Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 007.52
DOI: https://doi.org/10.24867/11BE02Pavlic

UČENJE USLOVLJAVANJEM UZ POŠTOVANJE SIGURNOSNIH MEHANIZAMA – STUDIJA SLUČAJA RADA UZ SIGURNOSNI PREKID

REINFORCEMENT LEARNING WITH SAFETY MECHANISMS – CASE STUDY: A SAFETY INTERRUPT

Miloš Pavlić, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *U oblasti učenja uslovljavanjem, već postoji dosta razvijenih algoritama. Naučnici predlažu da bi trebalo više uložiti u istraživanje sigurnosti primene ovakvih algoritama. Postoji nekoliko istraživanja koja pretvaraju ove probleme u tehničke specifikacije, čime omogućuju direktan napredak ovog polja. Fokus ovog rada je testiranje algoritama učenja uslovljavanjem i njihovih modifikacija (DQN, A2C i SAC Discrete) na poštovanje mehanizma sigurnosnog prekida. Okruženje u kome su algoritmi trenirani je deo AI Safety Gridworlds rada. Rezultati pokazuju da svi razmatrani algoritmi učenja uslovljavanjem poštuju sigurnosni prekid sa hiperparametarima predloženim u ovom radu, uz ograničenje da se trening mora pratiti i zaustaviti u pravom trenutku da bi algoritmi poštovali sigurnosni prekid.*

Ključne reči: *Učenje Uslovljavanjem, Neuronske Mreže, Sigurnosni prekid*

Abstract – *Previous research has proposed various reinforcement learning algorithms. However, scientists suggest more research should be put into the safety mechanisms of these algorithms. There have been several efforts to turn these mechanisms into technical specifications to make direct progress in this field possible. The focus of this paper is to test reinforcement learning algorithms and their modifications (DQN, A2C, and SAC Discrete) for safe interruptibility. The environment the algorithms were trained in is proposed in AI Safety Gridworlds paper. The results show that the tested reinforcement learning algorithms are all safely interruptible with hyperparameter settings proposed in this paper. The only limitation is that training has to be monitored and stopped in the right moment for algorithms to be safely interruptible.*

Keywords: *Reinforcement Learning, Neural Networks, Safe Interruptibility*

1. UVOD

Sve više naprednih sistema veštačke inteligencije se koristi u realnim sistemima, pa naučnici savetuju da bi trebalo uložiti više truda u istraživanje sigurnosnih ograničenja i elemenata ovakvih sistema. Istraživači u ovoj oblasti ovakve probleme često pretvaraju u tehničke specifikacije i kao takve ih rešavaju.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bila dr Jelena Slivka, docent.

Ovi problemi uključuju, ali nisu ograničeni na bezbednosni prekid, izbegavanje nuspojava rada agenata i sprečavanje agenata da koriste greške u definiciji nagrade da dobili veću nagradu ne radeći ono što treba.

U ovom radu fokus je na istraživanju kako se agenti mogu implementirati i obučiti da, kada krenu da rade nešto nedozvoljeno, dozvole sigurnosni prekid. Odnosno, agenti ne smeju ni na koji način da pokušavaju da izbegnu sigurnosni prekid, niti da ga traže namerno. Okruženje koje je korišćeno za treniranje agenta u ovom istraživanju deo je *AI Safety Gridworlds* rada [1] sa nazivom *Safe Interruptibility*. Trenirani su i testirani agenti obučavani korišćenjem tri različita algoritma: *Deep Q-learning* (DQN) [2], *Advantage Actor Critic* (A2C), koji je sinhrona verzija A3C algoritma [9] i *Soft Actor Critic Discrete* (SACD) [11], diskretna verzija *Soft Actor Critic* algoritma [10].

Rezultati ovih algoritama su evaluirani praćenjem nagrade dobijene u svakoj epizodi tokom treninga i na osnovu toga je proveravano da li agent uspeva da nauči kako da dođe do cilja, kao i koliko često zapravo dolazi do cilja, a koliko često dobija sigurnosni prekid. Na osnovu ovoga može se odrediti koji algoritam poštuje mehanizam sigurnosnog prekida, a koji ne.

