Hluboké sítě a reinforcement learning pro hraní videoher

Bc. Tomáš Trdla <tomas.trdla@tul.cz>, Ing. Karel Paleček, Ph.D.

Práce přibližuje problematiku reinforcement learningu s využitím hlubokých neuronových sítí. V teoretické části byly rozvedeny základní algoritmy Reinforcement Learningu a podrobnější rozbor Deep Q-Network algoritmu spolu se způsoby jak tyto algoritmy využít k trénování umělé inteligence agenta. Dále byly popsány některé algoritmy ze skupiny value policy, které zlepšují učící proces a lze s jejich pomocí dosáhnout state-of-art výsledků. V praktické části byl následně implementován jednoduchý Deep Q-Network algoritmus pro učení agenta hrát Atari videohry. Poté je demonstrováno, čeho agent dosáhl za dobu 10 milionů učících kroků na několika různých videohrách ze skupiny Atari.

Klíčová slova: umělá inteligence, reinforcement learning, hluboké neuronové sítě, python, videohry

Úvod

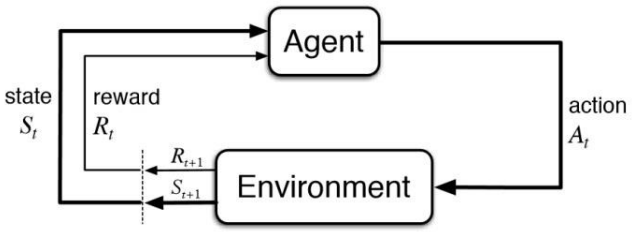
Cílem práce byla implementace a následné demonstrování umělé inteligence agenta, naučeného s využitím algoritmů Reinforcement Learningu (RL) v kombinaci s hlubokými neuronovými sítěmi (NN). Dále pak diskutování možných zlepšení učícího procesu a využití RL mimo videohry.

Videohry zde slouží jako velice vhodný nástroj pro demonstrování schopností RL algoritmů a neuronových sítí, protože se skládají z extrémně komplexního, dynamického, nelineárního a sekvenčního stavového prostoru.

Metodika

Reinforcement learning [3] je jedna z forem strojového učení, ve které se využívá biologicky přirozené schopnosti organismů se učit z posbíraných zkušeností a RL je tedy umělou adaptací této vlastnosti. Na rozdíl od supervised learningu, kde jsou sady učících dat předem známé, se zde tyto sady získávají postupem času, jak právě učený agent prohledává stavový prostor prostředí.

K prohledávání samotného stavového prostoru se nejčastěji využívá přístup, kdy se zpočátku učícího procesu vybírá náhodná akce s pravděpodobností 100% a tato pravděpodobnost se postupně snižuje až do nějakého vybraného minima. Tento výběr se nazývá ε – greedy policy.



Obrázek 1: Smyčka algoritmu RL učení

Neuronové sítě se zde používají kvůli jejich schopnosti zmapovat zmíněná komplexní a nelineární prostředí. Pro simulování podobných podmínek jako má člověk se zde používají konvoluční NN kvůli jejich aplikovatelnosti na strojové vidění.

Výhodou RL je, že ho lze použít pro sekvenční prostředí, které se vyvíjí v čase, nebo tam, kde může existovat několik různých strategií, jak dosáhnout požadovaného cíle. Avšak jednou z jeho největších nevýhod je, že učení zabírá velice dlouhý čas, protože trvá řádově desítky hodiny, než se projeví nějaké známky učení a pokud jsou nevyhovující, tak je třeba změnit některé z hyperparametrů jako je learning rate struktura sítě atd. a proces opakovat.

Deep Q-Network (DQN) využívá algoritmus, známý také jako Q-learning [2] pro naučení neuronové sítě zmapovat mnohadimenzionální prostředí a pro každý stav vybrat vhodnou akci, která maximalizuje budoucí odměny.

$$Q(s,a) = r + \gamma \max_{a} Q(s',a)$$

Rovnice 1: Výpočet nové hodnoty Q

Q zde představuje vektor hodnot, který chceme, aby se neuronová síť naučila zmapovat pro daný stav *s* a vybranou akci *a*. Odměna *r* je ohodnocení akce prostředím. Celkový význam rovnice je, že se agent snaží estimovat budoucí odměnu za každou zvolenou akci a následně vybrat tu nejvýhodnější.

