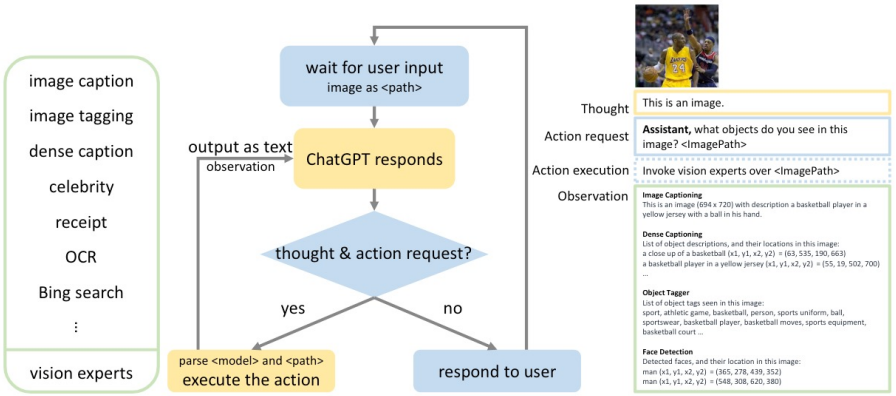
MM-ReAct (arXiv 2023)

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。

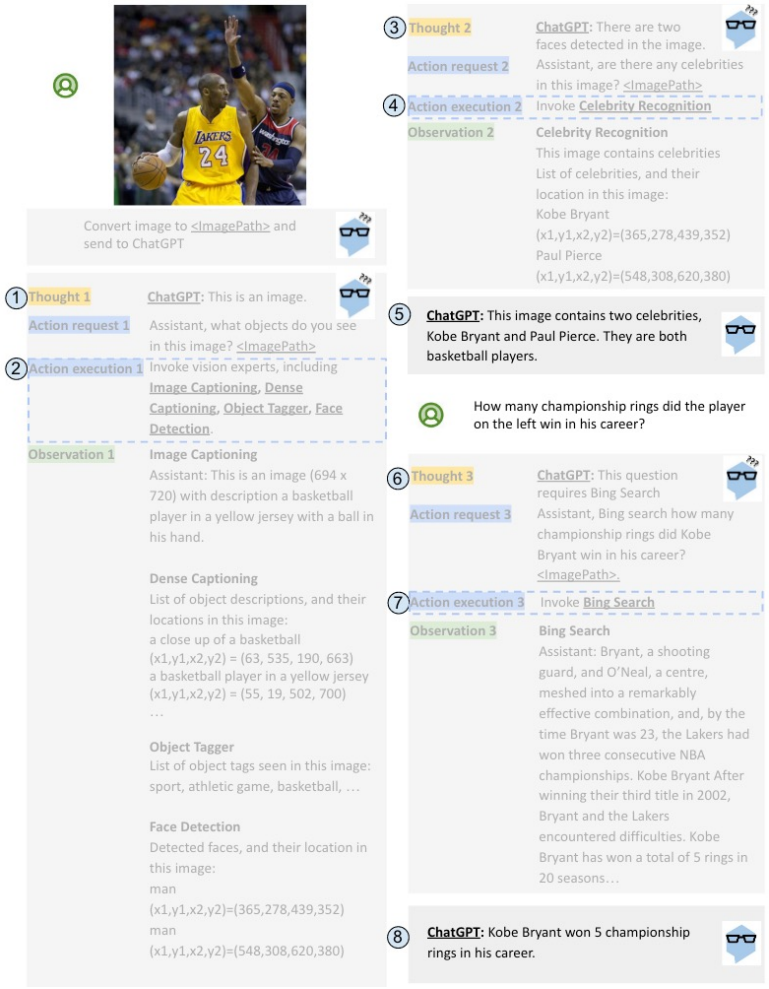
整体思路：MM-REACT定义了一系列高级视觉任务，并引入了一种文本提示设计，可以表示文本描述、文本化的空间坐标，以及以对齐文件名表示的图像和视频等密集视觉信号。



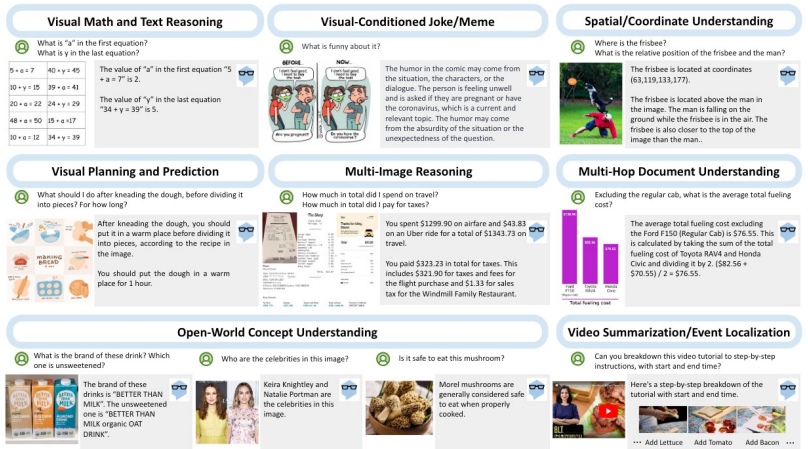
目的：汇聚众多视觉专家，为ChatGPT赋予视觉理解能力。

利用ReAct调用多个视觉模型为ChatGPT形成文本的caption。

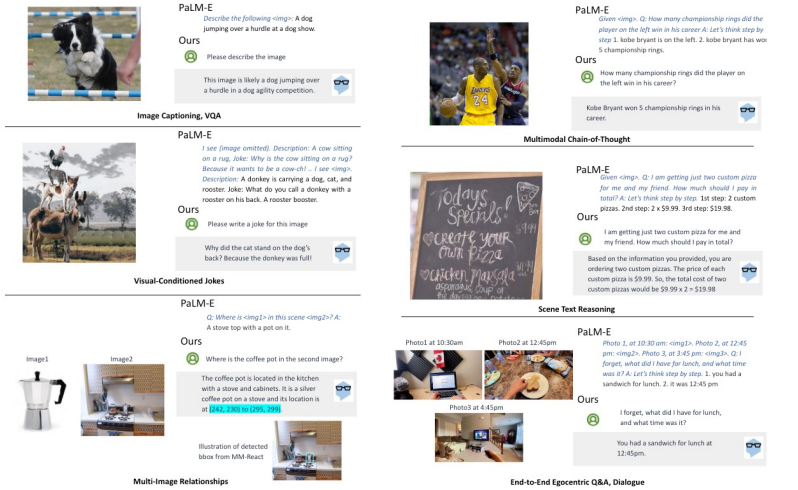
1. **用户输入**：由于ChatGPT无法直接处理图片与视频，因此在输入时将图片与视频用路径代替，并让ChatGPT调用视觉专家模型从不同角度理解图像内容；
2. **ChatGPT回复**：ChatGPT有两种回复，第一种是直接回复文本，第二种是调用视觉专家；
3. **视觉专家**：给定ChatGPT的操作请求，使用正则表达式解析专家名和文件路径，并调用相关操作；视觉专家的输出格式可能是多种形式的，为了方便ChatGPT理解，需要将其转化成文本形式。



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。



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。



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。

http://arxiv.org/abs/2303.11381