Open Access **Baghdad Science Journal** **P-ISSN: 2078-8665**
2021, 18(4) Supplement: 1350-1355 **E-ISSN: 2411-7986**

DOI: https://dx.doi.org/10.21123/bsj.2021.18.4(Suppl.).1350

Generative Adversarial Network for Imitation Learning from Single Demonstration

Tho Nguyen Duc* ***Chanh Minh Tran*** ***Phan Xuan Tan*** ***Eiji Kamioka***

School of Engineering and Science, Shibaura Institute of Technology, Japan.
*Corresponding author: nb20501@shibaura-it.ac.jp
E-mails: nb20502@shibaura-it.ac.jp , tanpx@shibaura-it.ac.jp , kamioka@shibaura-it.ac.jp

Received 15/10/2021, Accepted 14/11/2021, Published 20/12/2021

Abstract:

Imitation learning is an effective method for training an autonomous agent to accomplish a task by imitating expert behaviors in their demonstrations. However, traditional imitation learning methods require a large number of expert demonstrations in order to learn a complex behavior. Such a disadvantage has limited the potential of imitation learning in complex tasks where the expert demonstrations are not sufficient. In order to address the problem, a Generative Adversarial Network-based model is proposed which is designed to learn optimal policies using only a single demonstration. The proposed model is evaluated on two simulated tasks in comparison with other methods. The results show that our proposed model is capable of completing considered tasks despite the limitation in the number of expert demonstrations, which clearly indicate the potential of our model.

Keywords: Deep Learning, Few-shot Learning, Generative Adversarial Network, Imitation Learning, One-shot Learning.

Introduction:

Imitation learning, known also as learning from demonstration, has recently gained significant attention since it enables training autonomous agents in complex environments where reward functions are unavailable. The main goal of imitation learning is to imitate expert behaviors in their demonstrations by learning a mapping between observation states and actions. It is widely adopted in robotic and human-computer interaction fields [1], such as self-driving vehicles [2–4] and social robot interaction [5,6]. However, traditional imitation learning algorithms usually require a significant number of demonstrations in order to acquire complex behaviors from the expert, since it is challenging to train an agent using a few or only one demonstration.

Indeed, humans are capable of learning a new behavior by observing it produced by the expert just once. Inspired by this human ability, the authors in [7] proposed a reinforcement learning model to solve an Atari game using only one demonstration. The model is trained by starting from a carefully selected state in the demonstration. One main drawback of the model is that it requires reward signals for every timestep. In a game environment, these reward signals can be easily collected. However, in a typical imitation learning setting, reward functions are unavailable and can be difficult to be defined manually.

Therefore, in this paper, a model to imitate expert behaviors from a single demonstration is proposed. Moreover, the model leverages Generative Adversarial Network in order to learn optimal policies without having access to the reward function. The main contributions of this paper are as follows:

- A GAN-based model is presented for imitating expert behaviors from a single demonstration.
- A comprehensive evaluation is conducted, which demonstrates the potential of our proposed model.

The rest of the paper is organized as follows. First, the related works of the proposed model is introduced. Second, the imitation learning problem is formulated. Third, the proposed model is presented in detail. Forth, the proposed model is evaluated, and the results is analyzed. Finally, the paper is concluded.

Related Work:

Imitation learning has been successfully applied to train autonomous agents [1,4,5,8] in many fields. Behavioral Cloning (BC) and Inversed Reinforcement Learning (IRL) are two main approaches to imitation learning. Behavioral Cloning [9] utilizes supervised learning in order to mimic expert behaviors. Although BC is a straightforward method, it is vulnerable to the distribution shift between the training and testing data. In contrast, IRL [10] has succeeded in a wide range of tasks by first trying to recover a reward function from expert demonstrations and then leveraging it to find an optimal policy. However, IRL requires an extremely high computational cost since iterations of reinforcement learning are involved during the training phase [11–13]. In order to overcome this drawback, recent studies [14,15] have applied Generative Adversarial Network [16] to imitate expert behaviors by finding a mapping between states and actions. However, the above-mentions methods require a significant number of demonstrations during the training phase. Ideally, the agent can have the same ability as humans in which they can imitate expert behaviors from only one or a few demonstrations.

The work in [7] proposed a reinforcement learning model learning to play an Atari game using only one demonstration. A carefully selected state in the demonstration is input to the model at each training step in order to imitate the expert behaviors and avoid learning a sub-optimal solution. However, the model requires reward signals in every timestep. These reward signals can be easily collected in an Atari game environment. However, in a typical imitation learning setting, reward functions are unavailable and can be difficult to be defined manually.

