

Hatékony Megerősítéssel Tanulás Intelligens Közlekedési Rendszerekhez

Tézisfüzet:
Bálint Kővári



M Ű E G Y E T E M 1 7 8 2

Budapesti Műszaki és Gazdaságtudományi Egyetem
Közlekedésmérnöki és Járműmérnöki Kar
Közlekedés- és Járműirányítási Tanszék

Konzulens:
Tamás Bécsi, PhD
Közlekedés- és Járműirányítási Tanszék
Budapesti Műszaki és Gazdaságtudományi Egyetem

June 10, 2025

Fejezet 1

Bevezetés

A doktori disszertációmban a megerősítéses tanulás hatékonyságának növelésére fókuszálok az intelligens közlekedési rendszerek területén. A problémát hierarchikus módon közelítem meg, több szinten kezelve a hatékonyság javításának módszereit.

A megerősítéses tanulás jelentős figyelmet kapott figyelemre méltó eredményei miatt (Silver et al. [2017], Fawzi et al. [2022]), különösen az intelligens közlekedési rendszerekhez kapcsolódó alkalmazások terén Zhu et al. [2023]. Az egyik legnagyobb kihívást ugyanakkor a tanítás hatékonysága jelenti, mivel a megerősítéses tanulás gyakran sok iterációt igényel a kielégítő teljesítmény eléréséhez (Henderson et al. [2018]). Ez különösen problémás az Intelligens Közlekedési Rendszerek esetében, ahol egyszerre több cél – például a fenntarthatóság és közlekedési hatékonyság – összehangolására van szükség. Elsődleges célom a megerősítéses tanulás hatékonyabbá tétele olyan módszerek kidolgozásával, amelyek csökkentik a szükséges tanítási iterációk számát, miközben megtartják vagy akár javítják is a teljesítményt.

Az megerősítéses tanulás hatékonyságának szisztematikus javítása érdekében kutatásomat több szintre tagoltam:

- **Probléma felírás:** Először azt vizsgálom, hogy a szabályozási probléma megerősítéses tanulási felírása milyen mértékben befolyásolja az ágens tanulási hatékonyságát. A legnagyobb kihívást itt az jelenti, hogy olyan állapot-, akció- és jutalom-absztrakciókat definiáljak, amelyek pontosan lefedik a problémát, és elősegítik a hatékony tanulást.
- **Módszer fejlesztés az iteratív absztrakció keresés gyorsítására:** Kifejlesztettem olyan módszereket, amelyek támogatják a megerősítéses tanulási problémák megfogalmazását, különös tekintettel a jutalomfüggvények összehasonlítására. Mivel a jutalom irányítja a tanulást, a legalkalmasabb jutalomfüggvény kiválasztása anélkül, hogy ismétlődő próbálgatási ciklusokra lenne szükség, jelentősen növeli a hatékonyságot.
- **Skálázhatóság több-ágenses megerősítéses tanulással:** A változtatható sebességhatár-szabályozás területén vizsgálom a skálázhatóságot, több-ágensű megközelítést al-

kalmazva a különböző problémaméretet hatékony kezelésére. Ez olyan ágensek tanítását jelenti, amelyek képesek általánosítani tanulásukat eltérő problémaméretekre.

- Mintahatékony tanítóminta priorizáláson keresztül: Egy új, tanítóminta priorizálási technikát vezetnek be a megerősítéses tanulás mintahatékonyának növelésére. Ez a módszer problémától független, és a konvergencia eléréséhez szükséges tanítási iterációk számának csökkentésére koncentrál, így a megerősítéses tanulás gyakorlati alkalmazását is elősegíti valós rendszerekben.

Ahogy fentebb említettem, a disszertációm első kutatási iránya a probléma felírás hatásának vizsgálata az elérhető teljesítményre. Ehhez a közlekedési jelzőlámpa-vezérlés problémáját választottam, mivel ennek széleskörű szakirodalma van, és a megerősítéses tanuláshoz kapcsolódó tanulmányok elemzése után világossá vált, hogy a legtöbb publikáció új problémaformulációk létrehozására törekszik, míg csak kis részük fókuszál mélyebb módszertani kérdésekre. Ennek megfelelően elegendő eredmény állt rendelkezésre a probléma felírás adaptív jelzőlámpa irányításra gyakorolt hatásának vizsgálatához. A [Haydari et al. \[2021\]](#) tanulmány részletesen bemutatja az adaptív jelzőlámpa irányítás gyakori megközelítéseit.

