2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Efficient Learning of Socially Aware Robot Approaching Behavior Toward Groups via Meta-Reinforcement Learning

Chengxi Li[1], Ginevra Castellano[1] and Yuan Gao[1]

***Abstract*— Despite the recent rapid advancement of applying reinforcement learning in social robotics, the current state-of-the-art algorithms are still not able to adapt quickly when facing a similar but new social scenario. In this study, we examine meta-reinforcement learning (Meta-RL) methods in a task where a robot needs to learn to approach a group of human agents. For the robot to learn socially aware behavior quickly and effectively in multiple similar but different robot approaching scenarios, we incorporate proximal policy optimization (PPO) with the model-agnostic meta-learning frameworks, i.e., MAML and Reptile, in the training process. Our results show that the Meta-RL methods are feasible to be adopted in learning robot approaching behavior toward groups and they outperform the baseline PPO algorithm (train from scratch), measured by accumulated reward and time.**

I. INTRODUCTION

As intelligent robots have become more and more popular, they have become more involved in our daily lives, and this has led to more interactions and collaborations between humans and robots. In these interactions and collaborations, conformity to human social norms and expectations is an important principle. Currently, it is feasible to learn the social norm by adopting existing methods, however, the gained knowledge is task-specific and hard to transfer. This is unlike humans.

For humans, the ability of adapt to different situations efficiently is a significant advantage. To have better and more natural interaction and cooperation with humans, we should also try to enable robots to have such adaptation capabilities. However, learning algorithms facing new scenarios are often trained from scratch without considering any prior knowledge. In some cases, such a training method is tedious and resource-intensive due to objective constraints. Therefore, we want to be able to find a new approach that can use prior experience to improve the agent's ability to adapt to a new scenario.

In this work, our learning scenario is a robot agent approaching human groups. To be more specific, the robot needs to gain socially aware approaching behavior efficiently toward different formations of three-person human groups. Figure 1 shows an instance of our learning environment modelled using Unity3D. The silver agent is the robot agent who needs to learn to approach the human group which consists of three green human agents. The arrows around agents are the social forces applied to them according to Hall's proxemics theory [1] and Pedica's Social Force Model (SFM) [2].

Previous research [3] used Proximal Policy Optimization (PPO) [4] to train a two-person group. However, for our tasks, the diversity of the formation of three-person groups is indefinite and it is cumbersome to train from scratch for every group formation only using PPO. When the PPO is used to train the the robot agent in a specific scenario, the robot agent only learns a scenario-specific behavior. A method that could improve the learning efficiency should be considered. Inspired by the success of meta-learning in the field of data-efficient supervised learning, we utilize the meta-learning algorithms with deep reinforcement learning (DRL) methods for this purpose. The meta-learning methods we used are variants of classic algorithm Model-Agnostic Meta-Learning (MAML) [5], which are called First-Order MAML (FOMAML) and Reptile [6]. To sum up, our work aims to train a robot agent to gain a strong learning ability using Meta-RL in our learning scenarios so that the robot agent could use this ability to quickly adapt and generate approaching behavior when placed in a similar but unseen new scenario.

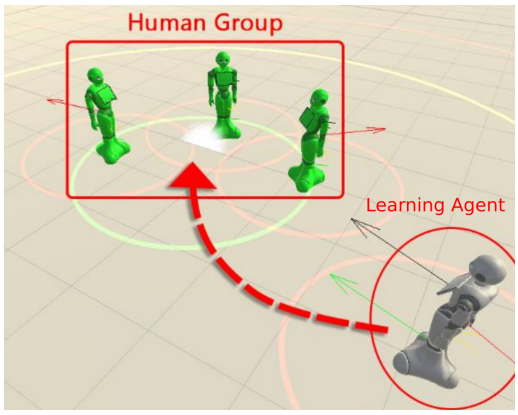

Fig. 1. The illustration of a learning scenario. The robot agent (silver agent) learns to approach a three-person human group (green agents) by using DRL methods. The arrows from robots indicate the social forces [2] applied to the agents. The light-colored circles on the ground represent different social spaces, namely personal, social, public spaces, proposed by Hall [1].

The contributions we made are as follows:

- We explore and test the feasibility of the reinforcement learning (RL) method PPO in learning socially aware robot approaching behavior toward different three-person groups.
- We build the virtual training environments of our scenarios for Meta-RL training.

[1]Department of Computer Science, Uppsala University, Sweden lcx.billy@gmail.com, {ginevra.castellano, alex.yuan.gao}@it.uu.se

- We propose to use the meta-learning method FOMAML, Reptile together with the PPO algorithm to train the robot agent in different scenarios. Our results show that the meta-learning based algorithms can outperform the baseline PPO algorithm (trained from scratch) in terms of accumulated reward and time in sampled testing scenarios.