On the other hand, our proposed model leverages GAN to imitate expert behaviors without the need of a reward function. Moreover, the proposed model is capable of learning from only one expert demonstration and can provide a competitive performance in such a challenging setting.

Problem Formulation:

In this paper, the imitation learning problem is described as a Markov Decision Process (MDP) with finite time horizon:

$$\mathcal{M} = (S, A, P, T) \quad (1)$$

where, S denotes the state space, A is the action space, $P: S \times A \rightarrow S$ represents the transition function, and T is the time horizon. It is important to note that a shaped reward function is unavailable in imitation learning. A policy $\pi: S \rightarrow A$ represents a mapping from observation states to actions. An expert demonstration $\mathcal{D} = \{(s_{\mathcal{D}}^t, a_{\mathcal{D}}^t): t \in [0, T]\}$ is a sequence of state-action pairs. Our main objective is to learn an optimal policy π^* given a single demonstration $\mathcal{D}$.

The Proposed Model:

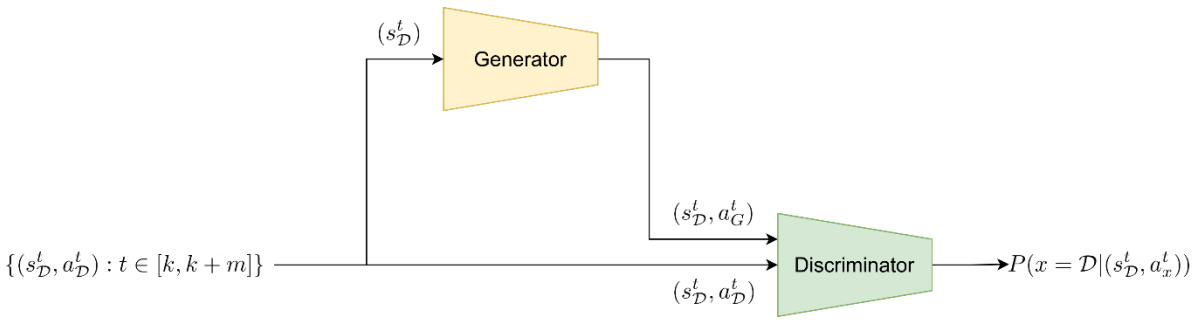


Figure 1. The architecture of the proposed model.

In this section, our proposed model is presented. The model leverages Generative Adversarial Network in order to learn optimal policies from a single expert demonstration $\mathcal{D}$. The architecture of the model is illustrated in Fig. 1. The model includes two deep feed-forward networks G and D.

The discriminator D is trained to distinguish between a state-action pair $(s_{\mathcal{D}}^t, a_{\mathcal{D}}^t)$ from the expert and a state-action pair $(s_{\mathcal{D}}^t, a_G^t)$ generated by the generator. Meanwhile, the generator G aims to produce an action a_G^t so that $(s_{\mathcal{D}}^t, a_G^t)$ looks as similar as possible to $(s_{\mathcal{D}}^t, a_{\mathcal{D}}^t)$. The model finds optimal policies by playing a min-max game with the discriminator trained together with the generator using the following objective function [14,16]:

$$max_{\,G}\, min_D\, \mathfrak{L}(G, D) \quad (2)$$

subject to $\mathfrak{L}(G,D)$

$$= \mathbb{E}[log\, D(s_{\mathcal{D}}^t, a_{\mathcal{D}}^t)] + \mathbb{E}\left[log\left(1 - D(s_{\mathcal{D}}^t, a_G^t)\right)\right] = \mathbb{E}[log\, D(s_{\mathcal{D}}^t, a_{\mathcal{D}}^t)] + \mathbb{E}\left[log\left(1 - D\left(s_{\mathcal{D}}^t, G(s_{\mathcal{D}}^t)\right)\right)\right] \quad (3)$$

The model acquires optimal policies by finding a saddle point, where:

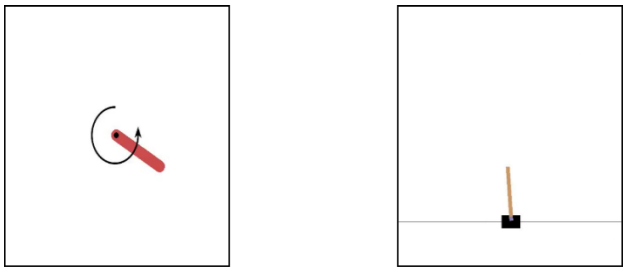
$$\hat{G} = argmax_G\, \mathfrak{L}(G, \hat{D}) \quad (4)$$
$$\hat{D} = argmin_D\, \mathfrak{L}(\hat{G}, D)$$

In order to train the model with only one demonstration, the demonstration is divided into multiple sub-demonstrations $\tau_k = \{(s_{\mathcal{D}}^t, a_{\mathcal{D}}^t): t \in [k, k+m)\}$ with the same length $0 < m \leq T$, where $k = 0,1,\ldots,(T-m+1)$ is the starting timestep. Within each training iteration, sub-demonstrations are feed into the model in random order to prevent overfitting and improve the stability of training.

