MorphVLM: Towards more Efficient and Robust Multimodal Vision Language Models

Chirag Khatri
ckhatri@usc.edu

Mihir Mangesh Pavuskar
pavuksar@usc.edu

Pothula Punith Krishna
pothulap@usc.edu

Prince Verma
princeve@usc.edu

Lavrenti Mikaelyan
mikaelya@usc.edu

Abstract

In the rapidly evolving realm of multimodal language models, the need for efficient fine-tuning methodologies becomes increasingly vital. This work concentrates on refining the utilization of pre-trained multimodal vision language models, specifically optimizing their performance under hardware resource constraints. Acknowledging that harnessing powerful models often requires substantial computational resources and extensive datasets, our objective is to democratize the fine-tuning process, ensuring accessibility and impact across diverse applications. Through targeted experiments in Visual Question Answering (VQAv2) and medical domain tasks using A-OKVQA and PubMedQA datasets, we navigate the delicate balance between performance and resource efficiency. Leveraging the OpenFlamingo framework, our work(F23, 2023) (F23, 2023) explores the potential of large pre-trained Visual Language Models (VLMs) through component substitution and domain adaptation experiments.

1 Introduction

In the landscape of multimodal language models, the quest for efficient fine-tuning methodologies becomes increasingly imperative. Our work centers on the enhancement of utilizing pre-trained multimodal vision language models, with a distinct focus on optimizing their performance under hardware resource constraints. The backdrop of this exploration lies in the recognition that leveraging powerful models often demands substantial computational resources and extensive datasets. Our aim, therefore, is to democratize the fine-tuning process, making it accessible for diverse applications. Through targeted experiments in Visual Question Answering (VQA)(Goyal et al., 2017) and medical domain tasks using the A-OKVQA(Schwenk et al., 2022) and PubMedQA(Jin et al., 2019) datasets, respectively, we navigate the balance between performance and resource efficiency. This work builds upon one of the popular VLM frameworks, Flamingo(Alayrac et al., 2022), utilizing its open-source implementations.

We explore the potential of leveraging a large pre-trained VLM through the following experiments:

Component Substitution: We will investigate the impact of replacing the VLM's pre-trained vision and language components with alternative pre-trained uni-modal constituents. This will provide insights into the contribution of individual components and how the performance changes as the component models are changed

Domain Adaptation: We will assess the VLM's adaptability by fine-tuning it on domain-specific datasets, and comparing its performance with other relevant state-of-the-art models.

Since the Flamingo implementation and the training dataset used are not publicly available, we utilized its open-source implementation OpenFlamingo (Awadalla et al., 2023). Along with OpenFlamingo, we also use the MultiModal-GPT(Gong et al., 2023) which builds upon it, with joint instruction-tuning on the language and vision models optimizing on dialogue based interaction.

2 Approach

Our approach aims to boost the Visual Language models in two ways. First, we tweak its inner workings by swapping out the Language Model with alternatives trained on specific domains, helping us understand how vision and language play a role in it's performance. At the same time, we test how well the VLM adapts by fine-tuning it on specific datasets and comparing it with other top models.

2.1 Domain-Specific Language Modeling Module (Expert LM) Selection

First, we select a Large Language Modeling (LLM) trained on the data from a specific area or domain.

By choosing an LLM that is already pre-trained and performs well on benchmark datasets, we capitalize on the existing domain expertise. This reduces the need to start from scratch and significantly accelerates the fine-tuning process.

2.2 Integration with Visual Language Model

The selected domain-specific LLM is then integrated into the VLM model. By replacing an LLM trained generally on a wide corpus with the domain-specific LLM, we integrate domain knowledge directly into the multimodal context of the Visual Language Model.

2.3 Introduction of Trainable LoRA Blocks

A critical enhancement to our approach involves the incorporation of Low-Rank Adaption (LoRA) (Hu et al., 2021) blocks. Low-Rank Adaptation is a parameter-efficient fine-tuning method for large language models. It works by approximating the updates to the LLM's weights with a low-rank matrix, which significantly reduces the number of trainable parameters, leading to faster training times as compared to traditional fine-tuning methods. The LoRA blocks are strategically added at the end of each Attention block and Feed Forward blocks in the language encoder within the OpenFlamingo architecture as shown in Figure 1.