II. Background

With the maturation and widespread use of humanoid robotics, there is a growing expectation for robots to be socially conscious, and with this comes new challenges and greater interest in human-robot interaction (HRI) research [7].

At this stage, there is a desire for robots to not only have a human-like appearance, but more importantly for them to have social thinking and produce behaviors more like humans. How to make robots behave like human beings in interactions can be said to be one of the hotspots in HRI research. In recent years, researchers in the field of HRI have explored various aspects in order to solve such problems, including but not limited to visual-based interaction [8] and language-based interaction [9]. However, due to the diversity of human consciousness, behavior, and social norms, it is difficult to explain human social interaction behavior in terms of a universal model. In turn, it is more difficult to use artificially developed models to derive a generic social strategy. Therefore, the approach of learning through interaction has received significant attention.

In recent years, DRL is a more widely used interaction-based learning method. Previous studies have shown that DRL can be used to handle a variety of classical tasks, such as human-machine control systems [10] and audio-visual gaze control [11]. For social HRI, RL methods have also been tried in a variety of social settings [12][13]. However, previous studies of interaction scenarios were only used to address specific tasks. When faced with similar but different tasks, a common approach is to train from scratch for each task and use transfer learning. However, such an approach limits the generalization of the learning model. Thus, Wang [14] first proposed the use of meta-learning methods to improve the efficiency and generalization capabilities of RL. Recently, the model-based parametric and gradient-based meta-learning method MAML [5] has achieved significant success in the field of supervised learning and RL. The method is based on optimizing the initial weights of deep neural networks. MAML is a framework compatible with tasks that can be optimized using gradients, thus it offers the opportunity to be combined with DRL algorithms.

III. Methods

In this section, the algorithms and learning scenario adopted in our work are introduced in detail.

A. Reinforcement Learning and Meta-learning

Reinforcement learning algorithms provide a framework for autonomous robots to learn how to make better decisions and actions by gaining rewards from their interactions with the environment [15]. In our work, the research goal is to use RL algorithms to build a robot agent that could learn socially aware approaching behavior towards three-person groups autonomously. We plant to adopt state-of-the-art RL method PPO to learn such approaching behavior in social scenarios. However, the composition of the three-person groups is diverse and it is cumbersome if we learn from scratch when meeting a new scenario. As for the different learning scenarios, they are a series of similar but not the same, such learning scenarios distribution is consistent with the learning objectives of meta-learning. Therefore, meta-learning based PPO methods are used to learn a generic model from the diverse of three-person groups to improve the learning efficiency. We chose the classical gradient-based meta-learning algorithms FOMAML and Reptile incorporate with PPO to adaptively learn the approaching behavior towards three-person groups. The details of PPO can be found in Schulman's paper [16]. The idea of FOMAML and Reptile can be accessed in Alex Nichol's paper. [6]

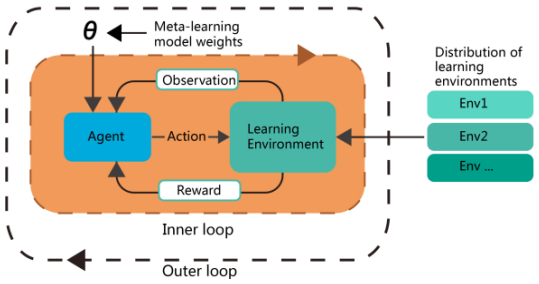

Fig. 2. General learning process of optimization procedure in FOMAML and Reptile. The learning process includes the two optimization loops, the inner loop is same as the learning process of normal PPO and the outer loop is used to update learning parameters of meta-learning based PPO.

Both FOMAML and Reptile are meta-learning algorithms based on gradient optimization. From a macro perspective, they aim to improve the overall learning ability of the model. In theory, they are concerned with having good initialization of parameters for PPO. Good initialization of parameters enable the PPO to achieve a better learning ability, allowing it to perform well even when trained on a small amount of data. The general learning process of FOMAML and Reptile are shown in the following Figure 2.

B. Learning Environment

In our work, the methods are developed based on RL. To verify the feasibility of our approach, we use Unity3D and ML-Agents platform [17] to design and build learning scenarios that can be compatible with RL methods. The overview of a learning scenario can be seen in Figure 1.

Specifically, we aim to train a robot agent to learn appropriate approaching behavior toward a three-person human group effectively in the learning scenarios. The robot agent interacts with a three-person group consisting of human agents in a learning scenario, and then we combine various factors such as the final state, social norm, and running steps

to judge the behaviors. The criteria for judging whether the interaction following social norms, we use the SFM, and the details of how to quantify the evaluation can be referred to as Yuan's paper [3].

IV. Experiments

The experimental evaluation is carried out to figure out the following research questions: (1) Could the meta-learning based PPO using FOMAML and Reptile learn faster than the normal PPO (train from scratch) when the agent is spawned in a new social scenario? (2) Could the meta-learning based methods PPO using FOMAML and Reptile acquire higher accumulated rewards than the normal PPO (train from scratch)?

