Анализа на техники од длабоко учење со поттикнување (DQN) во „Мунти 21“
Агентно-базирани системи

Александар Филиповски, 211047
Теодор Ангелески, 211080

Ноември 2024

1 Апстракт

Мунти 21, или попопуларно наречена Блекџек, е игра со карти која често се игра со одреден паричен влог. За разлика од голем дел на игрите во казино за кои може да се каже дека се целосно стохастички, правилата на игра на блекџекот дозволуваат развој на стратегии кои можат да го доведат играчот во поповолна ситуација.

Малиот простор на акции, како и малиот простор на обсервации, ја прават оваа игра погодна за примената на учење со поттикнување, за тренирање на агент кој ќе ја совлада околината.

Ова истражување ќе се фокусира на примената на алгоритми за длабоко учење со поттикнување врз околината Blackjack-v1 од Gymnasium, и ќе се извлечат одредени заклучоци во однос на поволноста на потезите во оваа средина.

2 Вовед во проблемот

Правилата на игра се следните. Типично, се игра со 6-8 шпилови карти (поедноставување во околината на Gymnasium е фактот дека играме со „бесконечен“ шпил на карти), и има еден играч, наречен дилер, кој ги дели картите на останатите. Играта започнува со обложување, по што секој играч и дилерот добиваат по две карти, при што картите на играчите се свртени нагоре, а дилерот има една откриена и една затворена карта. Картите од два до десет ја имаат својата вообичаена вредност, додека лицата – џандар, дама и поп – вредат по десет, а асот може да има вредност еден, или единаесет, зависно од тоа што е поповолно за играчот.

Откако ќе се поделат картите, играчот има неколку можности: да земе дополнителна карта, да остане со сегашниот збир, или да подели пар ако има две исти карти (во поедноставената околина од Gymnasium ова не е случајот, туку се разгледуваат единствено случаите кога играчот има една рака).

Дилерот игра според фиксни правила – мора да извлекува карти додека не достигне најмалку 17, а ако притоа надмине 21, играчите што не се прегореле победуваат. Ако играчот има збир поблиску до 21 од дилерот, победува и добива исплата од еден спрема еден (во околината од Gymnasium, овој исход се вика Win и има награда +1), освен ако има блекџек – ас и карта со вредност десет – што се исплатува еден и пол спрема еден (во околината од Gymnasium, овој исход се вика Natural, и има награда +1.5). Ако и играчот и дилерот имаат ист збир, велиме

дека играта е нерешена и облогот се враќа (draw, 0). Доколку играчот има збир на вредности на карти поголем од 21, велиме дека тој „изгорел“ (во околината од Gymnasium, овој исход се вика Lose, и има награда -1).

3 Сродни истражувања

Во *Machine Learning and Bioinspired Optimisation: CA2: Cooper, Francis, Maxwell, Tate, и Williams*, се истражува примената на алгоритми за токму оваа околина. Истражувачите применуваат DQN, DDQN, и оптимизација на хиперпараметри за да тренираат агенти, и ги споредуваат нивните резултати.

4 Опис на методологијата

За потребите на истражувањето, трениравме два агенти. Едниот е класичен DQN, со имплементацијата од stable-baselines3. Другиот е проширена варијанта од DQN, таканаречениот QRDQN кој користи квантилна регресија - DQN ја предвидува просечната награда која ќе се добие од акцијата, но притоа го намалува значењето на несигурностите во одлучувањето. QR-DQN подобрува врз оваа основа и користи квантилна регресија за да ги моделира и овие случаи. Имплементацијата на QRDQN која ја користиме во проектот е од stable baselines contrib проектот.

Со помош на Optuna, изнаоѓаме оптимални вредности за хиперпараметрите. Користиме 15 опити по 5000 timestamps.

```
learning_rate = trial.suggest_loguniform("learning_rate", 1e-5, 1e-2)
buffer_size = trial.suggest_int("buffer_size", 10_000, 500_000, step=10_000)
batch_size = trial.suggest_categorical("batch_size", [32, 64, 128, 256])
gamma = trial.suggest_float("gamma", 0.1, 0.99, step=0.05)
train_freq = trial.suggest_int("train_freq", 1, 20)
target_update_interval = trial.suggest_int("target_update_interval", 500, 5000, step=500)
exploration_fraction = trial.suggest_float("exploration_fraction", 0.0001, 0.1, log=True)
exploration_final_eps = trial.suggest_float("exploration_final_eps", 0.01, 0.1, step=0.01)
```