Az első fejezetben új típusú megközelítést dolgoztam ki a közlekedési lámpák vezérlésére modellfüggetlen megerősítéses tanulás alkalmazásával. A fő újítás egy új jutalmazási koncepció bevezetése, amely egyetlen csomópont sávjai közötti várakozási sorhosszok szórásán alapul. Ahelyett, hogy közvetlenül a hagyományos mutatók – mint az átlagos várakozási idő vagy az utazási idő – minimalizálására törekedtem volna, a sorhosszok közötti szórás csökkentésére koncentráltam. Ez a nézőpontváltás jelentősen javítja az ügynök képességét a forgalom hatékony kiegyensúlyozására, amellyel a hagyományos módszerek gyakran nem boldogulnak.

A jelenlegi közlekedési lámpavezérlési megoldások, mint például a fix ciklusú vezérlők vagy az olyan szabályozott rendszerek, mint a SUMO beépített algoritmusai, vagy nem elég alkalmazkodóképesek, vagy nem veszik figyelembe a sávok közötti egyenlőtlen forgalomeloszlást. A korábbi megerősítéses tanulás-alapú megközelítések általában olyan mutatók optimalizálására törekedtek külön-külön, mint a sorhossz vagy az átlagsebesség, ami oda vezethet, hogy egyes sávok kiürülnek, míg mások torlódnak. Az általam javasolt módszer közvetlenül a forgalmi egyensúlyhiány kezelésére irányul, így lényegénél fogva alkalmazkodóbb.

A módszer értékeléséhez megerősítéses tanulási ágenszt tanítottam be az új jutalmazási koncepció alapján. Az ügynökök teljesítményét összehasonlítottam fix ciklusú vezérlőkkel, a SUMO szabályozott vezérlőjével, valamint más, hagyományos jutalomfüggvényeket alkalmazó megerősítéses tanulási-módszerekkel. Az eredmények azt mutatták, hogy az általam javasolt megközelítés jelentősen felülmúlta a referencia módszereket mind a klasszikus mutatók (pl. várakozási idő, utazási idő, sorhossz), mind a fenntarthatósági mutatók (pl. CO₂-kibocsátás, üzemanyag-fogyasztás) tekintetében.

Ez a javulás annak köszönhető, hogy az állapotleírás, az akciótér és a jutalomfüggvény szoros integrációja lehetővé teszi az ügynök számára, hogy kontextusérzékenyebb döntéseket hozzon. A módszer emellett jobb általánosítási képességet és stabilitást mutatott

különböző forgalmi szituációkban, ami alátámasztja gyakorlati alkalmazhatóságát.

Azáltal, hogy a forgalmi terhelések kiegyensúlyozására összpontosít, ahelyett, hogy elszigetelt mutatók pusztán minimalizálására törekedne, megközelítem nemcsak a közlekedési hatékonyságot javítja, hanem csökkenti a környezeti terhelést is. Ezáltal ígéretes, korszerű megoldást kínál a városi közlekedésirányításban, működési és fenntarthatósági előnyöket biztosítva.

A fent bemutatott fejezet kimerítő tanítási folyamata után figyelmemet arra a problémára fordítottam, hogy miként lehet enyhíteni azt az iteratív tanítási folyamatot, amely minden egyes probléma felírás értékeléséhez és összehasonlításához szükséges.

A legmegfelelőbb jutalomfüggvény azonosítása ugyanis megköveteli, hogy a megerősítéssel tanuló ágenszt több lehetséges függvénnyel is betanítsuk, ami időigényes és erőforrásigényes folyamat. Ennek a problémának a megoldására egy új módszert vezettem be, amely a Monte-Carlo Fakereső (MCTS) algoritmust alkalmazza a jutalomfüggvények előzetes értékelésére, még a tényleges tanítás előtt (Kocsis and Szepesvári [2006]). Ez a megközelítés feleslegessé teszi a teljes körű betanítást, így jelentős számítási erőforrás takarítható meg.