In order to verify the feasibility of the proposed method, we sample three test scenarios for our experiment. The test scenarios contain three cases (i.e., single-side, uniform and asymmetric formations) and the group formations are shown in Fig 3. These three scenarios are selected as the single-side, uniform and asymmetric formations are typical formations of three-person groups.

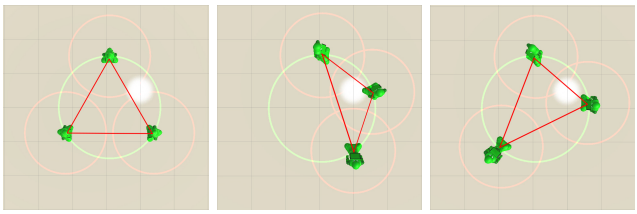

Fig. 3. The three test scenarios with different group formations, which we adopted in our experiments. They are uniform, single-side and asymmetric formations from left to right.

In our experiments, the normal PPO (train from scratch) method is adopted as a baseline. For meta-learning based PPO (using FOMAML and Reptile), they are pre-trained for 80 iterations in random scenarios to acquire a generalized model initialization.Then, we evaluated the performance of meta-learning based PPO and the baseline in terms of the accumulative reward. Figure 4 shows the learning curves of three test scenarios after training for 500 epochs. In each figure, the blue line represents the learning curve of the normal PPO method, while the yellow and green lines represent FOMAML and Reptile based PPO respectively. The red line represents a uniformly random policy. Each curve is smoothed by using a moving average of last 40 epochs. We could observe that in the first 100 epochs, the meta-learning based PPO methods are already able to learn a stable policy and acquire a higher accumulated reward than the normal PPO (train from scratch) in three different scenarios. However, meta-learning based PPO methods have a larger variance at the beginning in all the scenarios. One assumption is that this phenomenon is caused by the the learning agent's randomly initialized positions. After training for 500 epochs, the meta-learning based PPO methods have a much smaller variance than the normal PPO (train from scratch).

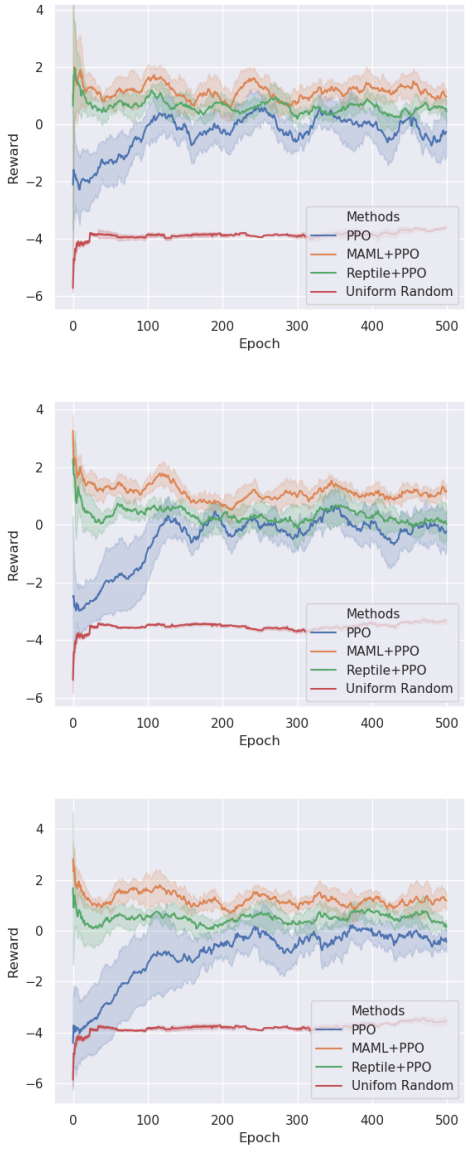

Fig. 4. Learning curves of MAML, Reptile and PPO for 500 epochs in testing group formations for 6 groups valdiate experiment. Uniform, single-side and asymmetric formation from top to bottom

Table IV shows the best results of each method in three different scenarios. For the scenario of uniform group formation, the FOMAML+PPO method achieves the highest accumulated reward 1.71 ± 0.28. Simultaneously, the FOMAML+PPO method also achieves the highest accumulated rewards in scenarios of asymmetric and single-side group formations, reaching $1.80(\pm 0.55)$ and $1.78(\pm 0.32)$ respectively.

As a consequence, we could conclude that, while the number of training episodes increases, the policies generated by meta-learning based PPO methods become more stable and they can gain a higher accumulated reward than the normal PPO (train from scratch) at the end.

