IntentionGAN: Multi-Modal Imitation Learning from Unstructured Demonstrations

Karol Hausman*, Yevgen Chebotar*, Stefan Schaal, Gaurav Sukhatme, Joseph J. Lim
University of Southern California, Los Angeles, USA

I. Introduction

Traditionally, imitation learning has focused on using isolated demonstrations of a particular skill [3]. The demonstration is usually provided in the form of kinesthetic teaching, which requires the user to spend sufficient time to provide the right training data. This constrained setup for imitation learning is difficult to scale to real world scenarios, where robots have to be able to execute a combination of different skills. To learn these skills, the robots would require a large number of robot-tailored demonstrations, since at least one isolated demonstration has to be provided for every individual skill.

In order to improve the scalability of imitation learning, we propose a framework that can learn to imitate skills from a set of unstructured and unlabeled demonstrations of various tasks.

As a motivating example, consider a highly unstructured data source, e.g. a video of a person cooking a meal. A complex activity, such as cooking, involves a set of simpler skills such as grasping, reaching, cutting, pouring, etc. In order to learn from such data, three components are required: i) the ability to map the image stream to state-action pairs that can be executed by a robot, ii) the ability to segment the data into simple skills, and iii) the ability to imitate each of the segmented skills. In this work, we tackle the latter two components, leaving the first one for future work.

In this paper, we present a novel imitation learning method that learns a multi-modal stochastic policy, which is able to imitate a number of automatically segmented tasks using a set of unstructured and unlabeled demonstrations. Our results indicate that the presented technique can separate the demonstrations into sensible individual skills and imitate these skills using a learned multi-modal policy.

II. Multi-modal Imitation Learning

The traditional imitation learning scenario considers a problem of learning to imitate one skill from demonstrations. The demonstrations represent samples from a single expert policy π_{E1}. In this work, we focus on an imitation learning setup where we learn from unstructured and unlabelled demonstrations of various tasks. The demonstrations come from a set of expert policies $\pi_{E_1}, \pi_{E_2}, \ldots, \pi_{E_k}$, where k can be unknown, that optimize different reward functions/tasks. We refer to this set of unstructured expert policies as a mixture of policies π_E. We aim to segment the demonstrations of these policies into separate tasks and learn a multi-modal policy that imitates all of them.

To be able to learn multi-modal policy distributions, we augment the policy input with a latent intention i distributed by a categorical or uniform distribution $p(i)$, similar to [1]. The goal of the intention variable is to select a specific mode of the policy, which corresponds to one of the skills presented in the demonstrations. The resulting policy can be expressed as $\pi^i(a|s,i) = p(i|s,a)\frac{\pi^i(a|s)}{p(i)}$.

We augment the trajectory τ to include the latent intention as $\tau_i = (s_0, a_0, i_0, ...s_T, a_T, i_T)$. The resulting reward of the trajectory with the latent intention is $R(\tau_i) = \sum_{t=0}^{T} \gamma^t R(s_t, a_t, i_t)$. $R(a,s,i)$ is a reward function that depends on the latent intention i as we have multiple demonstrations that optimize different reward functions for different tasks. The expected discounted reward is equal to: $\mathbb{E}_{\pi_\theta^i}[R(\tau_i)] = \int R(\tau_i)\pi_\theta^i(\tau_i)d\tau_i$ where $\pi_\theta(\tau_i) = p_0(s_0)\prod_{t=0}^{T-1} P(s_{t+1}|s_t,a_t)\pi_\theta^i(a_t|s_t,i_t)p(i_t)$.

Here, we show an extension of the derivation presented in [2] for a policy $\pi^i(a|s,i)$ augmented with the latent intention variable i, which uses demonstrations from a set of expert policies π_E. We are aiming at maximum entropy policies that can be determined from the latent intention variable i. Accordingly, we transform the original max-entropy inverse reinforcement learning (IRL) problem [5] to reflect this goal: $\max_R \left(\max_{\pi^i} H(\pi^i(a|s)) - H(\pi^i(a|s,i)) + \mathbb{E}_{\pi^i} R(s,a,i)\right) - \mathbb{E}_{\pi_E} R(s,a,i)$. This objective reflects our goal: we aim to obtain a multi-modal policy that has a high entropy without any given intention, but it collapses to a particular task when the intention is specified. Analogously to the solution for a single expert policy presented in [2], this optimization objective results in the optimization of the generative adversarial imitation learning network with the state-action pairs (s,a) being sampled from a set of expert policies π_E:

$$\begin{aligned}\max_\theta \min_w \; & \mathbb{E}_{i\sim p(i),(s,a)\sim\pi_\theta^i}[\log(D_w(s,a))] \\ & + \mathbb{E}_{(s,a)\sim\pi_E}[1 - \log(D_w(s,a))] \\ & + \lambda_H H(\pi_\theta^i(a|s)) - \lambda_I H(\pi_\theta^i(a|s,i)),\end{aligned} \tag{1}$$

where λ_I, λ_H correspond to the weighting parameters on the respective objectives. The resulting entropy $H(\pi_\theta^i(a|s,i))$ term can be expressed as

$$\begin{aligned}H(\pi_\theta^i(a|s,i)) &= \mathbb{E}_{i\sim p(i),(s,a)\sim\pi_\theta^i}(-\log(\pi_\theta^i(a|s,i)) \\ &= -\mathbb{E}_{i\sim p(i),(s,a)\sim\pi_\theta^i}\log(p(i|s,a)) + H(\pi_\theta^i(a|s)) - H(i),\end{aligned} \tag{2}$$

* Equal contribution

where $H(i)$ is a constant that does not influence the optimization. This results in the same optimization objective as for the single expert policy [2] with an additional term $\lambda_I \mathbb{E}_{i\sim p(i),(s,a)\sim\pi_\theta^i} \log(p(i|s,a))$ responsible for rewarding state-action pairs that make the latent intention inference easier. We refer to this cost as the latent intention cost and represent $p(i|s,a)$ with a neural network.