Módszerem lényege, hogy a Monte-Carlo Fakeresés segítségével előre megállapíthatóvá válik az egyes jutalomfüggvények teljesítmény szerinti rangsora egy adott probléma esetén. A jutalomfüggvények közvetlen beépítésével a Monte-Carlo Fakeresés struktúrájába az algoritmus képes szimulálni a lehetséges kimeneteket anélkül, hogy szükség lenne a teljes tanítási folyamatra. Így a Monte-Carlo Fakeresés nemcsak értékeli az egyes jutalomfüggvények várható teljesítményét, hanem azonosítja azokat a viselkedési mintázatokat is, amelyek a tanulás során kialakulhatnak. A megközelítés legnagyobb előnye, hogy lehetővé teszi a legígéretesebb jutalomfüggvény előzetes kiválasztását, ezáltal jelentősen csökkentve a megerősítéssel tanulásban szokásos próbálgatásos folyamatot.

A módszer hatékonyságát két különböző szabályozási feladaton keresztül demonstráltam: közlekedési lámpák vezérlésében és sávkövetési feladatban. A közlekedési lámpák irányítása esetén különböző jutalmazási stratégiákat hasonlítottam össze, például a sorhossz minimalizálását és az átlagsebesség maximalizálását, mind MCTS-sel, mind hagyományos megerősítéssel tanuló módszerekkel, mint például a Deep Q-Network (DQN). Az eredmények azt mutatták, hogy a Monte-Carlo Fakeresés pontosan megtudta állapítani a legjobban teljesítő jutalomfüggvényt, azonos eredményre jutva, mint a teljes tanítással elért kimenetek. Hasonlóképpen, a sávkövetési feladatban – ahol a cél a jármű sávon belül tartása – a Monte-Carlo Fakeresés anélkül tudta helyesen rangsorolni a jutalomfüggvényeket, hogy szükség lett volna az ágensek teljes tanítására.

A legfontosabb megállapítás az volt, hogy a Monte-Carlo Fakeresés által megállapított jutalomfüggvény-rangsor szorosan egybevágtott a tényleges tanítással kapott rangsorral, ami igazolja, hogy a Monte-Carlo Fakeresés megbízható előzetes értékelő eszközként működhet. Például a jelzőlámpa irányítás esetén az a jutalomfüggvény, amely minimalizálta a sávok közötti sorhossz szórását, következetesen felülmúlta a többi függvényt a CO₂-kibocsátás és az üzemanyag-fogyasztás tekintetében is, ahogyan azt a Monte-Carlo Fakeresés előre jelezte. Hasonlóképpen, a sávkövetési problémában az a jutalomfüggvény, amely a stabilitást és a sávközéphez való minimális eltérést hangsúlyozta,

helyesen lett azonosítva a leghatékonyabbként.

Módszerem egy kritikus szűk keresztmetszetet kezel a megerősítéses tanulás gyakorlati alkalmazásában, mivel csökkenti a szükséges tanítási futások számát. Ahelyett, hogy időigényes tanításokkal próbálnánk ki több jutalomfüggvényt, a Monte-Carlo Fakeresés egy hatékony előzetes szűrőfolyamatot biztosít, amely irányt mutat a kutatóknak a legígéretesebb konfigurációk felé. Ez az előrelépés felgyorsítja a megerősítéses tanulás-alapú megoldások fejlesztését, és csökkenti a számítási költségeket, ezáltal a megerősítéses tanulás alkalmazását hatékonyabbá teszi összetett problémák esetén.

Kutatásom azt mutatja, hogy a Monte-Carlo Fakeresés jelentősen növelheti a megerősítéses tanulás hatékonyságát azáltal, hogy előzetesen értékeli a jutalomfüggvényeket. Ez a módszer áthidalja az elméleti jutalomformuláció és a gyakorlati alkalmazás közötti szakadékot, lehetővé téve a megerősítéses tanulás gyorsabb és megbízhatóbb bevetését különböző szabályozási feladatokban, beleértve a forgalomirányítást és az önvezető járművek vezérlését is. Ez a megközelítés megnyithatja az utat a hatékonyabb erőforrásfelhasználású megerősítéses tanulási alkalmazások előtt az intelligens közlekedési rendszerekben.

