An Examination of Expansive Multimodal Models: Insights from an Educational Overview

Abstract

This document provides a summary of a presentation centered on extensive multimodal models, specifically their development to a level comparable to and potentially exceeding that of multimodal GPT-4. The exploration is divided into three sections. Initially, the context is established by discussing recent large-scale models akin to GPT, which are designed for vision and language processing. This sets the stage for exploring research in large multimodal models (LMMs) that are fine-tuned with instructions. Subsequently, the foundational aspects of instruction tuning in large language models are covered, which is a method that is further adapted to the multimodal domain. The final section demonstrates the creation of a basic version of multimodal models similar to GPT-4 using publicly available resources. Additionally, a review of newly developing areas in this field is presented.

1 Introduction

With the widespread integration of advanced language models into modern society, there's a burgeoning enthusiasm among scholars and scientists to create open-source large language models (LLMs) and to investigate their growth into large multimodal models (LMMs). This manuscript concentrates on leveraging LLMs for multimodal applications and training LMMs in a comprehensive manner, enabling them to process visual data and engage in conversation.

2 Background

2.1 Image-to-Text Generative Models

In their present configuration, LMMs predominantly function as image-to-text generators, accepting images as input and producing textual content as output. The architectural design of these models generally includes an image encoder for deriving visual characteristics and a language model for generating textual sequences. These visual and linguistic components can be interconnected through an adaptable module. Both the image encoder and the language model have the flexibility to be developed from the ground up or based on previously trained models.

The training methodology typically involves employing an auto-regressive loss on the generated text tokens. Within the Transformer framework, image tokens have the capability to interact with one another, and each text token is influenced by the preceding text tokens and all image tokens.

2.2 Case Studies

We will analyze several established LMMs to demonstrate how the architecture can be actualized across various models while adhering to the same auto-regressive training principle.

Case Study I: LMM Trained with Image-Text Pairs

Many LMMs are developed using extensive collections of image-text pairs. Notable models like Generative Image-to-Text Transformer (GIT) and Bootstrapping Language-Image Pre-training (BLIP2)

.

have set high standards across various datasets. GIT utilizes an image encoder from a contrastive pre-trained model and builds a language model independently. Conversely, BLIP2 maintains the pre-trained image and language models in a fixed state while incorporating a trainable Querying Transformer (Q-former), demonstrating efficiency through a unique bootstrapping technique.

Case Study II: LMM Trained with Interleaved Image-Text Sequences

Flamingo serves as an exemplary model in this category, incorporating pre-trained image and language models with the addition of new integrative components. It includes a Perceiver Sampler to streamline computational demands and a Gated Transformer to enhance stability during the early training phase. Flamingo is trained on a diverse mix of large-scale multimodal data sourced exclusively from the web, bypassing the need for conventionally annotated machine learning datasets. Post-training, Flamingo can adapt to vision-based tasks through few-shot learning without additional task-specific tuning.

A standout feature of Flamingo is its capability for multimodal in-context learning. When presented with image-text pairs as a demonstration, Flamingo can generalize to new, unseen tasks, such as visual math problems, without further training. It successfully interprets the patterns in task instructions from examples and applies this understanding to new images. Flamingo represents a significant advancement in multimodal learning, akin to the breakthroughs seen with GPT-3 in language processing.

2.3 OpenAI Multimodal GPT-4 and Research Gaps

Released in March 2023, OpenAI's GPT-4 showcases advanced capabilities in understanding and reasoning with visual data. Although specifics of the model remain undisclosed, its ability to facilitate new applications is evident from highlighted examples in technical reports. For instance, it can discern unusual elements within images and demonstrate sophisticated reasoning across text and images.

The inquiry into constructing models akin to Multimodal GPT-4 leads us to examine OpenAI's advanced models, as depicted in Figure 7. Key observations are: (i) GPT-2 serves as the auto-regressive equivalent in the era dominated by BERT's pre-training then fine-tuning paradigm. (ii) GPT-3, a 175-billion parameter model trained on extensive web text, showcases emergent properties such as in-context learning and chain-of-thoughts (CoT) reasoning without requiring further training. This model represents a shift from fine-tuning model weights to utilizing prompts for broader generalization and reduced adaptation costs. (iii) ChatGPT and InstructGPT emphasize the importance of models following instructions and aligning with human intentions by fine-tuning on high-quality instruction data and using a reinforcement learning framework. (iv) GPT-4 not only enhances previous models' language capabilities but also incorporates visual inputs for comprehension and reasoning.

3 Pre-requisite: Instruction Tuning in Large Language Models

Instruction-following is a concept that originated in the field of natural language processing (NLP). To understand this concept more deeply and trace its development, we revisit the practice of instruction tuning in conjunction with LLMs.

