MorphVLM: Towards more Efficient and Robust Multimodal Vision Language Models

Chirag Khatri
ckhatri@usc.edu

Prince Verma
princeve@usc.edu

Lavrenti Mikaelyan
mikaelya@usc.edu

Mihir Mangesh Pavuskar
pavuksar@usc.edu

Pothula Punith Krishna
pothulap@usc.edu

Abstract

This project aims to enhance the performance and versatility of the state-of-the-art Multimodal Vision Language Model, specifically focusing on the Flamingo model. We propose modifications to the model's data processing pipeline, with an emphasis on substituting its vision and language components with alternative pretrained uni-modal constituents. By investigating the model's adaptability and potential shortcomings, we aim to address the pressing challenge of holistic understanding in the multimedia domain. Our work further distinguishes itself by emphasizing the ethical evaluation of models, spotlighting bias detection and mitigation. This proposal also delineates datasets for fine-tuning and evaluation, and technical challenges anticipated.

1 Introduction

Domain: Enhancing Multimodal Vision Language Modeling (VLM)
Goal: Modifying modules within the Flamingo (Alayrac et al., 2022) model's data processing pipeline and meticulously studying the resultant effects on its performance both in few-shot and zero-shot settings. One potential way to achieve this is to replace its vision and language components with other pretrained uni-modal components. We also aim to explore the model's inherent biases against ethical datasets.

2 Related Work

The architecture of Flamingo consists of two pre-trained and frozen models: a vision model that perceives visual scenes and a large language model that performs reasoning over text. It also incorporates a Perceiver-based architecture, allowing it to handle high-resolution images or videos by producing a fixed number of visual tokens per input. It achieves few shot learning by interleaving support example pairs (image-text or video-text) followed by the query visual input to build a prompt. Flamingo's flexibility in processing a variable number of videos or images is showcased through open-ended evaluations using beam search for decoding.

Chinchilla (Hoffmann et al., 2022) is a 70B parameter large language model (LLM) developed by DeepMind. It was trained using the same compute budget as Gopher (Rae et al., 2022) (a 280B parameter LLM), however by using 4x more training data, Chinchilla was able to uniformly and significantly outperform Gopher on a wide range of downstream evaluation tasks.
This breakthrough suggests that model size is not the only factor that matters for performance, and it is possible to train highly effective LLMs without using massive amounts of compute resources.

3 Datasets

3.1 MSCOCO: A cornerstone for image captioning and object detection.

The Microsoft COCO (Lin et al., 2014) (Common Objects in Context) dataset is a large-scale object detection, segmentation, and captioning dataset. It contains 328,000 images with over 2.5 million labeled object instances and 91 object categories.

3.2 MMLU: A colossal multitask test encompassing multiple-choice queries from diverse knowledge domains.

The Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021) dataset is a benchmark designed to measure knowledge acquired during pretraining by evaluating models exclusively in zero-shot and few-shot settings. It covers 57 subjects across STEM, the humanities, the social sciences, history, law, and ethics.

3.3 SuperGLUE: A successor to GLUE, equipped with challenging language tasks and enhanced resources.

SuperGLUE (Sarlin et al., 2020) (Wang et al., 2019) is a 160,000-image dataset for evaluating visual matching algorithms. The dataset is divided into four main categories:
Indoor images such as living rooms, kitchens, and offices, Outdoor scenes, such as streets, parks, and forests, Aerial scenes, such as cities and landscapes, and Synthetically generated images with known correspondences.
Each category contains a variety of subcategories, such as day/night, cluttered/uncluttered, and occluded/unoccluded.

3.4 VLStereoSet: Investigating stereotypical bias in pretrained vision-language models.

VLStereoSet (Zhou et al., 2022) is a vision-language probing dataset to measure stereotypical bias in vision-language models. It consists of images each accompanied by three candidate captions: stereotypical, anti-stereotypical, and semantically meaningless.

4 Approach

We aim to achieve the goal through the following distinct contributions:
Component Enhancement: We aim to replace Flamingo's vision and language components with relatively smaller architectures to potentially decrease inference cost while maintaining performance.
Bias Evaluation: An additional, less-explored dimension is the ethical evaluation of models. Our work will emphasize understanding biases present in these models, a feature missing in most current research. This is the additional goal that we aim to pursue for further research.

4.1 Efficient Fine-tuning through LoRA

LoRA (Hu et al., 2021), or 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.

It is a powerful fine-tuning method for LLMs that improves sample efficiency, and reduces overfitting, while having minimal to no impact on inference latency. It shows promise for training large models on limited hardware resources and for fine-tuning them to new tasks. However, it is important to be aware of the potential drawbacks of LoRA, such as reduced adaptation capacity and increased training time.

4.2 Technical Challenges

4.2.1 Model Replication

The act of replicating results from a sophisticated model such as Flamingo demands precision. Since the model is not open sourced by DeepMind, we are looking at other open source implementations online. (Awadalla et al., 2023)

4.2.2 Data and Component Pipeline Modification

Altering modules in a well-established data processing pipeline can introduce many challenges. Determining the best configurations of the vision and language components in the model and gauging how they impact the model's performance will be intricate.

4.2.3 Model Fine-tuning

The enhancement of the pre-trained model components via fine-tuning requires adjustments in the architecture's weights. These refinements must be executed such that there are no issues with convergence.

4.2.4 Computational Resources

The fine-tuning and testing of different uni-modal models as the components requires a significant number of experiments. Performing that on limited access to GPU hardware is challenging

5 Division of Labor

Our group will undertake the following roles:

- Literature review, initial model result replication.
- Fine-tuning, benchmark testing.
- Component selection, testing.
- Bias evaluation, documentation.
- Curating ethical datasets, bias mitigation.

Note: This division may undergo revisions as the project evolves.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft coco: Common objects in context.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling language models: Methods, analysis insights from training gopher.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. Superglue: Learning feature matching with graph neural networks.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding.

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