Miután több szempontból elemeztem a Monte-Carlo Fakereső algoritmust, úgy döntöttem, hogy tovább folytatom a problémaformulálás vizsgálatát, ezúttal azonban a skálázhatóság oldaláról közelítve meg a kérdést. A skálázhatóság kulcsfontosságú kérdés a megerősítéses tanulás területén, különösen többügynökös rendszerekben, mivel ezeknél a problémák bonyolultsága jóval nagyobb. Az intelligens közlekedési rendszerekben számos több-ágensű probléma található, ezért célkitűzésemet úgy fogalmaztam meg, hogy olyan ágenszt szeretnék létrehozni, amely különböző problémaméreteket is képes kezelni, miközben magas szintű teljesítményt nyújt. Az alkalmazott probléma a változtatható sebességhatár-szabályozás (VSLC) volt, mivel ez nagy potenciállal rendelkezik, és a megerősítéses tanulás viszonylag új megközelítés ebben a témában.

A VSLC problémára egy új megközelítést dolgoztam ki több-ágensű megerősítéses tanulás alkalmazásával. A legfőbb hozzáadott érték a csúszóablak-alapú állapotleírási módszer, amely skálázhatóvá és a szabályozott autópályaszakasz hosszától függetlenné teszi a módszert. Ez a megfogalmazás lehetővé teszi, hogy a több-ágensű megerősítéses tanulóval tanított ágens hatékonyan működjön bármilyen hosszúságú autópályán, megoldva ezzel a korábbi módszerek skálázhatósági problémáját (Zheng et al. [2023], Kušić et al. [2020]).

A megoldandó probléma alapvetően az, hogy a jelenlegi VSLC megközelítések hatékonysága különböző hosszúságú autópályaszakaszokon nem kielégítő. A hagyományos módszerek vagy fix diszkrét, vagy folytonos sebességhatárokat használnak, és jellemzően minden egyes autópályahosszhoz külön tanítást igényelnek, ami a gyakorlatban megnehezíti és számításigényessé teszi az alkalmazásukat (Zheng et al. [2023], Zhang et al. [2023]). Az én módszerem azonban csak egyszer igényel tanítást, és később is hatékony marad, akkor is, ha hosszabb útszakaszokra alkalmazzák.

A legfontosabb újítás az állapotleírásban rejlik, amely egy csúszóablakot alkalmaz a szomszédos autópályaszakaszok felett, lehetővé téve, hogy minden ágens csak a helyi forgalmi viszonyokat érzékelje, anélkül hogy az egész autópályáról információra lenne szük-

sége. Ez a módszer biztosítja az egyenletes teljesítményt a szakasz hosszától függetlenül, és csökkenti a számítási terhelést, mivel nem szükséges újratanítani az ágenszt minden hosszhoz. Emellett az akciótér is úgy lett kialakítva, hogy a sebességhatárok fokozatos módosítását tegye lehetővé, ami elősegíti a forgalom egyenletes áramlását és csökkenti a lökéshullámokat.

A módszer hatékonyságát úgy demonstráltam, hogy az ágenszt egy 1 km-es autópályaszakaszon tanítottam, majd kiértékeltem 1 km-es, 3 km-es és 10 km-es szakaszokon. Az eredmények következetesen azt mutatták, hogy a módszerem jelentősen felülmúlja az alapg megoldásokat, beleértve a Motorway Control System (MCS) szabályozást és a szabályozás nélküli scenáriókat. Az ágens még hosszabb pályák esetén is stabil teljesítményt nyújtott, bizonyítva ezzel a módszer skálázhatóságát.

Összefoglalva, a módszerem egy skálázható, hatékony megoldást kínál a VSLC problémára, a több-ágensű megerősítéses tanulás és a csúszóablakos állapotleírás kombinálásával. Ez javítja a forgalom áramlását és csökkenti a környezeti terhelést, így ígéretes megközelítés lehet a valós közlekedésirányítási rendszerek számára.

