Optimizing Decision Parameters of Humanoid Robots using Deep Reinforcement Learning

Richard Pufe
Department of Electrical Engineering, Medical Engineering and Computer Science, Offenburg University
rpufe@stud.hs-offenburg.de

Abstract. This work investigates the use of deep reinforcement learning to enable humanoid Nao robots in the RoboCup 3D Soccer Simulation to autonomously decide when to switch between complex behaviors. Two main experiments were conducted. In the first, an agent was trained to learn the optimal moment to transition from walking towards the ball to executing a kick. The robot was randomly initialized at varying distances and orientations relative to the ball and trained using Proximal Policy Optimization to maximize accuracy in kicking the ball towards a target after approaching it. The resulting models achieved strong performance on par with the handcrafted baseline in simulated matches. The second experiment extended this setup by allowing the agent to also determine a favorable pre-kick position round the ball before deciding to switch. Despite the richer decision space, the resulting models performed significantly worse than the baseline, indicating the increased difficulty of jointly learning spatial positioning and timing.

Keywords: Deep Reinforcement Learning; Behavior Switching; Humanoid Robots.

1 Introduction

In simulated humanoid soccer, coordinated decision-making between locomotion and ball interaction is essential for effective play. Conventionally, the timing for transitioning from approaching to kicking the ball is determined through manually tuned thresholds or heuristic rules, which restrict adaptability to diverse situations. This work explores an end-to-end learning approach in which a robot autonomously learns the appropriate timing and later also the positioning, necessary to execute a successful kick.

Experiments were conducted in the RoboCup 3D Soccer Simulation SimSpark [1] using the Nao humanoid robot model [2]. The training task required the robot to approach a stationary ball and kick it towards a fixed target position. The robot was randomly initialized at various distances and orientations relative to the ball to encourage generalization. To stabilize training, the robot's spawning area was gradually increased. Initially, the robot was initialized within a narrow angular sector behind the ball to simplify early exploration. As training progressed, the initialization range expanded to the full 360°, requiring the policy to handle complex approach directions and orientation corrections.

Training utilized the Proximal Policy Optimization (PPO) [3] algorithm with continuous action spaces. Each control cycle, the agent received a 170-dimensional observation vector containing information from onboard sensors and internal states. This included joint angles and speeds, accelerometer and gyroscope readings, torso pose, locomotion steering outputs, foot force sensors and estimates of the robot's linear and angular velocities. Additionally, ball-related features such as its relative position and velocity were provided to inform spatial awareness. All features were normalized to a fixed range to stabilize learning and ensure comparable magnitudes across dimensions. The reward function was designed to encourage the robot to approach the ball efficiently and to switch to the kicking behavior at the appropriate distance to maximize the accuracy of the kick towards the target position.

Two learning configurations were investigated: In this first configuration, the model's output represented a binary decision indicating whether to switch from walking to kicking. As long as the network output remained below the switching threshold, the robot continued to execute a fixed walking behavior towards the ball. Once the model signaled a switch, the robot transitioned to a predefined kicking behavior. The learning process thus focused solely on identifying the optimal timing for the transition, without any influence on the locomotion or kick mechanics themselves. This setup allowed the policy to concentrate on temporal decision-making based on the evolving spatial relationship between the robot and the ball. The second configuration extended the output space of the model to include a relative position and approach angle around the ball in addition to the binary switching signal. These continuous outputs were interpreted as a desired offset position towards which the robot navigated using its walking

behavior. However, the model was free to trigger the switch to the kicking behavior at any time and it was not required to reach the proposed position before switching. Thus, as long as the switching output remained inactive, the robot continued to walk towards the indicated position and as soon as the model decided to switch, the transition to the kicking behavior occurred immediately. This setup allowed the policy to dynamically adjust both positioning intent and switching timing within the same episode, coupling spatial and temporal reasoning in a flexible but more complex learning process that proved difficult to optimize effectively.

2 Conclusion

The experiments demonstrated that deep reinforcement learning can learn to imitate handcrafted decision mechanisms for switching between walking and kicking behaviors in simulated humanoid soccer. In the first experiment, where the agent learned only the switching decision, the models achieved strong performance during training but still slightly lagged behind the handcrafted baseline in terms of overall accuracy. However, the implementation of a curriculum learning strategy, gradually increasing the range of initialization angles around the ball, proved essential to achieving near-baseline performance during full simulated matches (Table 1). This curriculum allowed the agent to generalize across a wider range of approach directions, improving robustness. One notable observation was that the learned model typically decided to switch earlier, requiring fewer control cycles before executing a kick compared to the baseline, which partially compensated for its slightly lower target accuracy.

Table 1. Simulated match outcomes. Each learned policy (Team A) competed against the handcrafted baseline (Team B) over multiple games. The baseline entry represents self-play, where both teams used the handcrafted decision mechanism. *Goals A/B* are per-match means and *Wins A/B* and *Ties* are counts over Games.

Version	Games	Goals A	Goals B	Wins A	Wins B	Ties
Baseline	400	2.788	2.858	158	159	83
Switch Model	500	2.938	3.022	182	208	110
Positioning Model	500	1.684	4.424	13	456	31

The second experiment, which extended the learning task to include both positioning and switching, introduced significantly higher complexity. None of the trained models were able to reliably learn both aspects well enough, resulting in performance that fell well below the baseline in both training evaluations and match simulations (Table 1). This highlights the increased difficulty of jointly learning spatial positioning and temporal decision-making within a single policy. Future improvements may require refined or extended curriculum strategies and more targeted fine-tuning to help the agent learn effective positional reasoning alongside switching behavior.

References

[1] J. Boedecker and M. Asada, "SimSpark - concepts and application in the RoboCup 3 d soccer simulation league," 2008. Accessed: Jul. 31, 2025. [Online]. Available: https://www.semanticscholar.org/paper/SimSpark-%E2%80%9 3-Concepts-and-Application-in-the-RoboCup-Boedecker-Asada/872b60cbbf89bda70488d3089958140 740a60a14.

[2] R. Gelin, "NAO," in *Humanoid Robotics: A Reference*, Springer, Dordrecht, 2019, pp. 147–168, ISBN: 978-94-007-6046-2. DOI: 10.1007/978-94-007-6046-2_14. Accessed: Jul. 31, 2025. [Online]. Available: https://link.springer. com/rwe/10.1007/978-94-007-6046-2_14.

[3] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, *Proximal policy optimization algorithms*, Aug. 28, 2017. DOI: 10.48550/arXiv.1707.06347. arXiv: 1707.06347[cs]. Accessed: Mar. 8, 2025. [Online]. Available: http://arxiv.org/abs/1707.06347.