大模型可以带来通用机器人

舒文 520021911192

近些年，OpenAI 开发的 ChatGPT 使大型模型成为了公众关注的焦点。这些模型在多个领域显示出了惊人的能力，比如 ChatGPT 在自然语言处理上的高效表现，以及 AI 绘画创作出的令人叹为观止的艺术作品。作为一种革命性工具，大型模型已经渗透到生产和日常生活的各个方面，开始在多个层面上协助人类。机器人技术也顺应这一趋势，成为热门领域之一。开发一种能够理解世界并与之互动，精确操控物体的机器人，一直是机器人领域长期以来的挑战。过去几年，虽然许多公司在机器人技术上投入了“AI 赋能”，但之前的 AI 技术智能水平有限，训练成本高昂，泛化能力不足。尽管在机器人视觉领域取得了一些进展，但在动作复杂、涉及物理交互和操作因果性的领域，成果并不令人满意。然而，大模型的出现带来了新的希望。考虑到大型模型在 NLP 和 CV 领域的突破，这种成功模式很可能也将促进机器人技术朝着这一目标迈进。在探索机器人技术时，是否可以简单地应用大模型的策略引起了进一步思考。虽然大型语言模型（LLM）和视觉语言模型（VLM）在感知和决策方面取得了显著进步，但它们对机器人的实际操作影响有限。那么，我们该如何提高机器人执行任务的可靠性呢？机器人技术的一个关键挑战在于其执行层面的局限性以及对高度可靠性的需求。这引出了一个问题：在 CV 和 NLP 领域取得成功的大模型范式是否同样适用于机器人技术？

一篇 Nishanth J. Kumar 的在 Twitter 上的文章[1]介绍了一场 CoRL2023 一个

Workshop：Deployable@CoRL2023 的一个辩论环节，在这场辩论会针对 Is scaling enough to deploy general-purpose robots?（大模型方法是否适用于机器人行业？）进行了激烈的讨论，其中很多论点对我们都很有冲击力和可以学习的地方，这里结合正方的几个观点和一些论文进行讨论。

一个观点是 Scaling 在 CV 和 NLP 领域被验证有非常好的效果[2, 3]，那么在机器人领域也同样可以取得好的效果。在这个方面，Brohan，Anthony 等人[4]介绍了一种名为 Robotics Transformer 的模型类别，它展示了有希望的可扩展模型属性。论文中验证了不同模型类别的泛化能力，作为数据规模、模型规模和数据多样性的函数。研究基于在真实机器人上执行真实任务的大规模数据收集。这项研究表明，通过从大型、多样化的任务非特定数据集中转移知识，现代机器学习模型可以在机器人学习领域以零样本或小型特定任务数据集实现高水平的性能；而 Yu Tianhe 等[5] 讨论了在机器人学习中使用大规模数据的重要性，并提出了一种名为 ROSIE 的方法，利用在计算机视觉和自然语言处理中广泛使用的文本到图像的基础模型来获得机器人学习的有意义数据。这种方法利用最先进的文本到图像扩散模型，在现有的机器人操纵数据集上执行积极的数据增强，包括使用文本指导对未见物体的操作、背景和干扰因素进行填充。实验表明，这样训练的操作策略能够解决完全未见的任务，并在新对象和干扰因素方面表现出更强的鲁棒性。通过在极大的数据语料库上训练大型模型，取得了前所未有的惊人效果以及“涌现”能力，GPT4-V 和 SAM 这样最新模型的效果有目共睹。训练大型模型的基本方法是通用的，它并不是针对 NLP 或 CV 领域独有的，既然它被证明在一些领域有效，那么它理应对其他领域（比如机器人）也有效，例如在 Microsoft Research 的 ChatGPT for Robotics 这篇工作中，研究者使用基于 GPT3.5 的

ChatGPT 生成机器人的高层控制代码，让操作者可以通过自然语言和机器人交流并控制机械臂、无人机、移动机器人等各种形态的机器人[6]。

目前已经有很多进展可以证明 Scaling 在机器人领域很可能会有效。基于 Brohan，Anthony 等人介绍的 RT-1 模型[4]，DeepMInd 近期完成了 RT-2 模型[7]，可以证明单一模型在大量机器人数据上训练可以得到泛化能力；Lake, Brenden M 等[8]提出了一种称为组合性元学习 (Meta-learning for Compositionality，MLC) 的新方法，该方法可以提高 ChatGPT 等工具进行组合泛化的能力。实验结果表明，MLC 方法不仅优于现有方法，还表现出人类水平的系统泛化（systematic generalization，SG）能力，在某些情况下甚至优于人类。组合泛化能力也是大型语言模型（LLM）有望实现通用人工智能（AGI）的基础。这项研究表明 AI 模型是可以具备较强的组合泛化能力的，具有里程碑意义；Russ Tedrkake 指出最近的 Diffusion Policies 论文也显示了令人惊讶的能力；Sergey Levine 指出他的团队在构建和部署导航用途的机器人通用基础模型方面取得了不错的进展。虽然这些工作都比较初期，使用相对较小的模型和相对较少的数据进行训练（和 GPT4 比），但似乎有些苗头是扩大这些模型和数据集会指向更好的机器人学习结果。