Performance Evaluation:

In this section, the performance of the proposed model is evaluated. The evaluation settings and results are presented in the following subsections.

Evaluation Settings



(a) Pendulum **(b)** CartPole

Figure 2. Visual rendering of two simulated environments used in the evaluation

Two simulated environments is considered:

- Pendulum [17]: The pendulum starts at a random position. The goal of the task is to keep the pendulum stays upright by swinging it up.
- CartPole [17,18] : A pole is attached to a cart. The goal is to keeps it stays upright by applying a force of +1 or -1 to the cart.

The visualizations of the two environments are shown in Fig. 2. For each environment, one demonstration is collected by training the Trust Region Policy Optimization (TRPO) [19] which is a reinforcement learning algorithm for optimizing the learned policies by using gradient descent. The TRPO is trained with direct access to the environment and the shaped reward. In addition, the performance of our proposed model is also compared with TRPO. This baseline set an upper bound for the performance of our proposed model. The proposed model and TRPO are run on a personal computer with an Intel i7-8750H @ 2.20GHz and 16GB RAM system.

Network Structure and Hyperparameters

The generator and discriminator are deep feed-forward networks with 2 hidden layers. Each hidden layer has 32 nodes. Adam [20] is used which is a stochastic gradient descent algorithm to optimize the proposed model during the training phase. The Adam method is provided with a learning rate of 0.0003.

Results:

Fig. 3 and 4 visualize the behaviors of policies learned by the evaluated models on Pendulum and CartPole environments, respectively. Observing from Fig. 3, the policies trained with TRPO can swing the pendulum up faster and keeps it stays vertical for a longer period of time than our proposed model. Accordingly, the policies learned by our proposed model have trouble applying a strong enough force to swing the pendulum upright at first. However, after the pendulum is upright, the learned policies can apply few light forces to maintain it vertically. For the CartPole environment in Fig. 4, it can be observed that both policies trained with TRPO and our proposed model can move the cart in order to prevent the pole from falling over.

Baghdad Science Journal
P-ISSN: 2078-8665
E-ISSN: 2411-7986

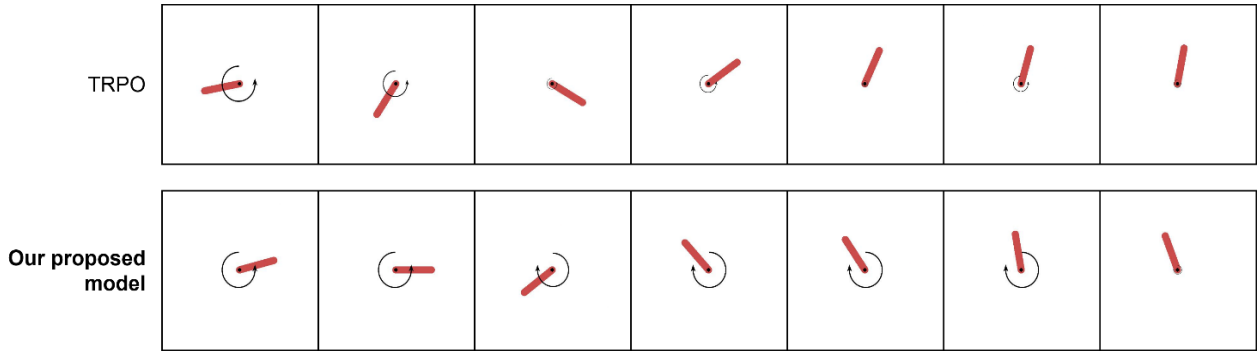


Figure 3. Execution of policies learned by the evaluated models on the Pendulum environment.

