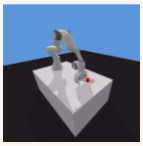

Deep Reinforcement Learning for Robotics Grasping with Multi-Modal Sensor Fusion

Peizhuo, Yu, Dandan, Zhang & Dr Benny, Lo.

HAMLYN SYMPOSIUM ON MEDICAL ROBOTICS 2021

Abstract

Vision and touch are important for a contact-related manipulation task, yet few approaches exploit the combination of both two sensing modalities. This project aims to explore applying reinforcement learning and multimodal representation learning methods in automating robotic picking and placing task.

Introduction

Robotic manipulation tasks require interaction with objects and visual and haptic feedback are complementary and concurrent during these tasks. Hence, combining these two sensing modalities has great potential in robot control. However, these two types of data have very different characteristics, making hand-designed features impossible [1].

Reinforcement Learning (RL) can generate adaptive behaviours for robots. But standard the method has encountered the problem of requiring many trial-and-error attempts to obtain an effective control policy [2]. One of the reason is that it learns to extract features from high dimensional data and a control policy simultaneously [3].

To boost the sample efficiency in reinforcement learning, the feature extraction part can be undertaken separately by a learning-based method, which is known as representation learning [3]. In addition, the representation learning method has the power of fusing multimodal data.

The ideal outputs of this project:

- Multimodal representation learning of visual and haptic data to accelerate reinforcement learning

Method

6-Dimensional (6D) Pose Estimation and Grasping:

- Creating a dataset with Pybullet simulator including images and objects' poses
- Training DenseFusion network for 6D pose estimation [4]
- Combining Pose Estimation and RL to robot picking and placing.

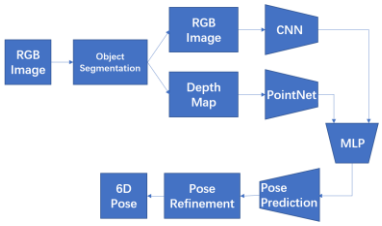

Figure 1 – Simplified DenseFusion Network

Multimodal Representation Learning:

- Creating Multimodal Dataset with Pybullet
- Using the neural network to learn fuse multimodal data to create a representation
- Using proximal policy optimization to obtain control policy
- Integrating obtained representation model into RL to get a policy for robotic picking and placing

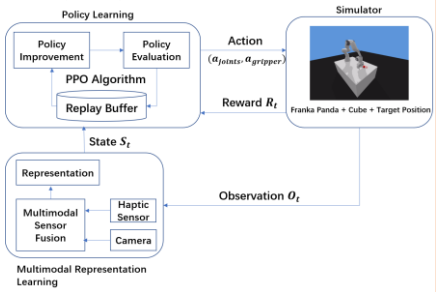

Figure 2 – General Framework of RL with multimodal representation

Preliminary Results

Created Dataset Collects the RGB image, segmentation mask and points cloud from depth map in the camera frame. For each category, 1200 frames are collected.

(A) (B) (C)

Figure 3 - (A) RGB image (B) Segmentation Mask (C) Point Cloud Representation Object's Pose in color green

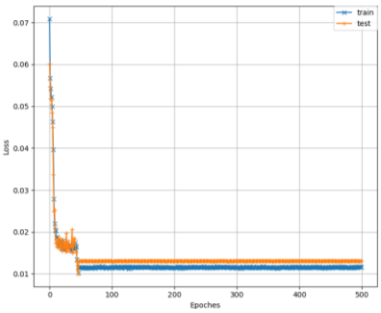

Figure 4 – Learning Curve for DenseFusion Network and loss is the average distance between the predicted point cloud and ground truth

Evaluate the newly trained network on the original test set in training and a new test set. The threshold for success is when the average distance between the predicted cloud and ground truth is less than 0.1 model's diameter.

Evaluation	Original Test Set	New Test Set
Success Rate (%)	84	81

Future Plan

6D Pose Estimation:

- Using prediction from pose estimation model and robot's proprioceptive data to train RL algorithm for robotic picking and placing
- Comparing results between RL with pose estimation and RL with fully observable states

Multimodal Representation Learning:

- Collecting visual ,haptic data and robot's proprioceptive data from the simulator
- Choosing domain-specific encoders to process different types of data
- Designing self-supervised learning objectives to train multimodal fusion model for avoiding manual annotation
- Adapting multimodal representation model to boost sample efficiency in reinforcement learning
- Running ablative studies to examine the effectiveness of multimodal representation.

References

[1] M. A. Lee et al., 'Making Sense of Vision and Touch: Learning Multimodal Representations for Contact-Rich Tasks', Available from: http://arxiv.org/abs/1907.13098
[2] L. Marzari et al, 'Towards Hierarchical Task Decomposition using Deep Reinforcement Learning for Pick and Place Subtasks', Available from: http://arxiv.org/abs/2102.04022
[3] A. Stooke et al, 'Decoupling Representation Learning from Reinforcement Learning', Available from : http://arxiv.org/abs/2009.08319.
[4] C. Wang et al., 'DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion, Available from: http://arxiv.org/abs/1901.04780.