Foundation Models for Robotics: Best Known Practices

Shaocong Xu, Hao Zhao
Tsinghua University, AIR
xushaocong@stu.xmu.edu.cn, zhaohao@air.tsinghua.edu.cn

Abstract

Artificial general intelligence (AGI) used to be a sci-fi word but recently the surprising generalization capability of foundation models have triggered a lot of attention to AGI, in both academia and industry. Large language models can now answer questions or chat with human beings, using fluent sentences and clear reasoning. Diffusion models can now draw pictures of unprecedented photo-realism, according to human commands and controls. Researchers have also made substantial efforts to explore new possibilities for robotics applications with the help of foundation models. Since this interdisciplinary field is still under fast development, there is no clear methodological conclusions for now. In this tutorial, I will briefly go through **best known practices** that have shown transformative capabilities in several sub-fields. Specifically, there are five representative paradigms: (1) Using foundation models to allow human-friendly human-car interaction; (2) Using foundation models to equip robots the capabilities of understanding vague human needs; (3) Using foundation models to break down complex tasks into achievable subtasks; (4) Using foundation models to composite skill primitives so that reinforcement learning can work with sparse rewards; (5) Using foundation models to bridge languge commands and low-level control dynamics. I hope these best known practices to inspire NLP researchers.

1 Introduction

This is a tutorial paper that summarizes my talk at CCL 2023, on the topic of *Foundation models for robotics: best known practices*. The concept of foundation models emerge in the community of natural language processing (NLP), like GPT (Brown et al., 2020). Current foundation models can learn language skills using few shots and notably without fine-tuning the model (known as in-context learning). This human-like behavior has not been demonstrated before and widely considered as an encouraging step towards artificial general intelligence (AGI). However, a long-existing problem still lingers around at the age of foundation models, which is known as the Moravec's paradox. Specifically speaking, high-level human intelligence like language and reasoning actually consumes relatively less computation while low-level control is much more complicated than one may think. This paradox is echoed by the current state of AI research: while language models are getting stronger and stronger, robots still struggle to move and manipulate, at least in the context of artificial general intelligence.

Despite the disappointing situation as described by the Moravec's paradox, we see a line of encouraging research efforts related to foundation models that have definitely expanded the scope of robotics research. As a disclaimer, this does not mean that the paradox is solved. Specifically, we observe several promising paradigms that successfully marry robotics with the recent progress of foundation models, among which five representative works are covered in this tutorial: (1) Using foundation models to allow human-friendly human-car interaction; (2) Using foundation models to equip robots the capabilities of understanding vague human needs; (3) Using foundation models to break down complex tasks into achievable sub-tasks; (4) Using foundation models to composite skill primitives so that reinforcement learning can work with sparse rewards; (5) Using foundation models to bridge languge commands and low-level control dynamics. Apart from these five topics, other smaller ones w.r.t. open-set understand-

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 24-29, Harbin, China, August 3 - 5, 2023.

ing (Liu et al., 2023) or anomaly detection (Tian et al., 2023) do exist although they are not covered.