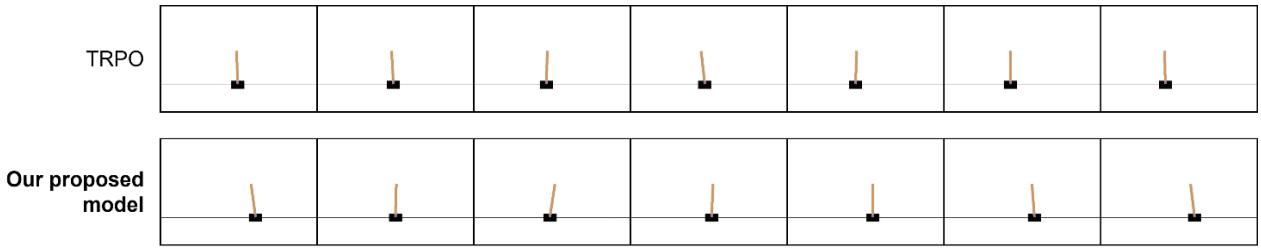


Figure 4. Execution of policies learned by the evaluated models on the CartPole environment.

Tables 1 and 2 tabulate the comparison results in terms of average cumulative reward and average training time between the proposed model and TRPO. It can be observed from Table 1 that TRPO outperforms our proposed model in terms of averaged cumulative reward on both environments. However, this result is expected since TRPO had direct access to states and the reward function of the environment in order to optimize their policies. On the other hand, the proposed model is trained using only one expert demonstration and without access to any reward signals, yet it can provide a competitive performance, especially in the CartPole environment. Moreover, according to Table 2, while TRPO takes more than 2 hours to finish training, the proposed model is about 5 times faster. Even though presenting lower averaged cumulative reward values, the proposed model is able to achieve a competing performance and requires a significantly smaller training time. These results clearly indicate the potential of our proposed model.

Table 1. The performance of the proposed model. These scores represent the cumulative rewards obtained from executing a learned policy in the simulator, averaged over 100 episodes.

Environment	TRPO	Our proposed model
Pendulum	-143.95 ± 84.36	-313.35 ± 137.58
CartPole	497.13 ± 28.56	317.97 ± 174.06

Table 2. The training time of the proposed model. These scores represent the minutes needed to train a model with 10^7 iterations, averaged over 3 times.

Environment	TRPO	Our proposed model
Pendulum	123.62 ± 2.43	23.31 ± 3.34
CartPole	148.42 ± 2.80	20.31 ± 3.27

Conclusion:

In this paper, a model was proposed that utilizes Generative Adversarial Network to imitate expert behaviors using only one demonstration. Despite such a challenging setting, the model successfully learns optimal policies on two simulated environments. In comparison with TRPO which is a Reinforcement Learning model, the proposed model provides a competitive performance with an extremely better training time. The results prove that the proposed model can be promisingly applied in imitation learning. In future work, our goal is to improve the performance of our proposed model on more complex imitation tasks.

Authors' declaration:

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are mine ours. Besides, the Figures and images, which are not mine ours,

have been given the permission for re-publication attached with the manuscript.
- The author has signed an animal welfare statement.
- Ethical Clearance: The project was approved by the local ethical committee in Shibaura Institute of Technology.

Author's contributions:

T.N.D., C.M.T., and P.X.T. conceived of the presented idea. T.N.D. and C.M.T. developed the theory and performed the computations. T.N.D., C.M.T., P.X.T., and E.K. verified the analytical methods. P.X.T. and E.K. supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.

References:

