Mein Titel

Untertitel

Bachelorarbeit

vorgelegt am 22. Februar 2023

Fakultät Wirtschaft

Studiengang Wirtschaftsinformatik

Kurs WWI2020F

von

Leon Henne

Betreuerin in der Ausbildungsstätte:	DHBW Stuttgart:
IBM Deutschland GmbH Sophie Lang Senior Data Scientist	Prof. Dr. Kai Holzweißig Studiendekan Wirtschaftsinformatik

Unterschrift der Betreuerin

Inhaltsverzeichnis

Abkürzungsverzeichnis

DHBW Duale Hochschule Baden-Württemberg

RL Reinforcement Learning

KPI Key Performance Indicator

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Problemstellung

Reinforcement Learning (RL) findet heutzutage bereits Anwendung in vielerlei Forschungsprojekten wie Deepmind AlphaStar oder OpenAI Five, aber auch in Produkten und Dienstleistungen wie AWSDeepRacer oder Metas Horizon open-source RL-Plattform.[1] RL ist im Bereich des maschinellen Lernens eine Herangehensweise zur Lösung von Entscheidungsproblemen.[2] Ein Software-Agent leitet dabei durchzuführende Aktionen aus seiner Umgebung ab, mit dem Ziel die kumulierte erhaltene Belohnung zu maximieren, währenddessen sich seine Umgebung durch alle Aktionen verändert.[3] Die Umgebungen beinhalten in ihrer einfachsten Form eine simulierte Welt, welche zu jedem Zeitschritt eine Aktion entgegennimmt, und den eigenen nächsten Zustand sowie einen Belohnungswert zurückgibt.[4] Da ein Problem beim Einsatz von RL Algorithmen Limitierungen sein können, Daten in der echten Welt zu sammeln und fürs Training zu verwenden, werden häufig hierfür Simulationsumgebungen eingesetzt.[5] Eine Limitierung können bspw. Sicherheitsaspekte sein, welche beim Training von Roboterarmen, oder sich autonom bewegenden Systemen auftreten, da die einzelnen physischen Bewegungen nicht vorhersehbar abschätzbar sind.[6] Simulationen nehmen damit als Testumgebung eine wichtige Rolle ein in der Entwicklung von Kontrollalgorithmen.[7] Insgesamt bedarf die erfolgreiche Anwendung von Reinforcement Learning demnach nicht nur effiziente Algorithmen, sondern auch geeignete Simulationsumgebungen.[8] Besonders schwierig, und daher sehr wichtig zu erforschen, ist es die Trainingsumgebung bestmöglich an die echte Welt anzupassen, sodass die Agenten für Roboter und autonome Fahrzeuge nach dem Training, mit generalisierten Policies in der Realität eingesetzt werden können.[9] In der Forschungsliteratur wird diese beschriebene Problematik als „Sim to real"-Transfer beschrieben.[10] Eine Methodik diesen Transfer zu begünstigen ist die Veränderung von visuellen oder dynamischen Parametern der Umgebung, was in der Forschungsliteratur als „Domain Randomization" referenziert wird.[11] Eine Domäne der echten Welt wird dabei eher selten ausschließlich von veränderten dynamischen Parametern und nur einer Person oder nur einer Organisation geprägt. Oftmals beeinflussen mehrere Parteien teilweise kooperierend aber auch teilweise konkurrierend den eigenen Erfolg, wie bspw. einen dem Wettbewerb unterliegenden Markt. Stellt man sich ein solches Szenario vor, ist es naheliegend, dass auch jene Einflüsse möglichst präzise in die Simulationsumgebung integriert sein müssen, um ein generalisierendes Modell erlernen zu können.

[1] Vgl. Li 2019, S. 4
[2] Vgl. Schuderer/Bromuri/van Eekelen 2021, S. 3
[3] Vgl. Schuderer/Bromuri/van Eekelen 2021, S. 3
[4] Vgl. Reda/Tao/van de Panne 2020, S. 1
[5] Vgl. Zhao/Queralta/Westerlund 2020, S. 737
[6] Vgl. Zhao/Queralta/Westerlund 2020, S. 738
[7] Vgl. Cutler/Walsh/How 2014, S. 2
[8] Vgl. Reda/Tao/van de Panne 2020, S. 8
[9] Vgl. Slaoui u. a. 2019, S. 1
[10] Vgl. Zhao/Queralta/Westerlund 2020, S. 738
[11] Vgl. Slaoui u. a. 2019, S. 2

Während bereits in Produkten wie Powertac nach Collins/Ketter 2022 die Simulation von Märkten entwickelt wurde, scheint der Einfluss von kompetitiven Simulationen auf die Robustheit von RL Algorithmen unerforscht.