2 Best Known Practices

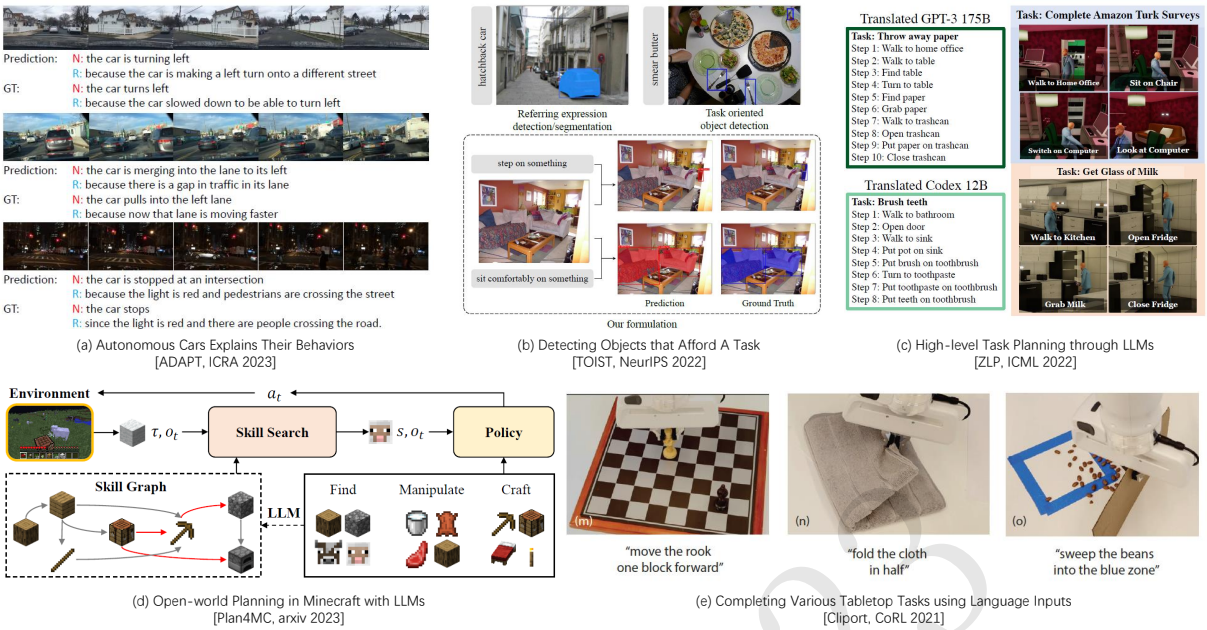

Figure 1: Foundation models including large language models (LLMs) and large vision-language models (VLMs) have enabled new capabilities in various robotics tasks, as demonstrated in this figure. These new progresses come in different paradigms and these five methods represent the **best known practices**. Specifically speaking, (a) ADAPT (Jin et al., 2023) allows autonomous driving cars to describe and explain their own behaviors using natural language. (b) TOIST (Li et al., 2022a) equips robots with the ability to find an object that affords any tasks specified by natural language. (c) ZLP (Huang et al., 2022) exploits large language models to break down high-level commands into step-by-step action primitives. (d) Plan4MC (Yuan et al., 2023) combines skill primitives using large language models so that agents can learn to accomplish diverse tasks in Minecraft with sparse rewards. (e) Cliport (Shridhar et al., 2022) separates the pathways for language and affordance so that the agents can accomplish manipulation tasks according to any natural language commands and transfer their skills to real worlds seamlessly.

As shown in Fig. 1, we demonstrate five paradigms that foundation models can benefit robotic perception, planning and control. It is far from an extensive list but reasonably representative.

2.1 Talking with agents

Talking with agent is an potential practices of using vision-language foundation models (VLMs). There are several representative works in this direction. Firstly, ADAPT (Jin et al., 2023) is an algorithm that allows autonomous cars to simultaneously describe their behaviors and explain the rationale behind. While robotaxis are now running smoothly in large cities, the experience is still far from that of taking a taxi driven by human beings. A notable difference is that human drivers can tell the passengers why they drive in a certain manner thus comforting the passenger. ADAPT can tell the passenger that the car is stopping because of red traffic lights and pedestrians ahead (Fig. 1-a last row), which enhances human trust on robotic systems.

Additionally, SPRING (Long et al., 2023) is an algorithm that enables agents to describe their responses not only based on the attributes of objects (the black jacket) in the scene, but also on the spatial

relations of the objects (the black jacket on the leftmost floor rack). Such method is particularly useful when the agent is faced with an extremely complex situation.

Furthermore, SQA3D (Ma et al., 2022) decomposes the question answering task into situation localization and answering. The response to the question 'What is the object situated behind the apple that is currently positioned in front of you?' is dependent on the agent's situation. Failure to account for this factor results in agent performance falling short of human expectations in real-world embodied environments. SQA3D allows the agent to learn how to adapt to various situations and execute different commands issued by humans.

Thus, the applications of VLMs for human-robot interaction have the potential to alleviate Moravec's Paradox. We hope that further research efforts can be dedicated to this area.

