Journal of Computer Science and Electrical Engineering
ISSN: 2663-1938
DOI: 10.61784/jcsee3020

MLMM : MULTI-MODAL LANGUAGE MODELS BASED MULTI-AGENTS COLLABORATED SOLVING CV PROBLEMS

YanQiao Ji
Liaoning Equipment Manufacturing Vocational and Technical College, Shenyang 110161, Liaoning, China.
Corresponding Email: 491318458@qq.com

Abstract: To enhance the system's ability to interact with and comprehend the visual world, we integrate vision capabilities into a large language model (LLM) framework, specifically the AutoGen framework. By employing a multi-agent conversation methodology, our work aims to mitigate errors in image generation and improve the system's output to better meet user expectations. Our work is based on the Meta Llama 3 family of pretrained and instruction-tuned generative text models, specifically optimized for dialogue applications, LLaVA for image recognization and the stable diffusion model for image generation. Our work is efforting on addressing vision-related problems and the potential for further enhancements with the support of more sophisticated models.
Keywords: Software engineering; Artificial intelligence; Deep learning

1 INTRODUCTION

The journal of Upubscience Publisher gives preference to manuscripts of high scientific level, which have not been published, and are written not only for specialists but also for the general public interested in the questions of related fields. A large language model (LLM) is a type of artificial intelligence model that uses machine learning techniques to analyze and generate human-like text. It is trained on a massive amount of text data, allowing it to learn language structure, grammar, and vocabulary, and can be used for various natural language processing tasks such as language translation, text summarizing, and question answering. As the rapid development of LLMs, it' s becoming a powerful back-end of developing intelligence agent technology because of the ability of planning, memorizing, tool using, and adaptation to novel observations in a multitude of real-world tasks[1, 2]. Vision plays an important role in recognizing and interacting the worlds for humans and many other animals. A central aspiration of Artificial Intelligence (AI) is to devise AI agents capable of emulating the efficient perception and generation of visual signals, thereby enabling the inference about and interaction with the visual world. Examples of this include the recognition of objects and actions within scenes, as well as the creation of sketches and images for communicative purposes. Establishing foundation models endowed with visual capabilities represents a thriving research area in the pursuit of this goal.
A enduring ambition within the field of artificial intelligence is the creation of versatile assistants capable of adhering to users' (multimodal) directives to accomplish an extensive array of real- world duties[3]. Recently, there has been a surge in the community's focus on the development of foundational models that exhibit spontaneous capabilities for multimodal comprehension and production within open-world settings [4]. Whereas the methodologies for deploying Large Language Models (LLMs), such as GPTs[5], to engineer universal assistants for linguistic tasks have been validated, the approaches for constructing all-purpose, multi-modal aides for computer vision and vision-language assignments are yet to be fully charted.
In our work, we mainly made the following contributions:

■We integrate vision capabilities into our system, enabling it to interact with and comprehend the visual world.

■Our work based on a multi-agent conversation methodology to mitigate errors in image generation and improve the system's ability to deliver images that better meet the user's expectations.

■ We implement an iterative exchange between the image-generating agent and the evaluator to refine the image generation process, which enhance the alignment between the system's output and the user's specific requirements through the collaborative dialogue between the agents.

2 RELATED WORKS

2.1 LLM based Multi-Agents Communication Framework

In this part, we explore the application of a proposed multi-agent framework to enhance the functional- ity of existing AI models. We focus on several prominent multi-agent collaboration frameworks[6, 7] and assess how these models operate and their limitations.
Auto-GPT[6], a milestone in AGI development, demonstrates autonomous capabilities through its integration of internet access, memory management, and advanced text generation. It can execute a wide range of tasks with minimal user input. Our framework can be mapped onto Auto-GPT's architecture, treating its main agent as a node in a multi-agent system. This agent interacts with various plugins, which can be seen as external tools or services, and can create and