Az eddig bemutatott eredmények és módszerek részben alkalmazáspecifikusak, vagy nem közvetlenül a megerősítéses tanulás alapvető hatékonyságát javítják. Ezért az utolsó tézisem céljaként egy olyan módszertan kidolgozását tűztem ki, amely a megerősítéses tanulás egyik alapvető gyengeségét, a mintahatékonyság hiányát kezeli. A mintahatékonyság többféleképpen javítható, én azonban a tanítási minták prioritizálását választottam, mivel ez egy könnyűsúlyú, mégis nagy hatással bíró módszer, amely jelentősen befolyásolhatja az ágens végső teljesítményét és konvergenciaidejét.

Egy új tanítóminta prioritizálási módszert dolgoztam ki a megerősítéses tanulás számára, amely jelentősen javítja a tanítás hatékonyságát. A fő hozzáadott érték egy innovatív megközelítés, amely az Upper Confidence Bound (UCB) elvét ötvözi meglévő prioritási technikákkal, kifejezetten az exploráció és az exploítáció közötti egyensúly javítása érdekében. Ez a módszer optimalizálja a mintavételi hatékonyságot, és gyorsabb konvergenciát tesz lehetővé a korszerű Prioritized Experience Replay (PER) módszerhez képest (Schulman et al. [2015]).

A megoldott probléma a megerősítéses tanulási ágensek tanítása során alkalmazott mintavételi stratégiák hatékonysághiánya volt. A hagyományos módszerek, különösen a PER, gyakran túlzottan előnyben részesítik a magas hibájú tapasztalatokat, ami túlzott exploítációhoz és az exploráció elhanyagolásához vezet. Ez az egyensúlyhiány lassabb konvergenciát és megnövekedett számítási költséget eredményez. Az általam javasolt módszer az UCB alkalmazásával dinamikusan egyensúlyozza az explorációt és exploítációt, figyelembe véve az egyes tapasztalatok felhasználási gyakoriságát és az időbeli különbség hibát. Ez a kettős fókusz lehetővé teszi az ügynök számára, hogy az informatív és alulfeltárt tapasztalatokat részesítse előnyben, fenntartva ezzel a tanulás hatékonyabb menetét.

Módszerem hatékonyságát négy problémán teszteltem az OpenAI Gym csomagból: CartPole, Acrobot, Taxi és CliffWalking. A megerősítéses tanulási ágenseket az általam javasolt UCB-alapú módszerrel és a PER-rel is betanítottam azonos feltételek mellett. Az eredmények azt mutatták, hogy módszerem következetesen gyorsabb konvergenciát

és magasabb kumulatív jutalmat ért el, mint a PER.

A PER, mint összehasonlítási alap, a megerősítéses tanulás területén a legelterjedtebb tanítóminta prioritizálási módszer, amely a tanítómintákat az időbeli különbség hibájuk alapján rangsorolja. Ez gyakran azt eredményezi, hogy az algoritmus egy szűk, túlpriorizált mintahalmazon tanul újra és újra. Ezzel szemben az általam javasolt UCB-alapú megközelítés kiegyensúlyozottabb mintavételi stratégiát kínál, mivel a prioritási értéket az explorációs ösztönzőkkel kombinálja, ezáltal változatosabb és informatívabb tanulási folyamatot biztosít.

Összefoglalva, a javasolt tanítóminta prioritizálási módszer a megerősítéses tanulás mintavételi stratégiáinak hatékonysághiányát kezeli, kiegyensúlyozottabb exploráció–exploitáció egyensúlyt kínálva. A konvergencia sebességének és a tanítás stabilitásának javulása azt mutatja, hogy a módszerem jelentősen csökkenti a számítási igényeket, és ezzel a megerősítéses tanulás gyakorlati alkalmazását is elősegíti összetett, valós problémák esetén.

