A Literature Review on learning from demonstration as well as social navigation

Name: Hao Yan Course: ECE 7970 Semester: Winter 2021

Overview

Robot Learning from Demonstration (LfD) is a paradigm for enabling robots to autonomously perform new tasks. Compared with the traditional way, which requires the users to analysis the kinematic or dynamic behavior of the robot, and then build the control algorithm either through programming or flowchart, the work in LfD takes the view that an appropriate robot control policy can be learned by observing the teacher's demonstration. The research goal for LfD is to facilitate the scalability and usability of robot, so that it can be utilized in a wider scenario.

Most LfD problems focuses on building a mapping between the observed world states and the actions of the robot, which enables the agent to select the best optimized action among its action space, based on the current observation. Also, the modern hardware like Lidar and high-resolution camera is able to collect higher volumes of data in more efficiently, and as the GPU develops into a popular platform used to provide powerful data processing ability, all these progresses make is possible for many AI applications such as reinforcement learning to be applied to the autonomous agent, including driver assistant vehicle, social assistant robot, and human-computer interactions.

And at the same time, there are several questions raised during the research [1] :

- the source of demonstration, which means the methods used to capture the actions where the robot learns from, and it can be achieved through different sensing methods;
- the future representation, which process the data from previous layer (sensor) and extract the key points from the data called feature;
- the construction of mapping, which will build the map between the world states and the robot action from the features using classification, regression or other learning-based approach;
- how to deal with undefined scenario, which will address the self-evolving problem of the robot, because the teacher cannot list all possible scenarios during demonstration, and the robot needs to behave properly when the environment is unseen.

Problem formulation

In a recent work done by Louie [2], the LfD system design mainly includes the following definitions:

- World state: all the states (e.g. robot state, object state, person state) relevant to a given task
- Robot behavior: a set of actions to provide a certain functionality (my understanding), behaviors of the robot modify the world state, and human demonstrated behaviors are often deterministic with one behavior always being executed in the specific state. A robot behavior could be "Greeting a User" and the behavior policy specifies that the robot should wave its right arm and say "Hi" for this behavior.
- Activity learning: activity learning module learns the world state to robot behavior mapping for an activity using the demonstrated trajectory
- Activity Trajectory: it is pairs from world state -> robot behavior during a teacher's demonstration of the activity.

- Policy: a policy is a function which maps the states observed by the agent (position, pose, velocity, conversation, facial expression, and etc.) to an action picked from the robot's action space (movement, robotic arm, audio prompt and etc.). The action policy is what the robot uses to determine which action to take in the current situation
- Activity (what is activity itself): activity, A, is defined by the multiple distinct states that it can be in: A = {a1, a2,…,ao}, and each activity state is the combination of user state, robot state, and time step.

Also, in a survey written by Ahmed [1], he gave the formal definition are also formalized:

- Demonstration: A demonstration is presented as a pair of input and output (x, y), where x is a vector of features describing the state at that instant and y is the action performed by the demonstrator.

In the Louie's paper [3], they divided the LfD system mainly into three blocks, which include the learning module, the interaction module, and the execution module. Also, they proposed a performance evaluation module, which is used to evaluate the performance of therapy given by the Social assistant robot (SAR):

- Learning module: this topic addresses the concern how to learn the bingo game from a teacher (volunteers), including how can the teacher communicate with the robot (GUI/speech), how to model an interaction (activity simulator, including the robot model, user model, activity model, and trajectory), and the mapping function from state to behavior;
- Interaction module: based on the world state (environment), give a reasonable physical behavior. This topic introduced how to identify the world states, and determine the behavior/action to deliver; also, the persuasive strategy is included in this module, to better influence the attitudes and behaviors.
- Execution monitoring module: observe the robot's behavior/action, identify the behaviors which failed to be delivered, as well as analyze the cause of the fault.
- Performance evaluation method: the performance of the learning sub-system, by comparing the effectiveness of teaching (how many times, consistency of different teachers), and the performance of the persuasion learning (the percentage of optimal persuasive strategy instances occurring during interactions with users).