manage other agents. The introduction of a "Supervisor Agent" in our framework could address issues like looping, while the concept of co-agents could enable collaboration among multiple Auto-GPT instances.
BabyAGI[8] operates on a triad of LLM chains for task generation, prioritization, and execution. Our framework can model BabyAGI as a network of specialized agents, enhancing its structure and modularity. The inclusion of a feedback loop could facilitate learning and improvement over time. Gorilla[7]is a fine-tuned LLaMA model with advanced document retrieval and API interaction capabilities, represents a significant advancement in language modelling. It can be represented within our framework by a single agent with plugins for handling APIs, allowing for flexibility and adaptability to changes in API documentation and functionality. Camel[9] proposed a novel communicative agent framework in which agents communicate with each other in a role-playing method. This framework leverages inception prompting as a means to steer chat agents towards the fulfillment of tasks while ensuring alignment with human objectives. They introduced an innovative communicative agent framework and facilitate further research in the domain of communicative agents and beyond.
Microsoft developed AutoGen[10] in the year 2023, which is an open-source architecture that em- powers developers to create LLM-driven applications through interconnected agents capable of dialogue to achieve objectives. These agents are adaptable and capable of interaction, functioning across different modes that integrate LLMs, human participation, and various tools. Developers can also wield the flexibility to stipulate the interaction protocols for agents using AutoGen. The framework accommodates both natural language and programming code to establish versatile com- munication protocols tailored for various applications. Acting as a universal platform, AutoGen facilitates the construction of a wide spectrum of applications, differing in complexity and LLM capabilities. Experimental research validates the framework's efficacy across numerous exemplar applications, spanning fields such as mathematics, programming, Q&A systems, operations research, digital decision-making, and entertainment.

2.2 Vision-Language Modeling

General vision-language modeling. Building on successes in large language and vision models, recent years have seen a growing interest in large vision-language models (VLMs)[11, 12]. VLMs are adapt at simultaneously understanding both visual and textual content. This multidimensional understanding has enabled VLMs to be effectively applied to a diverse array of tasks, including visual question answering [13], image captioning [14], optical character recognition [15], and object detection[16].
The integration of images into these models takes various forms. For instance, [17]enhance pre- trained language models with a mechanism that allows them to attend directly to a single context image. This approach enables the model to focus on specific visual information relevant to the language task at hand. Frozen[18] take another notable method, in which the vision encoder parameters are optimized through back-propagation, while the language model component remains frozen. This method allows for the fine-tuning of the visual processing capabilities without altering the linguistic knowledge already encoded in the language model. This separation of concerns ensures that the model's proficiency in understanding and generating language is preserved while its ability to interpret visual content is enhanced.
Our work is a demo based on AutoGen Framework and composed with large multi-modal model LLaVa[19] and image generation model stable diffusion. LLava stands for Large Language and Vision Assistant, an end-to-end trained, comprehensive multi-modal model that integrates a vision encoder with an LLM to achieve broad-spectrum understanding of visual and linguistic content. In order to promote ongoing research into the adherence to visual instructions, we have developed two evaluation benchmarks encompassing a variety of demanding and applied tasks.The Stable Diffusion model is a deep learning model used for generating images from text, developed by Stability AI in collaboration with other research institutions. It is a type of diffusion model, which is a class of deep learning models that aim to produce data with a specific structure from random noise. In the context of image generation, diffusion models work by starting with a corruption of the input data and learning to de-noise it over time, eventually producing a clean, structured output.

3 APPLICATION

In this section, we will introduce our work which is primarily built upon the foundational framework of AutoGen. Our system integrates vision capabilities, enabling it to interact with and comprehend the visual world. Due to time constraints, we will present a straightforward sequence to elucidate the comprehensive workflow of our research.
We assume a circumstance in which user need to generate a picture according to a certain theme. In the conventional approach, one might directly employ an image-generation model, which may yield sub optimal results, such as an output that does not adequately fulfill the user's vision. Our emphasis lies in leveraging a multi-agent conversation methodology to mitigate these types of errors.
Specifically, one agent is responsible for generating images, while another agent assumes the role of evaluator, who assesses and scores the produced images, as well as the prompt words employed, ensuring that the output meets the desired standard. Upon encountering discrepancies, the evaluator provides feedback to the image-generating agent, prompting it to make necessary adjustments. This iterative exchange continues until the evaluator deems the image and prompt words to be satisfactory, at which point the process concludes.