Fejezet 2

Új eredmények

2.1 1. Tézis

Új probléma felírást dolgoztam ki modellfüggetlen megerősítéses tanulási ágensek számára az egykereszteződéses közlekedési jelzőlámpa irányításban. Az új probléma felírás olyan jutalmazási elvet vezet be, amely a sávokhoz tartozó sorhosszok szórásán alapul, továbbá az állapotleírás is a sorhosszon alapul. A megerősítéses tanulási ágens, amelyet az új probléma felírással tanítottam, jobb teljesítményt ért el az alábbi mutatók tekintetében, mint a szakirodalomban található egyéb megközelítések: várakozási idő, utazási idő, sorhossz, üzemanyag-fogyasztás, CO₂-kibocsátás és NO_x-kibocsátás.

Kapcsolódó publikációk:

- (KPAB22) Kővári B, Pelenczei B, Aradi S and Bécsi T (2022), "Reward Design for Intelligent Intersection Control to Reduce Emission", IEEE Access. Vol. 10, pp. 39691 - 39699.
- (KTB21) Kővári B, Tettamanti T and Bécsi T (2021), "Deep Reinforcement Learning based approach for Traffic Signal Control", In Proceedings of The 24th Euro Working Group on Transportation Meeting (EWGT2021).
- (KSzBAG21) Kővári B, Szóke L, Bécsi T, Aradi S and Gáspár P (2021), "Traffic Signal Control via Reinforcement Learning for Reducing Global Vehicle Emission", Sustainability, MDPI., October, 2021. Vol. 13(11254), pp. 18.

2.2 2. Tézis

Bemutattam, hogy a Monte-Carlo Fakereső algoritmus képes Markov-döntési folyamatok különböző jutalomfüggvényeinek kiértékelésére. Ez a tulajdonság lehetővé teszi, hogy a Monte-Carlo Fakereső algoritmus kiváltsa a Markov-döntési folyamatok iteratív jutalomtervezési vagy összehasonlítási folyamatát, ami ágens tanítást igényel, mivel a Monte-Carlo Fakeresés előre képes meghatározni a különböző jutalomfüggvényekkel betanított ágensek várható teljesítmény szerinti sorrendjét. Egykereszteződéses közlekedési jelzőlámpa irányítási feladatot, valamint egy autonóm jármű oldalirányú vezérlési feladatát használtam annak bemutatására, hogy a Monte-Carlo Fakeresés képes a jutalomfüggvények kiértékelésére. Mindkét esetben a Monte-Carlo Fakeresés által adott rangsor megegyezett a tanított ágensek teljesítménye alapján kialakuló rangsorral.

Kapcsolódó publikációk:

- (KPAB24) Kővári B, Pelenczei B, Knáb IG and Bécsi T (2024), "Beyond Trial and Error: Lane Keeping with Monte Carlo Tree Search-Driven Optimization of Reinforcement Learning", Electronics. Vol. 13(11)
- (KPB22) Kővári B, Pelenczei B, and Bécsi T (2022) "Monte Carlo Tree Search to Compare Reward Functions for Reinforcement Learning." 2022 IEEE 16th International Symposium on Applied Computational Intelligence and Informatics (SACI). IEEE

2.3 3. Tézis

Egy új problémafelírást dolgoztam ki modellfüggetlen, többágensű megerősítéses tanulási ágensek számára az autópályaszakaszok változtatható sebességhatár-szabályozási problémájára. Az új problémafelírás egy olyan állapotleíró alkalmaz, amelyet csúzóablak-szerűen használtam fel. A szakirodalomban szereplő egyéb állapotleírási megközelítésekkel ellentétben a javasolt módszer nem igényli az egész szabályozott autópályaszakasz információját, így az állapotleírás függetlenné válik az irányított szakasz hosszától. A módszer méretfüggetlenségét úgy demonstráltam, hogy az ágenszt egy 1 km-es szakaszon tanítottam be, majd 1 km-es, 5 km-es és 10 km-es szakaszokon értékeltem ki. Az ágens minden esetben felülmúlta az alapmegoldásokat a várakozási idő, a CO₂-kibocsátás és a NO₂-kibocsátás tekintetében.