3.1 Instruction Tuning

Traditional Language Data

In the realm of natural language processing, the seq2seq format is frequently employed, where each data point comprises an input sequence and a corresponding output sequence. Typically, task instructions are implicitly understood rather than explicitly stated. Models trained on this data format often struggle to adapt to new tasks in a zero-shot manner because they lack the ability to interpret and generalize task instructions during testing.

Instruct Language Data

Recent advancements involve the explicit incorporation of task instructions during model training. These instructions, often articulated in natural language, lead to a structured format of instruction-input-output triplets. This enables the training of a single model capable of handling multiple tasks

with clear directives. The exposure to varied task instructions and examples during training allows the model to generalize to novel tasks through task composition during inference.

3.2 Self-Instruct and Open-Source LLMs

The collection of a wide array of high-quality instruction-following data can be achieved through two primary methods: human-human interaction and human-machine interaction. The former is resource-intensive, involving human task providers and annotators, while the latter involves machines or models performing the annotation tasks under human guidance.

Self-Instruct tuning represents a streamlined and potent method for aligning LLMs with human intent, utilizing instruction-following data produced by leading teacher LLMs. This technique, which leverages the in-context learning capability of LLMs, has significantly enhanced the zero- and few-shot generalization abilities of LLMs. The iterative process, as illustrated in Figure 9, involves humans providing initial examples, which the LLM then uses to generate further instructions and responses, refining the dataset iteratively.

4 Instructed Tuned Large Multimodal Models

This section describes the development of a minimal multimodal GPT-4 model using open-source tools, with a focus on the LLaVA model, and a similar approach in the MiniGPT-4 project.

4.1 Open-Source Prototypes: LLaVA / MiniGPT4

Inspired by successful concepts in NLP, we apply the self-instruct methodology from language processing to the vision-and-language domain. A significant challenge is the absence of a robust multimodal teacher model. Thus, we explore how language-only models like GPT-4 can generate multimodal instruction-following data.

4.1.1 Data Creation

Instead of directly inputting images into OpenAI GPT, symbolic sequence representations are used, as shown in Figure 12 (a). LLaVA utilizes captions and bounding boxes for several reasons: (1) GPT-4 is found to comprehend these representations effectively, unlike ChatGPT, which struggles with bounding box data; (2) these elements are crucial for an informative representation of the image.

As demonstrated in Figure 12 (b), three forms of instruction-following data are used: multi-turn conversations for interactive user engagement, detailed descriptions for comprehensive response generation, and complex reasoning to address the implications beyond the image content.

4.1.2 Network Architecture and Training

As shown in Figure 13, LLaVA's architecture is a specific implementation of the general image-to-text generative model framework discussed in Section 2 and Figure 3. LLaVA integrates a pre-trained CLIP ViT-L/14 visual encoder with the Vicuna large language model via a projection matrix. The training process involves two stages:

- **Stage 1: Pre-training for Feature Alignment.** Only the projection matrix is updated using a portion of the CC3M dataset, focusing solely on image captioning. - **Stage 2: End-to-End Fine-tuning.** Both the projection matrix and the LLM are fine-tuned to cater to various application scenarios.

4.1.3 Performance

Performance on Visual Chat

When fine-tuned on diverse multimodal instruction-following data, LLaVA demonstrates effectiveness in user-oriented applications. Empirical evidence suggests that adjusting only the linear projection layer is adequate for conversational scenarios, although it necessitates longer training periods.

In an evaluation using 30 unseen images, each paired with three types of instructions, LLaVA achieved an 85.1

Performance on Science QA

LLaVA, when fine-tuned on a scientific multimodal reasoning dataset, achieved a 90.92

Performance on OCR in the Wild

Despite not being explicitly trained on OCR data, LLaVA exhibits a surprising zero-shot OCR capability, as illustrated in Figure 16.

Emerging Topics

4.1.4 More Modalities (Beyond VL)

- **ChatBridge**: This model innovates by employing a Large Language Model as a linguistic mediator to connect different modalities [65]. - **PandaGPT**: A comprehensive model designed to adhere to instructions across various modalities [41]. - **SpeechGPT**: Enhances large language models by incorporating inherent cross-modal conversational capabilities [61]. - **X-LLM**: Advances large language models by conceptualizing multi-modalities as different languages [4].

Although there is considerable diversity in the types of models, the fundamental concept of integrating multiple modalities is consistent with the approach used in LMMs, which augment LLMs with visual capabilities.

4.1.5 Multitask Instruct with Established Academic Datasets/Tasks

- **MultiInstruct**: This initiative aims to enhance zero-shot learning across various modalities by employing instruction tuning [57]. - **mPlug-OWL**: Utilizes modularization to enrich large language models with multimodality, thereby improving their versatility [58]. - **InstructBLIP**: Develops general-purpose vision-language models by incorporating instruction tuning, making them adaptable to a wide range of tasks [6]. - **Multimodal-GPT**: A model that integrates vision and language to facilitate natural dialogues with users [13]. - **Instruction-ViT**: Introduces multi-modal prompts to enhance instruction learning within the Vision Transformer (ViT) architecture [54].

