**移动通信与编程作业四——大模型能否带来通用机器人**

**孙永鑫 520021911106**

**支持反方观点：大模型不能带来通用机器人**

1. **当下的机器人领域大模型的应用现状**

**1.1大模型能为机器人带来什么？**

研发能够深入了解世界并与之互动的机器人，是机器人领域长期以来的核心挑战。过去数年里，由于传统AI智能化程度较低，训练成本高，泛化能力弱，除了在计算机视觉领域有较多应用之外，在机器人这样动作连续且复杂、有较多物理交互和操作因果性的领域落地情况并不乐观。

基于Internet-Scale数据预训练的语言和图像大模型具备强大的语义理解、内容推理、图像识别、文本/图像生成、代码生成等能力[1]，这些能力对于需要在真实世界里执行多种任务的通用机器人来讲至关重要。

尽管大语言模型（LLMs）和视觉语言模型（VLMs）在机器人领域的应用取得了一定的成果，但其实践效果并不理想。主要原因在于这些模型缺乏对真实物理世界的深入了解，导致其推理输出难以在实际机器人场景中得到有效应用。此外，现有大模型主要关注语义推理和文本提示，而机器人实际需要的是具体的运动指令，如机械臂末端位置指令或底盘移动位置指令等，这两者之间存在明显的差异。

为了解决这些实际问题，来自不同研究机构和企业的人员开展了一系列研究工作：包括直接使用Transformer模型进行端到端的训练、利用机器人技能数据集对LLM进行微调、专注于解决机器人高层决策问题等等。这些努力催生了一系列具身大模型的创新应用。

**1.2几个经典的具身AI(Embodied Artificial Intelligence)大模型**

**1.2.1 PaLM-E: An embodied multimodal language model**

在文献[2]中，为了更好地将图像语言大模型的能力和知识迁移到机器人领域，PaLM-E融合了Google当时最新的大型语言模型 PaLM 和最先进的视觉模型 ViT-22B ，使用文本和其他多模态数据（主要来自于机器人的传感器，比如图像、机器人状态、场景环境信息等）代替之前的纯文本作为输入，输出以文本形式表示的机器人运动指令，进行端到端的训练，从而得到了用于机器人任务的多模态大模型。

下图简明展示了PaLM-E的机器人操作能力，例如，给机器人下达文字指令：“把抽屉里的薯片拿来给我”，模型会输出以下机器人的运动指令：

1.移动到抽屉旁边；

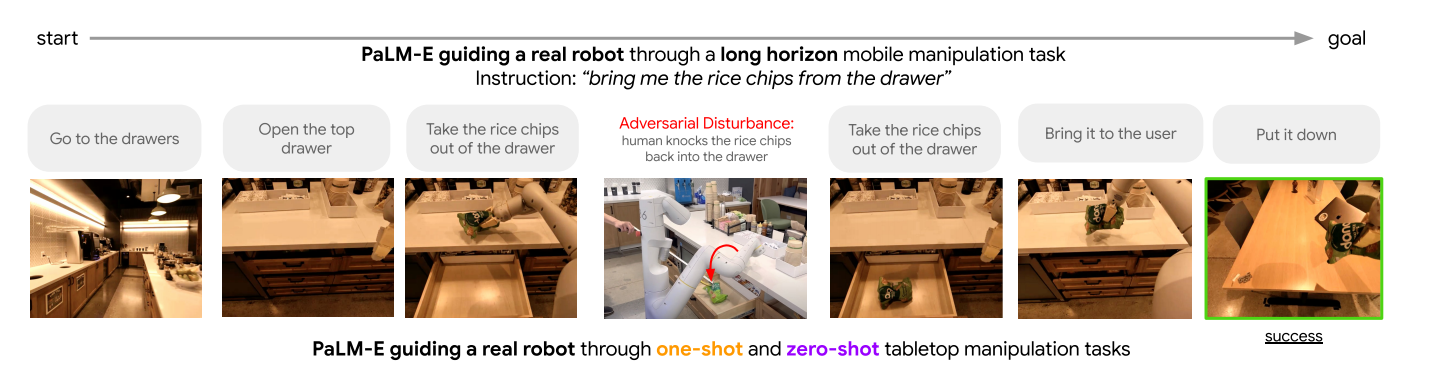
2.打开抽屉；

3.把薯片从抽屉里拿出来；

4.把薯片带到用户旁边；

5.放下薯片；

6.任务结束；



PaLM-E展示了将图像和语言大模型的知识迁移到机器人领域的一条可行之路，有了大模型的通识理解和推理能力，机器人可以将原有能力更好的泛化到新场景中。

**1.2.2 SayCan：Do As I Can, Not As I Say**

根据PaLM-E章节的描述，大模型能够将高级任务分解为多个在语义上符合逻辑的子任务。然而，由于大型语言模型普遍缺乏实际物理世界的经验，因此无法准确评估其输出对环境的影响，也无法获取关于机器人及其执行能力的具体信息。这就意味着，虽然大模型生成的子任务指令在逻辑上是合理的，但在实际场景中，机器人可能无法顺利执行这些任务。例如，当面对任务“请帮忙清理地上的瓜子皮”时，大模型可能会生成“使用吸尘器进行清洁”的建议。然而，如果机器人不具备使用吸尘器的技能，或者家中根本就没有吸尘器，那么这个看似合理的指令就无法得到有效执行。

