大模型在机器人中的应用

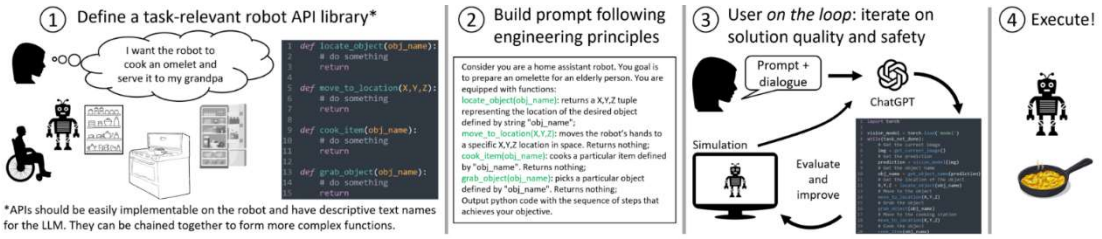
在人类文明的初始阶段，远远早于“人工智能”这个词汇被正式提出之时，出于对神权的敬仰和对于未知的想象，中外均已出现了智能机器人的思想。我国春秋时期名著《列子·汤问》中的《偃师献伎》有记载——偃师向王进献一具木偶，虽然木偶看上去在被偃师控制，但实际上有自己的意识，可以根据外界的变化做出智能的反映。与《列子》几乎同期，古罗马诗人奥维德在《变形记》中也有相似的思想体现：塞浦路斯国王皮格马利翁制作了一个象牙雕塑并祈求爱神维纳斯赋予这件雕塑生命，以作为他的妻子。事实上，从以上神话故事不难看出，研发可以学习了解这个世界并能灵活自主地与世界进行交互的仿人机器是古今中外共同的追求。

大模型时代前的大部分机器人与其说“人工智能”，不如说是一种可编程的专用设备。大规模的工业机械手臂在流水线上日复一日地重复着枯燥乏味的预编程动作，而为数不多的“自主决策”机器人的可泛化性也受到极大的局限——增加使用场景需要专业的开发人才耗费大量时间精力进行设计、编程。而随着大模型的出现，以往高成本、高门槛的人工智能开发，变成了“预训练大模型+特定任务微调”的形式，大幅提高了机器人的泛化能力，使得新功能的植入变得简单；同时配合语言、图像等大模型，智能机器人打破了文字-图片-动作之间的壁垒，使得各类任务的执行更为流畅，机器人真正可以从任务级的层面理解人的意图，并且自主尝试完成。可以说，大模型＋机器人是一个划时代的结合。

研究现状

在机器人决策与控制策略学习方面，大模型可以对模仿学习起到辅助作用，例如 Voltron[1]以语言作为条件，进行视觉重建下的局部空间表征，并将基于视觉的语言生成用于捕获语义表征。MimicPlay[2]提出了一种分层模仿学习算法，从人类游戏数据中学习潜在空间中的高级计划，从少量远程操作演示中学习低级运动命令，大语言模型的辅助下能够根据一个人类视频演示执行新任务。MUTEX[3]进一步探索了在视频、图像、文本和音频的多模态任务规范中学习统一策略，通过跨模态学习在单模态基线上显示出改进的策略性能；大模型还可以辅助强化学习，例如 Palo 等人[4]提出了通过集成大型语言模型和视觉语言模型来创建更统一的强化学习框架的方法。

在机器人的任务规划中，大语言模型提出了新的思路，例如 SayCan[5]使用大语言模型发布适配于指令集的指令，直接通过自然语言进行高级任务规划。ProgPrompt[6]的作者介绍了一种使用大语言模型直接生成动作序列的提示方法。微软团队提出使用 ChatGPT 生成机器人的高层控制代码[7]。



ChatGPT 控制机器人[7]

事实上，视觉大模型也对机器人的发展起到了决定性的作用，LM-Nav[8]通过语言大模型和视觉大模型结合，从人类的自然语言指令中提取具有视觉特征的参照物，提取其中的关键地标信息的文本，然后通过视觉-文本多模态模型找到出现对应地标图片，机器人可以根

据这些图片生成导航路径，自动寻路。VoxPoser[9]利用大型语言模型来生成代码，这些代码与视觉语言模型交互，以提取一系列 3D 功能图和约束图。PaLM-E[10] 是个多模态的大模型，不仅能理解文本，还能理解图片，因此可以让机器人自主执行任务——人类仅仅给出类似“将方块按颜色分类到角落”的自然语言指令，PaLM-E 可以自主进行任务拆解并推动各颜色方块到各个角落。

现有工作的局限性

即使当前大模型在各类机器人任务上表现优异，但是仍然存在很多不足。

首当其冲的便是缺少数据——机器人的操作、定位、导航等数据稀少，同时数据格式不尽相同、机器人的种类以及任务类型存在极大的多样性，如何使用这些数据执行自监督训练也是一个值得思考的问题。

其次是不确定性的量化问题，大模型本身还处于探索阶段，实例层面存在大量的不确定性（例如语言歧义等），同时还有分布层面的不确定性和分布移位问题，尤其是闭环的机器人部署引起的分布移位问题。

除以上外，对大模型训练的机器人进行安全评估是难以确定的。大模型源于神经网络，复杂的结构难以进行数学化的理解，因此在部署之前、更新过程中、工作过程中的安全问题预推演是一个重要的问题。

最后，大模型由于其庞大的结构，势必会影响机器人的实时性能，同时也会带来硬件上的高需求。如何应对基础模型推理时间长、如何降低推理演算成本也是限制当前大模型机器人的瓶颈问题。

我的观点

我认为大模型在机器人中的应用是人工智能领域最重要的进展之一。通过结合大模型的强大语言、图像理解和生成能力以及机器人的感知、决策和执行能力，我们可以实现更加智能化、自然化的机器人系统。

大模型极大地增强了机器人的语言交互能力，使得它们能够理解人类的自然语言，识别语义和语境，并以自然的方式进行回应。机器人因此能够更好地与人类进行沟通和协作，提高了用户体验和工作效率，机器人正在向人类所曾畅想的方向接近；大模型为机器人赋予了更强的智能决策能力，通过模型对大量数据的学习和分析，机器人可以从任务级的高度理解环境和任务要求，自主生成更为准确和智能的决策，在复杂的环境中进行自主导航、物体识别、路径规划等任务，提高工作的自动化程度和效率；此外，大模型还使机器人具有了较强的可泛化性能，其带来的知识获取和迁移的能力使得机器人能够不断地提升自己的能力，适应不断变化的需求和环境。

但是，大模型和机器人的适配还需要克服诸多挑战，例如模型的计算资源消耗、数据隐私保护等。只有克服了这些挑战，才能更好地发挥大模型在机器人中的作用，为人类生活和工作带来更多的便利和效益。

总而言之，大模型在机器人上应用是一个划时代的想法，使得人类更有望了结其古往今来对智能机器人的夙愿，相信在不久的将来，大模型和机器人都会有越来越多令人激动的发明、发现蓬勃而出。

[1] Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. In RSS, 2023.
[2] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. MimicPlay: Long-horizon imitation learning by watching human play. In CoRL, 2023
[3] Rutav Shah, Roberto Mart´ın-Mart´ın, and Yuke Zhu. MUTEX: Learning unified policies from multimodal task specifications. In CoRL, 2023.
[4] Norman Di Palo, Arunkumar Byravan, Leonard Hasenclever, Markus Wulfmeier, Nicolas Heess, and Martin Riedmiller. Towards a unified agent with foundation models. In Workshop on Reincarnating Reinforcement Learning at ICLR 2023, 2023.
[5] Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. Do as I can, not as I say: Grounding language in robotic affordances. In CoRL, pages 287–318. PMLR, 2023.
[6] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. ProgPrompt: Generating situated robot task plans using large language models. In ICRA, pages 11523–11530. IEEE, 2023.
[7] Vemprala S, Bonatti R, Bucker A, et al. Chatgpt for robotics: Design principles and model abilities[J]. Microsoft Auton. Syst. Robot. Res, 2023, 2: 20.
[8] Shah D, Osiński B, Levine S. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action[C]//Conference on Robot Learning. PMLR, 2023: 492-504.
[9] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. VoxPoser: Composable 3D value maps for robotic manipulation with language models. In CoRL, 2023.
[10] Driess D, Xia F, Sajjadi M S M, et al. Palm-e: An embodied multimodal language model[J]. arXiv preprint arXiv:2303.03378, 2023.