Skatinamojo mokymosi taikymas automobiliui apmokyti sėkmingai pravažiuoti duotu maršrutu

Oskaras Klimašauskas, Gintautas Dzemyda

Vilniaus universitetas, Matematikos ir informatikos fakultetas,
Duomenų mokslo ir skaitmeninių technologijų institutas,
Akademijos g. 4, LT-08412 Vilnius
oskaras.klimasauskas@mif.vu.lt

Santrauka. Šiame straipsnyje analizuojamas gilaus skatinamojo mokymosi (DRL) pritaikymas autonominiam automobilio valdymui simuliuotoje 2D lenktynių trasoje. Tyrimo metu lyginami du skirtingi įvesties duomenų tipai: spindulių pagrindu veikianti sistema, naudojanti atstumo iki trasos kraštų matavimus ir automobilio būsenos duomenis, bei vaizdo pagrindu veikianti sistema, apdorojanti aplink automobilį esančius vaizdus. Agentų mokymui naudojamas DDQN algoritmas, o jų veiklos efektyvumas vertinamas pagal nuvažiuotą atstumą. Eksperimentai atliekami su trimis skirtingomis mokymosi strategijomis: pradedant nuo trasos pradžios, pradedant atsitiktinėse trasos vietose ir mokant agentą dviejose skirtingose trasose. Tyrimo tikslas – nustatyti, kaip skirtingi įvesties duomenų tipai ir mokymosi strategijos veikia agentų mokymosi efektyvumą.

Raktiniai žodžiai: Mašininis mokymasis, Gilusis mokymasis, Skatinamasis mokymasis, Maršruto paieška, Demonstracinė aplinka.

1 Įvadas

Skatinamasis mokymasis yra galingas įrankis, leidžiantis agentams mokytis priimti optimalius sprendimus dinamiškose aplinkose. Tačiau skatinamojo mokymosi algoritmų efektyvumas labai priklauso nuo įvesties duomenų ir mokymosi parametrų. Šiame straipsnyje nagrinėjame gilaus skatinamojo mokymosi pritaikymą autonominio automobilio valdymui simuliuotoje aplinkoje. Pagrindinis dėmesys skiriamas dviejų skirtingų įvesties duomenų tipų – spindulių ir vaizdo – palyginimui. Spindulių pagrindu veikianti sistema teikia tikslius atstumo matavimus ir automobilio būsenos duomenis, o vaizdo pagrindu veikianti sistema apdoroja vizualinę informaciją iš aplinkos. Siekiame nustatyti, kurie įvesties duomenų tipai ir mokymosi strategijos leidžia agentams greičiausiai išmokti įveikti lenktynių trasą, pasiekti aukštą va-

DOI: https://doi.org/10.15388/LMITT.2025.12

žiavimo greitį ir efektyviausiai apibendrinti įgytas žinias naujoje, nematytoje trasoje.

Tyrimai [1] rodo, kad vizualinė informacija gali pasiekti panašų efektyvumą kaip ir struktūruoti duomenys, tačiau reikia:

- Ilgiau mokytis (dėl sudėtingesnio apdorojimo),
- Praleidinėti kadrus [2],
- Taikyti papildomą normalizaciją.

Tai rodo, jog įvesties pasirinkimas nėra tik techninis sprendimas – jis fundamentaliai keičia agento mokymosi eigą.

Be įvesties, agento veikimo efektyvumą taip pat reikšmingai lemia tai, kaip yra suformuota mokymosi aplinka. Mokymas pradedant nuo pastovios padėties leidžia greitai prisitaikyti prie konkretaus scenarijaus, tačiau agentas gali pernelyg užtrukti tyrinėdamas jau perprastą sekciją. Priešingai, atsitiktinių pradžios taškų arba skirtingų žemėlapių strategijos skatina aktyvesnį aplinkos pažinimą ir suteikiant įvairesnių mokymosi duomenų. Tokiu būdų agentui nereikia palaukti, kol išmoksta sekciją, kad gautų duomenų įvairovės. Tokie metodai kaip domeno atsitiktinės atrankos [3] ar mokymasis pagal mokymo programą [4] taip pat remiasi aplinkos įvairinimo idėja, siekiant išspręsti tyrinėjimo problemą. Tyrimai rodo, kad net nedidelis aplinkos sąlygų varijavimas ženkliai pagerina RL sistemų perkėlimo gebėjimus.

2 Metodologija

2.1 Aplinkos aprašymas

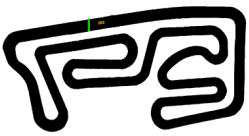
Eksperimentai buvo atlikti naudojant imituotą 2D juodai baltą lenktynių trasos aplinką, paremta 1 pav. [5] trasos forma, kurioje agentas (virtualus automobilis) pradeda kiekvieną epizodą trasos pradžioje kaip matome 2 pav. Agentas turi penkis veiksmų pasirinkimus, važiuoti pirmyn, atgal, sukti į kairę, sukti į dešinę, nieko nedaryt. Jis yra apdovanojamas proporcingai nuvažiuotam atstumuim ir baudžiamas už neveikimą link tikslo, 3 pav. vaizduoja didėjantį atstumą nuo starto būsenos. Susidūrimas su trasos ribas žyminčia balta siena yra griežtai baudžiamas, epizodas nutraukiamas. Epizodas taip pat nutraukiamas po 5000 pasirinktų agento veiksmų. Tyrimo metu buvo įgyvendintos trys mokymosi strategijos:

- Bazinis mokymas: Agentas kiekvieną epizodą pradeda toje pačioje pradinėje pozicijoje.
- Mokymas su atsitiktiniais startais: Agentas kiekvieną epizodą pradeda atsitiktinai parinktoje trasos vietoje.