2.2 Finding tools for any tasks

TOIST (Li et al., 2022a) is an algorithm that finds tools and segments their masks in visual inputs, according to any task specification, shown in Fig. 1-b. The task specification is given by natural language like a verb phrase *sit comfortably on*, so that it breaks the bottleneck of prior closed-set affordance understanding models. It leverages vision-language foundation models that understand well nouns but further distill the representation into pronouns and allow the understanding of verbs in an open-set manner. Using this model, intelligent robots can find tools that serve a human commander's vague goals without the need of exactly providing the tool's name.

The Touch-line transformer algorithm (Li et al., 2022b) is another representative work that can localize the objects both referred by natural language and indicated by the virtual touch line of human. Thus, This method endows VLMs with the capability to comprehend human gestures and accurately locate objects indicated by those gestures.

2.3 Breaking down high-level commands.

Just like the problem that TOIST addresses, a long-term goal of robotics is to fulfill high-level commands from human beings. While TOIST finds tools to serve a certain task, zero-shot language planner (ZLP (Huang et al., 2022) as shown in Fig. 1-c) can leverage the reasonging power of language foundation models to break down high-level commands into practical sub-tasks. For example, to achieve the task of brushing teeth, eight steps are needed and foundation models can be easily instructed to do this decomposition job in a zero-shot manner (shown in Fig. 1-d). Finally, a robot can conduct sub-tasks with pre-built motion primitives.

Similar to ZLP, SPLL (Sharma et al., 2021) utilizes the reasoning capabilities of language foundation models to decompose high-level language abstractions into several primitive low-level actions. This approach creatively leverages pre-trained language models to facilitate the acquisition of new skills by enabling the agent to reason about the underlying task structure and generate effective action plans.

2.4 Escaping the curse of sparse rewards

Reinforcement learning is widely considered as another promising path towards artificial general intelligence. If an agent can experience the world thousands of times in a realistic simulator, it can finally learn generic low-level control and high-level reasoning skills. However, apart from the realism problem, reinforcement learning usually suffers from the issue of sparse rewards when conducting complicated tasks. Plan4MC (Yuan et al., 2023) creatively combines low-level skill learning and the skill graphs pre-generated by language foundation models to effectively accomplish tasks under the guidance of very sparse rewards. ELLA (Mirchandani et al., 2021) leverages learned language abstractions to provide more accurate reward signals and enhance the agent's ability to learn and accomplish tasks effectively.

2.5 Versatile and clever hands that listen to your commands

While research works mentioned above focus on perception or relatively high-level planning, Cliport (Shridhar et al., 2022) is a method that successfully allows robot arms to accomplish diverse low-level control tasks according any human commands specified by natural language, as shown in Fig. 1-e. The architecture disentagles semantics and spatial information while conditioning two streams on language

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 24-29, Harbin, China, August 3 - 5, 2023.

representations generated by foundation models. Since the motion primitive is simplified as a bin-picking formulation, 2D affordance maps allow the network to achieve open-set tasks. Interestingly, this system exhibits good sim-to-real generalization capabilities and functions as a versatile and clever hand that listens to your arbitrary commands. PA (Shridhar et al., 2023) is another representative work that can effectively train a model to perform a total of 25 robotic manipulation tasks, including 18 RLBench tasks with 249 variations and 7 real-world tasks with 18 variations, using only a few natural language descriptions per task. This approach creatively leverages language-conditioned behavioral cloning to enable the agent to reason about task structures and generate effective action plans.