In order to better present our results, in addition to

Group form	PPO(from scratch)	FOMAML	Reptile
Uniform	0.63(±0.52)	1.71(±0.28)	1.18(±0.20)
Asymmetric	0.22(±0.49)	1.80(±0.55)	0.86(±0.41)
Single-side	0.65(±0.51)	1.78(±0.32)	0.74(±0.22)

the quantitative analysis, we also included qualitative presentations. The Figure 5 shows partial trajectories of the approaching behavior toward three-person group we obtained in a learning scenario by meta-learningbased PPO with FOMAML.

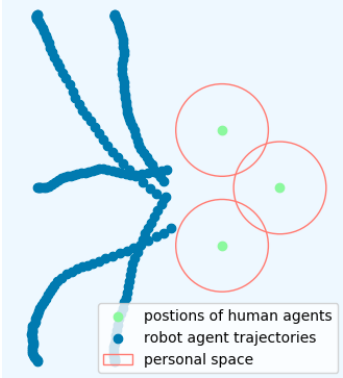

Fig. 5. Part of the trajectories are generated by meta-learning based PPO with FOMAML in the learning scenarios.

V. Discussion and Conclusions

In this work, we present a learning framework for improving the efficiency and effectiveness of social robotics learning using a meta-learning approach. And we design and implement virtual environments with multiple types of three-person groups to evaluate whether our approach is effective compared to PPO (train from scratch). Through experiments, we learned that the meta-learning approach outperforms PPO (training from scratch) both in terms of learning efficiency and final reward.

In the future, we will fine-tune our virtual environments and the experimental setups to be more similar to the real group formations. In terms of evaluation, we will 1) test our results in video-based perceptual studies. 2) implement our learned models on a physical Pepper robot so that we can test our methods in real scenarios.

Acknowledgement

This work was supported by the COIN project (RIT15-0133) funded by the Swedish Foundation for Strategic Research and by the Swedish Research Council (grant n. 2015-04378)

References

[1] Edward T Hall et al. “Proxemics [and comments and replies]”. In: *Current anthropology* 9.2/3 (1968), pp. 83–108.

[2] Claudio Pedica and Hannes Vilhjálmsson. “Social perception and steering for online avatars”. In: *International Workshop on Intelligent Virtual Agents*. Springer. 2008, pp. 104–116.

[3] Y. Gao et al. “Learning Socially Appropriate Robot Approaching Behavior Toward Groups using Deep Reinforcement Learning”. In: *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 2019, pp. 1–8.

[4] John Schulman et al. “Proximal policy optimization algorithms”. In: *arXiv preprint arXiv:1707.06347* (2017).

[5] Chelsea Finn, Pieter Abbeel, and Sergey Levine. “Model-agnostic meta-learning for fast adaptation of deep networks”. In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 1126–1135.

[6] Alex Nichol, Joshua Achiam, and John Schulman. “On first-order meta-learning algorithms”. In: *arXiv preprint arXiv:1803.02999* (2018).

[7] Kerstin Dautenhahn. “Socially intelligent robots: dimensions of human–robot interaction”. In: *Philosophical transactions of the royal society B: Biological sciences* 362.1480 (2007), pp. 679–704.

[8] Henny Admoni and Brian Scassellati. “Social eye gaze in human-robot interaction: a review”. In: *Journal of Human-Robot Interaction* 6.1 (2017), pp. 25–63.

[9] Rehj Cantrell et al. “Robust spoken instruction understanding for HRI”. In: *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2010, pp. 275–282.

[10] Hamidreza Modares et al. “Intelligent human–robot interaction systems using reinforcement learning and neural networks”. In: *Trends in control and decision-making for human–robot collaboration systems*. Springer, 2017, pp. 153–176.

[11] Stéphane Lathuilière et al. “Neural network based reinforcement learning for audio–visual gaze control in human–robot interaction”. In: *Pattern Recognition Letters* 118 (2019), pp. 61–71.

[12] Ahmed Hussain Qureshi et al. “Show, attend and interact: Perceivable human-robot social interaction through neural attention Q-network”. In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2017, pp. 1639–1645.

[13] Maja J Matarić. “Learning social behavior”. In: *Robotics and Autonomous Systems* 20.2-4 (1997), pp. 191–204.

[14] Jane X Wang et al. “Learning to reinforcement learn”. In: *arXiv preprint arXiv:1611.05763* (2016).

[15] Niko Sünderhauf et al. “The limits and potentials of deep learning for robotics”. In: *The International Journal of Robotics Research* 37.4-5 (2018), pp. 405–420.

[16] John Schulman et al. “Proximal Policy Optimization Algorithms”. In: *CoRR* abs/1707.06347 (2017). arXiv: 1707.06347. URL: http://arxiv.org/abs/1707.06347.

[17] Arthur Juliani et al. “Unity: A General Platform for Intelligent Agents”. In: *arXiv preprint arXiv:1809.02627* (2018).