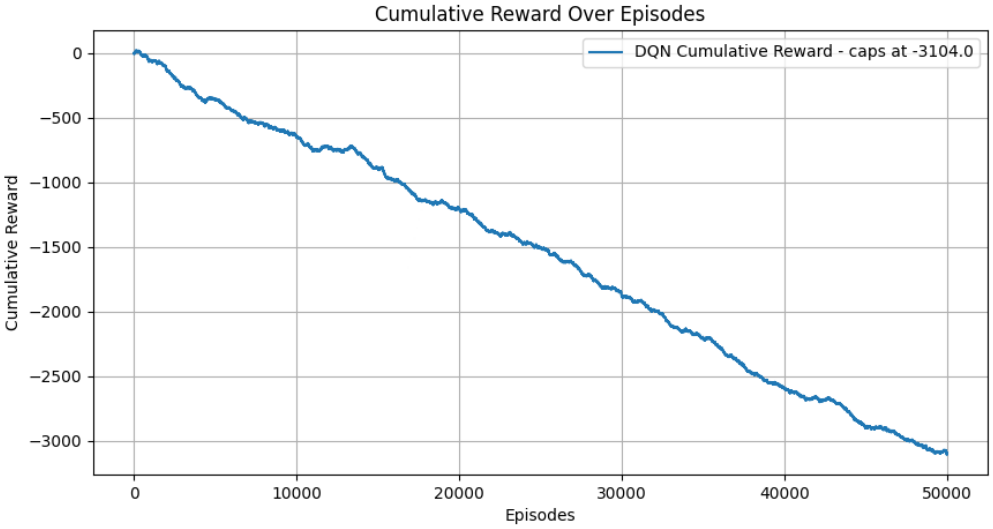
Добиените резултати ги користиме за да истренираме модел со 1M timestamps, кој потоа го тестираме со 50K рунди блекџек.

5 Резултати

За DQN агентот, ги добивме овие резултати:

Метрика	Резултат
Кумулативна награда	-3104
% победени	42.78%
% нерешени	8.97%
% изгубени	48.25%

Табела 1: Сумарни резултати од DQN за 50K епизоди

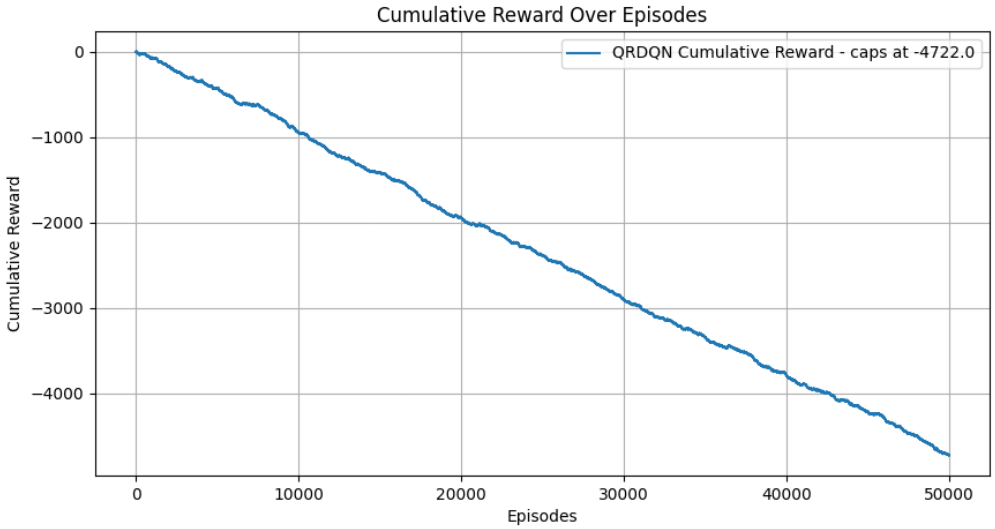


Слика 1: Кумулативна награда за DQN агентот

За QRDQN агентот, ги добивме следните резултати.

Метрика	Резултат
Кумулативна награда	-4722
% победени	38.51%
% нерешени	4.95%
% изгубени	56.53%

Табела 2: Сумарни резултати од QRDQN за 50K епизоди



Слика 2: Кумулативна награда за QRDQN агентот

6 Heatmap анализа

За подобра анализа на одлуките што ги носи агентот, користиме топлински мапи кои го прикажуваат изборот на акција во различни состојби во играта. Овие мапи ни овозможуваат да видиме дали агентот избира **Удирање (1) – Hit** или **Задржување (0) – Stick**, во зависност од збирот на играчот и откриената карта на дилерот.