The efficacy of this communication method lies in its ability to iterative refine the image generation process, with the goal of enhancing the alignment between the system's output and the user's specific requirements. By fostering a collaborative dialogue between the agents, the system is better equipped to deliver images that better meet the user's expectations.

3.1 Models Definition

As the back-end LLM of AutoGen, we use local-deployed LLaMA3-70b model using for conversation generation and summarization. Developed by Meta-AI, Meta Llama 3 family is a collection of pre- trained and instruction tuned generative text models. The Llama 3 instruction-tuned models have been specifically optimized for dialogue applications and have demonstrated superior performance against many open-source chat models in standard industry benchmarks. Moreover, in the development of these models, the optimization of helpfulness and safety has been prioritized to ensure their effective and responsible use. In the autoGen configuration file, setting the baseurl to the local host of ollama as shown in Algorithm 1.

```
[{
    "model": "llama3:70b",
    "api_key": "EMPTY",
    "tags": ["ollama"],
    "base_url": "http://localhost:11434/v1"
}]
```

Algorithm 1 Configure Setting of Ollama

In the task, there are two vision models will be used. One is responsible for reviewing and evaluating the picture. We deploy LLaVa-7b model locally using ollama framework. The other is focusing on generating an image according the message given by user. We employ the mature and dependable stable-diffusion-v1-5[20]model to transform text into imagery. This model is locally deployed through the diffuse library in Python, and it is selectively invoked by agents within the AutoGen framework.

3.2 Agent Creation

In AutoGen Framework, the core design principle is to optimize and integrate multi-agent workflows using multi-agent conversations, which greatly reduce the effort to create LLM-based applications on different LLMs in various frames and enlarge the re-usability of LLM-agents as well.
A conversable agent is an entity endowed with a particular role that facilitates the exchange of messages with other agents, enabling the initiation and continuation of conversations. These agents rely on the messages they send and receive to maintain an internal context. Additionally, they can be programmed with a suite of capabilities, which may be derived from large language models, integrated tools, or direct human input. These agents operate in accordance with predefined behavior patterns. The design of the agents is outline in Algorithm 2.

```
img_gen_assistant = AssistantAgent( name="te
    xt_to_img_prompt_expert",
    system_message="""
        You are a text to image AI model expert, you will use text_to_image_generation function to
            generate image with prompt provided, and also improve prompt based on feedback
            provided until it is 10/10.
        For image generation tasks, only use the function you have been provided with. Reply
            TERMINATE when the task is done.
    """,
    llm_config=llm_config_assistants,
)

img_critic_assistant = AssistantAgent( name="i
    mg_critic", system_message="""
        You are an AI image critique, you will use img_review function to review the image
            generated by the text_to_img_prompt_expert against the original prompt, and provide
            feedback on how to improve the prompt.
        For all tasks, only use the functions you have been provided with. Reply TERMINATE when
            the task is done.
    """,
   llm_config=llm_config_assistants,
```

Algorithm 2 Design of Agents

4 CONCLUSION AND DISSCUSSION

To address a vision-related problem, we have integrated a diverse range of models within our AutoGen framework. This integration enables a multi-agent system to possess both image generation and image comprehension capabilities. We anticipate that this collaborative approach can be readily extended to incorporate other models and functionalities, thereby addressing a wide array of challenges. Currently, the capabilities of the language models and the models utilized in our work are not yet optimal. We are confident that with the support of more sophisticated models, our system will be able to provide enhanced solutions. The relevant code can be found at this URL.

FUNDING

The project was supported by Research on the Construction of Virtual Simulation Training Base for Vocational Education (LZJG2023041).