Source of demonstration

The input for the LfD system will be the demonstration, which is considered as the first problem we need to conquer. There are mainly two problems to be solved: the choice of demonstrator, and the choice of the demonstration technique. Each problem has several options, so there will be certain amount of options to choose in determine the demonstration module for a specific LfD system [4]:

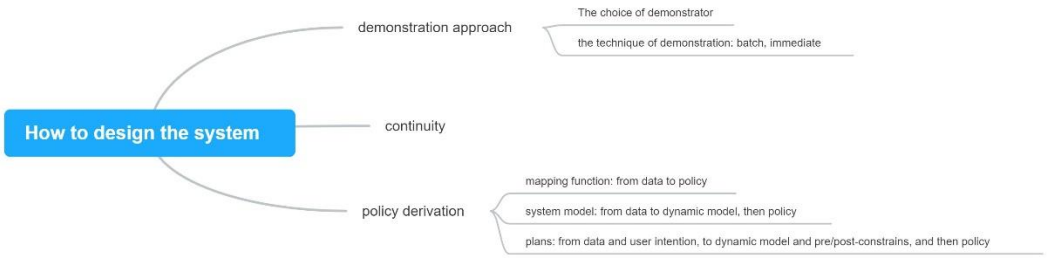

Fig.1 the LfD problem design choices

Most LfD work to date has made use of human demonstrators, although some techniques have also examined the use of robotic teachers, hand-written control policies and simulated planners.

The choice of demonstration technique refers to the strategy for providing data to the learner [4]. Similar to the training of neural networks, one approach is to provide batch data for learning purpose, in which case can be viewed as “offline” LfD, since the learning starts only when the interaction is done and all data are ready. Another choice of demonstration technique is “online” LfD, in which the interactive approaches allow the policy to be updated incrementally as training data becomes available, similar to how the agent was trained to play a game using reinforcement learning, provided by Maxim [5].

Feature representations

The feature is defined as the representation of the observable state, and the feature includes the information from the learner movement information, vocal information, manipulable objects, its surrounding environment and any information which may affect the policy mapping.

Given the high dimension of real world states [4], researchers need to extract the key information from the observed states, otherwise the policy learning may refuse to converge. So, there exists the necessity to consider the aspects of demonstrations presented to the learner.

For example, in the research from Hanbyul [6], they solve this problem by defining the social interaction as a signal prediction problem among individuals, rather than finding the mapping between the social signals and their semantics. In Igor’s work, they build the talking robot by using a sentiment analyzer assesses the sentiment of the text and gives as output a descriptor with information about the polarity of the sentiment (only consider positive/negative/neutral), and taking into account some simple head postures, voice parameters, and eyes colors as expressiveness enhancing elements. Ala’adin [7] designed a robot for Autism Spectral Disorder (ASD) children therapy by converting the extracted child verbal response from the audio data of the intervention session and converted to a spectrogram. The spectrogram was combined with the therapist’s last executed behavior to form a data tuple which was labeled with the current therapist behavior. Then the spectrogram and the last therapist behavior are sent to a DNN model to train the relationship between the demonstrator’s behavior and children’s reaction.

Learning from data

After we get the data from the environment, and extract the key information of the data, the next step will be build a mapping between the data and the desired output.

The most direct way for learning is direct imitation [1], which means learning a supervised model from the demonstration, where the action provided by the expert acts as the label for a given instance. Roman [8] mentioned a way to directly control the robot via a VR-based teleoperation system.

Traditional learning methods can be divided into classification and regression. Saleha [9] proposed a teaching coordinated strategy to soccer robots via imitation, in his research, the multiple sets of state-action data gathered from multiple players are combined in a manner that enables sharing of experiences gained by each object, high-level decisions are made from the classifier, and then the control strategy will be used to interact with the game.

