Reinforcement Learning mit Stable Baselines

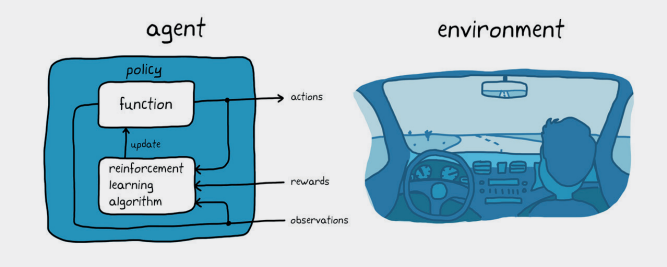

Abb. 1: Prinzip Reinforcement Learning

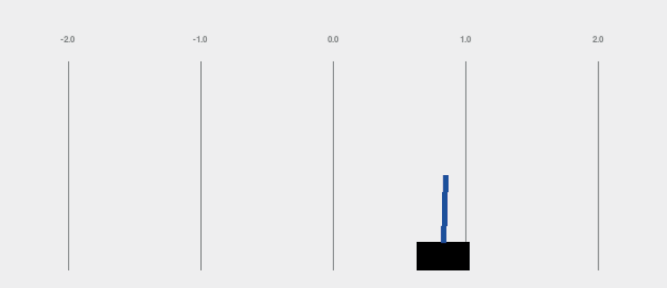

Abb. 2: Visualisierung inverses Pendel

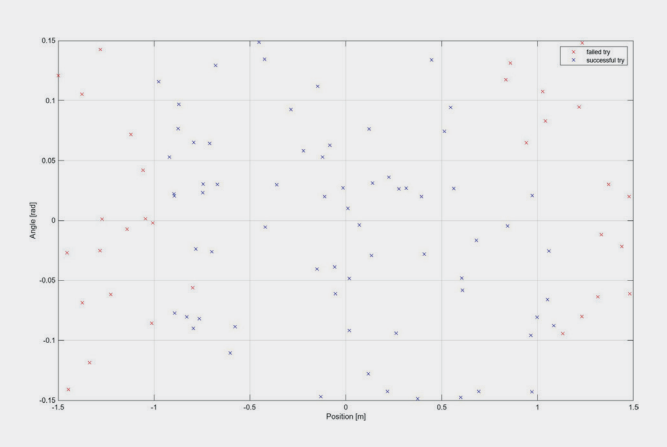

Abb. 3: Startzustände der Testdurchführungen Modell #2 mit PPO-Algorithmus

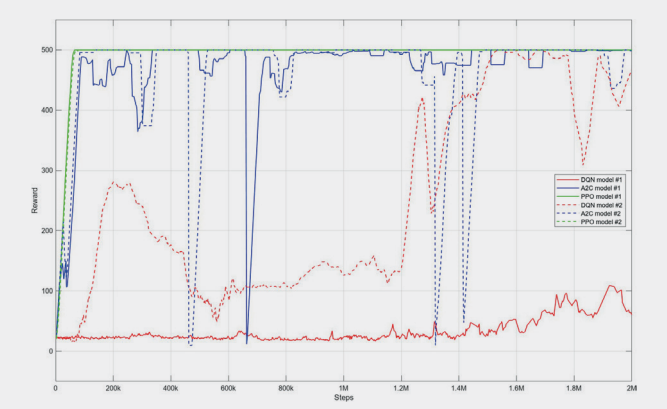

Abb. 4: Belohnung DQN-, A2C- und PPO-Algorithmus bei verschiedenen Parametern

Problemstellung

Reinforcement Learning wurde in den vergangenen Jahren immer wieder als Methode vorgeschlagen, um eine Regelung für ein System mit unbekannten Zustandsübergängen zu entwickeln. Oft ist das zugrundeliegende Modell nicht vorhanden und allenfalls nur eine Zielfunktion gegeben. Dafür soll die auf Python basierte Umgebung Stable Baselines3 angewendet und geprüft werden.

Lösungskonzept

Reinforcement Learning ist eine Methode von Machine Learning und erlernt Entscheidungsstrategien in einer Umgebung, um das Verhalten eines Agenten zu optimieren. Der Agent bestimmt aufgrund einer Beobachtung eine auszuführende Aktion und führt diese anschliessend in der Umgebung aus. Aufgrund dieser Aktion verändert sich der Zustand in der Umgebung und diese gibt dem Agent eine positive oder negative Belohnung zurück. Mit Stable Baselines3 werden virtuelle Regelungssysteme (z.B. inv. Pendel oder Mountain Car), ohne zugrundeliegendes Modell, eingelernt. Dies wird mit verschieden Algorithmen durchgeführt und soll als Ziel zu einer stabilen Regelung führen.

Realisierung

Am inv. Pendel wurden diverse Einlernprozesse, unter Verwendung verschiedener Parameter, mit den Algorithmen DQN, A2C und PPO durchgeführt. Bei DQN handelt es sich um einen Werteoptimierungsansatz und bei A2C sowie PPO um einen Policy-Optimierungsansatz. Die eingelernten Modelle wurden in der virtuellen Umgebung geprüft und die Resultate analysiert.

Zusätzlich wurde die Anwendung ‘Custom Environment’ in Stabel Baselines3 geprüft. Damit können spezifische Umgebungen selber implementiert werden. Am Beispiel vom inv. Pendel wurde eine selber implementierte Umgebung umgesetzt und diese mit der vorimplementierten verglichen.

Ergebnisse

Fürs inverse Pendel wird mit allen drei angewendeten Algorithmen eine mehrheitlich stabile Regelung erreicht. Mit dem A2C-Algorithmus sind alle durchgeführten Testdurchläufe erfolgreich, während mit dem DQN- und PPO-Algorithmus noch ca. 20 – 30% scheitern. Aufgrund der Analyse der Belohnungswerte, müssten auch für diese beiden Algorithmen mit längeren Einlernprozessen noch bessere Resultate erzielt werden.

Stable Baselines3 ist ideal für die gegebene Aufgabenstellung und bietet maximale Freiheit durch die Verwendung von ‹Custom Environment›. Die Einarbeitung erfordert zwar Aufwand, zahlt sich jedoch für umfangreiche Anwendungen definitiv aus.

Diplomand
Arnold Florian

Dozent
Prof. Dr. T. Hunziker

Themengebiet
Nachrichtentechnik/Signal Processing, Mechatronik/Automation/Robotik

Projektpartner
Intern