17 January 2025 ·

Qeios

Peer Review

Review of: "What's the Move? Hybrid Imitation Learning via Salient Points"

Martin Stetter[1]

1. Weihenstephan-Triesdorf University of Applied Sciences, Freising, Germany

In this paper, the authors present a novel mixed-strategy approach to imitation learning in the world of physical robots. The method takes advantage of the fact that complex long-horizon manipulation tasks can often be decomposed into a sequence of situations with either low accuracy requirements but high variation in action space or vice versa. The SPHINX system presented here learns two policies: a "dense" per-time step policy and a higher-level "waypoint" policy, which makes use of a linear controller to execute long-range movements. The term "waypoint" stands for a target end effector pose one complexity level above raw end effector states. In addition to existing approaches, SPHINX learns to predict when to switch forth and back between those policies with a high level of spatial generalization. For this, it learns to infer "salient points" within a so-far unseen scene while carrying out a given task. These are points that are likely to require high-accuracy action planning. For training, a demonstrator can mark salient points in a training scene using a custom web interface, from which the system learns to form saliency maps for each new task.

I do appreciate this work very much, due to the following reasons: (1) the work solidly builds upon existing similar approaches to the problem but adds the novel component of adaptive policy switching based on saliency maps; (2) for learning, state-of-the-art components are used, namely a transformer-type deep network for learning the waypoint policy (+ mode prediction) and the recently proposed diffusion policy approach (augmented with a mode prediction channel) for the dense policy; (3) a lot of engineering effort is invested to boost performance, e.g., a custom web interface for demonstration, multimodal input adapted to policy (3D point cloud for "waypt" vs. wrist image for "dense"), training set enhancement ("waypt"'s output is used to train "dense"); (4) the performance of SPHINX is benchmarked very carefully and close to exhaustively against the nearest existing approaches. For example, some competing approaches have been "modernized" in their methods for

 doi.org/10.32388/UWU05Z

fair comparability to SPHINX; (5) SPHINX really seems to work considerably better on a number of complex benchmark tasks than all of the competing methods.

I have little to comment on the existing work; instead, I feel that for many readers (myself included), it would be beneficial to learn more about the computational and demonstration details of that interesting approach even in the main text. For example: How many salient points have to be set usually for useful training? How long does demonstration take as compared to a dense-only approach or a hybrid approach without saliency points? What is the computational complexity (or training time on your machines) to train the transformer and the diffusion policy models? What are the hardware requirements? Finally, what software platform/platforms are used (PyTorch / Keras? Third? Several?)

I look forward to more exciting work from your lab!

Declarations

Potential competing interests: No potential competing interests to declare.