Robots Reach Human Level Tool Use? Approaching, But Still Far

Zhiyuan Zhang
Department of Automation
Tsinghua University
z-zy20@mails.tsinghua.edu.cn

Abstract

Tool use is always a great demonstration of human intelligence for long, as most of the animals do not possess this ability. However, recent advances in artificial intelligence shed light on improving robot tool use abilities to human level. In this paper, we analyze the possibilities of reaching human level tool use abilities by stitching current SOTA methods, propose reasonable solutions considering aspects that distinguished human tool use with other animals, and argue that we are still far away due to some fundamental problems of current deep learning models.

1 Possibility with current arts

With the rapid development of large Transformer-based multi-modal models[2], we have witnessed the progress in robot planning and control made by RT-2[1]. In this section, we identify abilities needed in robot tool use[3], and propose obvious directions to further extend RT-2 to tool use.

Previous research identified three skills needed for robot tool use: perception, manipulation, and high-level cognition skills. Here we illustrate how RT-2 gets these capabilities and how we can extend this to tool use. RT-2 is a multi-modal Transformer model, whose training data includes web-scale text and images like previous works. This training gave the model the basic visual perception and reasoning abilities. To include robot actions, researchers devised new datasets of robot action data, where robot commands (like translation, rotation) are expressed as tokens like any other large language model. The additional robot action data gave the model information about robot manipulation, and a co-fine-tuning process linked it to model's perception and reasoning abilities. See 1 for an example.

The first intuition of improving RT-2 is to also include tool use data into training. Still take 1 as an example, we may think of this as training data for task "Move the picture. Tools provided are: [cylinder, triangle, ...]". To put it more clearly, since visual perception and reasoning abilities originated from training Transformers on web-scale data, and robot manipulation commands can be added as tokens (in RT-2 it's rotation, translation and grip, we can extend it into describing tool use first, then giving commands to robot hand or grippers to manipulate tools).

2 Solvable Limitations

In this section we further consider extending training data for RT-2, and consider improvements based on several aspects that humans are specially better than animals [4].

2.1 Perception, hand-eye coordination

Some lines of evidence support that hand-eye coordination in humans are superior than chimpanzees [4], and in RT-2 we see examples that can benefit from better hand-eye coordination (see 2. To improve the hand-eye coordination, we may use only a small part of the output robot action steps, and adding more perception steps during the process, enabling the model know its position along the

Figure 1: RT-2 successfully moving coke to Taylor Swift

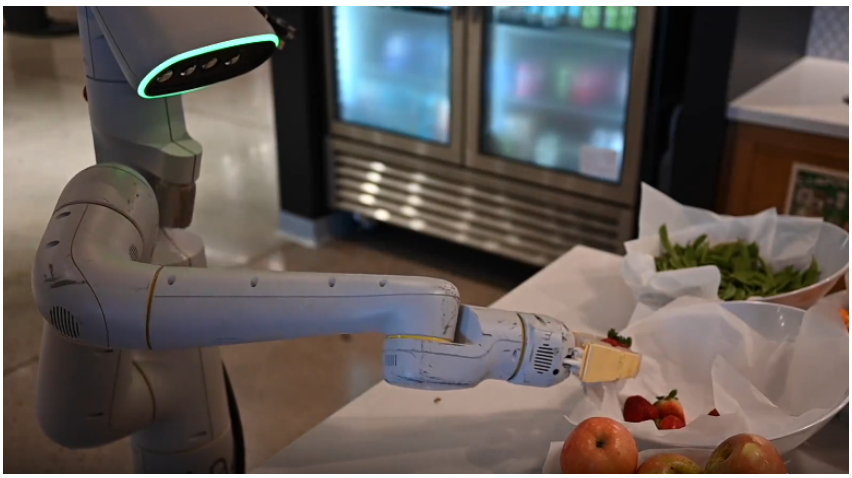

Figure 2: RT-2 moving the strawberry into the bowl, but pushing the bowl is unwanted

way. Or we can add training data for this, incorporating actions like "The target is blocked by my arm, so I need to change arm pose and look closer. [Command for calling robot module]". An alternative is we design a separate model for this estimation only and use it as a tool (like GPT4 plugins [2]).