1.2 Zielsetzung

Daher soll im Rahmen dieser Arbeit untersucht werden, ob die Integrierung eines RL basierten Gegenspielers in einer Simulation die Umgebung so beeinflussen kann, dass die erlernten Verhaltensmodelle, welche im Kontext von RL oftmals als Policies referenziert werden, robuster agieren unter den veränderten dynamischen Bedingungen und alternativen deterministischen Gegenspielern im Testszenario.

Dazu soll eine kompetitive Simulationsumgebung entwickelt werden, in welcher sich zwei konkurrierender Spieler in Form von Flugobjekten spielerisch gegenseitig bekämpfen. In der Simulation werden folgend Policies in drei verschiedenen Szenarien trainiert.

- Training mit regelbasiertem Gegenspieler unter gleichbleibenden Dynamikparametern
- Training mit RL basiertem Gegenspieler unter gleichbleibenden Dynamikparametern
- Training mit regelbasiertem Gegenspieler unter sich verändernden Dynamikparametern

Anschließend werden alle trainierten Policies in einer Reihe von Testszenarien untersucht. Jedes Testszenario verfügt dabei über festgelegte sich vom Training unterscheidende Dynamikparameter und jeweils leicht unterschiedliche Handlungspräferenzen des deterministischen Gegenspielers. Bei der Untersuchung werden jeweils die folgenden Variablen als Key Performance Indicator (KPI) betrachtet.

- durchschnittlich erzielte Belohnung
- Varianz der Belohnungen
- Anzahl an Abstürzen

Aus der Auswertung der Testszenarien kann der Effekt des RL basierten Gegenspielers auf die Robustheit mittels des Vergleichs mit dem regelbasierten Gegenspieler und der Domain Randomization evaluiert werden.

1.3 Forschungsfrage

Aus der beschriebenen Problemstellung und der für den Rahmen dieser Arbeit festgelegten Zielsetzung ergibt sich folgende Forschungsfrage:

Kann durch den Einsatz eines mittels RL trainierten Gegenspielers die Robustheit der gelernten Policy verbessert werden?

Zur Beantwortung der Forschungsfrage werden folgende Hypothesen aufgestellt und im Rahmen der Arbeit untersucht:

Hypothese 1: *Die durchschnittlich erzielte Belohnung ist under Verwendung der Policy aus dem Training mit RL basiertem Gegenspieler signifikant und zuverlässig höher als die Policy aus dem Training mit regelbasiertem Gegenspieler.*

Hypothese 2: *Die Varianz der Belohnung ist under Verwendung der Policy aus dem Training mit RL basiertem Gegenspieler signifikant und zuverlässig geringer als die Policy aus dem Training mit regelbasiertem Gegenspieler.*

Hypothese 3: *Die Anzahl von Abstürzen ist under Verwendung der Policy aus dem Training mit RL basiertem Gegenspieler signifikant und zuverlässig geringer als die Policy aus dem Training mit regelbasiertem Gegenspieler.*

1.4 Forschungsmethodik

Als Forschungsmethodik soll im Rahmen dieser Arbeit ein quantitatives Laborexperiment nach Recker 2021 durchgeführt werden. Hierbei wird häufig nach dem hypothetisch-deduktives Modell vorgegangen, in welchem Hypothesen formuliert, empirische Studien entwickelt, Daten gesammelt, Hypothesen anhand dessen evaluiert und gewonnene Erkenntnisse berichtet werden.[12] Eine Möglichkeit der Untersuchung der Ursache- und Wirkungsbeziehung stellt das Laborexperiment dar.[13] Dabei wird die kontrollierte Umgebung der Simulation erschaffen, in welcher dessen Aufbau die unabhängige Variable darstellt. Die Metriken anhand die Performance und die Robustheit der trainierten Policies gemessen werden, bilden im Experiment die abhängigen Variablen.

Neben dem quantitativen Laborexperiment wird die Literaturrecherche nach Webster, J./Watson 2002 durchgeführt.

