Vision-Guided Imitation Learning Using Action Chunk Transformers

Navaneet†, Manisha Lingala†, Sangmoon Lee†, Hongseok Yoo*‡
(†Kyungpook National University, ‡Kyungwoon University)

Abstract—This paper introduces the use of the Action Chunking with Transformers (ACT) technique in single-arm robotic manipulation for vision-guided pick-and-place tasks. ACT employs a Conditional Variational Autoencoder (CVAE) to predict sequences of actions, termed "action chunks," which are groups of actions predicted together to achieve more complex tasks efficiently. Unlike traditional methods that rely solely on joint position data and predict individual actions, our approach integrates visual data to enrich the learning context and enhance execution precision. We acquired the expert data by providing manual demonstrations of the task, allowing the model to learn from real-time, complex action sequences. By predicting these action chunks instead of single actions, the ACT model adapts from dual-arm to single-arm configurations, enhancing control strategies and demonstrating significant improvements in the robot's speed, precision, and reliability. This substantiates the paper's title, "Vision-Guided Imitation Learning Using Action Chunk Transformers," highlighting the critical role of vision in advancing robotic control systems. Further project details are available on our website: https://sainavaneet.github.io/ACTfranka.github.io/

Keywords: Imitation learning, Action Chunk Transformer, Robotic manipulation, Joint position prediction, Autoencoder

I. INTRODUCTION

The ACT algorithm's initial success with dual-arm configurations highlighted its capability for refined manipulation and coordination, forming a robust foundation for further exploration [1][2]. Our research adapts this sophisticated algorithmic approach to a single-arm setup, aiming to achieve comparable levels of precision and efficiency in simpler configurations. This endeavor not only explores the algorithm's adaptability across different robotic platforms but also enhances our understanding of its potential scalability.

Inspired by significant advancements in robotic capabilities through imitation learning, such as those demonstrated in the "Mobile ALOHA" project [2], we aim to apply these techniques to enhance single-arm robotic manipulation. The Mobile ALOHA project utilized a low-cost teleoperation system to successfully implement complex bimanual mobile manipulation tasks, driven by imitation learning from human demonstrations [2]. This project's approach to combining mobility with manipulation skills in a bimanual context lays the groundwork for adapting these strategies to the more limited but equally complex domain of single-arm robots.

Our adaptation involves refining the ACT model to better align with the operational specifics of the Franka robot, particularly by incorporating vision data into the learning process. By leveraging insights from both the initial dual-arm applications and subsequent innovations in vision-based mobile manipulation, we enhance the algorithm's framework to suit single-arm manipulation tasks more effectively. This integration of visual data is critical for demonstrating the algorithm's versatility and for expanding the capabilities of robotic automation in various industrial and service settings, highlighting our significant contributions to precision and efficiency in robotic control systems.

II. Methodology

In this study, we implement the ACT model on a

*Navaneet, Manisha Lingala and Sangmoon Lee are with School of Electronic and Electrical Engineering, Kyungpook National University, Daegu 41566, Republic of Korea (e-mail: sainavaneet@knu.ac.kr; lingala.manisha@knu.ac.kr; moony@knu.ac.kr).

Hongseok Yoo is with Kyungwoon University, Republic of Korea
(e-mail: dexweaver.hsyoo@gmail.com)

Franka Emika Panda robot to enhance its capability in performing precise manipulation tasks. The ACT model aims to leverage the sequential nature of actions in manipulation tasks, improving the robot's efficiency and adaptability by optimizing action execution through chunking. We executed the implementation both in a simulated environment and on the actual robot.

A. System Setup

Simulated Environment Setup : We use the Gazebo robotics simulator to model the Franka Emika Panda robot with high fidelity, including virtual sensors and actuators. This allows us to develop and test control algorithms under simulated real-world physics. The simulation runs on Ubuntu 20.04 with ROS, seamlessly integrating our Python-based control algorithms and leveraging Python's machine learning libraries to optimize performance.

Real-World Environment Setup: In the real-world setup, the Franka Emika Panda robot is controlled via the Franka Control Interface (FCI) [3] in a laboratory setting with hardware mirroring the simulated sensors and actuators. Using ROS on Ubuntu 20.04 ensures software consistency and facilitates the transfer of scripts from simulation to reality. Real-time sensor feedback enables the robot to perform high-precision tasks with accurate and adaptable actions.

Fig 1: This image shows the complete real-world setup of the task.

B. Data Collection

We conducted data collection using the Franka interface with ROS, employing subscribers to monitor joint angles and camera feeds. Each episode began by placing the robot in free move mode. Throughout the pick-and-place task, we manually manipulated the robot, recording data from start to finish. This approach mirrors the method used in the Mobile Aloha ACT algorithm, although our setup differs as it incorporates only a single arm, rather than a dual-arm system. Therefore, we acted as the demonstrator, guiding the robot's movements and capturing its actions along with the joint angles. We executed 50 manual demonstrations, varying the positions of the cube. For each demonstration, we recorded data from 8 joint positions—7 from the robot and 1 from the gripper, to determine whether it should be open or closed. We utilized a single camera, allowing us to collect visual data from each demonstration.

C. Model Training

We train the ACT model to predict future action sequences based on the robot's current sensory inputs, primarily focusing on the robot's joint positions for forthcoming timesteps. By mimicking a human operator's anticipation of actions guided by real-time observations, the model enhances the precision and adaptability of the Franka Emika Panda robot. During testing, we implement the most effective policy derived from validation results, concentrating on minimizing error accumulation that could lead the robot into untrained states. Model architecture integrates a CVAE with transformers for both the encoder and decoder. The encoder, adopting a BERT-like structure, inputs current joint positions and a target action sequence from our demonstrations, initiating with a "[CLS]" token, to encode both current and future states.

In the decoding phase, informed by current observations and a "style variable" z, the model predicts subsequent action sequences. This process is supported by the integration of ResNet image encoders and transformer architectures that handle and synthesize data from various sources, including camera images and joint positions. Observational data encompasses one 480x640 RGB images alongside the robot's 8 DoF in joint positions—7 from the arm and 1 from the gripper. Our action space consists of an 8-dimensional vector of these joint positions, with the model outputting a tensor representing the sequence of actions. This output is processed through a network that flattens image features and merges them with position embeddings. Employing cross-attention mechanisms in the decoder [4], the model generates predictions for the robot's subsequent movements by leveraging the processed encoder outputs, using L1 loss for action sequence

reconstruction to ensure heightened precision.

III. Results and discussions

After training, we evaluated the policy by predicting actions for the robot and publishing these actions for execution. We observed that the robot successfully performed tasks in positions that were not explicitly included in the training dataset. This outcome demonstrates the model's robustness and its ability to generalize to new scenarios, underscoring the effectiveness of our training approach in equipping the robot with adaptable and reliable manipulation capabilities.

Fig 2: The image above depicts the entire pick-and-place task performed by the robot. As illustrated, the robot successfully picks up the cube from its initial position and places it accurately at the designated location.

Acknowledgments

This research was supported by Defense Innovation Cluster Project funded by the Korea Research Institute for defense Technology planning and advancement & Defense Innovation Cluster Division Gyeongbuk·Gumi Programs Group (No.DC2023SD)

References

[1] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," RSS, 2023.

[2] Zipeng Fu, Tony Z. Zhao and Chelsea Finn, "Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation," in Proc. *Conference on Robot Learning (CoRL)*, 2024.

[3] Franka control Interface. https://frankaemika.github.io/docs/

[4] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," ArXiv, abs/1706.03762, 2017.