- (KKEBA24) Kővári B, Knáb IG, Esztergár-Kiss D, Bécsi T and Aradi S (2024), "Distributed highway control: a cooperative reinforcement learning-based approach", IEEE ACCESS, vol. 12, pp. 104463-104472
- (KKB23) Kővári, B., Knáb, I.G., Bécsi, T. (2025). Variable Speed Limit Control for Highway Scenarios a Multi-agent Reinforcement Learning Based Approach. In: Proceedings of the 2nd Cognitive Mobility Conference. COGMOB 23 2023. Lecture Notes in Networks and Systems, vol 1345. Springer, Cham.

2.4 4. Tézis

Egy új, problémától független tanítóminta-prioritási módszert dolgoztam ki modellfüggetlen, értékalapú megerősítéses tanulási ügynökök számára. A módszer minden tanítóminta kiválasztási valószínűségét az időbeli különbséghiba és az adott minta frissítésszámának kombinációjával határozza meg. Ez a megközelítés a mintaprioritáson keresztül az exploráció és az exploítáció egyensúlyának javítását célozza. A javasolt módszer gyorsabb konvergenciát és nagyobb kumulatív jutalmat eredményezett, mint a széles körben alkalmazott Prioritized Experience Replay (PER), amit a CartPole, Acrobot, Taxi és CliffWalking környezetekben végzett kiterjedt értékelések is alátámasztottak. Kapcsolódó publikációk:

- (KPB23) Kóvári B, Pelenczei B and Bécsi T (2023), "Enhanced Experience Prioritization: A Novel Upper Confidence Bound Approach", IEEE Access. Vol. 11, pp. 138488-138501.

Fejezet 3

Jövöbeli kutatási irányok

A jövöbeli kutatásaim fő iránya a megerősítéses tanulás (Reinforcement Learning, RL) mintahatékonyosságának növelésére irányul, különös tekintettel a redundáns tanítási minták problémájára. Korábbi munkámban már bemutattam egy módszert, amely képes megbecsülni az egyes tanítási minták információtartalmát, azonban továbbra is kihívást jelent annak eldöntése, hogy mely minták használata egyáltalán indokolt. Az RL-ben minden egyes interakciót általában eltárolunk egy memóriában, és később felhasználjuk a tanításhoz, de ezek közül sok minta redundáns, és akadályozza az ügynököt abban, hogy a valóban értékes adatokat hatékonyan hasznosítsa.

Ennek kezelésére két fő megközelítésem dolgozom. Az első az aktív tanulás (Active Learning, AL) és a megerősítéses tanulás kombinálását foglalja magában. Az AL-t hagyományosan a számítógépes látásban (Computer Vision) használják, ahol segít azonosítani azokat a címkézetlen mintákat, amelyek a legnagyobb mértékben járulhatnak hozzá a modell teljesítményéhez, ha címkézve vannak. Az RL-re alkalmazva ez a koncepció azt jelentené, hogy a memóriából nem minden egyes interakció kerülne felhasználásra, hanem csak a leginformatívabb tapasztalatokat választanánk ki.

A második megközelítés a mintagyűjtés fázisában történő megfelelő mértékű explorációra összpontosít. Az RL-ben az ügynök próbálkozás-alapú stratégiája gyakran egyenetlen lefedettséget eredményez az állapottérben, ami tanulási hézagokat okoz. Ennek ellensúlyozására olyan módszer kifejlesztését tervezem, amely aktívan figyelembe veszi a tanítási minták eloszlását az állapottérben, így biztosítva a kiegyensúlyozott mintasűrűséget. Ez lehetővé tenné az ügynök számára, hogy hatékonyabban tanuljon az alulfeltárt régiókra összpontosítva.

Ezek a módszerek nemcsak az RL területén alkalmazhatók, hanem más területeken is ígéretesek, különösen a mozgástervezésben, ahol az irányított exploráció optimalizálhatja az állapottér lefedettségét. Ezenkívül más mélytanulási területeken, például a számítógépes látásban is hasznos lehet a kiegyensúlyozott adatrepresentáció elérése, mivel ez javíthatja a modell teljesítményét és csökkentheti a tanítási időt azáltal, hogy kiküszöböli a redundáns mintákat.