1. Hussein A, Gaber MM, Elyan E, Jayne C. Imitation learning: A survey of learning methods [Internet]. Vol. 50, ACM Computing Surveys. Association for Computing Machinery; 2017 [cited 2021 May 23]. Available from: https://dl.acm.org/doi/abs/10.1145/3054912
2. Pan Y, Cheng CA, Saigol K, Lee K, Yan X, Theodorou EA, et al. Imitation learning for agile autonomous driving. Int J Rob Res. 2020 Oct 14;39(2–3):286–302.
3. Xu Z, Sun Y, Liu M. ICurb: Imitation learning-based detection of road curbs using aerial images for autonomous driving. IEEE Robot Autom Lett. 2021 Apr 1;6(2):1097–104.
4. Kebria PM, Khosravi A, Salaken SM, Nahavandi S. Deep imitation learning for autonomous vehicles based on convolutional neural networks. IEEE/CAA J Autom Sin. 2020 Jan 1;7(1):82–95.
5. Doering M, Glas DF, Ishiguro H. Modeling interaction structure for robot imitation learning of human social behavior. IEEE Trans Human-Machine Syst. 2019 Jun 1;49(3):219–31.
6. Al-Tameemi MI. RMSRS: Rover Multi-purpose Surveillance Robotic System. Baghdad Sci J. 2020 Sep 8;17(3(Suppl.)):1049–1049.
7. Salimans T, Chen R. Learning Montezuma's Revenge from a Single Demonstration. 2018 Dec 8 [cited 2021 Jun 14]; Available from: http://arxiv.org/abs/1812.03381
8. Cai P, Sun Y, Chen Y, Liu M. Vision-Based Trajectory Planning via Imitation Learning for Autonomous Vehicles. In: 2019 IEEE Intelligent Transportation Systems Conference, ITSC 2019. Institute of Electrical and Electronics Engineers Inc.; 2019. p. 2736–42.
9. Ly AO, Akhloufi M. Learning to Drive by Imitation: An Overview of Deep Behavior Cloning Methods. IEEE Trans Intell Veh. 2021 Jun 1;6(2):195–209.
10. Fernando T, Denman S, Sridharan S, Fookes C. Deep Inverse Reinforcement Learning for Behavior Prediction in Autonomous Driving: Accurate Forecasts of Vehicle Motion. IEEE Signal Process Mag. 2021 Jan 1;38(1):87–96.
11. Wang Z, Hong T. Reinforcement learning for building controls: The opportunities and challenges. Appl Energy. 2020 Jul 1;269:115036.
12. Arulkumaran K, Deisenroth MP, Brundage M, Bharath AA. Deep reinforcement learning: A brief survey. IEEE Signal Process Mag. 2017 Nov 1;34(6):26–38.
13. Pakzad AE, Manuel RM, Uy JS, Asuncion XF, Ligayo JV, Materum L. Reinforcement Learning-Based Television White Space Database. Baghdad Sci J. 2021 Jun 20;18(2(Suppl.)):0947–0947.
14. Ho J, Ermon S. Generative Adversarial Imitation Learning. In: Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2016.
15. Zuo G, Chen K, Lu J, Huang X. Deterministic generative adversarial imitation learning. Neurocomputing. 2020 May 7;388:60–9.
16. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Networks. Commun ACM. 2020 Oct 22;63(11):139–44.
17. Brockman G, Cheung V, Pettersson L, Schneider J, Schulman J, Tang J, et al. OpenAI Gym. 2016 Jun 5 [cited 2021 Jun 14]; Available from: http://arxiv.org/abs/1606.01540
18. Barto AG, Sutton RS, Anderson CW. Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems. IEEE Trans Syst Man Cybern. 1983;SMC-13(5):834–46.
19. Liu S, Feng Y, Wu K, Cheng G, Huang J, Liu Z. Graph-Attention-Based Casual Discovery With Trust Region-Navigated Clipping Policy Optimization. IEEE Trans Cybern. 2021 Oct 20;1–14.
20. Ilboudo WEL, Kobayashi T, Sugimoto K. Robust Stochastic Gradient Descent With Student-t Distribution Based First-Order Momentum. IEEE Trans Neural Networks Learn Syst. 2020;

Open Access **Baghdad Science Journal** **P-ISSN: 2078-8665**
2021, 18(4) Supplement: 1350-1355 **E-ISSN: 2411-7986**

شبكة الخصومة التوليدية للتعلم التقليد من مظاهرة واحدة

ثو نجوين دوك * **تشان مين تران** **فان شوان تان** **إيجي كاميوكا**

كلية الهندسة والعلوم، معهد شيبور للتكنولوجيا، اليابان.

الخلاصة:
التعلم التقليد هو طريقة فعالة لتدريب وكيل مستقل لإنجاز المهمة عن طريق تقليد سلوكيات الخبراء في مظاهراتهم .ومع ذلك، تتطلب طرق التعلم التقليدية التقليدية عددا كبيرا من مظاهرات الخبراء من أجل تعلم سلوك معقد .حدد هذا العيب محدودا إمكانية التعلم التقليد في المهام المعقدة حيث لا تكون مظاهرات الخبراء كافية .من أجل معالجة المشكلة، يقترح النموذج المستند إلى الشبكة المصنوعة من الشبكة المصممة على تصميم سياسات مثالية باستخدام مظاهرة واحدة فقط .يتم تقييم النموذج المقترح على مهمتين محاكاة مقارنة بطرق أخرى .تظهر النتائج أن نموذجنا المقترح قادر على إكمال المهام المدروسة على الرغم من القيد في عدد مظاهرات الخبراء، والذي يشير بوضوح إلى إمكانات نموذجنا.

الكلمات المفتاحية :التعلم العميق، القليل من التعلم، شبكة الخصومة التوليدية، تعلم التقليد، التعلم دفعة واحدة.