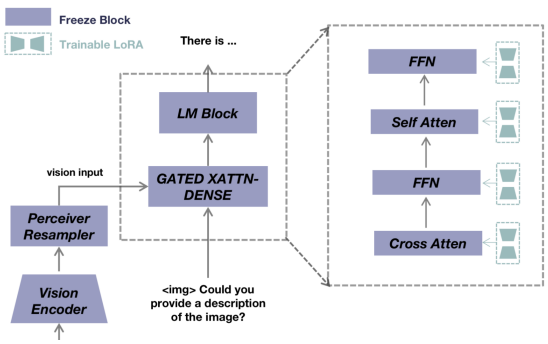

Figure 1: Trainable LoRA blocks added to the existing OpenFlamingo architechture

2.4 Fine-Tune Only the LoRA Trainable Blocks

Specifically, we freeze the parameters of the VLM model, with the exception of the newly introduced trainable LoRA blocks. This preserves the domain-specific knowledge encoded in the LLM, preventing it from being overwritten during fine-tuning. By narrowing the fine-tuning efforts to these blocks, we optimize computational resources and time. We choose several datasets for fine-tuning as described in the following section.

3 Experiments

3.1 Datasets

A-OKVQA (Augmented Outside Knowledge Visual Question Answering): Questions in A-OKVQA are challenging, conceptually diverse, require knowledge outside of the image, and differ from existing knowledge-based visual question answering datasets, in the sense that they cannot be answered by simply querying a knowledge base.
PubMedQA Dataset for Medical domain specific Fine-Tuning: The PubMedQA dataset serves as a specialized resource for fine-tuning machine learning models in the medical domain. It is curated from a collection of 19717 scientific publications from PubMedQA pertaining to diabetes.
VQAv2 Dataset for Model Evaluation: The VQAv2 dataset offers a comprehensive and diverse set of challenges for evaluating the performance of multimodal models. In our experiments, the VQAv2 Validation dataset serves as the evaluation benchmark for the Flamingo model.

3.2 Experiment 0: Baseline Model Evaluation on VQAv2

In this initial experiment, we establish a baseline 0-shot performance on the VQAv2 Val dataset using the OpenFlamingo model, **MPT (Team, 2023) 1B RedPajama (Computer, 2023) Language Model** and **ViT-L-14** (CLIP)(Radford et al., 2021) Vision Model pre-trained on Multimodal C4 dataset (Zhu et al., 2023).

The chosen configuration achieved a baseline accuracy of 45.5%. The achieved accuracy provides a reference point for evaluating the impact of subsequent experiments and modifications to the model.

3.3 Experiment 1: Fine-Tuning on A-OKVQA with OpenFlamingo

In this experiment, we opt to fine-tune the OpenFlamingo model on the **A-OKVQA dataset**using the **OpenLLaMA 3B** (Geng and Liu, 2023) Language Model. The rationale behind this choice stems from the improved performance exhibited by OpenLLaMA 3B on general question-answering benchmark datasets when compared to the previously utilized RedPajama LM.

Hyperparameter Tuning: We conduct a thorough exploration of hyperparameters - learning rate, warm-up steps, and learning rate scheduling. They control the magnitude of weight updates during training and collectively, these hyperparameters significantly influence the model's convergence and generalization capabilities. The results of these experiments are highlighted in Figure 2

We inserted LoRA layers at different points in the model architecture. By varying the placement of these layers, we aim to discern optimal configurations that result in improved performance on the A-OKVQA dataset. This experimentation provides insights into the importance of attention mechanisms in fine-tuning for specific tasks and the results can be seen in Figure 3.

The results inform subsequent iterations, guiding the refinement of OpenFlamingo for optimal performance in visual question answering tasks specific to the A-OKVQA dataset.