6.1 Декодирање на состојбата и следење на одлуките

Состојбата на агентот се извлекува од закодираното набљудување, кое ги вклучува:

- **Збирот на играчот**
- **Откриената карта на дилерот**
- Дали играчот има **Usable Ace карта**, т.е. дали асот во моментот се брои како 11.

Топлинските мапи визуелизираат **веројатности за одлуки** во различни состојби:

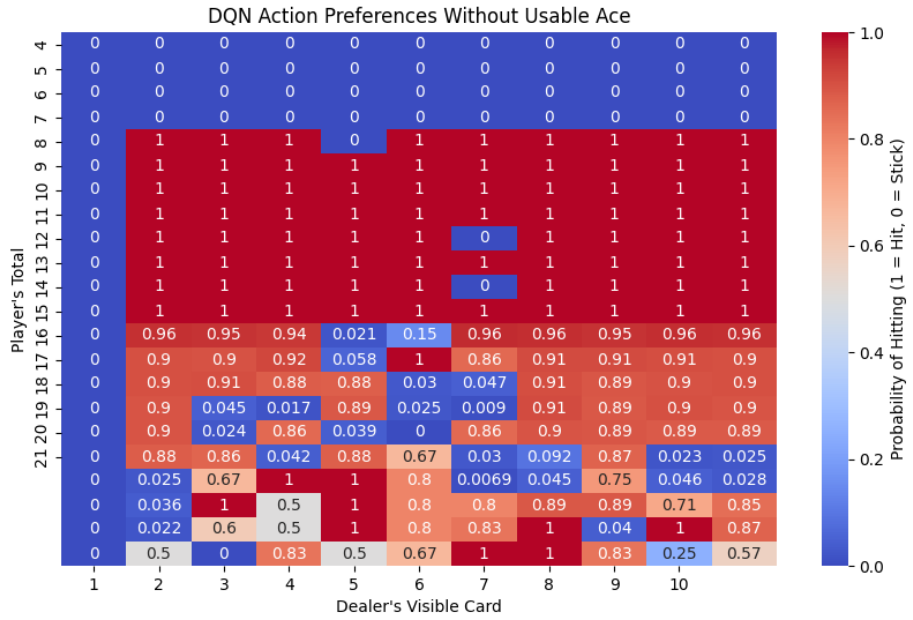
- **Црвените зони** покажуваат дека агентот **почесто се задржува (0)**.
- **Сините зони** покажуваат дека агентот **почесто удира (1)**.
- **Вредности блиски до 0.5** покажуваат **дека агентот има одредена несигурност**.

Генерираме две одделни топлински мапи:

1. **"Без Usable Ace карта"** (Usable Ace = 0) – кога играчот нема ас или асот мора да се брои како 1.
2. **"Со Usable Ace карта"** (Usable Ace = 1) – кога играчот има ас што може да се брои како 1 или 11.

7 Очекувана стратегија vs. Однесување на агентот

7.1 Анализа на топлинската мапа "Без Usable Ace карта"



Слика 3: No Usable Ace Heatmap

Кога играчот нема Usable Ace карта, стратегијата се заснова на едноставно сумирање на броевите на картите, каде што ризикот од преминување над 21 е значаен. Оптималната стратегија е:

