Towards Building Human-like Smart Agents in Modern 3D Video Games (Student Abstract)

Zhihang Sun[1,2], Shuhan Qi[1,2,*], Xinhao Huang[1], Xinyu Xiao[1], Jiajia Zhang[1], Xuan Wang[1], Peixi Peng[3,*]

[1]Harbin Institute of Technology, Shenzhen, China
[2]Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies, Shenzhen, China
[3]Peking University, Beijing, China
22S151156@stu.hit.edu.cn, shuhanqi@cs.hitsz.edu.cn, pxpeng@pku.edu.cn

Abstract

In recent years, reinforcement learning has been widely applied in the field of games. However, most studies focus on assisting agents to achieve victory, with less attention paid to whether the agents exhibit human-like characteristics. In order to build human-like agents with high performance, we propose a method for learning the strategies of human players in modern three-dimensional video games. Our method utilizes a hierarchical framework, learning basic behaviors and intentions of human players at the lower level through imitation learning, and generalized policies at the high level through reinforcement learning. Compared with other existing methods, our method demonstrates significant advantages in learning human-like strategies in complex environments.

Introduction

Reinforcement learning has enabled game agents to exhibit exceptional decision-making abilities. However, while these agents can efficiently complete tasks, their behavior often deviates from human-like actions, such as unnecessary colliding with obstacles. Such behaviors may not hinder task completion, but may affect the gaming experience in scenarios where agents need to interact with human players. It also has difficulty extending from the gaming domain to real-world applications.

Previous research has identified two primary approaches for building human-like agents. The first is reinforcement learning with reward shaping, where agents receive penalties for non-human-like behaviors to encourage human-like policy learning(Lample and Chaplot 2017). The second is imitation learning, where agents learn directly from human game-play data (Farhang et al. 2024), resulting in human-like behaviors.

As environments become more complex, reinforcement learning faces challenges in designing and tuning reward functions, which require extensive expert knowledge to define suitable guidance for every scenario. At the same time, imitation learning often struggles to generalize beyond the provided demonstrations, limiting the agent's performance in new situations and hindering its ability to develop complex strategies.

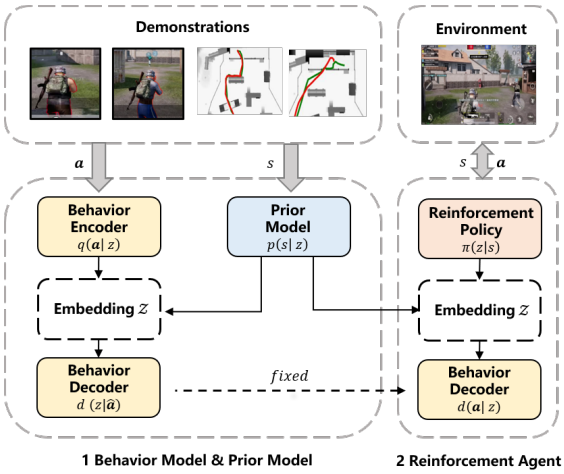

Figure 1: Our method first train a behavior model and a prior model through imitation learning, which models the fundamental behavior pattern and intention of the experts. The reinforcement learning is then applied to train the high-level policy based on the learned behavior model and prior model.

Inspired by (Pertsch, Lee, and Lim 2021), we proposes a method that combines reinforcement learning with imitation learning for developing human-like agents with high decision-making abilities in 3D video games. The method employs a hierarchical framework, where the lower level applies imitation learning to understand basic behaviors and intentions of human players, and the higher level applies reinforcement learning to learn generalized policies. Compared with other existing methods, our method demonstrates significant advantages in learning human-like strategies.

Methods

To build human-like agents with high performance, we develop a hierarchical learning framework that combines imitation learning and reinforcement learning as shown in Figure 1. The lower level learns a fundamental behav-

*Corresponding authors.

ior model to represent the expert's fundamental behaviors (fixed-length action sequences $\mathbf{a}$) as latent vectors and to reconstruct behaviors from these latent vectors. These latent vectors z, referred to as "intentions", represent the reasons behind the expert's adoption of specific action sequences. The learned behaviors often exhibit human-like characteristics, as they come from directly imitating humans. The behavior model is trained through evidence lower bound (ELBO)(Kingma, Welling et al. 2019) which minimize the reconstruction loss of actions with a regularization constraint. The behavior encoder $q(z|\mathbf{a})$ generates intention vectors from action sequences. It is parameterized as a Gaussian distribution by $\mathcal{N}(\mu_q, \sigma_q)$ and is regularized to match the standard Gaussian distribution $\mathcal{N}(0, I)$. This regularization is controlled by a coefficient β:

$$\mathcal{L} = \sum_{i=1}^{T} (\mathbf{a_i} - \hat{\mathbf{a}}_\mathbf{i})^2 + \beta D_{KL}(\mathcal{N}(\mu_q, \sigma_q) \| \mathcal{N}(0, I)). \quad (1)$$

Additionally, a state-limited prior model $p(z|s)$ is trained, which determines the expert's intention to take a certain action sequence in a certain state. The goal of the prior model is to learn a reasonable mapping from the state space $\mathcal{S}$ to the intention embedding space $\mathcal{Z}$. Specifically, the input to the expert prior model is the state in the demonstrations, and the output is the intention vector with the same dimension as the embedding space extracted by the behavior model. The prior model is jointly trained during the training of the behavior model. The learning goal is to reduce the difference between the distribution of the encoder output and the prior model output, which is achieved by minimizing the KL divergence as the following loss function:

$$L = D_{KL}(N(\lfloor \mu_q \rfloor, \lfloor \sigma_q \rfloor) \| N(\mu_p, \sigma_p)). \quad (2)$$

