Projekat 2 Duboko učenje

RL agent za RAF-Deep-RPG

Mateja Vasić
Sofija Todorović
Mihajlo Madžarević

I Analiza problema

Neophodno je napraviti “reinforcement learning” agenta koji će prevazilaziti nivoe RAF-Deep-RPG 2D igre koju je napisao Milan Bojić.

Bekend dostupan na: https://github.com/MilanBojic1999/greed_island

Frontend dostupan na: https://github.com/MilanBojic1999/raf-rpg-front

Prilikom analiziranja igre primetili smo sledeće:

- Igra sadrži šest nivoa dimenzija 26x13 različitih postavki kao i različite karaktere od kojih se neki pomeraju po mapi.
- Zbog razlike u nivoima potencijalno polja na koja smo mogli da pristupimo u prethodnim mapama nećemo moći pristupiti u novim mapama.
- U 5 nivoa se kapija i prodavac nalaze u gornjem delu mape, dok u nivou 5. se prodavac nalazi na donjem delu mape, a kapija na gornjem.
- U početku agent neće znati sredinu dovoljno i koji korak treba optimalno odraditi.

Napravili smo sledeće zaključke:

- Broj stanja možemo definisati kao širina mape puta dužina mape (26x13=338). Dobre strane ovakvog pristupa su manji broj stanja, jednostavnost implementacije kao i brže učenje i svest o celom observabilnom polju (celoj mapi). Problemi ovakvog pristupa su uračunavanja nepristupačnih stanja određene mape zbog pristupačnosti u prethodnim mapama, potencijalno sugerisanja poteza koji nas dovodi na neželjeno polje (uskok se nalazi na polju na kojem se nije nalazio u prethodnim pristupanjima datom polju).
- Zbog prethodno napomenutog zaključka kao i analiza koje smo izvršili, osmislili smo još jedan pristup mogućim stanjima agenta. Mogu se posmatrati samo četiri pristupačna polja oko agenta, i u tom slučaju bi broj mogućih stanja bio broj potencijalnih stvari na poljima stepenovan na broj pristupačnih polja oko agenta (približno 9^4=6561).

Postoji šest polja i tri aktera koji se mogu naći oko agenta. Dobre strane ovakvog pristupa su bolja lokalnost problema (agent zna tačno kako da postupi za svaku moguću situaciju), samim time i bolje prilagođavanje na novim mapama. Kod ovakvog pristupa je moguće iskoreniti problem skakanja na neoptimalna polja. Loše strane su veća kompleksnost većeg broja mogućih stanja, manjak informacija o globalnim pozicijama aktera na mapi i samim time potencijalnim dužim lutanjem agenta. Agent sa ovako definisanim stanjima bi trebalo da bolje izbegava koračanje na neželjena polja, ali bi „lutao“ dok ne nađe rešenje.

- Agent bi trebalo da se sa prvo navedenim pristupom stanja lakše snalazi u pronalaženju kapije i prodavca postojećih nivoa zbog njihovog pozicioniranja. Potencijalno će se teže snalaziti sa 5. nivom zbog pozicije prodavca na donjem delu mape.
- Prilikom testiranja različitih pristupa pravljenja agenta, shvatili smo da u početnim fazama agent nema svest koju akciju bi trebalo najbolje izvršiti pa je bolje favorizovati istraživanje (nasumično tumaranje po mapi kako bismo dobili svest šta je dobro, a šta ne), učenje (nova saznanja u početku treba drastičnije da menjaju postojeća) i sagledavati buduće nagrade u odnosu na trenutne (diskaunt rejt). Ove parametre treba terati u obrnute krajnosti kako trening nepreduje (smanjivati ih).

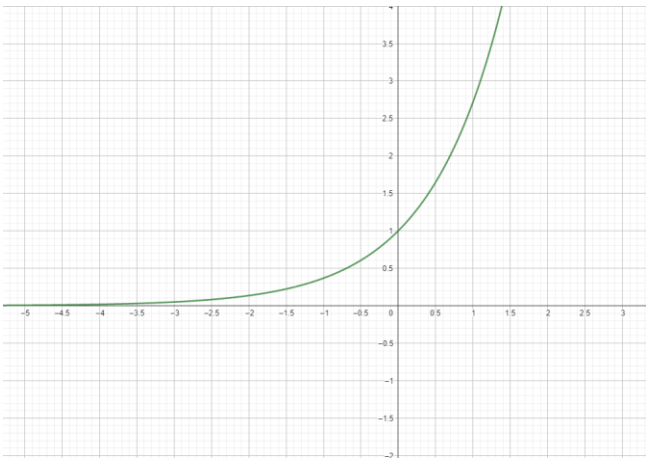
II Dizajn agenta

Za dizajn agenta korištena je “reinforcement learning” tehnika. Za našeg agenta koristili smo Q-learning tehniku radi jednostavnosti. Za popunjavanje Q-tabele koristili smo Bellmanovu jednačinu optimalnosti.

$$q^{new}(s,a) = (1-\alpha)\underbrace{q(s,a)}_{\text{old value}} + \overbrace{\alpha\left(R_{t+1} + \gamma \max_{a'} q(s',a')\right)}^{\text{learned value}}$$

Slika 1. Bellmanova jednačina optimalnosti.

