HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face

Yongliang Shen[1*], **Kaitao Song**[2*], **Xu Tan**[2], **Dongsheng Li**[2], **Weiming Lu**[1], **Yueting Zhuang**[1]
Zhejiang University[1], Microsoft Research Asia[2]
`{syl, luwm, yzhuang}@zju.edu.cn`, `{kaitaosong, xuta, dongsli}@microsoft.com`

Abstract

Solving complicated AI tasks with different domains and modalities is a key step toward advanced artificial intelligence. While there are abundant AI models available for different domains and modalities, they cannot handle complicated AI tasks. Considering large language models (LLMs) have exhibited exceptional ability in language understanding, generation, interaction, and reasoning, we advocate that LLMs could act as a controller to manage existing AI models to solve complicated AI tasks and language could be a generic interface to empower this. Based on this philosophy, we present HuggingGPT, a framework that leverages LLMs (e.g., ChatGPT) to connect various AI models in machine learning communities (e.g., Hugging Face) to solve AI tasks. Specifically, we use ChatGPT to conduct task planning when receiving a user request, select models according to their function descriptions available in Hugging Face, execute each subtask with the selected AI model, and summarize the response according to the execution results. By leveraging the strong language capability of ChatGPT and abundant AI models in Hugging Face, HuggingGPT is able to cover numerous sophisticated AI tasks in different modalities and domains and achieve impressive results in language, vision, speech, and other challenging tasks, which paves a new way towards advanced artificial intelligence[2].

1 Introduction

Large language models (LLMs) [1, 2, 3, 4, 5, 6], such as ChatGPT, have attracted enormous attentions from both academia and industry, due to their remarkable performance on various natural language processing (NLP) tasks. Based on large-scale pre-training on massive text corpora and reinforcement learning from human feedback (RLHF), LLMs can produce superior capability in language understanding, generation, interaction, and reasoning. The powerful capability of LLMs also drives many emergent research topics (*e.g.*, in-context learning [1, 7, 8], instruction learning [9, 10, 11, 12], and chain-of-thought prompting [13, 14, 15, 16]) to further investigate the huge potential of LLMs, and brings unlimited possibilities for us to build advanced artificial intelligence systems.

Despite these great successes, current LLM technologies are still imperfect and confront some urgent challenges on the way to building an advanced AI system. We discuss them from these aspects: 1) Limited to the input and output forms of text generation, current LLMs lack the ability to process complex information such as vision and speech, regardless of their significant achievements in NLP tasks; 2) In real-world scenarios, some complex tasks are usually composed of multiple sub-tasks, and thus require the scheduling and cooperation of multiple models, which are also beyond the capability of language models; 3) For some challenging tasks, LLMs demonstrate excellent results in zero-shot

[*] The first two authors have equal contributions. This work was done when the first author was an intern at Microsoft Research Asia.

[2] `https://github.com/microsoft/JARVIS`.