Multimodal In-Context-Learning

- **OpenFlamingo**: An open-source initiative that replicates the Flamingo model by DeepMind, trained on the extensive Multimodal C4 dataset, which includes images interleaved with text [2]. - **Otter**: This model stands out for its in-context instruction tuning capabilities, allowing it to adapt to new tasks based on the context provided in the instructions [18]. - **M3IT**: A comprehensive dataset designed for multi-modal multilingual instruction tuning, facilitating the development of models that can understand and generate content across different languages and modalities [22]. - **MetaVL**: Focuses on transferring the in-context learning ability from language models to vision-language models, enabling them to perform tasks based on contextual examples without prior training [30].

Parameter-Efficient Training

- **LLaMA-Adapter V2**: A parameter-efficient visual instruction model that demonstrates how to effectively adapt large language models for visual tasks with minimal parameter adjustments [10]. - **LAVIN**: Another parameter-efficient model that showcases efficient tuning strategies for vision-language tasks, emphasizing minimal computational resources [27]. - **QLoRA**: Introduces a method for efficient fine-tuning of quantized LLMs, significantly reducing the memory footprint required for training large models [7].

4.1.6 Benchmarks

- **Hidden Mystery of OCR in Large Multimodal Models**: Investigates the unexpected proficiency of LMMs in optical character recognition (OCR) without explicit training in this area [25]. - **Evaluating Object Hallucination**: Addresses the challenge of object hallucination in large vision-language models, providing a framework for assessing and mitigating this issue [23]. - **Adversarial Robustness of Large Vision-Language Models**: Examines the resilience of LMMs against adversarial attacks, which is crucial for their deployment in security-sensitive applications [64]. - **LAMM**: Introduces a language-assisted multi-modal instruction-tuning dataset, along

with a framework and benchmark for evaluating the performance of LMMs [59]. - **LVLM-eHub**: Presents a comprehensive evaluation benchmark for assessing the capabilities of large vision-language models across a variety of tasks [56].

4.1.7 Applications

- **PathAsst**: Reimagines the field of pathology by integrating a generative AI assistant, showcasing the potential of LMMs in specialized domains [42]. - **PMC-VQA**: Focuses on visual instruction tuning for medical visual question answering, demonstrating the applicability of LMMs in healthcare [63]. - **LLaVA-Med**: A model trained to assist in biomedicine, highlighting the use of LMMs for generating responses to open-ended research questions based on biomedical images [19].

5 How Close Are We to Reaching or Surpassing OpenAI's Multimodal GPT-4?

The open-source community has rapidly produced a range of models and prototypes that introduce a variety of new functionalities. For instance, LLaVA and Mini-GPT4 are leading the way in the creation of multimodal chatbots, replicating some of the functions described in OpenAI's GPT-4 technical documentation. Additionally, GILL has broadened the capabilities of LMMs to include comprehensive image generation, a feature not currently present in GPT-4. From the standpoint of introducing basic versions of new multimodal features, the open-source community is seemingly on par with OpenAI's Multimodal GPT-4, taking initial steps toward developing a versatile multimodal assistant.

Nevertheless, there remains a significant disparity when it comes to enhancing a particular functionality, such as the visual reasoning seen in LLaVA. The technical documentation from OpenAI provides examples of complex visual tasks that necessitate models capable of processing numerous high-resolution images and extended sequences, in addition to delivering responses that require specialized knowledge. This demands significantly greater computational power and more sophisticated language models, which are generally not accessible to most individuals.

6 Conclusion

This paper has outlined the foundational aspects and advanced functionalities of large multimodal models (LMMs). It has revisited the concept of instruction tuning in large language models (LLMs) and demonstrated the steps to construct a basic model akin to LLaVA and MiniGPT4 with open-source tools. Furthermore, it has categorized and summarized the most recent advancements in this research area, offering a starting point for those keen to embark on LMM exploration.

The paper also proposes future directions for community-driven efforts. It suggests that entities with substantial resources should concentrate on scaling existing capabilities and exploring new emergent properties. Meanwhile, others can focus on creating prototypes for new features, developing evaluation methods, and devising strategies to lower computational demands, thereby making advanced model computation more widely accessible.

Acknowledgments

We express our gratitude to all the researchers who have contributed to the papers on LLMs and LMMs, which have been instrumental in the creation of this tutorial. While we aimed to cover the relevant literature up to June 19, 2023, the rapid evolution of LMM research may mean that some contributions have been unintentionally omitted. We apologize for any such oversights.