III. Experiments

Reacher The actuator is a 2-DoF arm attached at the center of the scene. There are two targets placed at random positions throughout the environment. The goal of the task is, given a data set of reaching motions to random targets, to discover the dependency of the target selection on the intention and learn a policy that is capable of reaching different targets based on the specified intention input.

Walker-2D The Walker-2D is a 6-DoF bipedal robot consisting of two legs and feet attached to a common base. The goal of this task is to learn a policy that can switch between three different behaviors dependent on the discovered intentions: running forward, running backward and jumping. We use TRPO [4] to train single expert policies and create a combined data set of all three behaviors that is used to train a multi-modal policy using our imitation framework.

Humanoid Humanoid is a high-dimensional robot with 17 degrees of freedom. Similar to Walker-2D the goal of the task is to be able to discover three different policies: running forward, running backward and balancing, from the combined expert demonstrations of all of them.

The performance of our method in all of these setups can be seen in our supplementary video: http://sites.google.com/view/nips17intentiongan.

We first evaluate the influence of the latent intention cost on the Reacher task. For these experiments, we use a categorical intention distribution with the number of categories equal to the number of targets.

To demonstrate the development of different intentions, in Fig. 1 (left) we present the Reacher rewards over training iterations for different intention variables. When the latent intention cost is included, (Fig. 1-1), the separation of different skills for different intentions starts to emerge around the 1000-th iteration and leads to a multi-modal policy that, given the intention value, consistently reaches the target associated with that intention. In the case of the standard imitation learning GAN setup (Fig. 1-2), the network learns how to imitate reaching only one of the targets for both intention values.

We also seek to further understand whether our model extends to segmenting and imitating policies that perform different tasks. In particular, we evaluate whether our framework is able to learn a multi-modal policy on the Walker-2D task. The results are depicted in Fig. 2 (left). The additional latent intention cost results in a policy that is able to autonomously segment and mimic all three behaviors and achieve a similar performance to the expert policies (Fig. 2-1). Different intention variable values correspond to different expert policies: 0 - running forwards, 1 - jumping, and 2 - running backwards. The imitation learning

Fig. 1. Rewards of different Reacher policies for 2 targets for different intention values over the training iterations with (1) and without (2) the latent intention cost.

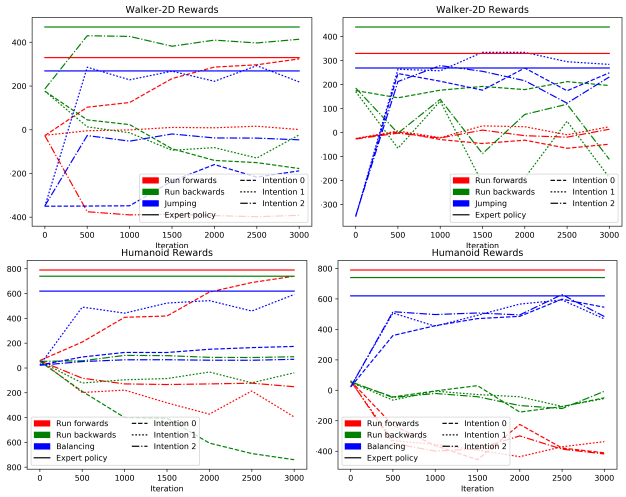

Fig. 2. *Top:* Rewards of Walker-2D policies for different intention values over the training iterations with (1) and without (2) the latent intention cost. *Bottom:* Rewards of Humanoid policies for different intention values over the training iterations with (3) and without (4) the latent intention cost.

GAN method is shown as a baseline in Fig. 2-2. The results show that the policy collapses to a single mode, where all different intention variable values correspond to the jumping behavior, ignoring the demonstrations of the other two skills.

To test if our multi-modal imitation learning framework scales to high-dimensional tasks, we evaluate it in the Humanoid environment. Fig. 2 (right) shows the rewards obtained for different values of the intention variable. Similarly to Walker-2D, the latent intention cost enables the neural network to segment the tasks and learn a multi-modal imitation policy. In this case, however, due to the high dimensionality of the task, the resulting policy is able to mimic running forwards and balancing policies almost as well as the experts, but it achieves a suboptimal performance on the running backwards task (Fig. 2-3). The imitation learning GAN baseline collapses to a uni-modal policy that maps all the intention values to a balancing behavior (Fig. 2-4).

IV. Conclusions

We present a novel imitation learning method that learns a multi-modal stochastic policy, which is able to imitate a number of automatically segmented tasks using a set of unstructured and unlabeled demonstrations. The presented approach learns the notion of intention and is able to perform different tasks based on the policy intention input.

REFERENCES

[1] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets, 2016.

[2] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *CoRR*, abs/1606.03476, 2016.

[3] Stefan Schaal. Is imitation learning the route to humanoid robots? *Trends in cognitive sciences*, 3(6):233–242, 1999.

[4] John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. In Francis R. Bach and David M. Blei, editors, *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1889–1897. JMLR.org, 2015.

[5] Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In Dieter Fox and Carla P. Gomes, editors, *AAAI*, pages 1433–1438. AAAI Press, 2008.