3.4 Experiment 2: Fine-Tuning on PubMedQA with OpenFlamingo

In this experiment, we fine-tune the OpenFlamingo model on the PubMedQA dataset using two distinct Language Models. Initially, Microsoft BioGPT(Luo et al., 2022) is chosen for its performance on benchmark medical question-answering datasets. The objective is to harness the domain-specific knowledge embedded in BioGPT for optimal adaptation to the medical context. Subsequently, OpenLLaMA was selected as an alternative LM for fine-tuning, and the model was trained on a subset of the A-OKVQA dataset to facilitate LM adjustment to vision inputs.

Following fine-tuning with both BioGPT and OpenLLaMA LMs, we plan to conduct a comparative analysis of the model's convergence speed and performance. The aim was to discern any variations in how the models adapt to the medical dataset under the influence of different language models.

Due to the large memory requirement of both PubMedQA and OpenLLaMA 3B, we were not able to proceed with fine-tuning OpenLLaMA on PubMedQA

4 Results and Discussion

Figure 2 highlights the results of hyper-parameter fine-tuning on A-OKVQA. As observed, a few warmup steps (20 in this experiment) with cosine learning rate scheduler gives the fastest convergence of all, where as exponential scheduler results in lower overall loss given enough iterations. However, providing a few warmup steps seems beneficial regardless of the scheduler chosen. This experiment was conducted on a sample of the dataset (8000 data points) and we believe that this would generalize over larger datasets as well.

Language Model	Fine-tune Data	Metric	Value
MPT 1B RedPajama (Mosaic ML)	Pre-trained (Multimodal) (C4)	VQAv2 (0 shot) Accuracy	45.5
OpenLLaMA 3B (OpenLM)	Pre-trained	VQAv2 (0 shot) Accuracy	38.2
OpenLLaMA 3B (OpenLM)	COCO Val + A-OKVQA	VQAv2 (0 shot) Accuracy	46.1
BioGPT 1.5B (GPT2)	PubMedQA + A-OKVQA	Time to converge (mins)	105
BioGPT 1.5B (GPT2)	A-OKVQA	Time to converge (mins)	65

Table 1: Finetuning results across experiments

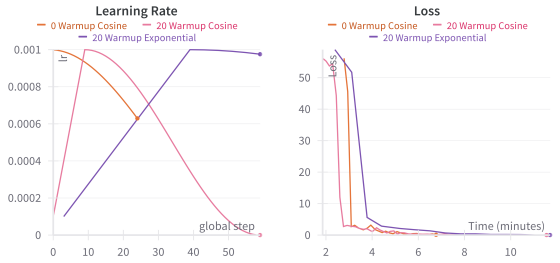

Figure 2: *left*: Change in learning rate with steps across lr schedulers and warmup *right*: Loss convergence over time

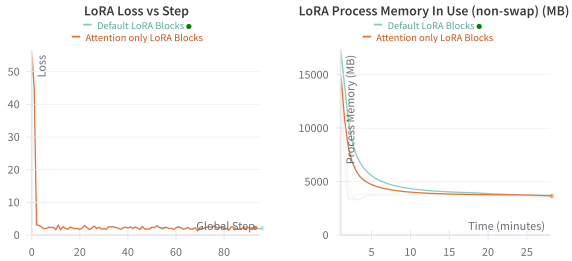

Figure 3: *left*: Loss over steps with different LoRA architectures. *right*: Process Memory usage over steps with different LoRA architectures

In Figure 3, we can observe how changing the number of LoRA blocks affects the training time and resource usage on A-OKVQA data and OpenLLaMA 3B model. Initially, we added trainable

LoRA blocks to all feed-forward blocks and attention blocks including the self attention layers of the language model and cross attention between the language encoder and perceiver-resampler. But upon removing the LoRA blocks from the feed-forward layers reducing the total number of trainable parameters from 14.8M to 11.4M, the convergence speed and overall loss seem unaffected whereas the memory usage, although initially less expensive, reaches the same level in both configurations