譬如，对于任务“我把瓜子皮撒地上了，能不能帮忙清理一下？”，大模型输出“使用吸尘器清洁”是符合逻辑的，但实际上执行这项任务的机器人可能还不会使用吸尘器，甚至家里根本就没有吸尘器。

因此，SayCan系统[3]将决定机器人如何执行任务的决策过程分为两个部分：Say和Can。Say部分利用大型语言模型（LLM）生成可供选择的运动指令，而Can部分则根据当前环境条件和机器人能力评估这些指令的可行性。这两部分通过值函数（Value Function）进行整合，共同决定最终应用于实际执行的指令。通过这种方式，SayCan系统确保了机器人在实际操作中的准确性和可靠性。

**1.2.3 RT-2: Robot Transform - 2**

RT-2的主要目标是利用互联网规模数据训练的视觉语言模型（VLM）进行端到端的机器人控制，旨在提升机器人操作的泛化能力和语义推理能力。[4]

与传统的从头训练Transformer模型的方法不同，RT-2选择直接采用现有的VLM模型作为基础模型，并通过更适合机器人任务的fine-tuning方法对其进行微调。简而言之，RT-2利用机器人数据集进行fine-tuning，并确保输出的文本形式为机器人位置指令，类似于VLM（如PaLM-E、PaLI-X）。

此外，为了更好地适应机器人的实际应用，RT-2专注于利用大模型的优点直接生成机器人底层（Low-level）运动指令。通过这种方式，RT-2旨在实现更高效、准确的机器人控制，从而提高机器人在复杂环境中的适应性和操作性能。

RT-2使用机器人技能数据集对已有的LLMs或者VLMs进行微调，可以快速地利用VLM的海量通识能力，大幅提升机器人的任务执行成功率和泛化能力。

# **1.2.4 RoboCat: 自我生成训练数据集**

当前训练机器人大模型/具身大模型面临的主要挑战之一是高质量训练数据的匮乏。在RT-1、RT-2以及PaLM-E等研究中，大部分机器人数据集是通过人工示范和远程操作的方式收集的，这种方法既费时又费力，效率低下。如果能自动生成机器人的训练数据，将极大地节约时间和人力成本。

RoboCat基于DeepMind的最新多模态通用大模型Gato创建，除了具备场景识别、任务推理拆解、机器人控制等功能外，其一大创新在于能够自动生成机器人训练数据集，从而更快地提升其能力。[5]

整个训练过程分为两个阶段：首先使用预先准备的机器人数据集对Gato模型进行初步训练；然后进入自改进训练环节。自改进训练包括以下五个步骤：

1.对于每个新任务或机械臂，通过人工控制机械臂，收集100-1000个演示过程。

2.使用这些演示数据对RoboCat进行微调，为每个新任务或机械臂生成一个新的分支。

3.每个新分支的机器人自主在新任务或机械臂上进行10000次练习，以生成更多训练数据。

4.将人工示教数据和机器人自主生成的数据合并到RoboCat现有的数据集中。

5.使用更新后的数据集训练新版本的RoboCat。

通过这种自改进的训练方式，RoboCat能够不断优化自身的能力，提高对新任务和环境的适应性，从而实现快速进步和持续成长。

目前大模型在机器人领域的现状可以概况为： 理解看到是什么东西、大小、形状、冷热、颜色，以及针对这些物体的基本操作如捡起/扔掉/移动，会简单的数学和逻辑推理用于推进任务。

虽然论文将这些都归为Manipulation，但距离常规理解的操作（从拧螺丝钉到组装宜家家具）还有较大差距。譬如，在RT-2中展示了一些失败的案例，移动笔的时候未考虑圆形的笔会滚动或控制周期太慢无法快速响应笔的运动；移动香蕉的时候未考虑长条形状物体的重心，机器人推动香蕉时无法得到期望的运动等等。

1. **大模型解决机器人专业学习问题难**

# **2.1高质量训练数据缺乏**

在训练机器人执行任务时，我们需要利用多种传感器来感知环境状态，并通过实际动作完成任务。因此，训练用于机器人的大模型需要大量与环境交互的数据集。然而相比图像和自然语言处理领域可以从网上大量获取训练数据或通过人类标注快速低成本地获取数据，用于训练机器人学会新任务的高质量数据非常匮乏。主要原因如下：

1.与CV和NLP相比，机器人执行任务通常需要花费更多时间，这导致收集数据的效率很低；

2.CV和NLP训练过程中只需要处理虚拟信息，而机器人训练过程中会对周围环境造成影响，可能会破坏环境、损坏任务对象，最终造成财务损失，而且这种损失在训练完成之前是不可避免的；