In the work done by Igor [10], they introduced using Generative Adversarial Network (GAN), the GAN network could have been trained using poses instead of movement segments, but then, the outputs would be single poses that should be afterwards concatenated to generate talking movements. However, this approach would produce fewer smooth gestures.

Evaluation

Since the learning of demonstration considers a high dimension input vector, as well as complicated output action space, the evaluation of such algorithm is challenging, and establishing a clear standard comparison criteria and automated benchmark tests can be difficult to achieve.

Right now, the evaluation can be categorized as quantitative evaluation and qualitative evaluation. The quantitative evaluation will assign scores to each executed policy, and the qualitative evaluation requires human to assign scores to the performance delivered by the robot, based on whether the robot action is natural and comfort.

Roman [11] conducted a user study with therapists from the ABA clinic for children with ASD, to evaluate the performance of the given therapy, based on if the therapy is planned, designed and implemented in a good way. They also compare their VR teleoperation system with the default robot control software, Choregraphe. The questionnaire include 5 degrees, ranging from efficiency to usability of these two systems.

How reinforcement learning can help

Learning from demonstration (LfD) and reinforcement learning (RL) are two common ways for a robot to learn the logic behind a task task. The LfD problem relies on the demonstration from the teacher, while reinforcement learning is the training of machine learning models to make a sequence of decision. The agent learns to achieve a goal in an uncertain, potentially complex environment. In reinforcement learning, an artificial intelligence faces a game-like situation.

Starting from a random policy, the agent in reinforcement learning problem modifies its own neural networks weights based on the rewards function output, which takes the environment states and executed policy from previous and current steps. Reinforcement learning can be used on its own to learn a policy for a variety of robotic applications. However, if a policy is learned from demonstration, reinforcement learning can be applied to fine-tune the parameters [1].

Compared with using the demonstrations to teach the robot, Brys [12] uses demonstrations as a bias for learning, which would relax the assumption of demonstration optimality, while speeding up learning and preserving convergence guarantees.

Changan [13] proposed a method to forecast future trajectories of all the humans to avoid collisions; as well as the robot not only needs to understand the behavior of humans but also interact with them to obtain high rewards. (penalty is set as the duration when robot-human distance smaller than the threshold). To achieve this, the author extracts feature for pair- wise interactions between the robot and each human and captures the interactions among humans via local maps. Then, aggregate the interaction features to infer the relative importance of neighboring humans with respect to their future states.

In another paper from Changan [14], the author proposed a transformer-based model learning the association between how objects look and how they sound to solve this problem. A goal descriptor captures both spatial and semantic properties of the target, and the persistent multimodal memory to reach the goal even when the sound (acoustic event) stops. Then the policy network generates the action based on the output from observation encoder and goal descriptor.

Yufan [15] introduced several popular model based methods, including their pros and cons. For learning based, it specifically mentioned that using Inverse RL to learn from human demonstration. Then it detailed introduced the states including observable and unobservable, robot itself and its surrounding agents. Then, a value function is defined in continues time. Also, compared with giving the comprehensive definition of social norms, it will be easier to define what is not acceptable in social interaction. For this research problem, it will be a union of passing/overtaking/crossing uncomfortable zones. Also, details about training multiagent value network are also given: feed the robot & nearby agents' states to a neural network (rectified linear unit (ReLU)) , which will generate value function, and then the value fucntion will be compenstated with penalty rewards (unacceptable social norms, or say, uncomfortable zone).

Yufan [16] developed a decentralized multiagent collision avoidance algorithm, which is based on value iteration rather than Q iteration: because unlike previous works that focus on discrete, finite action spaces, the action space here is continuous and the set of permissible velocity vectors depends on the agent’s state (preferred speed).