- Mokymas dviejose trasose: Agentas treniruojamas pakaitomis dviejose skirtingose lenktynių trasose, pradedant nuo fiksuotų pradinių pozicijų kiekvienoje trasoje.

Simuliacija kiekvienai strategijai truko 2000 epochų.

1 pav. Pavyzdinė trasa



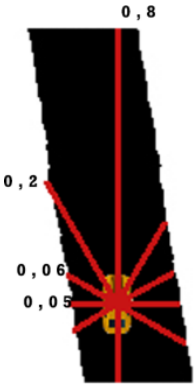
2 pav. Imituota juodai balta trasa

3 pav. Atstumų žemėlapis

2.2 Įvestys

Spinduliais grįsta įvestis

Šis įvesties metodas naudojo tiesiogiai iš simuliacinės aplinkos gaunamus duomenis. Agentui buvo pateikiami dešimt atstumo iki artimiausių trasos ribų matavimų, atliktų skirtingais kampais aplink automobilį, 4 pav vaizduoja jų pasiskirstymą. Be to, kaip įvesties savybės buvo naudojami ir vidiniai automobilio būsenos rodikliai: greitis, atbulinis greitis, teigiamas šoninis poslinkis (dreifas į dešinę) ir neigiamas šoninis poslinkis (dreifas į kairę). Taigi, spinduliais grįstos įvesties vektoriaus dydis buvo 14.



4 pav. Atstumą matuojantys spinduliai

Vaizdu grįsta įvestis

Vaizdu grįstas įvesties metodas rėmėsi vizualine informacija, gaunama tiesiogiai iš simuliacinės aplinkos, nereikalaujant tiesioginių simuliacijos būsenos duomenų. Agentas gavo aplink automobilį esančio vaizdo kadrus. Pradinis 125x125 pikselių spalvotas vaizdas buvo transformuojamas: pasukamas taip, kad automobilis visada būtų orientuotas vertikaliai į viršų, sumažinamas iki 50x50 pikselių ir konvertuojamas į nespalvotą formatą. Siekiant įvertinti automobilio greitį ir pagreitį, agento įvestį sudarė keturi tokie vienas po kito einantys vaizdo kadrai, jų vizualizaciją matome 5 pav. Vaizdu grįstos įvesties tenzoriaus dydis buvo [4, 50, 50]. Dėl kompiuterinių resursų apribojimų, buvo fiksuojamas tik kas ketvirtas simuliacijos žingsnio vaizdo kadras, todėl vaizdu grįstas metodas gavo mažesnį mokymosi duomenų kiekį nei spindulių metodas.

5 pav. Keturi vaizdo kadrai

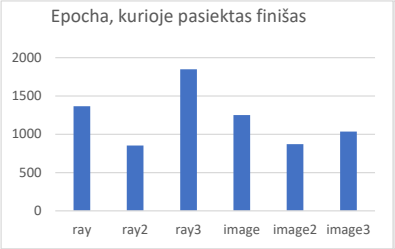
2.3 Algoritmai

Mokymui buvo naudojamas patobulintas Q-mokymosi algoritmas – Dvigubas Gilus Q-Tinklas (DDQN) [6]. Visuose eksperimentuose tarp skirtingų įvesties metodų buvo naudojami identiški hiperparametrai. Tačiau, atsižvelgiant į tai, kad vaizdo įvestis apima vizualinius duomenis, vaizdo informacijos apdorojimui buvo integruotas konvoliucinių neuronų tinklų (CNN) sluoksnis [7].

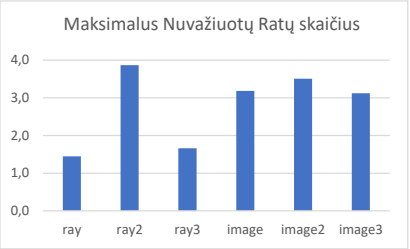
3 Rezultatai

Lentelėse „ray“ nurodo spindulių tipo duomenis, o „image“ – vaizdo. Skaičius 2 nurodo jog tai antroji strategija su atsitiktiniais startais, skaičius 3 – dviejų trasų. Eksperimento rezultatai rodo, kad atsitiktinių pradinių pozicijų naudojimas (ray2, image2) pagerino maksimalų pasiektų ratų skaičių vienoje epochoje, kaip matome 6 pav. ir žiūrint į 7 pav. sumažino epochų skaičių,

reikalingą trasai įveikti. Ray2 agentas pasiekė 3,9 rato per vieną epochą, o image2 – 3,5 rato, palyginti su atitinkamai 1,4 ir 3,2 rato fiksuotoje starto pozicijoje. Be to, šie modeliai pademonstravo geresnį apibendrinimą nematytoje trasoje, ką rodo 8 pav., ypač image2, kuris sugebėjo pilnai apvažiuoti prieš tai nematytą trasą, o ray2 agentas – 73 % trasos. Šie rezultatai rodo, kad atsitiktinės pradinės pozicijos skatina agentą įgyti bendresnius vairavimo įgūdžius ir geriau prisitaikyti prie nežinomų sąlygų.

