Combating False Negatives in Adversarial Imitation Learning (Student Abstract)

Konrad Żołna,[*,1,2] **Chitwan Saharia,**[*,2,3] **Léonard Boussioux,**[*,2,4,5]
David Yu-Tung Hui,[2] **Maxime Chevalier-Boisvert,**[2] **Dzmitry Bahdanau,**[2,6] **Yoshua Bengio**[2,7]

[1]Jagiellonian University[†], [2]Mila, [3]IIT Bombay, [4]MIT,
[5]École CentraleSupélec, [6]Element AI, [7]CIFAR Senior Fellow
{konrad.zolna, chitwaniit}@gmail.com, leobix@mit.edu

Abstract

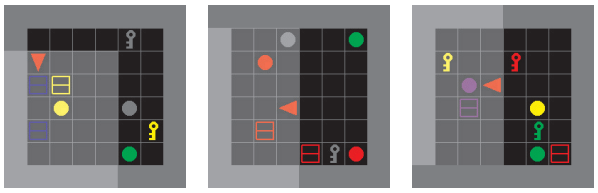
We define the False Negatives problem and show that it is a significant limitation in adversarial imitation learning. We propose a method that solves the problem by leveraging the nature of goal-conditioned tasks. The method, dubbed Fake Conditioning, is tested on instruction following tasks in BabyAI environments, where it improves sample efficiency over the baselines by at least an order of magnitude.

Introduction

Progress in grounded language learning using Deep Reinforcement Learning is impeded by the necessity of handcrafting reward functions (Luketina et al. 2019). Instead of using a reward to judge the quality of performance, Imitation Learning (IL) trains an agent using expert demonstrations to mimic a presented policy. The simplest version of IL, Behavioral Cloning (BC) (Bain and Sammut 1995), trains a policy to regress expert actions from demonstrations in a supervised setup.

Another IL method, Generative Adversarial Imitation Learning (GAIL) (Ho and Ermon 2016), has yielded some success by jointly learning reward functions and training policies. GAIL trains a discriminator to differentiate agent from expert trajectories, which simultaneously acts as a reward function. Hence, the agent tries to act more and more like the expert in order to fool the discriminator and get a higher reward.

While GAIL works well in the starting phase of the learning procedure, we observed that once the agent is able to solve the given task, its performance tends to be unstable. We show that this is due to the fact that the discriminator has to classify successful episodes from the agent as fake samples, even though they are very similar to expert demonstrations. This problem is negligible during the starting phase when the agent executes a random policy, but as the policy improves the number of such successful trajectories labeled as non-expert increases. We refer to this phenomenon as the *False Negatives* (FN) problem, as successful agent trajectories are falsely labeled as negative examples.



(a) GoToLocal *go to the yellow key*

(b) PickupLoc *pick up a box in front of you*

(c) PutNextLocal *put the green ball next to the red key*

Figure 1: Instructions and initial states for three BabyAI tasks. In (b), a memory in a reward model is necessary to determine success which depends on the initial position.

We investigate the FN problem within the case of jointly learning to understand language which specifies task objectives. We focus on three levels of increasing difficulty (see Figure 1) on the BabyAI platform (Chevalier-Boisvert et al. 2018) and show that a naive application of GAIL is not able to solve any level due to the aforementioned problem.

We propose a method that leverages the multi-goal nature of the setup to solve the FN problem. In particular, we replace the agent's true instruction with a random one when training the discriminator, thereby ensuring that the new goal and trajectory pair truly is a negative example, even when the agent's trajectory is an example of successful instruction execution. We show that this technique enables the agent's performance to approach 100% success rate of instruction following, while a naive application of GAIL fails.

Conditioned Recurrent Discriminator

We observed that equipping the discriminator with memory is necessary to model the true environment reward (see Figure 1(b) for example). Hence, the discriminator is implemented as a recurrent neural network. Its input is not a single state-action pair (s_i, a_i), as in the original GAIL implementation, but a full trajectory, i.e. $\tau = ((s_0, a_0), \ldots, (s_t, a_t))$. To address the goal-conditioned nature of the problem, the discriminator has to be conditioned on the instructions c.

[*]Equal contribution. Listed in random order.

[†]ul. Łojasiewicza 6, 30-348 Kraków, Poland, +48 12 664 66 29

Table 1: Success rate percentage for different models. For each task we consider three expert demonstrations set sizes. The largest one is the minimal necessary demonstration set needed to solve the tasks using BC. Two smaller subsets are tested (8 times and 64 times smaller). A task is considered solved if the agent achieves more than 99% success rate (bold values).