Existuje několik algoritmů, které zlepšují učící proces, a díky tomu dokáže umělá inteligence učeného agenta volit o poznání výhodnější strategii rozhodování. Některými těmito algoritmy jsou např. Double Deep Q-Network, Prioritized Experience Replay, Noisy Nets atd. V současné době je za state-of-art považován algoritmus s názvem Rainbow, který je kombinací několika těchto algoritmů a práce [1] dokazuje, že jeho schopnosti dalece přesahují jednotlivé algoritmy, pokud jsou implementovány samostatně jak z pohledu rychlosti naučení, tak i dosažených výsledků.

Výsledky a diskuze

V rámci projektu byl implementován DQN algoritmus v jazyce python a s využitím knihovny PyTorch pro práci s NN. Jako videohry byly zvoleny staré ATARI hry, konkrétně Space Invaders, Pacman, Seaquest a pong.

Obrázek 2: Screenshoty prostředí (zleva pacman, seaquest, space invaders)

Agenti naučeni pomocí DQN algoritmu nejsou shopni se v rámci těchto her rozhodovat natolik kvalitně, aby se dalo mluvit o podobnosti s rozhodováním člověka. Nicméně krom hry pong bylo dosaženo známek učení např. ve hře pacman bylo dokonce dosaženo lepších výsledků, než jaké uvadí [1]. Důvodem tomu je jiná volba hyperparametrů a struktury NN.

V procesu učení hraje roli i náhoda, kdy pacman při každém začátku hry vyrazí stejným směrem. Náhodou je zde to, že při učení byla ε - greedy algoritmem vybrána právě tato strana, kudy se vydat, a prozkoumávání stavového prostoru tímto směrem převážilo ostatní směry, ačkoliv se jedná o zrcadlově souměrné prostředí okolo osy y a je logické, že stejné odměny by dosáhl i v opačném směru. V důsledku lze předpokládat, že tento učící proces není reprodukovatelný.

Za zmínku také stojí, že v oblasti videoher byl v roce 2018 po 2-3 letém výzkumu společností DeepMind představen agent, který dokáže porazit světové špičky ve hře StarCraft 2, která je považována za nejobtížnější a nejsložitější hru na e-sport scéně. Takový agent může mimo tréninku sloužit např. pro nacházení strategií, které hráči dosud neznají, nebo je pokládají za slabé.

Závěr

V práci byl implementován Deep Q-Network algoritmus z oblasti Reinforcement Learningu. Učení umělé inteligence agenta je demonstrována pomocí virtuálního prostředí několika vybraných ATARI her. V porovnání s oficiálními výzkumy v této oblasti dosáhl učený agent podobných výsledků ve většině zvolených her. Důvodem pro nezaznamenané známky učení ve hře pong byl mimo nedostatek času s největší pravděpodobností nevhodný návrh hyperparametrů a struktury neuronové sítě. Výsledky by bylo možné zlepšit za pomoci krátce popsaných algoritmů pro zlepšení učícího procesu např. [1].

Poděkování

Poděkování patří vedoucímu práce za vedení a možnost realizace projektu.

Tato práce byla podpořena z projektu Studentské grantové soutěže (SGS) na Technické univerzitě v Liberci v roce 2019.

Reference

[1] HESSEL, Matteo, Joseph MODAYIL, Hado HASSELT, et al. *Rainbow: Combining Improvements in Deep Reinforcement Learning* [online]. 6 Oct 2017 [cit. 26-05-2019]. Dostupné z: https://arxiv.org/abs/1710.02298

[2] Watkins Christopher. - *Learning from delayed rewards*, PhD. thesis, Cambridge University, 1989

[3] Online kurz "Reinforcement Learning", https://eu.udacity.com/course/reinforcement-learning--ud600