3 Tutorial Outline

Part I: Introduction (20 min)

- The development of large language models
- The definition of the Moravec's paradox
- Promising paradigms ellivate the Moravec's paradox
  - Talking with agents
  - Finding tools for any tasks
  - Breaking down high-level commands
  - Escaping the curse of sparse rewards
  - Versatile and clever hands that listen to your commands

Part II: Best Known Practices (60 min)

- Talking with agent
- Finding tools for any tasks
- Breaking down high-level commands
- Escaping the curse of sparse rewards
- Versatile and clever hands that listen to your commands

Part III: Conclusion (10 min)

4 Reading List

1. ADAPT: Action-aware Driving Caption Transformer (Jin et al., 2023);
2. TOIST: Task Oriented Instance Segmentation Transformer with Noun-Pronoun Distillation (Li et al., 2022a);
3. Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents (Huang et al., 2022);
4. Plan4MC: Skill Reinforcement Learning and Planning for Open-World Minecraft Tasks (Yuan et al., 2023);
5. CLIPort: What and Where Pathways for Robotic Manipulation (Shridhar et al., 2022);

5 Instructor

Hao Zhao is an assistant professor at Tsinghua University, where he leads a research group focused on computer vision fields related to robotics, particularly 3D scene understanding. Prior to joining Tsinghua University, he was a research scientist at Intel Labs China and a joint postdoc affiliated with Peking University. He obtained his Ph.D. and Bachelor's degrees from the Department of Electronic Engineering at Tsinghua University. Additionally, he is proud to have served as a former leader of Skyworks, the largest robotics club at THU, where he contributed significantly to the growth of the club and fostered a culture of innovation and excellence.

In addition to his academic work, Hao Zhao has also been involved in entrepreneurship, co-launching more than 10 startups in various fields such as social networks, web development tools, unmanned aerial vehicles, intelligent delivery, smart grid, VR devices, virtual human, cloud design, autonomous driving, and smart manufacturing since 2009.

His homepage can be found at https://sites.google.com/view/fromandto

Acknowledgements

This tutorial substantially benefits from the suggestions of Prof. Hao Dong in Peking University.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR.

Bu Jin, Xinyu Liu, Yupeng Zheng, Pengfei Li, Hao Zhao, Tong Zhang, Yuhang Zheng, Guyue Zhou, and Jingjing Liu. 2023. Adapt: Action-aware driving caption transformer. *arXiv preprint arXiv:2302.00673*.

Pengfei Li, Beiwen Tian, Yongliang Shi, Xiaoxue Chen, Hao Zhao, Guyue Zhou, and Ya-Qin Zhang. 2022a. Toist: Task oriented instance segmentation transformer with noun-pronoun distillation. *Advances in Neural Information Processing Systems*, 35:17597–17611.

Yang Li, Xiaoxue Chen, Hao Zhao, Jiangtao Gong, Guyue Zhou, Federico Rossano, and Yixin Zhu. 2022b. Understanding Embodied Reference with Touch-Line Transformer, October.

Xinyu Liu, Beiwen Tian, Zhen Wang, Rui Wang, Kehua Sheng, Bo Zhang, Hao Zhao, and Guyue Zhou. 2023. Delving into shape-aware zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2999–3009.

Yuxing Long, Binyuan Hui, Fulong Ye, Yanyang Li, Zhuoxin Han, Caixia Yuan, Yongbin Li, and Xiaojie Wang. 2023. Spring: Situated conversation agent pretrained with multimodal questions from incremental layout graph. *arXiv preprint arXiv:2301.01949*.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. 2022. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*.

Suvir Mirchandani, Siddharth Karamcheti, and Dorsa Sadigh. 2021. Ella: Exploration through learned language abstraction. In *Neural Information Processing Systems*.

Pratyusha Sharma, Antonio Torralba, and Jacob Andreas. 2021. Skill induction and planning with latent language. In *Annual Meeting of the Association for Computational Linguistics*.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. 2022. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, pages 894–906. PMLR.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. 2023. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR.

Beiwen Tian, Mingdao Liu, Huan-ang Gao, Pengfei Li, Hao Zhao, and Guyue Zhou. 2023. Unsupervised road anomaly detection with language anchors. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7778–7785. IEEE.

Haoqi Yuan, Chi Zhang, Hongcheng Wang, Feiyang Xie, Penglin Cai, Hao Dong, and Zongqing Lu. 2023. Plan4mc: Skill reinforcement learning and planning for open-world minecraft tasks. *arXiv preprint arXiv:2303.16563*.