3.机器人保有量还太少，可用于收集训练数据的机器人就更少，进一步加重了数据收集的难度。机器人有不同的实体，有各种各样的形状、大小和形态。机器人实体的多样性意味着可能需要针对机器人的类型来收集数据，这使得数据收集问题变得更加困难。

4.机器人执行任务时，面临的环境和交互内容模态更多，更加复杂，因此需要的数据集规模也比CV和NLP领域要大。

虽然上文中提到的RoboCat可以自主生成机器人训练数据集，但受限于仿真模型的精细度、物理模型准确度、感知数据的准确性等因素，仿真数据与真实数据之间存在一定的差异，使得在仿真环境中只能训练一些对精度要求不高或者弱接触的场景。虚实迁移(Sim-to-Real)等相关领域的研究者也一直在试图努力缩小虚实数据的差异，从而最终达到通过虚拟场景进行大规模的数据收集和训练的目的。

# **2.2 安全性可靠性无法保障**

**2.2.1任务生成的安全性**

在实际环境中使用机器人的一个难点在于，机器人的动作会改变环境本身，而环境的改变又会影响机器人的下一步动作，能否在新的环境里更新任务并顺利执行到底，对于机器人能否被应用到非结构化场景中至关重要。

例如，有一个机器人在给厨房帮忙用蒸箱蒸馒头，蒸之前开蒸箱门往里面放是不需要考虑太多因素的，只要别发生碰撞即可。但是蒸好之后需要打开蒸箱拿出来的时候就需要考虑旁边是否有人，因为蒸箱打开的时候高温蒸汽喷出会对旁边的人造成烫伤，机器人是否能认识到这一点并在生成“打开刚刚使用过的蒸箱”任务动作时，考虑高温蒸汽对人的影响，就是很基本的安全要求了。

“等人离远点再打开蒸箱门”或者“播放语音提醒人员注意高温蒸汽”这两个动作对于机器人来讲很简单，但是能否在合适的时候生成这两个动作指令，需要机器人具备生活的常识。

虽然大模型具备很强的通识能力，但是如何保证每一次生成的任务都符合彼时情境的安全规范，仍然是一个需要持续优化的问题。

**2.2.2操作动作的安全性**

在很多需要专业技能的领域机器人还要关注细微工艺动作是否符合安全规范，譬如在机器人手术中，机器人生成的磨削骨头或者切割软组织的某个细微动作是否符合手术手法要求，是否会对病人造成额外伤害，也是需要慎重考虑的问题。

**2.2.3数据信息的安全性**

首先，模型创建者通过网络抓取增长数据集将受到法院解释服务条款的制约，如美国计算机欺诈和滥用法案(CFAA)，该法案将“未经授权”访问服务器视为犯罪。其次，训练集中包含的许多数据含有版权，受到知识产权法的保护。最后，一些训练数据集可能会违反隐私法。例如，美国伊利诺斯州允许个人起诉不当收集或使用生物特征数据(如视网膜或虹膜扫描、指纹、声纹、手或脸的几何形状扫描等等)。加州消费者保护隐私法案(CCPA)赋予个人“被遗忘的权利”，这引发了模型创建者是否需要从模型中“删除”训练数据的问题。

1. **未来机器人领域大模型的应用展望：有用机器人≠完美机器人**

AI背景的工程师和机器人背景的工程师在应用大模型时有很大的区别：AI背景的工程师试图打造一个具备超强泛化能力的通用机器人AI系统，完成机器人所需的从感知-决策-规划-控制的全流程任务；而机器人背景的则更多把大模型当做工具，期望使用具备一定通用智能的大模型系统让机器人快速学会某些特定技能。

从具体的机器人技能学习角度来看，基于数学模型或者针对特定应用专门训练一个小模型（引入人工示范和主动干预），大概率是比预训练大模型加微调更合适、效率更高、效果更好的方式。

从机器人应用角度来看，大模型是加速器和工具，不是目的。 大模型在机器人的作用仍然是在决策、任务拆解、常识理解上。大模型不具备、也不适合做具备实时性和精确性要求的机器人规划和控制任务。大模型负责常识、决策、推理，模型控制+专家知识负责专业技能，二者各自负责各自擅长的，结合使用是目前看到相对较快的落地方式。

**参考文献：**

1. Zhang C, Zhang C, Zheng S, et al. A complete survey on generative ai (aigc): Is chatgpt from gpt-4 to gpt-5 all you need?[J]. arXiv preprint arXiv:2303.11717, 2023.
2. Driess D, Xia F, Sajjadi M S M, et al. Palm-e: An embodied multimodal language model[J]. arXiv preprint arXiv:2303.03378, 2023.
3. Ahn M, Brohan A, Brown N, et al. Do as i can, not as i say: Grounding language in robotic affordances[J]. arXiv preprint arXiv:2204.01691, 2022.
4. Brohan A, Brown N, Carbajal J, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control[J]. arXiv preprint arXiv:2307.15818, 2023.
5. Bousmalis K, Vezzani G, Rao D, et al. RoboCat: A Self-Improving Foundation Agent for Robotic Manipulation[J]. arXiv preprint arXiv:2306.11706, 2023.