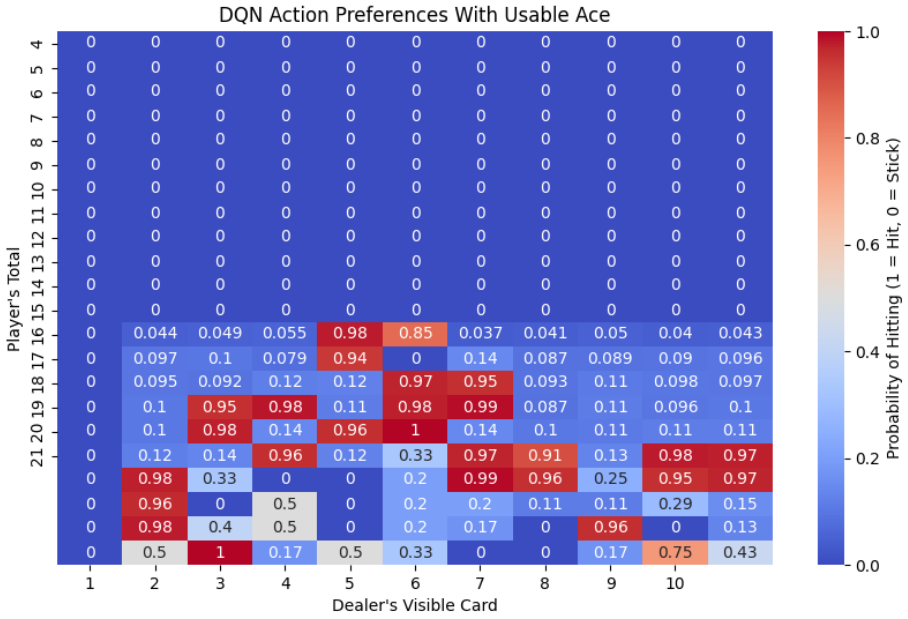
- **Задржување на 17+** за да се избегне ризик од преминување.

- **Удирање на 11 или помалку**, бидејќи нема можност за преминување.
- **Удирање на 12-16 доколку дилерот има силни карти (7-ас)**, бидејќи задржувањето скоро секогаш води до пораз.
- **Задржување на 12-16 доколку дилерот има слаби карти (2-6)**, бидејќи дилерот има висок ризик од преминување.

Однесување на агентот (Анализа на DQN)

- **Правилни стратегии:** Агентот правилно се задржува на 17+, секогаш удира на 11 или помалку и често удира на 12-16 против силни карти.
- **Проблеми:** Агентот премногу удира на 16 против дилер 2-6, што не е во согласност со основната стратегија.

7.2 Анализа на топлинската мапа "Со Usable Ace карта"



Слика 4: Usable Ace Heatmap

Кога играчот има **Usable Ace карта**, стратегијата се заснова на сумирање на картите земајќи го предвид фактот дека имаме ас кој може да ја смени својата вредност во 1, доколку е потребно, и имаме поголема флексибилност. Оптималната стратегија вклучува:

- **Поагресивно удирање** на збир под 18.
- **Удирање на збир 18 доколку дилерот има карти 9, 10 или ас**, бидејќи дилерот има силна рака.

- **Задржување на збир 19+**, бидејќи овие збирови веќе се доволно силни.

Однесување на агентот (Анализа на DQN)

- **Правилни стратегии:** Агентот правилно се задржува на збир 19+, агресивно удира на збир 13-17 против дилер со силни карти.
- **Проблеми:** Агентот понекогаш се задржува на 18 наместо да удира во случај кога дилерот има силни карти.

8 Споредба на двете топлински мапи:

Според ова, топлинските мапи треба да покажат јасна разлика во однесувањето:

- **Кога има Usable Ace карта** играчот треба да удира поагресивно.
- **Кога нема Usable Ace карта** играчот треба да биде повнимателен со удирањата.

Во нашиот случај:

- Агентот генерално го следи правилниот тренд – почесто удира со Usable Ace карта.
- Но, постојат неконзистентности кај 17-18, што укажува на нецелосно научена стратегија.

9 Заклучок

Од резултатите, забележуваме дека DQN е соодветен алгоритам за тренирање на агент за оваа околина. Со стапка на победени рунди од 42.78%, srезултатот кој го добивме е близок до традиционалната стратегија за блекџек (42.5%), што укажува дека агентот успеал да научи оптимални стратегии за игра.

Можеме да забележиме дека QRDQN не е соодветен DQN алгоритам за тренирање на агент за оваа околина. Со стапка на победени рунди од само 38.51%, резултатот кој го добивме е послаб од традиционалната стратегија за блекџек со 42.5%, меѓутоа сепак значително подобар од случајно избирање на чекори. Можни следни чекори би биле вршење на подетално испитување во однос на оптималните хиперпараметри, или пак повеќе тренирање на моделот. Како друга опција, би можеле да анализираме дали DQN агентот следи веќе познати хевристики за играње блекџек (дали стои на 17 во ситуации каде тоа е оптимално).

За околината Blackjack-v1 може да заклучиме дека е реалистичен пример на игра блекџек во која стратегиите на „броење карти“ се забранети. Стратегијата на играчот е ограничена од стохастичноста на играта, но сепак може да се постигне релативно висока стапка на победени игри со помош на параметрите кои ни се видливи.