在机器人领域，数据、计算能力和基础模型的发展正成为一个不可忽视的趋势。这一观点主要源于 Rich Sutton 的一篇具有深远影响的论文，其中强调了在 AI 研究历史上，那些相对简单且能够随数据量增加而扩展的算法往往胜过复杂且不可扩展的算法。Karol Hausman 曾在早期演讲中提出一个形象的比喻，他将数据和计算能力的提升比作技术进步的巨浪。这种浪潮不论人们是否准备好，都将带来更多的数据和更强的算力。面对这样的趋势，人们可以选择主动适应，也可以选择忽视。对于机器人行业来说，迎合这股浪潮意味着要采用在 NLP 和 CV

领域已经证明有效的大规模预训练方法，将其应用于机器人技术中。

虽然在现实世界的工业和家庭环境中，机器人的应用需要达到 99.X%的高准确性和可靠性，但现阶段的机器人学习算法尚未能满足这一要求。当前没有大型模型能够保证其输出结果的百分百准确无误。在软件应用中，这种不完美可能是可以接受的，例如 ChatGPT 偶尔提供的有趣或不准确的答案可能只会引起用户的一笑。然而，对于需要与现实世界互动的机器人而言，即便是微小的错误也可能导致严重甚至不可逆转的后果。如果不能保证整合了大型模型的机器人的安全性和可靠性，那么这些研究将无法实现其真正的价值和应用。

总的来看，将大模型应用到机器人领域是十分有前景有希望的，尽管需要真正落地还有很多工作要做，但我们应当对技术的积极的进步抱有一颗乐观的心态，并为之不懈努力奋斗。希望在不久的将来，大模型不仅能应用到机器人领域，还能在更多的领域发挥自己的能量。

参考文献

[1] KUMAR N J. Will Scaling Solve Robotics?: Perspectives From Corl 2023, 2023. https://nishanthjkumar.com/Will-Scaling-Solve-Robotics-Perspectives-from-CoRL-2023/.

[2] ALABDULMOHSIN I, NEYSHABUR B, ZHAI X. Revisiting Neural Scaling Laws in Language and Vision [J/OL] 2022, arXiv:2209.06640[https://ui.adsabs.harvard.edu/abs/2022arXiv220906640A. 10.48550/arXiv.2209.06640

[3] RIQUELME C, PUIGCERVER J, MUSTAFA B, et al. Scaling Vision with Sparse Mixture of Experts [J/OL] 2021, arXiv:2106.05974[https://ui.adsabs.harvard.edu/abs/2021arXiv210605974R. 10.48550/arXiv.2106.05974

[4] BROHAN A, BROWN N, CARBAJAL J, et al. RT-1: Robotics Transformer for Real-World Control at Scale [J/OL] 2022, arXiv:2212.06817[https://ui.adsabs.harvard.edu/abs/2022arXiv221206817B. 10.48550/arXiv.2212.06817

[5] YU T, XIAO T, STONE A, et al. Scaling Robot Learning with Semantically Imagined Experience [J/OL] 2023, arXiv:2302.11550[https://ui.adsabs.harvard.edu/abs/2023arXiv230211550Y. 10.48550/arXiv.2302.11550

[6] VEMPRALA S, BONATTI R, BUCKER A, KAPOOR A. ChatGPT for Robotics: Design Principles and Model Abilities [J/OL] 2023, arXiv:2306.17582[https://ui.adsabs.harvard.edu/abs/2023arXiv230617582V. 10.48550/arXiv.2306.17582

[7] BROHAN A, BROWN N, CARBAJAL J, et al. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control [J/OL] 2023, arXiv:2307.15818[https://ui.adsabs.harvard.edu/abs/2023arXiv230715818B. 10.48550/arXiv.2307.15818

[8] LAKE B M, BARONI M. Human-like systematic generalization through a meta-learning neural network [J]. Nature, 2023: 1-7.