Then the higher level of the hierarchical framework learns policy model to generate intention vectors, which serve as the input to the behavior model for the generation of action sequences. The policy model only needs to learn how to combine and apply fundamental behaviors, which significantly reduces the the scope of strategy search and simplifies reward design. Reinforcement learning enables the high level to explore strategies beyond those of the expert to tackle complex tasks. Common model-free reinforcement learning algorithms, such as Proximal Policy Optimization (PPO)(Schulman et al. 2017), can be used to train the policy model to maximize the rewards r. During the training process, the trained prior model guides the training of the policy model. Specifically, the KL divergence between the output of the prior model and the output of the high-level strategy is minimized, controlled by a coefficient α. This constraint ensures that the high-level strategy remains close to the expert behavior distribution. The optimization function is as follows:

$$J = E_\pi[\sum_{t=0}^{T} \hat{r_t} - \alpha D_{KL}(\pi(z|s), \mathcal{N}(\lfloor \mu_p \rfloor, \lfloor \sigma_p \rfloor))]. \quad (3)$$

Experiments

We choose a popular video game "PUBG Mobile" as the experimental platform. The game features a 3D, partially observable environment with a 2,344-dimensional state space and a 76-dimensional action space. The agent's performance is evaluated based on two criteria: its decision-making ability and the human-likeness of its behavior. For decision-making, we choose metrics $win\ rate^+$, number of opponents defeated ($kill^+$), and successful hits on opponents (hit^+). We assess human-like behaviors through a subjective analysis conducted by 30 experienced game players and game developers. These participants all have extensive experience with "PUBG Mobile" and are familiar with typical human player behaviors in this game. They were invited to observe the agents' gameplay and categorize categorize behaviors into eight categories, such as moving backward without reason ($reversing^-$) and colliding with obstacles or walls ($against\ shelter^-$). The plus or minus sign after a metric indicates that the metric is a positive metric or a negative metric.

Metric	BC	IBC	PPO	Hier-RL (Ours)
$win\ rate^+$	82.4%	87.9%	95%	**100%**
$kill+$	38.1	38.2	39.1	**39.3**
hit^+	**86.3**	85.1	77.5	80.0
$reversing^-$	1210.0	910.2	353.3	**23.6**
$against\ shelter^-$	189.4	133.1	57.9	**36.2**
$stuck^-$	173.7	153.9	5.8	**3.9**
$stuck\ at\ home^-$	256.2	235.4	148.9	**7.4**
$not\ fighting\ sprint^-$	13.1	19.2	70.6	**115.9**
$fighting\ move^+$	790.0	830.1	1110.8	**1331.2**
$fighting\ prone^+$	77.2	80.3	131.4	**151.7**
$valid\ sliding^+$	0.01	0.01	31.3	**48.9**

Table 1: Performance Comparison in PUBG Mobile.

We select the following algorithms as baseline algorithms, including Behavior Cloning (BC) (Pomerleau 1988), Implicit Behavior Cloning (IBC) (Florence et al. 2022), and PPO with human-like reward shaping. We refer to our method in this paper as Hier-RL. In the adversarial scenarios, we conduct 1,000 rounds of combat tests against built-in agents, and the results are shown in Table 1. Agents trained with our Hier-RL method exhibit more human-like behavior while maintaining a high level of decision-making. Compared to other methods, the number of inappropriate behaviors, such as backward movement and hitting obstacles, is significantly reduced. This improvement is attributed to the low-level behavior model, which learns human-like behavior from demonstration data. The experimental results demonstrate that, compared to using imitation learning or reinforcement learning alone, our method produces agents with superior strategies and human-like behaviors.

Discussion and Future Work

Our work explores human-like decision-making in 3D video game scenarios by applying an efficient hierarchical learning approach that learns human-like strategies while enhancing model generalization. Our work is still in its early stages, but we hope it provides a new perspective for developing human-like agents. In future work, we will further address the challenges related to offline training data requirements in imitation learning and focus on enhancing the adaptability of the algorithm to different scenarios.

Acknowledgments

This research was funded by Guangdong Key Laboratory (2022B1212010005), NSFC(No.62372139), NSF of Guangdong (No.2024A1515030024).

References

Farhang, A. R.; Mulcahy, B.; Holden, D.; Matthews, I.; and Yue, Y. 2024. Humanlike Behavior in a Third-Person Shooter with Imitation Learning. In *2024 IEEE Conference on Games (CoG)*, 1–4.

Florence, P.; Lynch, C.; Zeng, A.; Ramirez, O. A.; Wahid, A.; Downs, L.; Wong, A.; Lee, J.; Mordatch, I.; and Tompson, J. 2022. Implicit behavioral cloning. In *Conference on Robot Learning*, 158–168. PMLR.

Kingma, D. P.; Welling, M.; et al. 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4): 307–392.

Lample, G.; and Chaplot, D. S. 2017. Playing FPS games with deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Pertsch, K.; Lee, Y.; and Lim, J. 2021. Accelerating reinforcement learning with learned skill priors. In *Conference on robot learning*, 188–204. PMLR.

Pomerleau, D. A. 1988. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.