Ezen kutatási irányok követésével nemcsak az RL hatékonyságát kívánom javítani, hanem egy olyan keretrendszer kidolgozását is célul tűztem ki, amely más mélytanulási

területeken is alkalmazható, és hatékonyabb tanítási folyamatokat tesz lehetővé különféle alkalmazási területeken.

Fejezet 4

Az új eredményekhez kapcsolódó kutatások

- (KPAB22) Kővári B, Pelenczei B, Aradi S and Bécsi T (2022), "Reward Design for Intelligent Intersection Control to Reduce Emission", IEEE Access. Vol. 10, pp. 39691 - 39699.
- (KTB21) Kővári B, Tettamanti T and Bécsi T (2021), "Deep Reinforcement Learning based approach for Traffic Signal Control", In Proceedings of The 24th Euro Working Group on Transportation Meeting (EWGT2021).
- (KSzBAG21) Kővári B, Szőke L, Bécsi T, Aradi S and Gáspár P (2021), "Traffic Signal Control via Reinforcement Learning for Reducing Global Vehicle Emission", Sustainability, MDPI., October, 2021. Vol. 13(11254), pp. 18.
- (KPAB24) Kővári B, Pelenczei B, Knáb IG and Bécsi T (2024), "Beyond Trial and Error: Lane Keeping with Monte Carlo Tree Search-Driven Optimization of Reinforcement Learning", Electronics. Vol. 13(11)
- (KPB22) Kővári B, Pelenczei B, and Bécsi T (2022) "Monte Carlo Tree Search to Compare Reward Functions for Reinforcement Learning." 2022 IEEE 16th International Symposium on Applied Computational Intelligence and Informatics (SACI). IEEE
- (KKEBA24) Kővári B, Knáb IG, Esztergár-Kiss D, Bécsi T and Aradi S (2024), "Distributed highway control: a cooperative reinforcement learning-based approach", IEEE ACCESS, vol. 12, pp. 104463-104472
- (KKB23) Kővári, B., Knáb, I.G., Bécsi, T. (2025). Variable Speed Limit Control for Highway Scenarios a Multi-agent Reinforcement Learning Based Approach. In: Proceedings of the 2nd Cognitive Mobility Conference. COGMOB 23 2023. Lecture Notes in Networks and Systems, vol 1345. Springer, Cham.

(KPB23) Kővári B, Pelenczei B and Bécsi T (2023), "Enhanced Experience Prioritization: A Novel Upper Confidence Bound Approach", IEEE Access. Vol. 11, pp. 138488-138501.

Hivatkozások

- Fawzi, A., Balog, M., Huang, A., Hubert, T., Romera-Paredes, B., Barekatin, M., Novikov, A., R Ruiz, F. J., Schrittwieser, J., Swirszcz, G., et al. (2022). Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53. [1](#)
- Haydari, A., Zhang, M., Chuah, C.-N., and Ghosal, D. (2021). Impact of deep rl-based traffic signal control on air quality. In *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*, pages 1–6. IEEE. [2](#)
- Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., and Meger, D. (2018). Deep reinforcement learning that matters. In *Thirty-Second AAAI Conference on Artificial Intelligence*. [1](#)
- Kocsis, L. and Szepesvári, C. (2006). Bandit based monte-carlo planning. In *European conference on machine learning*, pages 282–293. Springer. [3](#)
- Kušić, K., Dusparic, I., Guériau, M., Gregurić, M., and Ivanjko, E. (2020). Extended variable speed limit control using multi-agent reinforcement learning. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–8. IEEE. [4](#)
- Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. (2015). Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. [5](#)
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359. [1](#)
- Zhang, Y., Quinones-Grueiro, M., Barbour, W., Zhang, Z., Scherer, J., Biswas, G., and Work, D. (2023). Cooperative multi-agent reinforcement learning for large scale variable speed limit control. In *2023 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 149–156. IEEE. [4](#)
- Zheng, S., Li, M., Ke, Z., and Li, Z. (2023). Coordinated variable speed limit control for consecutive bottlenecks on freeways using multiagent reinforcement learning. *Journal of advanced transportation*, 2023(1):4419907. [4](#)

Zhu, Z., Lin, K., Jain, A. K., and Zhou, J. (2023). Transfer learning in deep reinforcement learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[1](#)