FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI

STUDENTSKÁ VĚDECKÁ KONFERENCE 2016

Řízení dialogového systému s využitím zpětnovazebního učení

Adam Chýlek[1]

1 Úvod

Dialogové systémy umožňují přirozenou komunikaci člověka a stroje. Pravidla, kterými se řídí, je možné trénovat metodami strojového učení. Zpětnovazební (též posilované) učení se zakládá na myšlence, že se biologičtí agenti (zvířata, lidé) učí především interakcí s okolním prostředím. Dostatečným získáváním zkušeností (opakováním) jsme schopni odvodit a předvídat očekávanou odměnu při následování určité posloupnosti akcí (strategie).

Kromě agenta a prostředí, příp. modelu prostředí jsou dalšími prvky posilovaného učení pravidla, odměny a očekávané odměny. V každém čase, po provedení akce, předá prostředí agentovi hodnotu odměny. Ta hodnotí nový stav systému. Výše odměny pak definuje, zda akce vedle k tomu, že je systém ve stavu horším pro agenta (nízké hodnoty) nebo ve stavu přívětivějším (vyšší hodnoty). To lze přirovnat k bolesti, resp. potěšení u biologických systémů.

Cílem agenta ovšem není maximalizace okamžité odměnu za provedenou akci, ale maximalizace tzv. návratnosti za celou dobu provádění akcí. Některé akce totiž mohou vést k nízké okamžité odměně, ale v budoucnu k vyšší návratnosti. Agent tedy musí být schopen odhadnout výši návratnosti v budoucnu a výběr akcí založit právě na tomto odhadu.

2 Formální popis

Formálně pak můžeme hovořit o interakcích agenta a prostředí v diskrétních časových okamžicích $t = 0, 1, 2, \ldots$. V každé časovém okamžiku agent získává stav prostředí $S_t \in \mathcal{S}$, kde $\mathcal{S}$ je množina všech možných stavů prostředí. Agent pak vybírá akci $A_t \in \mathcal{A}(S_t)$, kde $\mathcal{A}(S_t)$ je množinou všech možných akcí za stavu S_t. V dalším časovém okamžiku získává agent odměnu $R_{t+1} \in \mathcal{R} \subset \Re$ a nový stav prostředí S_{t+1}. Předpokládáme, že stav systému v sobě sdružuje veškeré informace o minulosti. Jedná se o tzv. Markovův rozhodovací proces. Dále předpokládáme, že posloupnost akcí povede k cíli v konečném počtu časových okamžiků.

Jednou z možných metod odhadu očekávané návratnosti posilovaného učení je využití neuronových sítí. Jejich využití v oblasti hraní her na základě obrazového vstupu je demonstrováno v Mnih (2013).

Autoři používají sníženou návratnost v čase t definovanou jako $G_t = \sum_{t'=t}^{T} \gamma^{t'-t} R_{t'}$, kde T je čas dosažení cíle a $\gamma \in <0, 1>$ faktor snížení odměny. Dále definují optimální hodnotu kriteriální funkce $Q^*(s_t, A_t)$, která je maximální návratností při znalosti posloupnosti akcí a stavů $s_t = S_1, A_1, S_2, A_2, \ldots, A_{t-1}, S_t$ a při provedení akce A_t, jako $Q^*(s_t, A_t) = \max_\pi \mathrm{E}[G_t | s_t, A_t]$. Při znalosti sekvence s_{t+1} v následujícím okamžiku a znalosti $Q^*(s_{t+1}, A_{t+1})$ pro všechny možné akce pak $Q^*(s_t, A_t) = \mathrm{E}_{s_{t+1}}[R + \gamma \max_{A_{t+1}} Q^*(s_{t+1}, A_{t+1}) | s, A]$.

