arxiv.org/abs/2508.03923

CoAct-1: Computer-using Agents with Coding as Actions

未开源

Linxin Song [1], Yutong Dai [2], Viraj Prabhu [2], Jieyu Zhang [3], Taiwei Shi [1], Li Li [1], Junnan Li [2], Silvio Savarese [2], Zeyuan Chen [2], Jieyu Zhao [1], Ran Xu [2], and Caiming Xiong [2]

[1]University of Southern California, [2]Salesforce Research, [3]University of Washington

简介

本文提出CoAct-1，一种基于orchestrator-worker架构的multi-agent computer use方案。CoAct-1包含3个agent：Orchestrator是main agent，负责拆解用户任务，然后分配子任务给合适的sub-agent执行；GUI Operator用于执行GUI的点击与输入类型的子任务；Programmer则通过生成Python或Bash脚本来完成系统级或复杂的子任务。CoAct-1的核心思想是在GUI agent基础上，将“生成代码”作为一种新的action，对于一些要求精准操作的action就生成相应的代码去执行。这种设计对于程序员来说很容易理解，有的时候与其手动点击界面，不如直接写一段脚本在终端执行，既高效又可靠。

背景

本文属于computer use方向的工作，目的是构建一个接收用户任务(query)后能够自主操作电脑完成任务的agent系统。此类agent可以称为GUI agent、computer use agent或者OS agent。

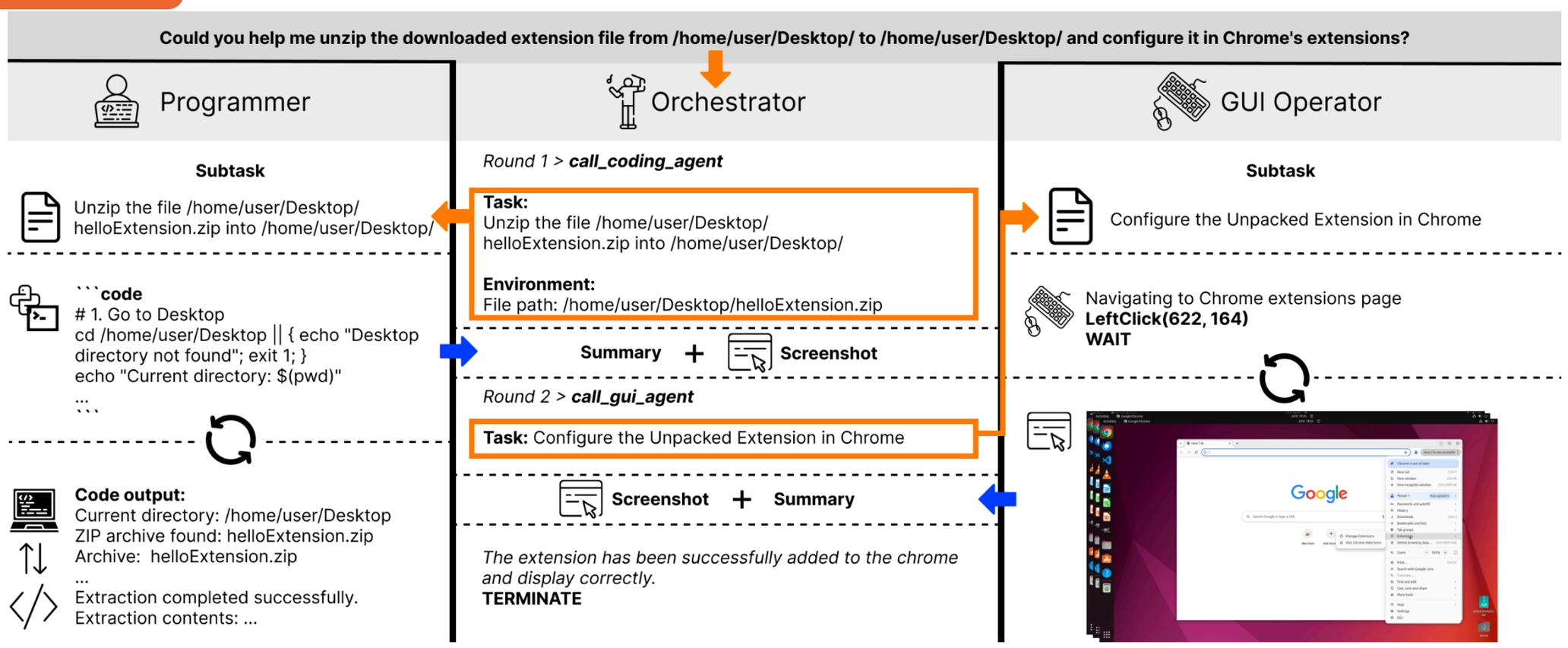
GUI agent背后是多模态模型(mllm)，通过视觉理解和GUI层级的操作完成任务，但是效果不够好，近期有工作为agent加入了planner机制，能提升一定的效果。本文进一步提出，在GUI操作基础上再增加生成代码作为action的机制，有些用鼠标操作不太精确的子任务就通过写代码执行来完成，可以类比前端和后端相互合作执行用户的任务。

实验设置

- 评估数据集：OSWorld benchmark
- orchestrator和programmer背后的llm: o3、o4-mini；GUI operator背后的模型：computer-use-preview

Agent Model	Approach Type	Success Rate
15 steps		
o3 (OpenAI, 2025b)	General Model	9.10
UI-TARS-72B-DPO (Qin et al., 2025)	Specialized Model	24.00
UI-TARS-1.5-7B (Qin et al., 2025)	Specialized Model	25.70
OpenAI CUA 4o (OpenAI, 2025b)	Specialized Model	26.00
Jedi-7B w/ o3 (Xie et al., 2025b)	Agentic Framework	26.80
Claude 3.7 Sonnet (Anthropic, 2025)	General Model	27.10
Claude 4 Sonnet (Anthropic, 2025)	General Model	31.20
Agent S2 w/ Gemini-2.5-Pro (Agashe et al., 2025)	Agentic Framework	34.64
Agent S2.5 w/ o3 (Agashe et al., 2025)	Agentic Framework	39.00
CoAct-1 w/ o3 & o4mini & OpenAI CUA 4o	Agentic Framework	**39.81**

CoAct-1架构



思考

这是我读的第一篇computer use方向的论文，整体框架和deep research中的一些方案类似，最大的区别就是以mllm为主。

另外，我点开论文中的链接，是没有开源代码的，不清楚作者后续是否有开源计划。