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Yao, S, Zhao, J, Yu, D, et al. ReAct: Synergizing Reasoning and Acting in Language Models, 2023. DOI: https://doi.org/10.48550/arXiv.2210.03629.
[2] Wang, L, Ma, C, Feng, X, et al. A Survey on Large Language Model based Autonomous Agents. Frontiers of Computer Science, 2023, 18. DOI: https://doi.org/10.48550/arXiv.2308.11432.
[3] Li, C, Gan, Z, Yang, Z, et al. Multimodal Foundation Models: From Specialists to General- Purpose Assistants. Foundations and Trends® in Computer Graphics and Vision, 2023, 16(1-2): 1-214. DOI: https://doi.org/10.48550/arXiv.2309.10020.
[4] Li, C, Liu, H, Li, L, et al. Elevater: A benchmark and toolkit for evaluating language- augmented visual models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh, eds., Advances in Neural Information Processing Systems, 2022, 35, 9287-9301. DOI: https://doi.org/10.48550/arXiv.2204.08790.
[5] OpenAI. GPT-4 Technical Report. 2023.
[6] Yang, H, Yue, S, He, Y. Auto-GPT for Online Decision Making: Benchmarks and Additional Opinions, 2023. DOI: https://doi.org/10.48550/arXiv.2306.02224.
[7] Patil, SG, Zhang, T, Wang, X, et al. Gorilla: Large Language Model Connected with Massive APIs, 2023. DOI: https://doi.org/10.48550/arXiv.2305.15334.
[8] Nakajima, Y. BabyAGI, 2023.
[9] Li, G, Hammoud, H A A K, Itani, H, et al. CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. In Thirty-Seventh Conference on Neural Information Processing Systems, 2023, 2264, 51991 - 52008. DOI: https://doi.org/10.48550/arXiv.2303.17760.
[10] Wu, Q, Bansal, G, Zhang, J, et al. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation, 2023. DOI: https://doi.org/10.48550/arXiv.2308.08155.
[11] Li, LH, Yatskar, M, Yin, D, et al. Visualbert: A simple and performant baseline for vision and language, 2019. DOI: https://doi.org/10.48550/arXiv.1908.03557.
[12] Hu, X, Gan, Z, Wang, J, et al. Scaling up vision-language pretraining for image captioning. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, 17959-17968. DOI: 10.1109/CVPR52688.2022.01745.
[13] Zhou, LW, Hamid Palangi, Zhang L, et al. Unified vision-language pre-training for image captioning and VQA. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(07): 13041-13049. DOI: https://doi.org/10.1609/aaai.v34i07.7005.
[14] Gan, Z, Li, L, Li, C, et al. Vision-language pre-training: Basics, recent advances, and future trends. Foundations and Trends® in Computer Graphics and Vision, 2022, 14(3-4): 163-352. DOI: https://doi.org/10.1561/0600000105.
[15] Li, M, Lv, T, Chen, J, et al. Trocr: Transformer-based optical character recognition with pre-trained models. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(11): 13094-13102. DOI: https://doi.org/10.1609/aaai.v37i11.26538.
[16] Chen, T, Saxena, S, Li, L, et al. Pix2seq: A language modeling framework for object detection. 2022. DOI: https://doi.org/10.48550/arXiv.2109.10852.
[17] Alayrac, J-B, Donahue, J, Luc, P, et al. Flamingo: a visual language model for few-shot learning. 2022. DOI: https://doi.org/10.48550/arXiv.2204.14198.
[18] Tsimpoukelli, M, Menick, J, Cabi S, et al. Multimodal few-shot learning with frozen language models. NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems, 2021, 16, 200-212.
[19] Liu, H, Li, C, Wu, Q, et al. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine, eds., Advances in Neural Information Processing Systems, 2024, 36, 34892-34916. DOI: https://doi.org/10.48550/arXiv.2304.08485.
[20] Rombach, R, Blattmann, A, Lorenz, D, et al. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, 10684-10695. DOI: https://doi.org/10.48550/arXiv.2112.10752.