Conclusion

In this survey, we review and discussed the procedures to build a LfD research problem. A key challenge in this field is how to make the robot recognize the undefined (poorly defined) scenario. Because the high complexity of real world, the teacher cannot upload all possible scenarios, the robot may execute undesired action when the world state is partially new. To solve this problem, we reviewed recent developments in reinforcement learning, to help the robot perform indirect learning from the given reward function, and refine its policy based on the feedback from the surrounding environment. Another way to solve this problem is by allow collaborative development of robot learning system [18], which allows multiple teacher to train the same agent.

Reference

[1] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, “Imitation learning: A survey of learning methods,” *ACM Comput. Surv.*, vol. 50, no. 2, 2017.

[2] W. Y. G. Louie and G. Nejat, “A Social Robot Learning to Facilitate an Assistive Group-Based Activity from Non-expert Caregivers,” *Int. J. Soc. Robot.*, vol. 12, no. 5, pp. 1159–1176, 2020.

[3] W. Y. G. Louie and G. Nejat, “A learning from demonstration system architecture for robots learning social group recreational activities,” *IEEE Int. Conf. Intell. Robot. Syst.*, vol. 2016-Novem, pp. 808–814, 2016.

[4] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, “A survey of robot learning from demonstration,” *Rob. Auton. Syst.*, vol. 57, no. 5, pp. 469–483, 2009.

[5] Maxim Lapan, *Deep Reinforcement Learning Hands-On 2nd Edition*. 2018.

[6] H. Joo, T. Simon, M. Cikara, and Y. Sheikh, “Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction,” *arXiv*, 2019.

[7] A. Hijaz, J. Korneder, and W. G. Louie, “In-the-Wild Learning from Demonstration for Therapies for Autism Spectrum Disorder.”

[8] “TO DELIVER ROBOT-MEDIATED INTERVENTIONS TO INDIVIDUALS WITH,” 2021.

[9] S. Raza, S. Haider, and M. A. Williams, “Teaching coordinated strategies to soccer robots via imitation,” *2012 IEEE Int. Conf. Robot. Biomimetics, ROBIO 2012 - Conf. Dig.*, pp. 1434–1439, 2012.

[10] I. Rodriguez, A. Manfré, F. Vella, I. Infantino, and E. Lazkano, “Talking with Sentiment: Adaptive Expression Generation Behavior for Social Robots,” *Adv. Intell. Syst. Comput.*, vol. 855, pp. 209–223, 2019.

[11] R. Kulikovskiy, M. Sochanski, A. Hijaz, M. Eaton, J. Korneder, and W. G. Louie, “Can Therapists Design Robot-Mediated Interventions and Teleoperate Robots Using VR to Deliver Interventions for ASD ?”

[12] T. Brys, A. Harutyunyan, H. B. Suay, S. Chernova, M. E. Taylor, and A. Nowé, “Reinforcement learning from demonstration through shaping,” *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2015-January, pp. 3352–3358, 2015.

[13] C. Chen, Y. Liu, S. Kreiss, and A. Alahi, “Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning,” *Proc. - IEEE Int. Conf. Robot. Autom.*, vol. 2019-May, pp. 6015–6022, 2019.

[14] C. Chen, Z. Al-Halah, and K. Grauman, “Semantic Audio-Visual Navigation,” 2020.

[15] Y. F. Chen, M. Everett, M. Liu, and J. P. How, “Socially aware motion planning with deep reinforcement learning,” *IEEE Int. Conf. Intell. Robot. Syst.*, vol. 2017-Septe, pp. 1343–1350, 2017.

[16] Y. F. Chen, M. Liu, M. Everett, and J. P. How, “Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning,” *Proc. - IEEE Int. Conf. Robot. Autom.*, pp. 285–292, 2017.

[17] M. K. Lee and L. Takayama, “‘Now, I have a body’: Uses and social norms for mobile remote presence in the workplace,” *Conf. Hum. Factors Comput. Syst. - Proc.*, pp. 33–42, 2011.

[18] B. Rossen and B. Lok, “A crowdsourcing method to develop virtual human conversational agents,” *Int. J. Hum. Comput. Stud.*, vol. 70, no. 4, pp. 301–319, 2012.