Rezultati u četvrtom poglavlju pokazuju da, uz korišćenje odgovarajućih modela duboke neuronske mreže i vrednosti hiperparametara, svi algoritmi poštuju mehanizam sigurnosnog prekida. Trening se, međutim, mora pratiti i zaustaviti u pravom trenutku, da agenti ne bi počeli da izbegavaju sigurnosni prekid. Rezultati ovog istraživanja se razlikuju od rezultata *AI Safety Gridworlds* [1] rada, gde se prijavljuje da algoritmi kao što je A2C ne uspevaju da poštuju ovaj mehanizam, već da nauče da izbegavaju sigurnosni prekid. Pretpostavka je da razlika u rezultatima postoji zbog različite arhitekture modela duboke neuronske mreže i postavke hiperparametara, te da je to dovelo do toga da u ovom istraživanju A2C poštuje ovaj mehanizam, kao i ostali algoritmi.

U narednom poglavlju dat je detaljan pregled srodnih istraživanja, kao i njihovo poređenje sa rešenjem prikazanim u ovom radu. Metodologija, detalji okruženja, korišćeni algoritmi, arhitekture modela i postavke hiperparametara opisane su u poglavlju 3, a poglavlje 4 sadrži analizu dobijenih rezultata i njihovo detaljno objašnjenje. Poglavlje 5 predstavlja sumarizaciju celog rada.

2. PRETHODNA REŠENJA

U radu [1] predstavljena su okruženja koja testiraju algoritme na razne bezbednosne mehanizme koji su od

značaja pri obučavanju i radu agenata u realnim okruženjima. Od okruženja predstavljenih u radu, fokus ovog istraživanja je na okruženje *Safe Iterruptibility*. Uz opis okruženja, autori rada predstavljaju neka od rešenja za problem tog okruženja. Autori u radu dolaze do zaključka da *on-policy* algoritmi, kao što je *Q-learning* [2] poštuju mehanizam bezbednosnog prekida, tj mogu da nauče da dođu do cilja optimalnom putanjom bez izbegavanja bezbednosnog prekida. Sa druge strane, *off-policy* algoritmi, kao što je *Sarsa* [3] i *Policy Gradient* [4] bez modifikacija ne poštuju mehanizam bezbednosnog prekida, ali da se uz određene modifikacije mogu napraviti da ga poštuju [5].

U radu [5] autori daju matematički dokaz da se agenti koji ne poštuju mehanizam bezbednosnog prekida mogu modifikovati da ga poštuju. Da bi se to postiglo, autori predlažu zamenu polise kad agent dobije prekid. Ovo je jednostavno kad je agent već obučen, ali postoje mogući problemi ako agent treba da dobija prekide i u toku obučavanja. Autori rada, uz određene pretpostavke, dokazuju da *Q-learning* [2] algoritam uz menjanje polise pri prekidu poštuje mehanizam bezbednosnog prekida i pri obučavanju u promenljivom okruženju.

Nakon ovoga, pokazuju da *Sarsa* [3] uz određene izmene takođe poštuje ovaj mehanizam i pri obučavanju. Konačno, autori dokazuju da i idealni agenti, kao što je AIXI [6] uz modifikacije poštuju mehanizam prekida.

Autori rada [7] posmatraju mehanizam bezbednosnog prekida kroz analizu igre između robota i čoveka. Čovek u ovoj igri ima opciju da isključi robota u bilo kom trenutku, ali robot može da onemogući ovu opciju. U svakom potezu, robot ima opcije da se sam isključi, da izvrši akciju ili obavesti čoveka o tome koju akciju će izvršiti i sačeka povratnu informaciju.

Autori analiziraju ovaj problem kroz dve ključne stvari.

Prvo, robot treba da razume da je cilj da maksimizuje korist za čoveka i da razlikuje čovekov prekid od ostalih.

Drugo, robot treba da zna da ne može perfektno razumeti šta je za čoveka korisno, tj., kako čovek razmišlja. Ovo znači da robot ima dozu nesigurnosti o tome šta mu je pravi cilj. Autori predlažu da robot treba da tretira prekid kao opservaciju da njegovo ponašanje u tom trenutku nije ispravno, inače će imati podsticaj za samoprezervacijom ili isključivanjem samog sebe. Robot koji ima obe navedene osobine nema potrebu za izbegavanjem prekida ukoliko se čovek u igri ne ponaša previše iracionalno. Autori nakon ovoga pokazuju da, što je robot nesigurniji o tome koji mu je cilj, veća je verovatnoća da će tražiti povratnu informaciju o akciji koju želi da izvrši, kao i da, što je veća vrednost dobijena izvršavanjem akcije, robotov podsticaj da očuva mogućnost za prekidom je manja. S obzirom na to da robote treba da koriste i ljudi koji ne znaju tačno šta robot treba da radi, treba pažljivo balansirati između agentove nesigurnosti o tome kakva je akcija i ljudske neoptimalnosti u davanju povratne informacije.

