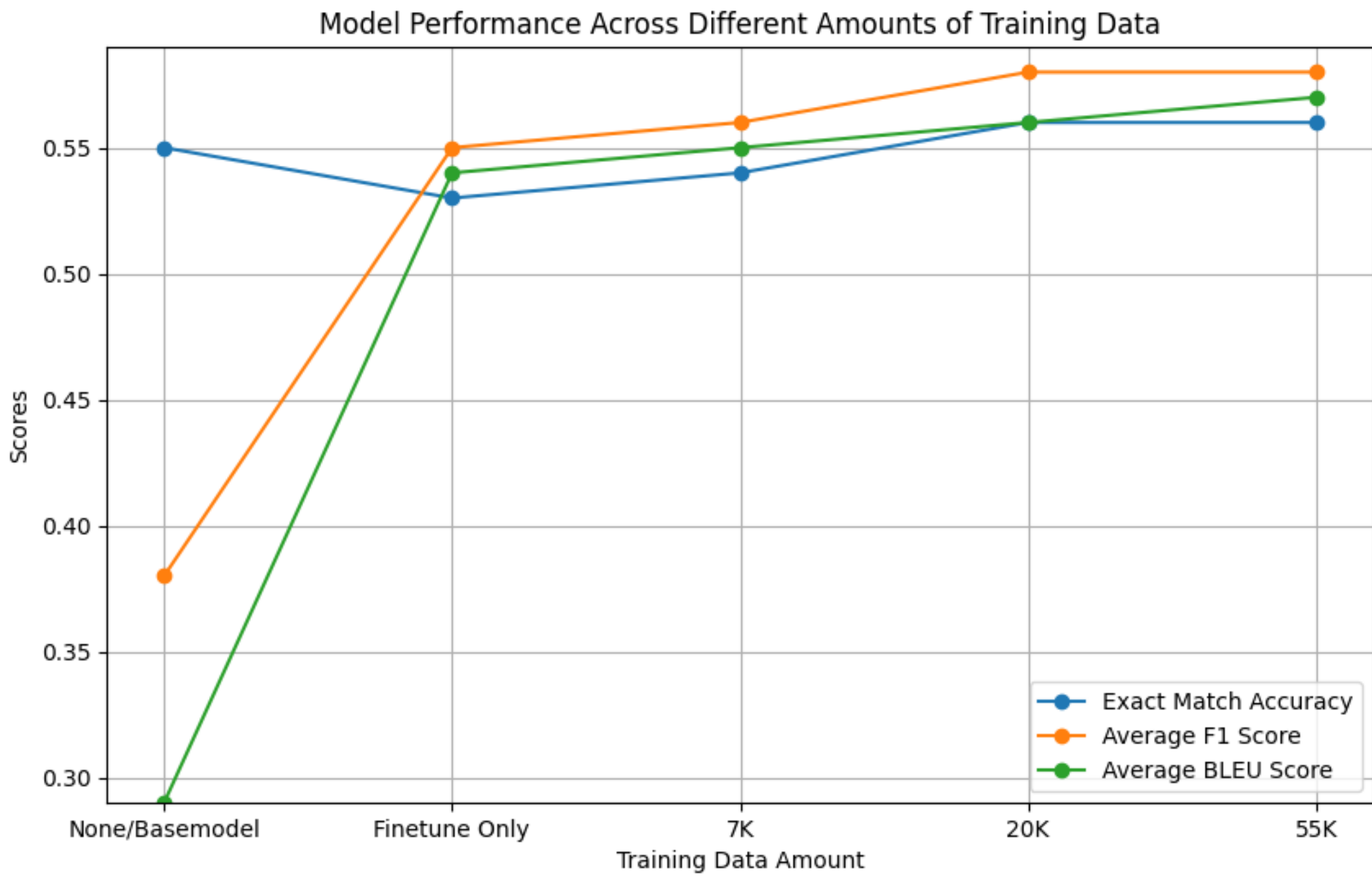


Figure 2. Model performance on the VQA-RAD dataset improves with increasing amounts of pretraining data from the ROCO dataset (stage 1), followed by fine-tuning on VQA-RAD (stage 2). The base model and fine-tune only model, which do not utilize ROCO pretraining, show the lowest performance. As the amount 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.

Evaluation

- Exact Match:** Measures the model's precision in generating verbatim responses by comparing predicted answers to the ground truth.
- BLEU Score:** Assesses the linguistic quality and naturalness of the generated responses by comparing them to typical human answers.
- F1 Score:** Evaluates the overlap between tokenized predicted and true answers, providing insights into the model's ability to retrieve relevant information while minimizing irrelevant details.
- Human Evaluation:** A human evaluator assesses the model's answers for coherence, relevance, and medical accuracy, focusing on responses that require professional judgment.

Results

| 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 3. 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.

Conclusion

- The two-stage fine-tuning approach effectively adapts the IDEFICS2 8B model to the medical domain, enabling precise answers and identification of subtle medical nuances in visual question answering tasks.
- Ablation studies and prompting experiments provide insights into the model's reliance on visual information and the potential of domain-specific prompting to enhance accuracy and relevance.
- Qualitative analysis, including human evaluation, offers a comprehensive understanding of the model's strengths and weaknesses, guiding future research and development efforts.
- Results highlight the potential of fine-tuned multimodal models in assisting medical professionals, but challenging cases and occasional incorrect answers indicate the need for further improvements.
- Future work should explore additional medical datasets, advanced prompting strategies, and few-shot learning techniques to enhance the model's understanding and adaptability to new medical imaging modalities and question types.

References

- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.
- Tim Dettmers.