基于 GUI Agent 的 ID 与 OOD 在数据集层面划分的思考

Yang

2025 年 7 月 7 日

1 Related Work

根据https://arxiv.org/pdf/2505.12842知 GUI Agent 的 ID 与 OOD 的划分可以定义为如下两个层次：

- Internalization：在熟悉的环境中工作，但是操作的功能之前未见过
- Extrapolation：在未见过的或者说动态变化的环境中工作

希望基于此知识做出的改进：实现对数据集的划分 (如 AndroidControl)，数据集的具体划分：

$$\text{数据集}\begin{cases}\text{训练}\\ \text{测试}\begin{cases}ID\\ OOD\end{cases}\end{cases}$$

判断我们定义的划分是否正确的方法是使用 GEM 模型：在按照我们所定义的划分方式对数据集进行划分之后存在 D_{ID}，由 GEM 模型作出 $(s_i, x_i) \to e_i \to \mu \to d_i$，如果训练集中的 $|l_e(s,x) - \mu|_2 \in \cup_{j=1}^{m}[\mu_j - 3\sigma_j, \mu_j + 3\sigma_j]$ 则认为是 ID，否则为 OOD。先按照我们的划分方式为训练集中的数据贴上 tag 再与 GEM 模型的判断结果进行对比，统计定义的可信度 (但是 GEM 模型的自身误差的叠加也需要考虑)

2 Thought

2.1 对数据集的分析

数据集中通常包含的元素：

- high level instrution

- URL 图形界面
- low level action

往往一个 high level 会对应多个 low level。high level 和 low level 的文本描述显然是区分 ID 和 OOD 的重要指标，另外不同的 URL 也可能会导致 OOD 问题 (比如交换了一些图标的位置导致具体 action 时坐标差异较大，或者使得某些图标消失)

2.2 对于 ID 和 OOD 分类的标准的思考

我认为论文中提及的 embedding distance 是一个很重要的指标，与已知 embedding space 的集中分布的区域。在训练数据训练之后，模型内部会产生一个 embedding space，利用 *GEM* 模型可以构建一个混合高斯分布，通过数据的距离是否落在区间内判断 ID 与 OOD，基于此从直观上做出一些分类：

- 向 Bob 发送邮件通知会议和向 Alice 发送邮件表示感谢 → action 几乎一致，instrution 只有对象和目的的区别，这两者主要影响输入的文本而不影响操作 → 属于 ID
- 用京东购买 XX 商品和用淘宝购买 XX 商品 → action 在逻辑上是一致的，比如点击搜索框，输入商品名称，点击详情页，点击购买。主要差异在于 URL 分布的差异 → 待定，稍后会说明
- 用京东购买 XX 商品和在京东上收藏某个店铺 → URL 界面是一致的，但是无论是 instruction 还是 action 都有很大的差别，比较明显的属于 OOD 的第一类 *to* 理论上应该属于 OOD
- 用京东购买商品和发送邮件 → 差距巨大，属于 OOD 第二类 → 属于 OOD

上述示例描述前者属于训练数据，后者属于测试数据。对于上述分类中待说明的部分进行详细说明：淘宝与京东这个例子在界面上具备较高的相似性 (搜索框，商品详情页和购买按键之间位置较为相似)，因此可以认为两张图形界面的相似度很高，在这一部分事实上很难做出非常准确的划分，依赖于模型对图形的提取和识别能力。但是在数据集层面能做的是比较测试数据与已有数据之间的余弦距离，如果余弦距离低于某一阈值则认为两张图片之间的差异性很大，是 URL 造成的 OOD，这一部分可以结合 *GEM* 模型的判断结果和余弦距离之间的关系来划定阈值的取值。

2.3 划分标准初拟

将一个数据分成两个部分：text(instruction 和 action) 和 visual(URL)，对于 text 可以进行与任务相关的关键词的提取 (如购物，发邮件，下载，卸载，播放音乐) 与 ID 进行对比，如果在 ID 的分布列表中可以进入上述分类的分类一和分类二的情形，因为

我认为他们执行的任务在逻辑上是比较相似的，因为实现某一功能需要完成的基本操作是高度相似的。对于 visual 部分可以将测试数据与已知数据作余弦距离计算，通过设定好的阈值判断是否是 OOD 或是 ID。

对于 text 中关键词提取的关键词列表的形成统计的思考：可以考虑借鉴 Synthesis 的思路，从应用本身出发，比如购物和收藏店铺是购物 APP 可以进行的操作，因此可以作为两个不同的元素，倘若只是列为收藏则可能会和音乐 APP 的收藏冲突，所以具体操作需要视操作环境决定，可以考虑从 action 中打开的图标作为提示补充说明 (但可能也比较麻烦，有待商榷)