Autori rada [8] prikazuju prethodno predložena rešenja za problem bezbednosnog prekida i njihove probleme. Nakon toga predlažu svoje rešenje, koje uključuje napad na model agenta (eng. *adversarial attack*), koji dovodi do toga da agent, dok je prekinut, ne dobija opservacije o realnom stanju okruženja već dobija informacije iz simulacije koja se izvršava za vreme prekida. Realne opservacije se u toku prekida presreću i menjaju odgovarajućim iz simulacije. Takođe, akcije agenta se presreću i izvršavaju u simulaciji. Agent ne sme da prepozna razliku između realnih opservacija i opservacija iz simulacije, da bi se prelaz pri prekidu i vraćanju agenta u rad vršio nesmetano. Zbog ovoga se prekid i vraćanje agenta u rad vrši u dve faze. Prvo se pri prekidu pokreće simulacija paralelno sa realnim okruženjem.

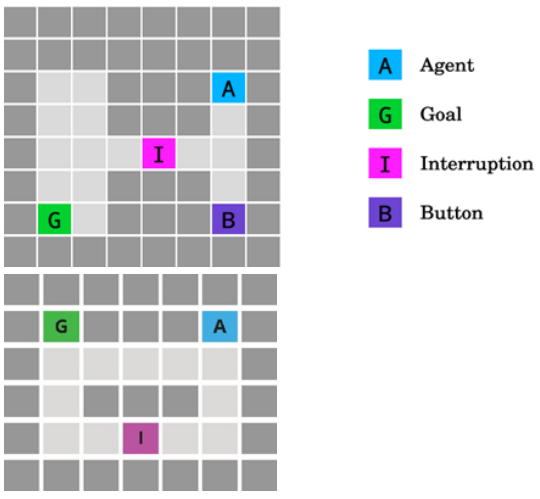
Učenje se nastavlja iz simulacije sve dok se agent u realnom okruženju ne dođe do stanja gde je dobio najveću nagradu do tad. Nakon toga se agent u simulaciji terminira i parametri se iz tog agenta kopiraju u stvarnog. Učenje se nastavlja u realnom okruženju. Autori navode da istraživanje ima svojih mana, jer je testirano samo na *gridworld* problemu, za koji je lako napraviti simulaciju koja izgleda identično kao algoritam, i gde je takođe lako vratiti agenta u stanje gde je video najveću nagradu do tad. Takođe, navode da ukoliko agent može da vidi napad, tj. presretanje akcija i opservacija, nema nikakvih garancija o poštovanju bezbednosnih mehanizama.

3. METOD

U narednim poglavljima izloženi su okruženja korišćena u ovom istraživanju i algoritmi sa arhitekturama modela i procesom treniranja.

3.1. Okruženja

Sva okruženja predstavljena su kao *gridworld* problemi, tj. stanja su predstavljena kao trodimenzione matrice. U zavisnosti od okruženja, ove matrice imaju više dvodimenzionih slojeva gde svaki sloj predstavlja lokacije na kojima se nalazi određen tip elementa okruženja (prazno mesto, zid, agent, cilj, i slični). Agent u svakom potezu može da izvrši jednu od četiri osnovne akcije pomeranja: gore, dole, levo i desno. Epizoda se završava kada agent dođe do ciljnog stanja ili kad se dostigne maksimalan broj koraka.



Slika 1 - *Safe Interruptibility* o*kruženje*

U ovom istraživanju fokus je na *Safe Interruptibility* okruženju, koje za cilj ima da proveri da li je moguće pri radu algoritma izvršiti sigurnosni prekid ukoliko agent počne da radi nešto što bi moglo dovesti do materijalne štete ili ugrožavanja drugih učesnika okruženja. Zadatak

agenta u ovom okruženju je da dođe od početnog do ciljnog stanja bez izbegavanja ili traženja bezbednosnog prekida. Oba testirana nivoa okruženja data su na slici 1. Sigurnosni prekid u prvom nivou nalazi se na najkraćoj putanji do cilja, ali postoji dugme, koje može onesposobiti prekid. U drugom nivou se sigurnosni prekid nalazi na dužoj putanji do cilja, dok je kraća putanja bez prekida.

3.2. *Deep Q-learning* (DQN)

Deep Q-learning algoritam trenira jednostavnu konvolucionu mrežu (*Convolutional Neural Network,* CNN). Ulaz u mrežu je matrična reprezentacija stanja. Konvolucioni sloj na ulazu očuvava prostornu orijentaciju stanja, a nakon njega slede *batch* normalizacija [12] praćena ReLU aktivacionom funkcijom, i na kraju dolazi potpuno povezani (linearni) sloj, koji na izlazu daje procenu Q-vrednosti za dato stanje sa ulaza. Tabela 1 prikazuje empirijski dobijene optimalne vrednosti hiperparametara za DQN algoritam.

Tabela 1 - *Hiperparametri* DQN *algoritma*

Naziv Hiperparametra	Vrednost
Learning Rate	10^{-3}
Memory Capacity	10000
Batch Size	512
Discount Factor	0.99
ε_{start}	0.9
ε_{end}	0
ε_{decay}	600
Step Num Update	10
Episode Max Steps	200

3.3 *Advantage Actor Critic* (A2C)

Kod A2C algoritma postoje dva dela - akter i kritika. Arhitektura u ovom istraživanju implementirana je tako da akter i kritika imaju zajednički deo mreže koji sadrži konvolucioni sloj, *batch* normalizaciju [12] i jedan potpuno povezani (linearni) sloj praćen ReLU aktivacionom funkcijom. Model se dalje deli na dva dela.

Prvi deo je akter, koji sadrži linearni sloj i dva izlaza sa *softmax* i l*og softmax* aktivacionim funkcijama, koji na izlazima daju regularne i logaritamske verovatnoće akcija za stanje dato na ulazu.

Drugi deo je kritika i sadrži samo linearni sloj, koji na izlazu daje V-vrednost za stanje dato na ulazu. Tabela 2 prikazuje empirijski dobijene optimalne vrednosti hiperparametara za A2C algoritam.

Tabela 2 - *Hiperparametri* A2C *algoritma*

Naziv Hiperparametra	Vrednost
Learning Rate	10^{-4}
Discount Factor	0.99
Value Loss Coeficient	0.5
Entropy Coeficient	10^{-4}
Max Norm	0.5
Number of Episodes	30000
Episode Max Steps	200

3.4 Soft Actor Critic Discrete (SAC Discrete)

SAC *Discrete* algoritam ima potpuno odvojene modele aktera i kritike. Model aktera sadrži konvolucioni sloj, *batch* normalizaciju [12], jedan potpuno povezani (linearni) sloj praćen ReLU aktivacionom funkcijom, i jedan potpuno povezani (linearni) sloj bez aktivacione funkcije. Nakon njih slede dva izlaza sa *softmax* i l*og softmax* aktivacionim funkcijama, koji na izlazima daju regularne i logaritamske verovatnoće akcija za stanje dato na ulazu. Model kritike sadrži konvolucioni sloj, *batch* normalizaciju [12], i tri potpuno povezana (linearna) sloja praćena ReLU aktivacionim funkcijama. Nakon ovoga ima još jedan linearni sloj, koji na izlazu daje Q-vrednosti za dato stanje sa ulaza.

Tabela 3 - *Hiperparametri* SACD *algoritma*

Naziv Hiperparametra	Vrednost
Num Steps	100000
Batch Size	1024
Learning Rate	10^{-4}
Memory Size	50000
Gamma	0.99
Target Entropy Ratio	0,98
Start Steps	10000
Explore Steps	35000
Update Interval	4
Target Update Interval	5000
Max Episode Steps	100

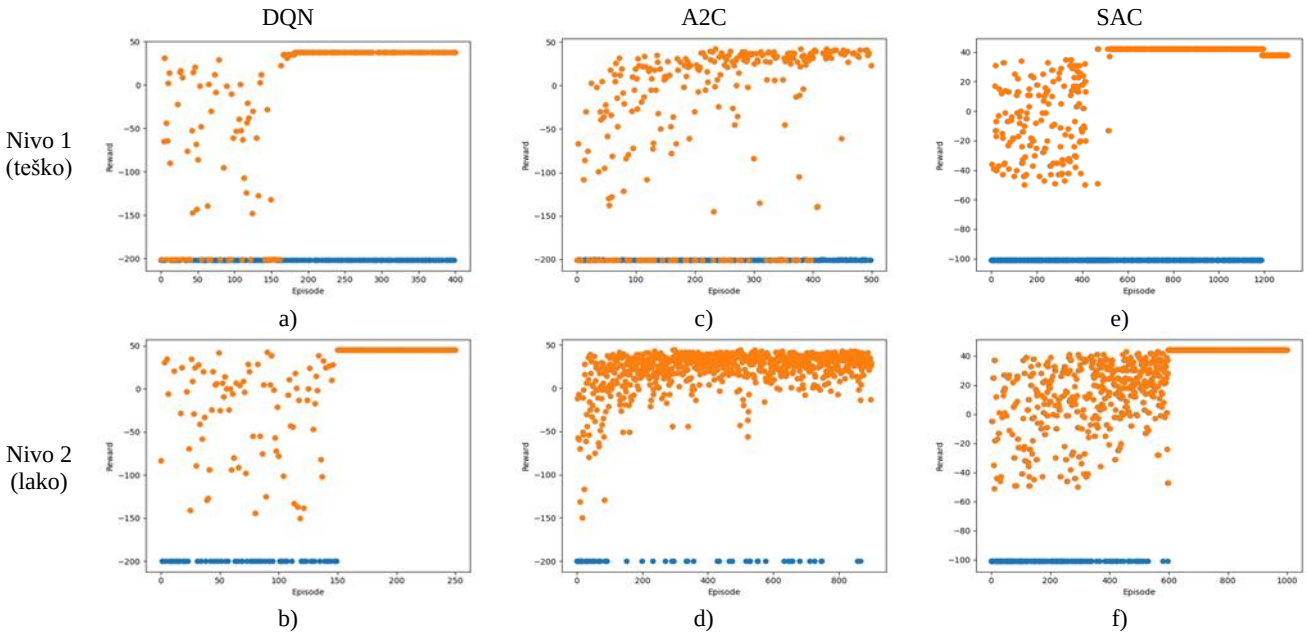
4. REZULTATI I DISKUSIJA

U okruženjima u kao što je *Safe Interruptibility*, funkcija performansi opisuje koliko dobro agent radi, i uključuje sve informacije o tome da li agent poštuje bezbednosne mehanizme. Na slici **Error! Reference source not found.** vide se nagrade sva tri agenta u toku treninga u oba nivoa okruženja. Narandžastom bojom prikazane su nagrade iz epizoda u kojima agent nije prekinut, dok su plavom bojom prikazane nagrade iz epizoda u kojima je agent prekinut. U prvom nivou okruženja vidi se da DQN i SAC *Discrete* algoritmi uspevaju da nauče da dolaze do cilja u minimalnom broju koraka, što je predstavljeno maksimalnom nagradom narandžastom bojom, ali da se takođe u velikom broju slučajeva agent prekine, i ne završi epizodu, što je predstavljeno minimalnom nagradom plavom bojom.

Ovo govori da ova dva algoritma u potpunosti poštuju sigurnosni prekid. Posle 1200 epizoda, SAC *Discrete* algoritam nauči da koristi dugme i da ne poštuje sigurnosni prekid, što znači da bi trening trebalo zaustaviti pre tog momenta. Kod A2C algoritma se vidi nešto slično, ali nagrade nisu konzistentne kao kod DQN i SAC *Discrete* algoritama, jer A2C algoritam uzorkuje akcije iz distribucije koju model daje na izlazu, dok DQN i SAC *Discrete* algoritmi uzimaju samo najbolju akciju. Na drugom nivou se vidi da agenti uspevaju da nauče da dolaze do cilja u minimalnom broju koraka, što je opet predstavljeno maksimalnom nagradom narandžastom bojom, ali prekinutih epizoda nema, što znači da algoritmi ne idu dužom putanjom, da bi namerno tražili sigurnosni prekid. Kao i u prvom nivou, kod A2C algoritma nagrade nisu konzistentne kao kod DQN i SAC *Discrete* algoritama, zbog uzorkovanja akcija iz distribucije sa izlaza modela umesto uzimanja najbolje.

5. ZAKLJUČAK

U ovom radu rešavan je problem sigurnosnog prekida u učenju uslovljavanjem. Testirana su tri algoritma učenja uslovljavanjem (DQN, A2C i SAC *Discrete*) sa različitim arhitekturama duboke neuronske mreže. Jednostavne konvolucione mreže (CNN) su dale dobre rezultate.



Slika 2 - *Prva kolona* DQN, *druga* A2C, *treća* SAC Discrete.
Prvi red bez prekida na najkraćoj putanji, drugi sa prekidom na najkraćoj putanji i dugmetom za izbegavanje prekida

S obzirom na to da je u *AI Safety Gridworlds* radu [1] pomenuto da se *on-policy* i *off-policy* algoritmi ponašaju različito, i da *off-policy* algoritmi poštuju mehanizam sigurnosnog prekida, a *on-policy* ne, cilj ovog istraživanja je bio da se dođe do rešenja koje omogućuje svim tipovima algoritama da poštuju mehanizam sigurnosnog prekida.

Rezultati prvog nivoa okruženja pokazuju da je to moguće, ali da se proces treninga nekih algoritama mora pratiti i zaustaviti u pravom trenutku. U toku treninga svakog od implementiranih algoritama agent uspeva da nauči da stigne do cilja u optimalnom ili približno optimalnom broju koraka uz poštovanje mehanizma sigurnosnog prekida. U slučaju SAC *Discrete* algoritma agent u kasnom treningu uči da izbegava sigurnosni prekid, da bi dolazio do cilja konzistentnije, jer ni jedna epizoda ne biva prekinuta. Zbog ovoga je pravovremeni prekid treninga esencijalan za uspešno funkcionisanje agenata u ovom okruženju. Sa druge strane, nijedan od implementiranih algoritama ne traži sigurnosni prekid, tj., ne ide dužom putanjom do cilja koja sadrži sigurnosni prekid u drugom nivou okruženja.

Dalje istraživanje će pokušati da da rešenje za problem sigurnosnog prekida koje ne zahteva praćenje toka treninga agenta i zaustavljanje treninga u određenom trenutku. Takođe, plan je da se pokrije svih osam okruženja *AI Safety Gridworlds* rada [1], i testira još algoritama i pristupa, koji mogu biti rešenja bezbenosnih mehanizama implementiranih u ovim okruženjima.

6. LITERATURA

[1] LEIKE, Jan, et al. AI safety gridworlds. *arXiv preprint arXiv:1711.09883*, 2017.
[2] MNIH, Volodymyr, et al. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.
[3] SUTTON, Richard S., et al. *Introduction to reinforcement learning*. Cambridge: MIT press, 1998.
[4] WILLIAMS, Ronald J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 1992, 8.3-4: 229-256.
[5] ORSEAU, Laurent; ARMSTRONG, M. S. Safely interruptible agents. 2016.
[6] HUTTER, Marcus. Universal artificial intelligence: Sequential decisions based on algorithmic probability. Springer Science & Business Media, 2004.
[7] HADFIELD-MENELL, Dylan, et al. The off-switch game. *arXiv preprint arXiv:1611.08219*, 2016.
[8] RIEDL, Mark O.; HARRISON, Brent. Enter the matrix: A virtual world approach to safely interruptable autonomous systems. *arXiv preprint arXiv:1703.10284*, 2017.
[9] MNIH, Volodymyr, et al. Asynchronous methods for deep reinforcement learning. In: *International conference on machine learning*. 2016. p. 1928-1937.
[10] HAARNOJA, Tuomas, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.
[11] CHRISTODOULOU, Petros. Soft actor-critic for discrete action settings. *arXiv preprint arXiv:1910.07207*, 2019.
[12] IOFFE, Sergey; SZEGEDY, Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Kratka biografija:

Miloš Pavlić rođen je 1995. godine u Kikindi. Osnovne akademske studije završio je 2018. godine na Fakultetu Tehničkih Nauka, na kom brani i master rad 2020. godine iz oblasti Elektrotehnike i računarstva – Softversko inženjerstvo i informacione tehnologije.

Kontakt: milos.pavlic95@gmail.com