[12] Vgl. Recker 2021, S. S.89f.
[13] Vgl. Recker 2021, S. 106

1.5 Aufbau der Arbeit

Insgesamt gliedert sich die Arbeit nach einem Schema von Holzweißig 2022 Die Arbeit beginnt mit einem einleitenden Kapitel in welchem Motivation, Problemstellung, Zielsetzung und Forschungsmethodik erläutert sind. Anschließend wird im zweiten Kapitel der aktuelle Stand der Forschung zu den relevanten Konzepten der Problemstellung wiedergegeben. Im dritten Kapitel wird die Forschungsmethodik durchgeführt, indem die Simulationsumgebung als Messinstrument entwickelt wird sowie verschiedene Messszenarien erläutert und entsprechende Daten gesammelt werden. Daraufhin sind im folgenden vierten Kapitel die gesammelten Messdaten auszuwerten und aufgestellte Hypothesen zu überprüfen. Im Zuge dessen kann ebenso die Forschungsfrage anhand der Annahme oder Ablehnung der Hypothesen beantwortet werden. Abschließend wird im letzten Kapitel ein Fazit zu den erzielten Forschungsergebnissen dargelegt und ein Ausblick auf weitere Forschung gegeben.

2 Diskussion des aktuellen Stands der Forschung und Praxis

2.1 Aufbau der Literaturrecherche

Artikel	Konzepte			

3 Durchführung des Laborexperiments

4 Ergebnisse des Laborexperiments

5 Reflexion und Forschungsausblick

Anhang

Anhangverzeichnis

Anhang 1: Interview Transkripte

Anhang 1/1: Interview Transkript: Mitarbeiter eines Unternehmens

Literaturverzeichnis

Collins, J./Ketter, W. (2022): Power TAC: Software architecture for a competitive simulation of sustainable smart energy markets. In: *SoftwareX* 20, S. 101217. ISSN: 2352-7110. DOI: https://doi.org/10.1016/j.softx.2022.101217. URL: https://www.sciencedirect.com/science/article/pii/S2352711022001352.

Cutler, M./Walsh, T. J./How, J. P. (2014): Reinforcement learning with multi-fidelity simulators. In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. DOI: 10.1109/icra.2014.6907423.

Holzweißig, K. (2022): Wissenschaftliches Arbeiten. In: 6.6.

Li, Y. (2019): Reinforcement Learning Applications. DOI: 10.48550/ARXIV.1908.06973. URL: https://arxiv.org/abs/1908.06973.

Recker, J. (2021): Scientific research in information systems: A beginner's guide. Second Edition. Progress in IS. Cham: Springer International Publishing. ISBN: 9783030854362. URL: https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=6789173.

Reda, D./Tao, T./van de Panne, M. (2020): Learning to Locomote: Understanding How Environment Design Matters for Deep Reinforcement Learning. In: *Motion, Interaction and Games*. Hrsg. von Daniele Reda/Tianxin Tao/Michiel van de Panne. New York, NY, USA: ACM, S. 1–10. DOI: 10.1145/3424636.3426907.

Schuderer, A./Bromuri, S./van Eekelen, M. (2021): Sim-Env: Decoupling OpenAI Gym Environments from Simulation Models. In: *International Conference on Practical Applications of Agents and Multi-Agent Systems*. Springer, Cham, S. 390–393. DOI: 10.1007/978-3-030-85739-4{\textunderscore}39. URL: https://link.springer.com/chapter/10.1007/978-3-030-85739-4_39.

Slaoui, R. B./Clements, W. R./Foerster, J. N./Toth, S. (2019): Robust Domain Randomization for Reinforcement Learning. In: *CoRR* abs/1910.10537. arXiv: 1910.10537. URL: http://arxiv.org/abs/1910.10537.

Webster, J./Watson, R. T. (2002): Analyzing the Past to Prepare for the Future: Writing a Literature Review. In: *MIS Q.* 26.2, S. xiii–xxiii. ISSN: 0276-7783.

Zhao, W./Queralta, J. P./Westerlund, T. (2020): Sim-to-Real Transfer in Deep Reinforcement Learning for Robotics: a Survey. In: *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. DOI: 10.1109/ssci47803.2020.9308468.

Erklärung

Ich versichere hiermit, dass ich meine Bachelorarbeit mit dem Thema: *Mein Titel* selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Ich versichere zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

(Ort, Datum) (Unterschrift)