Radi napomenutih zaključaka iz sekcije 1. stepen istraživanja, učenja i diskaunt rejt se smanjuju kako trening teče. Za ove potrebe smo se koristili sledećom formulom e^x čiji se graf vidi na slici 2.

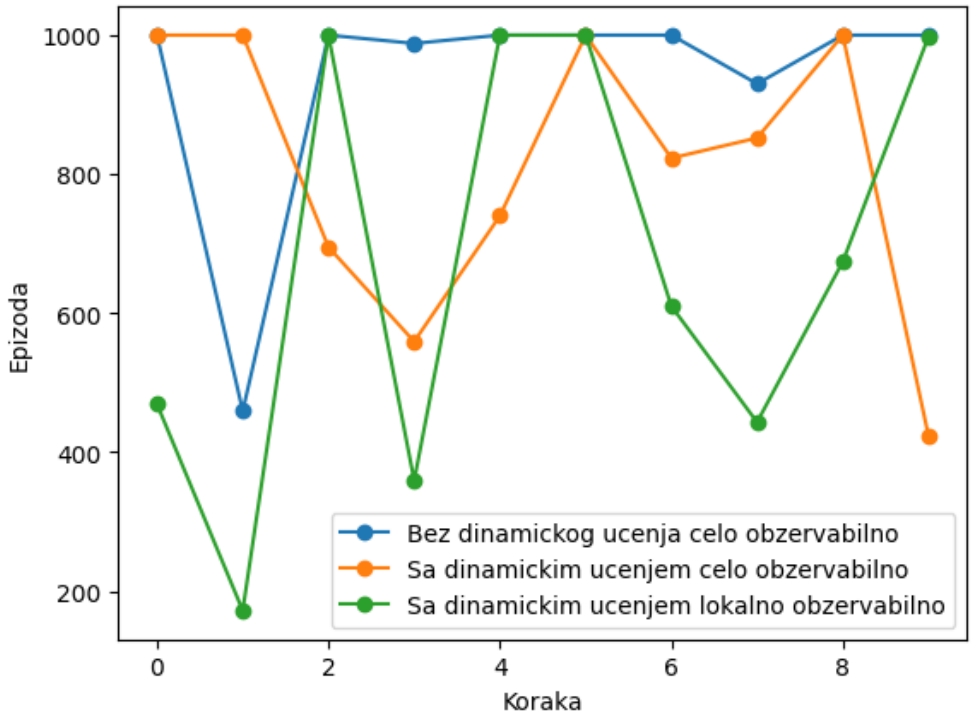


Slika 2. $y = e^x$.

Kao što se vidi sa grafa na slici 2. kako se vrednost za x (negativan redni broj epizode pomnozen konstantom smanjenja) smanjuje, tako i y (parametar koji smanjujemo) opada i konvergira ka 0. Ovo je idealno za prilagođavanje pomenutih parametara tokom epizoda agenta. Pored pomenutog, stepen istraživanja je na početku jednak 1 kako bi agent učio s obzirom da je Q-tabela prazna. Za krajnji izbor obzervabilnog polja uzeli smo samo četiri polja oko agenta (iznad, ispod, levo i desno od agenta).

III Rezultati

Prilikom korišćenja cele mape kao obzervabilnog polja bez adaptivnog stepena učenja i diskaunt rejta na 10 epizoda smo dobili samo 3 pobede sa prosekom oko 800 koraka po epizodi. Uvođenjem adaptivnog stepena učenja i diskaunt rejta sa prilagođavanjem početnog stanja parametara smo dobili bolje rezultate. Dobili smo 6 pobeda od 10 epizoda sa prosekom oko 680 koraka po epizodi. Puštanjem datog agenta na 100 epizoda dobili smo 30% pobeda. Promenom obzervabilnog polja na samo polja koja se nalaze oko agenta smo na 10 epizoda dobili 7 pobeda sa prosekom od oko 530 koraka. Ovo je ujedno i naše krajnje rešenje koje je na 100 epizoda imalo 38% pobeda. Krajnji rezultat bez treninga koji smo ostvarili je doneo 6 od 10 pobeda sa prosekom od oko 490 koraka. Na slici 3. se vide upoređeni rezultati pomenutih rešenja.



Slika 3. Poređenje rezultata agenta.

IV Zaključak

Pretpostavljamo da zahvalnost uspešnosti modela sa lokalno obzervabilnim poljima pripada činjenici da se mape dinamički menjaju. Još jedan od faktora uspešnosti je postepena degradacija stepena istraživanja, učenja i diskaunt rejta. Smatramo da bi neuronska mreža mogla biti potencijalna nadogradnja ovog modela ili pažljivije eksperimentisanje na različitim nivoima sa parametrima Q-učenja, kao i bolje definisanje stanja Q-tabele. Dodavanjem načina da agent sagleda gde se na mapi nalaze prodavac i kapija sa lokalnom obzervabilnošću bi mogao biti dobar pristup. To se može sagledati u poteškoćama agenta da nađe iste prilikom treninga. Pristup totalne obzervabilnosti može biti zbunjujuć po agenta iz razloga pomenute dinamike same igrice kao i različitim pozicijama polja u različitim nivoima. Jedna od stavki koja ide u korist ovakvom pristupu je činjenica da se prodavac i kapija na većini nivoa nalaze u gornjim delovima mape.