POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	**Exploration strategies for reinforcement learning with function approximation**
Jméno autora:	**Alena Moravová**
Typ práce:	diplomová
Fakulta/ústav:	Fakulta elektrotechnická (FEL)
Katedra/ústav:	Department of Computer Science
Oponent práce:	Mgr. Rudolf Kadlec, PhD
Pracoviště oponenta práce:	IBM Czech Republic

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání **náročnější**

Hodnocení náročnosti zadání závěrečné práce.

Obtížnost zadání odpovídá zvyklostem. Staví na v minulosti dobře studovaném problému a vyžaduje pokus o vylepšení některého ze stávajících standartních řešení.

Splnění zadání **splněno**

Posuďte, zda předložená závěrečná práce splňuje zadání. V komentáři případně uveďte body zadání, které nebyly zcela splněny, nebo zda je práce oproti zadání rozšířena. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.

Práce splňuje všechny body zadání.

Zvolený postup řešení **správný**

Posuďte, zda student zvolil správný postup nebo metody řešení.

Postup řešení logicky navazuje (přehled algoritmů RL a přístupů k exploration problému, návrh vlastního přístupu, evaluace přístupů v dvou prostředích) a shledávám jej správným.

Odborná úroveň **A - výborně**

Posuďte úroveň odbornosti závěrečné práce, využití znalostí získaných studiem a z odborné literatury, využití podkladů a dat získaných z praxe.

Řešení práce vyžadovalo jak znalosti z průběhu studia tak sledování poslední literatury.

Formální a jazyková úroveň, rozsah práce **B - velmi dobře**

Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku.

K úpravě práce mám několik drobných výhrad:

1. Sekce „2. Background“ je složena z dvou podsekcí (1. RL a 2. Deep Neural Networks (DNNs)) přičemž podsekce o DNNs v hloubce detailu silně zaostává za první podsekcí o RL.
2. Na podsekci „Policy based methods“ z přehledu RL metod se už ve zbytku práce nenavazuje. V experimentální sekci se používá „value based“ algoritmus DQN. Čtenář může tuto sekci přeskočit.
3. Na několika málo místech jsou špatně použity citace, jedná se ale spíše o překlepy:
    a. U článku Osband et. al. chybý rok vydání.
    b. „… Deep Reinforcement Learning by (Tang et al., 2017) …“ by mělo být „ … Deep Reinforcement Learning by Tang et al. (2017) …“ (závorka jen kolem roku)

Výběr zdrojů, korektnost citací **A - výborně**

Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení závěrečné práce. Charakterizujte výběr pramenů. Posuďte, zda student využil všechny relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně

odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.
Práce cituje nejnovější relevantní literaturu a dokonce na ní i staví v případě rozšíření metody od Bellemare et al. (2017).

Další komentáře a hodnocení *Vyjádřete se k úrovni dosažených hlavních výsledků závěrečné práce, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, publikačním výstupům, experimentální zručnosti apod.*
Vložte komentář (nepovinné hodnocení).

III. CELKOVÉ HODNOCENÍ, OTÁZKY K OBHAJOBĚ, NÁVRH KLASIFIKACE

Shrňte aspekty závěrečné práce, které nejvíce ovlivnily Vaše celkové hodnocení. Uveďte případné otázky, které by měl student zodpovědět při obhajobě závěrečné práce před komisí.

Práci jako celek považuji za zdařilou a mám k ní jen menší připomínky.

Nově navržená strategie explorace („uncertainty“) podle empirických výsledků nefunguje lépe než její alternativy. To není samo o sobě vzhledem k možnostem diplomové práce špatně. Bylo by ale hezké do práce zahrnout obsáhlejší diskusi, která by navrhla konstruktivní vylepšení této strategie.

V sekci 5.3 OpenAI Cart-pole mi schází informace o tom jak je přesně zadefinována odměna na kterou se optimalizuje, není pak jasné jakého výsledku dosáhne optimální strategie.

Závěrem mám na autorku jeden dotaz:

- Podle sekce 3.5.1 epsilon greedy agent s epsilon=1 v doméně cart pole dosáhl optimálních výsledků. Takový agent by ale neměl používat žádnou naučenou znalost protože s pravděpodobností 1 vybírá s uniformním rozložením náhodnou akci--- jde vlastně o „random baseline“. To je ale v rozporu s tím, že podle grafu 5.12 (a) se agent něco naučil (v epoše 0 má reward cca 3, pak se ale dostane až na optimální hodnotu). Prosím o vysvětlení těchto nesrovnalostí.

Předloženou závěrečnou práci hodnotím klasifikačním stupněm **A - výborně.**

Datum: 19.1.2018 Podpis: