多模态大语言模型在机器人操作中的应用综述

张三

January 10, 2025

Abstract

随着人工智能技术的飞速发展，多模态大语言模型（MLLMs）在机器人操作领域的应用日益广泛。本文综述了近年来多模态大语言模型与机器人操作结合的研究进展，包括 PaLM-E、ManipVQA、ManipLLM 等模型的应用，以及视觉-语言规划在机器人任务中的应用，如 GPT-4V 在机器人视觉-语言规划中的应用和 LEO 在 3D 世界中的具身通用代理。此外，本文还探讨了模仿学习与数据生成技术，包括 MimicPlay、Vid2Robot、Universal Manipulation Interface、DexCap、HIRO Hand、RoboGen 和 MimicGen 等系统的研究进展。最后，本文讨论了自适应聚类变换器在端到端目标检测中的应用，以及多模态大语言模型在机器人操作领域面临的技术挑战和未来发展方向。

1 引言

近年来，随着人工智能技术的飞速发展，多模态大语言模型（Multimodal Large Language Models, MLLMs）在自然语言处理、计算机视觉等领域取得了显著的成就。特别是在机器人操作领域，MLLMs 的应用为机器人提供了更高级的理解和执行自然语言指令的能力，极大地推动了机器人技术的发展。然而，尽管 MLLMs 在理解和生成自然语言方面表现出色，但在处理涉及复杂物理交互和精确操作的机器人任务时，仍面临诸多挑战。这些挑战主要包括对机器人操作环境的理解、物理概念的把握、以及从多模态输入中提取有用信息的能力。

为了解决这些问题，研究者们提出了多种创新的方法和技术，旨在将 MLLMs 与机器人操作系统更紧密地结合起来。这些方法不仅包括改进模型架构和训练策略，还涉及到从视觉、语言到动作的多模态数据的融合与理解。通过这些努力，机器人能够更好地理解人类的指令，更准确地执行复杂

的操作任务，从而在工业生产、家庭服务、医疗护理等多个领域展现出巨大的应用潜力。

本文旨在综述近年来多模态大语言模型在机器人操作领域的应用进展，探讨这些技术如何帮助机器人更好地理解和执行任务，以及面临的挑战和未来的发展方向。通过对相关研究的回顾和分析，本文希望能够为读者提供一个全面的视角，理解 MLLMs 在机器人操作中的重要作用，以及它们如何推动机器人技术向更高层次发展。

2 多模态大语言模型与机器人操作的结合

多模态大语言模型（Multimodal Large Language Models, MLLMs）在机器人操作领域的应用，标志着人工智能技术向更加复杂和实用的方向迈进。这些模型通过整合视觉、语言和其他传感器数据，为机器人提供了理解和执行自然语言指令的能力，极大地扩展了机器人的应用范围和效率。本节将详细介绍几种关键的多模态大语言模型及其在机器人操作中的应用。

2.1 PaLM-E: 一种具身多模态语言模型

PaLM-E 模型是一种创新的具身多模态语言模型，它通过将现实世界的连续传感器模态直接整合到语言模型中，建立了词汇与感知之间的直接联系。这种模型的输入是多模态句子，这些句子交织了视觉、连续状态估计和文本输入编码。PaLM-E 模型通过端到端的训练，与预训练的大型语言模型结合，能够处理包括序列机器人操作规划、视觉问答和图像描述在内的多种具身任务。PaLM-E-562B 模型，拥有 5620 亿参数，不仅在机器人任务上表现出色，还在 OK-VQA 等视觉语言任务上达到了最先进的性能，同时随着规模的增加，其通用语言能力也得到了保持。

2.2 ManipVQA: 注入机器人可操作性和物理基础信息的多模态大语言模型

ManipVQA 框架提出了一种新颖的方法，通过视觉问答（VQA）格式向多模态大语言模型注入以操作为中心的知识。这种方法包括工具检测、可操作性识别和物理概念的广泛理解。通过精心策划的交互对象图像数据集，ManipVQA 挑战了机器人在工具检测、可操作性预测和物理概念理解方面的能力。通过统一的 VQA 格式和精心设计的微调策略，ManipVQA 有效地将机器人特定知识与 MLLMs 固有的视觉推理能力结合起来，展示了在机器人模拟器和各种视觉任务基准测试中的强大性能。

2.3 ManipLLM: 面向对象中心机器人操作的具身多模态大语言模型

ManipLLM 引入了一种创新的机器人操作方法，利用多模态大语言模型的强大推理能力来增强操作的稳定性和泛化能力。通过微调注入的适配器，ManipLLM 保留了 MLLMs 固有的常识和推理能力，同时赋予它们操作的能力。这种方法的核心在于引入的微调范式，包括对象类别理解、可操作性先验推理和面向对象的姿态预测，以激发 MLLM 在操作中的推理能力。在推理过程中，ManipLLM 利用 RGB 图像和文本提示来预测末端执行器的姿态，并在建立初始接触后，引入主动阻抗适应策略以闭环方式规划即将到来的路径点。在现实世界中，ManipLLM 设计了测试时适应（TTA）策略，使模型能够更好地适应当前现实世界的场景配置。在模拟器和现实世界中的实验显示了 ManipLLM 的优异性能。

3 视觉-语言规划在机器人任务中的应用

近年来，随着多模态大语言模型（MLLMs）和视觉-语言模型（VLMs）的快速发展，它们在机器人任务中的应用越来越广泛。特别是在视觉-语言规划领域，这些模型展现出了巨大的潜力。通过将视觉信息与语言理解相结合，机器人能够更好地理解环境，进行有效的任务规划和执行。

3.1 Look Before You Leap: 揭示 GPT-4V 在机器人视觉-语言规划中的力量

在《Look Before You Leap: Unveiling the Power of GPT-4V in Robotic Vision-Language Planning》一文中，作者提出了一种新型的机器人视觉-语言规划方法 ViLa。该方法利用视觉-语言模型（VLMs）生成一系列可执行的步骤，直接整合感知数据到其推理和规划过程中，从而实现对视觉世界中常识知识的深刻理解，包括空间布局和物体属性。ViLa 支持灵活的多模态目标指定，并自然地融入视觉反馈。通过在真实机器人和模拟环境中的广泛评估，ViLa 展示了其在开放世界操作任务中的有效性，超越了现有的基于大语言模型（LLMs）的规划器。

3.2 LEO: 3D 世界中的具身通用代理

《An Embodied Generalist Agent in 3D World》介绍了 LEO，一个在 3D 世界中表现出色的具身多模态通用代理。LEO 通过两阶段训练：3D 视觉-语言（VL）对齐和 3D 视觉-语言-动作（VLA）指令调优，实现了在 3D

世界中的感知、接地、推理、规划和行动。LEO 的引入解决了现有模型在 3D 输入和任务上的限制，通过大规模数据集的收集和精心设计的 LLM 辅助管道，LEO 在 3D 字幕、问答、具身推理、导航和操作等广泛任务中展现了卓越的能力。LEO 的成功不仅为未来的具身通用代理开发提供了宝贵的见解，也为实现更高级的机器人智能铺平了道路。

4 模仿学习与数据生成

模仿学习与数据生成在机器人技术领域扮演着至关重要的角色，尤其是在提高机器人操作技能和任务执行能力方面。通过模仿学习，机器人能够从人类示范中学习复杂的操作技能，而数据生成技术则为机器人学习提供了丰富和多样化的训练数据。这些技术的发展不仅加速了机器人学习过程，还提高了学习效率和效果。

4.1 MimicPlay: 通过观看人类游戏进行长视距模仿学习

MimicPlay 是一种创新的模仿学习框架，它通过观看人类自由与环境交互的视频来学习长视距任务。这种方法的核心在于利用人类游戏数据中的丰富物理交互信息，即使人类和机器人的形态不同，这些信息也能有效指导机器人策略学习。MimicPlay 采用分层学习框架，从人类游戏数据中学习潜在计划，以指导低级别的视觉运动控制。这种方法在 14 个长视距操作任务的系统评估中显示出优越的性能，包括任务成功率、泛化能力和对干扰的鲁棒性。

4.2 Vid2Robot: 使用跨注意力变换器的端到端视频条件策略学习

Vid2Robot 是一种端到端的视频条件策略学习方法，它通过观察人类演示视频来学习机器人操作任务。这种方法利用大规模的视频-机器人轨迹对数据集，学习人类和机器人动作的统一表示。Vid2Robot 使用跨注意力变换器层在视频特征和当前机器人状态之间产生动作，以执行与视频中展示的相同任务。通过辅助对比损失，Vid2Robot 能够更好地对齐提示和机器人视频表示，从而产生更优的策略。

4.3 Universal Manipulation Interface: 无需野外机器人的野外机器人教学

Universal Manipulation Interface（UMI）是一种数据收集和策略学习框架，它允许直接从野外人类演示中转移技能到可部署的机器人策略。UMI 通过手持夹具和精心设计的接口实现便携、低成本和信息丰富的数据收集。UMI 的策略接口设计考虑了推理时延迟匹配和相对轨迹动作表示，使得学习到的策略硬件无关，并可在多个机器人平台上部署。UMI 框架解锁了新的机器人操作能力，允许通过仅改变每个任务的训练数据来实现零样本泛化的动态、双手、精确和长视距行为。

4.4 DexCap: 可扩展和便携的动捕数据收集系统

DexCap 是一种便携的手部动作捕捉系统，旨在通过从人类手部动作数据中学习来赋予机器人人类般的灵巧性。DexCap 提供基于 SLAM 和电磁场的精确、抗遮挡的手腕和手指运动跟踪，以及环境的 3D 观察。利用这些丰富的数据集，DexIL 采用逆运动和基于点云的模仿学习来无缝复制人类动作。DexCap 还提供可选的人类在环校正机制，以在策略执行期间进一步改进任务性能。

4.5 HIRO Hand: 一种用于手把手模仿学习的可穿戴机器人手

HIRO Hand 是一种创新的可穿戴灵巧手，它集成了专家数据收集和灵巧操作的实施。HIRO Hand 使操作者能够利用自己的触觉反馈来确定适当的力、位置和动作，从而更准确地模仿专家的动作。通过开发非学习和基于视觉行为克隆的控制器，HIRO Hand 成功实现了抓取和手中操作能力。

4.6 RoboGen: 通过生成模拟释放无限数据用于自动化机器人学习

RoboGen 是一种生成式机器人代理，它通过生成模拟自动学习多样化的机器人技能。RoboGen 利用基础模型和生成模型的最新进展，自动生成多样化的任务、场景和训练监督，从而以最小的人类监督扩展机器人技能学习。RoboGen 的自我引导的提出-生成-学习循环使机器人代理能够提出有趣的任务和技能，生成相应的模拟环境，并学习策略以获取提出的技能。

4.7 MimicGen: 使用人类示范进行可扩展机器人学习的数据生成系统

MimicGen 是一种系统，用于从少量人类示范中自动合成大规模、丰富的数据集。通过将示范适应到新上下文，MimicGen 从约 200 个人类示范中生成了超过 50K 的示范，涵盖了 18 个任务。机器人代理可以通过模仿学习在这个生成的数据集上有效训练，以在长视距和高精度任务中实现强大的性能。MimicGen 数据的有效性和效用与收集额外的人类示范相比具有优势，使其成为扩展机器人学习的有力且经济的方法。

5 技术进展与挑战

在机器人操作领域，技术的进步和面临的挑战是多方面的。特别是在多模态大语言模型（MLLMs）与机器人操作的结合方面，近年来取得了显著的进展。然而，这些进展也伴随着一系列挑战，特别是在提高模型的泛化能力、减少计算资源消耗以及增强模型对复杂环境的适应能力方面。

5.1 自适应聚类变换器在端到端目标检测中的应用

自适应聚类变换器（Adaptive Clustering Transformer, ACT）是一种创新的变换器变体，旨在减少高分辨率输入的计算成本。在机器人操作中，目标检测是一个关键任务，它要求模型能够准确地识别和定位环境中的对象。传统的目标检测方法，如 Faster-RCNN，虽然性能优异，但在计算资源消耗方面存在较大问题。ACT 通过使用局部敏感哈希（Locality Sensitive Hashing, LSH）自适应地聚类查询特征，并利用原型-键交互近似查询-键交互，从而将自注意力机制中的二次复杂度 $\mathrm{O(N^2)}$ $O(NK)$ K

ACT 的引入为机器人操作中的目标检测任务提供了一种高效的解决方案。通过减少对高计算资源的依赖，ACT 使得在资源受限的机器人平台上实现实时目标检测成为可能。此外，ACT 的可扩展性和灵活性也为处理复杂和动态环境中的目标检测任务提供了新的可能性。然而，尽管 ACT 在减少计算成本方面取得了显著成效，但在处理极端复杂场景和高度动态环境时，其性能仍有待进一步提高。此外，如何在不牺牲检测准确率的前提下进一步优化 ACT 的计算效率，也是未来研究的一个重要方向。

总之，自适应聚类变换器在端到端目标检测中的应用展示了在机器人操作领域实现高效、准确目标检测的潜力。随着技术的不断进步和优化，预计 ACT 及其衍生技术将在未来的机器人操作系统中发挥更加重要的作用。

6 结论与未来方向

本文综述了多模态大语言模型在机器人操作中的应用，包括 PaLM-E、ManipVQA、ManipLLM 等模型，以及视觉-语言规划在机器人任务中的应用，如 Look Before You Leap 和 LEO。此外，还探讨了模仿学习与数据生成领域的最新进展，包括 MimicPlay、Vid2Robot、Universal Manipulation Interface、DexCap、HIRO Hand、RoboGen 和 MimicGen 等技术。这些技术的发展不仅推动了机器人操作能力的提升，也为机器人学习提供了新的方法和工具。

尽管取得了显著进展，但仍面临诸多挑战。首先，多模态大语言模型在处理复杂、动态环境中的实时决策和规划时，仍存在效率和准确性的问题。其次，模仿学习和数据生成技术虽然在减少对大量人类示范数据的依赖方面取得了进展，但如何进一步提高生成数据的质量和多样性，以及如何更有效地利用这些数据进行机器人学习，仍是亟待解决的问题。此外，机器人操作中的安全性和伦理问题也需得到更多关注。

未来研究方向包括但不限于：开发更加高效和鲁棒的多模态大语言模型，以更好地支持复杂环境下的机器人操作；探索新的模仿学习和数据生成方法，以提高机器人学习的效率和效果；加强机器人操作安全性和伦理问题的研究，确保技术的健康发展。随着技术的不断进步和跨学科合作的深入，预计未来机器人操作将更加智能、灵活和安全，为人类社会带来更多便利和价值。

References

[1] Driess, Danny and Xia, Fei and Sajjadi, Mehdi SM and Lynch, Corey and Chowdhery, Aakanksha and Ichter, Brian and Wahid, Ayzaan and Tompson, Jonathan and Vuong, Quan and Yu, Tianhe and others. *Palm-e: An embodied multimodal language model*. arXiv preprint arXiv:2303.03378, 2023.

[2] Huang, Siyuan and Ponomarenko, Iaroslav and Jiang, Zhengkai and Li, Xiaoqi and Hu, Xiaobin and Gao, Peng and Li, Hongsheng and Dong, Hao. *Manipvqa: Injecting robotic affordance and physically grounded information into multi-modal large language models*. arXiv preprint arXiv:2403.11289, 2024.

[3] Li, Xiaoqi and Zhang, Mingxu and Geng, Yiran and Geng, Haoran and Long, Yuxing and Shen, Yan and Zhang, Renrui and Liu, Jiaming and

Dong, Hao. *Manipllm: Embodied multimodal large language model for object-centric robotic manipulation*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18061–18070, 2024.

[4] Hu, Yingdong and Lin, Fanqi and Zhang, Tong and Yi, Li and Gao, Yang. *Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning*. arXiv preprint arXiv:2311.17842, 2023.

[5] Huang, Jiangyong and Yong, Silong and Ma, Xiaojian and Linghu, Xiongkun and Li, Puhao and Wang, Yan and Li, Qing and Zhu, Song-Chun and Jia, Baoxiong and Huang, Siyuan. *An embodied generalist agent in 3d world*. arXiv preprint arXiv:2311.12871, 2023.

[6] Wang, Chen and Fan, Linxi and Sun, Jiankai and Zhang, Ruohan and Fei-Fei, Li and Xu, Danfei and Zhu, Yuke and Anandkumar, Anima. *Mimicplay: Long-horizon imitation learning by watching human play*. arXiv preprint arXiv:2302.12422, 2023.

[7] Jain, Vidhi and Attarian, Maria and Joshi, Nikhil J and Wahid, Ayzaan and Driess, Danny and Vuong, Quan and Sanketi, Pannag R and Sermanet, Pierre and Welker, Stefan and Chan, Christine and others. *Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers*. arXiv preprint arXiv:2403.12943, 2024.

[8] Chi, Cheng and Xu, Zhenjia and Pan, Chuer and Cousineau, Eric and Burchfiel, Benjamin and Feng, Siyuan and Tedrake, Russ and Song, Shuran. *Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots*. arXiv preprint arXiv:2402.10329, 2024.

[9] Wang, Chen and Shi, Haochen and Wang, Weizhuo and Zhang, Ruohan and Fei-Fei, Li and Liu, C Karen. *Dexcap: Scalable and portable mocap data collection system for dexterous manipulation*. arXiv preprint arXiv:2403.07788, 2024.

[10] Wei, Dehao and Xu, Huazhe. *A wearable robotic hand for hand-over-hand imitation learning*. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 18113–18119, 2024.

[11] Wang, Yufei and Xian, Zhou and Chen, Feng and Wang, Tsun-Hsuan and Wang, Yian and Fragkiadaki, Katerina and Erickson, Zackory and

Held, David and Gan, Chuang. *Robogen: Towards unleashing infinite data for automated robot learning via generative simulation*. arXiv preprint arXiv:2311.01455, 2023.

[12] Mandlekar, Ajay and Nasiriany, Soroush and Wen, Bowen and Akinola, Iretiayo and Narang, Yashraj and Fan, Linxi and Zhu, Yuke and Fox, Dieter. *Mimicgen: A data generation system for scalable robot learning using human demonstrations*. arXiv preprint arXiv:2310.17596, 2023.

[13] Zheng, Minghang and Gao, Peng and Zhang, Renrui and Li, Kunchang and Wang, Xiaogang and Li, Hongsheng and Dong, Hao. *End-to-end object detection with adaptive clustering transformer*. arXiv preprint arXiv:2011.09315, 2020.