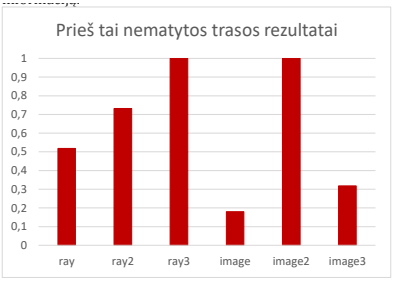


6 pav. Metodų palyginimas pagal finišo pasiekimą



7 pav. Metodų palyginimas pagal nuvažiuotus ratus epochoje

Mokymas dviejose trasose taip pat labiau pagerino apibendrinimą, ypač ray3 agentui, kuris taip pat sugebėjo visiškai įveikti nematytą trasą. Tačiau šis metodas padidino mokymo trukmę: ray3 agentui prireikė 1848 epochų, o image3 – 1036 epochų, kad pasiektų vienos trasos įveikimą. Vaizdu pagrįsti modeliai buvo stabilesni treniruotėse, tačiau jų našumas nematytoje trasoje buvo mažesnis nei spindulių pagrindu veikiančių modelių. Viena iš galimų priežasčių – dėl prieš tai minėto kadrų praleidimo, vaizdo tipo agentai gavo mažiau mokymosi duomenų, kas galėjo neigiamai paveikti jų gebėjimą apibendrinti informaciją.



8 pav. Nematytos trasos nuvažiuota dalis

4 Išvados

Remiantis šiame tyrime atlikta spindulių ir vaizdo tipo įvesčių metodų analize autonominio automobilio valdymo uždavinyje, galima daryti šias išvadas:

1. Vaizdu grįstas metodas, ypač derinant su atsitiktinių pradinių pozicijų strategija, pasižymėjo gebėjimu apibendrinti įgytas žinias ir sėkmingai įveikė visiškai nematytą lenktynių trasą. Šis pasiekimas yra reikšmingas, atsižvelgiant į tai, kad agentas mokymosi metu neturėjo tiesioginės prieigos prie vidinių automobilio būsenos duomenų, tokių kaip greitis ar poslinkis, o veikė tik apdorodamas vizualinę informaciją iš aplinkos ir taip pat gavo mažiau mokymosi duomenų dėl kompiuterio resursų apribojimų.
2. Spinduliais grįstas metodas, ypač naudojant atsitiktinių pradinių pozicijų strategiją, parodė žymiai geresnį našumą treniruočių metu ir padidino agento gebėjimą apibendrinti įgytas žinias, lyginant su fiksuota pradine pozicija. Vis dėlto, treniravimo strategija, apimanti mokymąsi dviejose skirtingose lenktynių trasose (ray3), užtikrino dar geresnį apibendrinimą ir leido agentui sėkmingai įveikti nematytą trasą, nors tokiai strategijai pareikalavo ilgesnio mokymosi laiko apvažiuoti vieną trasą kurioje mokinosi.

Tyrimas atskleidė, kad tinkamas įvesties duomenų tipo pasirinkimas ir mokymosi strategijos pritaikymas turi didelę įtaką gilaus skatinamojo mokymosi agentų našumui autonominio valdymo.

Literatūra

[1] A. Zhang, R. McAllister, R. Calandra, Y. Gal ir S. Levine, Learning invariant representations for reinforcement learning without reconstruction, 2021 International Conference on Learning Representations.

[2] M. V., K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra ir A. Riedmiller, Playing Atari with Deep Reinforcement Learning, arXiv:1312.5602 2013.

[3] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba ir P. Abbeel, Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World, 2017. IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, 2017.

[4] Y. Bengio, J. Louradour, R. Collobert ir J. Weston, Curriculum learning, 2009 ICML ,09: Proceedings of the 26th Annual International Conference on Machine Learning.

[5] Adobe Stock, [Tinkle]. Available: https://stock.adobe.com/lt/images/aerial-top-view-race-kart-track-track-for-auto-racing-top-view-car-race-asphalt-and-curve-grand-prix-street-circuit-aerial-view-asphalt-race-track/566864463. [Kreiptasi 06 04 2025].

[6] H. v. Hasselt, A. Guez ir D. Silver, Deep Reinforcement Learning with Double Q-learning, 2015 Proceedings of the AAAI conference on artificial intelligence. Vol. 30. No. 1.

[7] A. Krizhevsky, I. Sutskever ir G. E. Hinton, ImageNet Classification with Deep Convolutional, Curran Associates, Inc., 2012, pp. (Vol. 25, pp. 1097-1105).