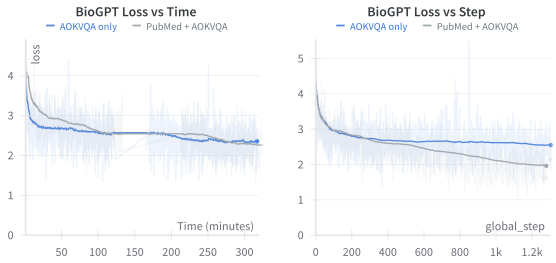

Figure 4: *left*: Loss convergence over time *right*: Loss convergence over step

Figure 4 focuses on how the BioGPT LM converges with different datasets. We initially train it on domain specific PubMedQA dataset as well as AOKVQA which provides visual-language data. And in the second experiment, we remove the medical data and focus only on finetuning BioGPT LM on visual language data. The vision-language only data seems to provide faster convergence with time as compared to domain specific data. However, when we observe loss against iterations, we see that the combined dataset performs better. This is because the VL only dataset has fewer data-points as compared to the combined dataset

5 Challenges

As computational resources and training datasets continue to grow, scalability and resource efficiency become increasingly important. Future works should explore techniques for scaling multimodal models efficiently, both in terms of model size and training data. We utilized the following hardware resources in our experiments:

Google Colab Pro: It provides access to Nvidia V100, A100, T4 GPUs. Among these only the A100 GPU had enough RAM (40GB) to accommodate the largest language model OpenLLaMA 9B model, access to the A100 GPUs is not guaranteed. To work around this, we used the V100 GPUs for evaluation of our fine-tuned models.

NVIDIA GeForce RTX 4080 GPU: Our team utilized an RTX 4080 GPU to mitigate the cost and reliability challenges associated with cloud GPUs. While most fine-tuning experiments were successfully conducted on this machine, we encountered out-of-memory issues when working with the PubMedQA data and OpenLLaMA model. We are currently running the experiment on a smaller sample of the PubMedQA dataset.

6 Conclusion

In this project, we addressed the challenges of fine-tuning multimodal vision language models under hardware constraints and domain-specific contexts. Leveraging an RTX 4080 GPU and OpenFlamingo framework, we conducted experiments on A-OKVQA and PubMedQA datasets.

Our hyperparameter tuning revealed optimal settings for faster convergence, with a focus on learning rate and warm-up steps. Additionally, the introduction of trainable LoRA blocks showcased the impact of attention mechanisms on fine-tuning efficiency.

Experiment 1 involved fine-tuning on A-OKVQA with OpenFlamingo, employing OpenLLaMA 3B LM. This aimed to enhance performance in medical question-answering. Experiment 2 focused on PubMedQA, exploring adaptability with Microsoft BioGPT and OpenLLaMA LMs. Both experiments included domain-specific language models and hyperparameter tuning.

Despite resource challenges, our efforts provide insights into resource-efficient fine-tuning, ethical considerations, bias detection, and vision model enhancement. Future work should emphasize scalability, bias mitigation, and vision model advancements for comprehensive multimodal understanding.

While our experiments primarily focused on language model adaptations, future works should equally prioritize the enhancement of vision models within the multimodal framework. Future works should also focus on developing techniques to systematically identify and mitigate biases within multimodal models including addressing biases present in both vision, language components, and biases that may emerge during fine-tuning. We identified the VLStereoset(Zhou et al., 2022) dataset to be used to detect bias in the model output.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models.

Together Computer. 2023. Redpajama: an open dataset for training large language models.

Group 47 CSCI 544 NLP F23. 2023. Nlp vlm project - finetuning and evaluation.

Xinyang Geng and Hao Liu. 2023. Openllama: An open reproduction of llama.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge.

MosaicML NLP Team. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms. Accessed: 2023-05-05.

Kankan Zhou, Eason Lai, and Jing Jiang. 2022. VLStereoSet: A study of stereotypical bias in pre-trained vision-language models. Association for Computational Linguistics.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023. Multimodal c4: An open, billion-scale corpus of images interleaved with text.