2.2 Function representation

Primates do not attach certain functionalities with their tools, which limits their generalization ability for tool use [4]. But this problem may be largely addressed by the vast common knowledge of large vision-language models like GPT-4 [2], as they can utilize language descriptions to first extract possible tool functionalities using language, then guide the following actions.

2.3 Social learning, or teaching

Humans can learn tool use rapidly from a teacher [4]. For models like RT-2, we may use few shot prompting tricks to realize this as this trick already showed amazing performance gain [5]. Heuristics like "teacher must pick the tool with unique or best attribute, like weight" [6] can be easily expressed to models like RT-2.

2.4 Causality

Though the model cannot explicitly model causality, we can still easily add training data (sequence of actions and their results) or use few shot prompting tricks (some examples given to the model illustrating how to utilize causal effects of the tools) to give the model a more sense of causality, or easily examine if its causality inferences were correct.

3 Fundamental limitations

However, even considering these possible solutions, we are still far from building a robot that can achieve human level tool use. These limitations are common for deep learning models.

3.1 Generalization

We cannot expect any kind of generalization from these models [1], even though we observe their strong performances. For example, there are multiple layers of compositions in training data: object appearance, robot type, tool type, and so on. It's impossible to include them all in the dataset, while we humans easily transfer knowledge of tool use.

3.2 Hallucination

Another fundamental problem of using LLMs is hallucination [2]. As models hallucinate, there are too many stages where we could fail. We may found the models hallucinating tools or objects that are not in the image or hallucinating tool use that is impossible, and we cannot trust the model to perform tasks for us.

3.3 Continual learning

Humans learn new tools rapidly, but for tools or functions that are way too out of distribution, we cannot expect the model to learn quickly through trickes mentioned before , like few shot prompting. Instantiating a new round of training requires big data, and fine-tuning may lead to unknown performance change of validated tool use scenarios.

4 Conclusion

Tool use is a complex skill that sets humans apart from most other species. Recent models like RT-2 show promise in bridging the gap between human and robot abilities, but they aren't quite there yet. While these models are good at tasks like object recognition and basic actions, they still far from human level tool use.

We discussed possible ways to improve these models, including better hand-eye coordination, function representation, and social learning. While these solutions seem doable, they aren't enough to overcome some core challenges that deep learning models face commonly today. First, these models cannot guarantee desired generalization. Second, they sometimes make mistakes by hallucinating things, which can be risky for robot using tools in real life. Lastly, they don't adapt well to new tasks without going through a whole new round of training.

To summarize, while we're making progress, we're still far from having robots that can use tools as effectively as humans do.

References

[1] Anthony Brohan, Noah Brown, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, MontseGonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-WeiEdward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. 1, 3

[2] OpenAI OpenAI. Gpt-4 technical report. Mar 2023. 1, 2, 3

[3] Meiying Qin, Jake Brawer, and Brian Scassellati. Robot tool use: A survey. *Frontiers in Robotics and AI*, Jan 2023. doi: 10.3389/frobt.2022.1009488. URL http://dx.doi.org/10.3389/frobt.2022.1009488. 1

[4] Krist Vaesen, Lluı´sBarcelo´-Coblijn, Antoni Gomila, SarahR Beck, Jackie Chappell, IanA Apperly, Nicola Cutting, Guido Gainotti, RalphL Holloway, Hans Ijzerman, Francesco Foroni, PierreO Jacquet, Alessia Tessari, Ferdinand Binkofski, AnnaM Borghi, GuyA Orban, Giacomo Rizzolatti, Mathias Osvath, Tomas Persson, Peter Ga, EricM Patterson, Janet Mann, SimonM Reader, StevenM Hrotic, ACKNOWLEDGMENT S, Antonio Rizzo, Gijsbert Stoet, LawrenceH Snyder, DanielJ Weiss, KateM Chapman, JasonD Wark, and DavidA Rosenbaum. The cognitive bases of human tool use. 1, 2

[5] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision), 2023. 2

[6] Yixin Zhu, Yibiao Zhao, and Song-Chun Zhu. Understanding tools: Task-oriented object modeling, learning and recognition. In *CVPR*, 2015. 2