Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Jan Uhlík
Název práce	Cooperative Multi-Agent Reinforcement Learning
Rok odevzdání	2021
Studijní program	Informatika **Studijní obor** Umělá inteligence

Autor posudku	Milan Straka **Role** Oponent
Pracoviště	Ústav formální a aplikované lingvistiky

Text posudku:

Hluboké zpětnovazební učení dosáhlo v uplynulé dekádě znatelných úspěchů (hraní her jako Go, šachy či videohry zcela bez lidských trénovacích dat, navigace robotů, návrhy neuronových architektur). V několika posledních letech se aktivně zkoumá také možnost využití algoritmů úspěšných v jednoagentním prostředí (SaRL) v případě víceagentních systémů (MaRL). Cílem diplomové práce bylo navrhnout nové algoritmy pro využití ve víceagentních systémech a porovnat je se stávajícími.

V první kapitole diplomové práce jsou zavedeny teoretické základy pro jednoagentní a víceagentní zpětnovazební prostředí. Vzhledem k tomu, že víceagentní rozšíření MDP byly definovány relativně nedávno, jsou v různých publikacích používány různé varianty a různá značení, takže definice s unifikovaným značením pro jednoagentní a víceagentní prostředí navržené v diplomové práci jsou prvním přínosem.

Druhá kapitola nabízí zdařilou rešerši existujících algoritmů pro jednoagentní prostředí. Autor popisuje sekvenci stále pokročilejších algoritmů, které na sebe navazují, až k TD3-FORK, což je algoritmus z října 2020. Popis existujících algoritmů pro víceagentní prostředí pak následuje v kapitole třetí.

Nově navržený algoritmus je prezentován v kapitole 4. Jedná se o úpravu algoritmu TD3-FORK pro víceagentní prostředí, s dvěma možnými rozšířeními (soft varianta dle Wei et al. (2018) a minimum dvou kritiků v policy update dle Fakoor et al. (2020)).

Kapitoly 5 a 6 popisují reimplementaci a porovnání všech algoritmů popsaných v diplomové práci, a to jak v případě spojitých tak diskrétních akcí. Kromě podrobných výsledků (7 stran v kapitole 6, 14 stran v appendixu) obsahuje práce i důkladný popis hyperparametrů a výsledky několika vhodně navržených heuristik.

Práci považuji za velmi zdařilou, věnující se aktivně zkoumanému a vysoce relevantnímu tématu. Navržení nového konzistentního značení vyžadovalo značný přehled v dané rozvíjející se oblasti, a vzniklý text tak může sloužit jako studijní text v pokročilých kurzech zpětnovazebního učení. Objektová reimplementace všech algoritmů názorně demonstruje společné rysy i rozdíly a vyhodnocovací platforma dovoluje snadné přidání dalších experimentů a vizualizaci chování natrénovaných agentů. V neposlední řadě, navržený algoritmus dosahuje velmi dobrých výsledků.

Práce je psána srozumitelnou angličtinou. Přestože obsahuje více jazykových chyb, než kdyby byla psána česky, tyto chyby nebrání porozumění, takže volbu angličtiny považuji za výhodu, díky které může mít práce i mezinárodní dosah.

Poznámky k textu práce:

- [str 5]: „… [based on the current state of the environment and the chosen action] the agent receives the information about the new state together with a *reward*, which signals whether it was a wise decision“ – akce se nemusí projevit okamžitě, tj. zda byl dobrý nápad akci provést se může projevit až mnohem později, byť text naznačuje, že se to agent dozví okamžitě
- [str 6]: v textu je algoritmus definován jako *model-based*, pokud má přístup k dynamice MDP; běžně se nicméně jako *model-based* klasifikují algoritmy, které využívají i natrénovaný model prostředí (koneckonců autor sám označuje TD3-FORK jako model-based).
- [str 8]: byť je definice délky trajektorie formálně správně, přijde mi neintuitivní, protože trajektorie (s0, a0, s1) má dle definice délku 0, zatímco já bych očekával 1, tj. počet počet akcí v trajektorii. Tato varianta by dovolila zavést délku „prázdné“ trajektorie (s0) a v následných definicích by stačilo v definicích R změnit spodní index na 1.
- [str 9]: poznámka pod čarou není myslím potřeba, rovnice (1.3) platí i pro finite-horizon situaci, protože čas musí být také součástí stavu, aby prostředí vyhovovalo definici (PO)MDP.
- [str 15]: podmínka na R v případě kooperativního a kompetitivního prostředí před sekcí 1.2.3 je nejednoznačná (není jasné, jestli má být vynechaná kvantifikace přes stavy a akce na začátku nebo na konci kvantifikátorů); použití funkce *sign* místo existenčního kvantifikátoru pro *k* by asi bylo přímočařejší.
- [str 26]: target network se upravuje pomocí exponential moving average, ne Polyak averaging (vzoroec je správně, jen název není korektní).
- [str 26]: „We also need to ensure that this function is differentiable w.r.t. action *a*. By using NN, this assumption is fulfilled.“ Toto tvrzení přímočaře platí jen v případě spojitých akcí; v případě diskrétních akcí (které jsou v DP také uvažovány) je potřeba použít nějakou spojitou relaxaci kategorické distribuce.
- [str 26]: Použitá spojitá relaxace kategorické distribuce (ST Gumbel-Softmax) by si zasloužila podrobnější popis, nepovažuji ji za běžnou znalost.
- [str 28]: Poznámka pod čarou není potřeba, termín TD4 je zavedený i v hlavním textu.
- [str 33]: „The naive implementation causes an exponential growth in these spaces with a number of agents“ – myslím, že stejně jako lze faktorizovat akce agentů, je možné faktorizovat i vstupní pozorování (tj. reprezentace vstupu pak nebude one-hot a bude růst lineárně s počtem agentů na vstupu); ve spojitém případě se používá stejná faktorizace, takže očekávám, že by fungovala korektně i v případě diskrétních stavů.
- [str 37]: „and done signal for each agent $\boldsymbol{d}$.“ Formalismus ukončování episod by si asi zasloužil podrobnější popis. Jedna z možností je přidat do definic prostředí funkci, která pro daný stav řekne, zda je terminální či ne (v případě MDP je takový test triviální, ale v případě POMDP ho nemusí být možné provést pouze z pozorování, proto je další funkce potřeba).
- [str 40]: střední hodnota přes výsledky $\boldsymbol{\mu_\theta}^{-i}$ nedává formálně smysl (protože je to funkce a ne distribuce); použití replay-bufferu by víc odpovídala střední hodnota přes možné chování ostatních agentů (tj. jednotlivé vzorky střední hodnoty by byly nějaké $\bar{\boldsymbol{\mu}}^{-i}$).

Práci doporučuji **k obhajobě.**

Práci nenavrhuji **na zvláštní ocenění.**

Datum 11. červen 2021 **Podpis**