Právě k aproximaci $Q(s_t, A_t; \theta) \approx Q^*(s_t, A_t)$ této funkce autoři využívají neuronových sítí. Trénování parametrů sítě θ probíhá iterativně minimalizací posloupnosti ztrátových funkcí

[1] student doktorského studijního programu Aplikované vědy a informatika, obor Kybernetika, e-mail: chylek@students.zcu.cz

$L_i(\theta_i) = \mathrm{E}_{s,A}\left[(y_i - Q(s, A; \theta_i))^2\right]$. K optimalizaci ztrátové funkce lze volit gradientní metody. Během prvních iterací učení je vhodné volit akce A_t náhodně s pravděpodobností ϵ, která se s počtem iterací snižuje.

3 Řízení dialogu s informaci o odjezdech a příjezdech vlaků

Systém slouží jako potvrzení konceptu, že lze teorii zpětnovazebního učení použít i pro část řízení dialogového systému - volbu posloupnosti akcí systému. Definováno bylo 7 možných akcí, které systém může provést: dotaz na nástupní stanici, dotaz na cílovou stanici, dotaz na obě stanice, potvrzení uživatelem vyřčené informace, potvrzení uživatelem vyřčené informace následované dotazem na nástupní stanici, potvrzení uživatelem vyřčené informace následované dotazem na cílovou stanici a přečtení nalezených výsledků.

Na rozdíl od hraní her je vhodné chování uživatele modelovat, neboť trénování z interakce s reálným uživatelem je časově náročné. Proto byl expertně vytvořen model reakce uživatele na akce systému a generování odměn na základě znalostí z vývoje dialogového systému ze stejné domény.

Stav prostředí, který slouží jako vstup pro neuronovou síť, je vektor uchovávající typy sémantických entit obsažených v promluvě uživatele a informaci o stavu systému (vyplněné sloty). Neuronová síť má 2 plně propojené skryté vrstvy (164 a 150 neuronů) vstupní vrstva má neuronů 17 a výstupní vrstva má 7 neuronů (lineární aktivační funkce). Aktivační funkce skrytých vrstev jsou ReLU a v době trénování za každou následuje dropout vrstva s pravděpodobností 0,2. Optimalizace je prováděna algoritmem Adam. Po každé akci je provedeno přetrénování sítě. Z historie posledních 100 n-tic (stav, akce, odměna, příznak konce dialogu, stav po provedení akce) je pro přetrénování náhodně vybírána dávka 50 n-tic. Jedna epocha je jedním celým dialogem, trénování probíhalo 500 epoch.

4 Závěr

Natrénovaná síť byla testována na 10 000 dialozích za použití stejného modelu uživatele jako při tréninku a jako míra pro vyhodnocení byl zvolen poměr dokončených dialogů. Dialog se považoval za nedokončený, pokud přesáhl počet 10 akcí systému nebo pokud se systém pokusil provést akci, aniž by pro ni měl k dispozici dostatek informací. Tento poměr činil $93\ \%$.

Několik z testovacích dialogů bylo kontrolováno člověkem seznámeným s doménou a průběh těchto dialogů odpovídal průběhu dříve vytvořených expertních dialogových systémů. Můžeme tedy konstatovat, že tento koncept získávání strategie řízení je vhodný pro využití s dialogovými systémy. V budoucnu bude třeba se zaměřit především na vhodné modelování chování uživatele, kde je zapotřebí buď výborných znalostí experta na danou doménu nebo velké množství dat, které umožní model vytvořit.

Další výzvou je použití této metody učení pro systémy inkrementální, kdy se o reakci systému rozhoduje inkrementálně po menších jednotkách, než jsou promluvy (např. rámce audia, fonémy, slova).

Poděkování

Příspěvek byl podpořen grantovým projektem SGS-2016-039.

Literatura

Mnih, V., et al., 2013. *Playing Atari with Deep Reinforcement Learning*. Dostupné z: http://arxiv.org/abs/1312.5602.

Sutton, R., Barto, A., 1998. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge.