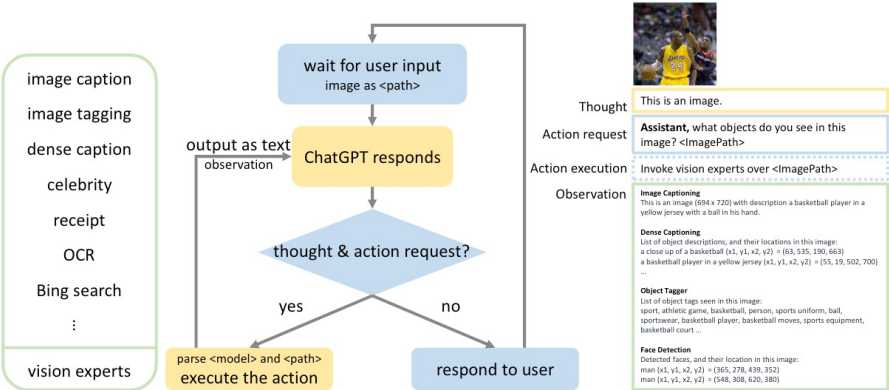


MM-ReAct (arXiv 2023)

Motivation: 现有视觉语言模型的能力有效，在一些具体任务上没有特定任务模型好，本文融合不同的特定任务模型完成多任务下的视觉推理和行动。

整体思路： MM-REACT定义了一系列高级视觉任务，并引入了一种文本提示设计，可以表示文本描述、文本化的空间坐标，以及以对齐文件名表示的图像和视频等密集视觉信号。



- 目的：** 汇聚众多视觉专家，为ChatGPT赋予视觉理解能力。利用ReAct调用多个视觉模型为ChatGPT形成文本的caption。
- 1. 用户输入：** 由于ChatGPT无法直接处理图片与视频，因此在输入时将图片与视频用路径代替，并让ChatGPT调用视觉专家模型从不同角度理解图像内容；
 - 2. ChatGPT回复：** ChatGPT有两种回复，第一种是直接回复文本，第二种是调用视觉专家；
 - 3. 视觉专家：** 给定ChatGPT的操作请求，使用正则表达式解析专家名和文件路径，并调用相关操作；视觉专家的输出格式可能是多种形式的，为了方便ChatGPT理解，需要将其转化成文本形式。

Convert image to <imagePath> and send to ChatGPT

Thought 1 ChatGPT: This is an image.

Action request 1 Assistant, what objects do you see in this image? <imagePath>

Action execution 1 Invoke vision experts, including Image Captioning, Dense Captioning, Object Tagger, Face Detection

Observation 1 Image Captioning Assistant: This is an image (694 x 720) with description a basketball player in a yellow jersey with a ball in his hand.

Dense Captioning List of object descriptions, and their locations in this image: a close up of a basketball (x1,y1,x2,y2) = (63, 535, 190, 663) a basketball player in a yellow jersey (x1,y1,x2,y2) = (55, 19, 502, 700) ...

Object Tagger List of object tags seen in this image: sport, athletic game, basketball, person, sports uniform, ball, sportswear, basketball player, basketball moves, sports equipment, basketball court

Face Detection Detected faces, and their location in this image: man (x1,y1,x2,y2) = (365, 278, 439, 352) man (x1,y1,x2,y2) = (548, 308, 620, 380)

Thought 2 ChatGPT: There are two faces detected in the image. Assistant, are there any celebrities in this image? <imagePath>

Action request 2 Invoke Celebrity Recognition

Action execution 2 Celebrity Recognition This image contains celebrities List of celebrities, and their location in this image: Kobe Bryant (x1,y1,x2,y2)=(365,278,439,352) Paul Pierce (x1,y1,x2,y2)=(548,308,620,380)

Thought 3 ChatGPT: This image contains two celebrities, Kobe Bryant and Paul Pierce. They are both basketball players.

Thought 4 ChatGPT: This question requires Bing Search Assistant, Bing search how many championship rings did Kobe Bryant win in his career? <imagePath>

Action request 3 Invoke Bing Search

Action execution 3 Bing Search Assistant: Bryant, a shooting guard, and O'Neal, a centre, meshed into a remarkably effective combination, and, by the time Bryant was 23, the Lakers had won three consecutive NBA championships. Kobe Bryant After winning their third title in 2002, Bryant and the Lakers encountered difficulties. Kobe Bryant has won a total of 5 rings in 20 seasons...

Thought 5 ChatGPT: Kobe Bryant won 5 championship rings in his career.

一个MM-ReAct的实例，在初始阶段基于ChatGPT调用不同视觉专家生成图片的描述，然后结合用户的问题，让ChatGPT进一步选择调用专家还是生成回复，从而得到最终的答案。

Visual Math and Text Reasoning What is "a" in the first equation? What is "b" in the last equation? The value of "a" in the first equation "5 + a = 7" is 2. The value of "b" in the last equation "34 + b = 39" is 5.

Visual-Conditioned Joke/Meme What is funny about it? The humor in the comic may come from the situation, the characters, or the dialogue. The person is feeling unwell and is asked if they are pregnant or have the coronavirus, which is a current and relevant topic. The humor may come from the absurdity of the situation or the unexpectedness of the question.

Spatial/Coordinate Understanding Where is the frisbee? What is the relative position of the frisbee and the man? The frisbee is located at coordinates (63.13, 133.17). The frisbee is located above the man in the image. The man is falling on the ground while the frisbee is in the air. The frisbee is also closer to the top of the image than the man.

Visual Planning and Prediction What should I do after kneading the dough, before dividing it into pieces? For how long? After kneading the dough, you should put it in a warm place before dividing it into pieces, according to the recipe in the image. You should put the dough in a warm place for 1 hour.

Multi-Image Reasoning How much in total did I spend on travel? How much in total did I pay for taxes? You spent \$1299.90 on airfare and \$43.83 on an Uber ride for a total of \$1343.73 on travel. You paid \$323.23 in total for taxes. This includes \$323.90 for taxes and fees for the flight purchase and \$1.33 for sales tax for the Windmill Family Restaurant.

Multi-Hop Document Understanding Excluding the regular cab, what is the average total fueling cost? The average total fueling cost excluding the Ford F350 (Regular Cab) is \$76.55. This is calculated by taking the sum of the total fueling cost of Toyota RAV4 and Honda Civic and dividing it by 2. (\$82.56 + \$70.55) / 2 = \$76.55.

Open-World Concept Understanding What is the brand of these drinks? Which one is unsweetened? The brand of these drinks is "BETTER THAN MILK". The unsweetened one is "BETTER THAN MILK organic OAT drink".

Video Summarization/Event Localization Can you breakdown this video tutorial to step-by-step instructions, with start and end time? Here's a step-by-step breakdown of the tutorial with start and end time: ... Add Lettuce Add Tomato Add Bacon ...

通过为ChatGPT配备不同的视觉专家，它可以有效解决以上8种任务：视觉数学计算、理解视觉笑话、空间和坐标理解、视觉规划与预测、多图推理等任务。

PaLM-E Describe the following image: A dog jumping over a hurdle at a dog show.

Image Captioning, VQA Describe the following image: A dog jumping over a hurdle at a dog show.

Visual-Conditioned Jokes What is the cat doing in this scene? A: A stove top with a pot on it.

Multi-Image Relationships Where is the coffee pot in the second image? The coffee pot is located in the kitchen with a stove and cabinets. It is a silver coffee pot on a stove and its location is at (884, 330, 930, 700).

Scene Text Reasoning Photo 1, at 10:30 am: Photo 2, at 12:45 pm: Photo 3, at 4:45 pm: Photo 4, at 10:30 am: Photo 5, at 12:45 pm: Photo 6, at 4:45 pm: I forgot, what did I have for lunch, and what time was it? You had a sandwich for lunch at 12:45pm.

采用了一些Case与PaLM-E进行对比，通过对比可知MM-ReAct的效果更优。