A Framework for Modeling Cognitive Processes in Intelligent Agents Using Behavior Trees

Kejia Wan
National University of Defense Technology, changsha, China

Yuntao Liu
Academy of Military Sciences, Beijing, China

Hengzhu Liu
National University of Defense Technology, Changsha, China

Xinhai Xu
Academy of Military Science, Beijing, China

Jinlong Tian
National University of Defense Technology, Changsha, China

Xianglong Li
Academy of Military Sciences, Beijing, China

Hao Tang
NUDT, Changsha, China

Abstract

Advances in deep multi-agent reinforcement learning learning (MARL) enable sequential decision making for a range of exciting multi-agent applications. The black-box characteristic of MARL restricts the safe and scalable application of decision models in practical deployment. However, existing interpretability methods for deep reinforcement learning models are are not suitable for addressing challenges posed by multi-agent environments and often inadequate in generating logical sequential decisions. We present an innovative framework called BT4MARL, which introduces the behavior tree structure to explainable MARL. The proposed method clusters state space by aggregating temporally related states and divides agents into several groups in the new state. Based on these clustered states and agents, we constructs behavior tree structures. In this way, we use an exploration technique based on pairing a combined behavior tree with the target model. We empirically show that our framework is effective in four benchmark MARL domains. Moreover, the results of a user study show that the generated explanations significantly improve performance and satisfaction. This work represents a significant stride towards addressing the challenges of explainability and performance in MARL applications.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*