Model	GoToLocal			PickupLoc			PutNextLocal		
	$\frac{1}{64}$	$\frac{1}{8}$	1	$\frac{1}{64}$	$\frac{1}{8}$	1	$\frac{1}{64}$	$\frac{1}{8}$	1
Behavioral Cloning	67.3	90.8	**99.9**	65.2	95.9	**100.0**	26.8	77.9	**99.9**
Baseline GAIL	64.6	86.7	98.4	59.5	89.9	98.7	30.7	77.2	97.2
Oracle Filtering	**99.5**	**99.4**	–	94.9	**99.5**	–	84.2	**99.0**	–
Fake Conditioning	85.8	**99.5**	–	89.4	98.5	–	94.8	**99.5**	–

The conditioned recurrent discriminator loss is the following:

$$L_D(\theta) = \mathbb{E}_{(c,\tau)\sim B_{agent}} - \log(1 - D_\theta(c,\tau)) \quad (1)$$
$$+\mathbb{E}_{(c,\tau)\sim B_{expert}} - \log(D_\theta(c,\tau)),$$

where B_{agent} and B_{expert} represent the agent trajectories and expert demonstration, respectively.

Using whole trajectories to train the results in a more stable training procedure. We hypothesize that this is due to the partial observability of the environment which makes short sub-trajectories of B_{expert} elements hard to discriminate from unsuccessful agent trajectories.

False Negatives

The streams of positive and negative examples that are input to the GAIL discriminator can become very similar as the agent gets better due to FN problem. The discriminator, playing the role of reward model in GAIL, can no longer assume that all successful trajectories are expert and has to detect idiosyncratic features in expert demonstrations that are not necessarily related to solving a given task.

Oracle Filtering To analyze the impact of FN problem on the performance of GAIL training, we use the environment's true reward signal to determine the successful trajectories and filter them out resulting in B_{agent} consisting of only unsuccessful trajectories. We call this *Oracle Filtering*.

Fake Conditioning Since the Oracle Filtering technique requires access to environment rewards, it cannot be applied in practice and is used to diagnose FN problem only. We propose a technique that does not need environment rewards to rectify this fundamental limitation. Our technique is aimed at tasks where the policy is goal-conditioned. In our case, we assume that the task is conditioned on language instructions but the technique is general and can be applied in any multi-goal setup.

Firstly, we maintain a set of possible language instructions S which is initialized from all the unique instructions in the expert demonstrations and updated with the instructions collected when the agent interacts with the environment. Then, for each trajectory (c,τ) sampled from the agent buffer B_{agent}, we replace it with $(\tilde{c},\tau)$, where $\tilde{c} \sim S\backslash\{c\}$. Hence, in this case, the discriminator's loss is given as follows:

$$L_D(\theta) = \mathbb{E}_{(c,\tau)\sim B_{agent}} - \log(1 - D_\theta(\tilde{c},\tau)) \quad (2)$$
$$+\mathbb{E}_{(c,\tau)\sim B_{expert}} - \log(D_\theta(c,\tau)).$$

We call this technique *Fake Conditioning*. It is motivated by the fact that the success of a trajectory is conditioned on the instruction. Therefore, for each trajectory generated by the agent, through replacing the instruction with a random one, we can ensure that the instruction and trajectory pair is not a false negative anymore, even when the agent's trajectory is successfully conditioned on the original instruction.

Experiments

The performance of IL algorithms depends on the number of expert demonstrations. Chevalier-Boisvert et al. (2018) report the minimal number of demonstrations needed to solve each BabyAI task using BC. To make a fair comparison, we use exactly the same policy architecture. We report all our experimental results in Table 1.

Oracle Filtering significantly improves the performance and can solve levels with orders of magnitude fewer expert demonstrations. This decisive improvement is achieved by only filtering out successful agent trajectories, which experimentally proves that the FN problem is the main limiting factor for GAIL.

Fake Conditioning proves to be very effective in solving all tasks using an order of magnitude fewer demonstrations than BC. Even when 64 times fewer demos are used, its performance ($\approx 90\%$) is satisfactory and much better than the GAIL baseline ($\approx 50\%$).

Our experiments show that addressing the FN problem is crucial for achieving good performance when training by adversarial imitation learning.

Acknowledgments

Konrad Żołna is supported by the National Science Center, Poland (2017/27/N/ST6/00828, 2018/28/T/ST6/00211).

References

Bain, M., and Sammut, C. 1995. A framework for behavioural cloning. In *Machine Intelligence*.

Chevalier-Boisvert, M.; Bahdanau, D.; Lahlou, S.; Willems, L.; Saharia, C.; Nguyen, T. H.; and Bengio, Y. 2018. BabyAI: A platform to study the sample efficiency of grounded language learning. In *ICLR*.

Ho, J., and Ermon, S. 2016. Generative adversarial imitation learning. In *NeurIPS*.

Luketina, J.; Nardelli, N.; Farquhar, G.; Foerster, J. N.; Andreas, J.; Grefenstette, E.; Whiteson, S.; and Rocktäschel, T. 2019. A survey of reinforcement learning informed by natural language. *arXiv*.