Explorácia pomocou internej motivácie a samokontrolovaného učenia

Matej Pecháč a Igor Farkaš

Fakulta matematiky, fyziky a informatiky
Univerzia Komenského v Bratislave
Email: {matej.pechac,igor.farkas}@fmph.uniba.sk

Učenie posilňovaním (reinforcement learning, RL) predstavuje významnú kategóriu metód strojového učenia, ktoré boli úspešne použité na riešenie rôznych sekvenčných úloh, kde agent interaguje s prostredím, napríklad v robotike alebo pri hraní počítačových hier. Jedným z kľúčových konceptov RL je vnútorná motivácia, ktorá umožňuje agentovi zlepšiť svoje správanie a dosahovať tak vyššie odmeny (kritérium pri RL).

Vnútorná motivácia (intrinsic motivation, IM) je definovaná (Ryan a Deci, 2000) ako vykonávanie činnosti pre prirodzené uspokojenie, a nie pre nejaký oddeliteľný následok (alebo inštrumentálnu hodnotu). IM bola operačne definovaná rôznymi spôsobmi, podloženými psychologickými teóriami, ktoré však poukazujú na určitú neistotu v tom, čo IM presne znamená. V kontexte RL, ak je motivácia generovaná v rámci štruktúr, ktoré sú súčasťou agenta, znamená to, že ide o internú motiváciu. Na to môže agent využívať hneď niekoľko prístupov, ktoré sa delia na dve veľké kategórie (Oudeyer a Kaplan, 2009): IM *založená na znalosti* a IM *založená na kompetencii*. Prvá vedie agenta k preskúmaniu každého možného stavu prostredia, druhá vedie zasa k nadobudnutiu všetkých možných zručností (napr. cieľom podmienené stratégie), ktoré v danom prostredí existujú. Podľa toho, akým mechanizmom dochádza ku generovaniu IM v prípade metód založených na znalosti, ich rozdelujeme na 3 prelínajúce sa kategórie: IM založená na predikčnej chybe, IM založená na detekcii novosti, IM založená na informačných konceptoch. Modely prvej kategórie obsahujú nejaký predikčný modul a jeho chyba slúži ako zdroj motivačného signálu. Modely druhej kategórie často využívajú koncept počítania návštevnosti stavu a z toho odvodzujú motiváciu agenta. Modely tretej kategórie sa snažia maximalizovať získanú informáciu z prostredia a napr. miera poklesu neistoty vo vzťahu k prostrediu alebo k agentovi je mierou motivácie.

V našej práci sme navrhli a otestovali novú triedu motivačných modelov *Self-supervised Network Distillation* (SND) založených na algoritmoch samokontrolovaného učenia (self-supervised learning) a využívaní destilačnej chyby ako detekcie novosti. Preto je ich možné zaradiť do rodiny metód IM založených na znalosti a pohybujú sa kdesi na rozhraní kategórie IM založenej na predikčnej chybe a detekcii novosti. Prvým takýmto modelom bol *Random Network Distillation* (RND) model (Burda a spol., 2018), ktorý sa stal základom našich modelov a v tomto kontexte sa javí ako ich špeciálny prípad. Naša metóda používa dva modely, *cieľový model*, ktorý poskytuje cieľové reprezentácie (vektorov príznakov) a *učiaci sa model*, ktorý sa ho snaží napodobniť (destilovať znalosti). Oba modely používajú ako svoj vstup reprezentáciu stavu v každom kroku diskrétneho času. Rozdiel vo výstupoch medzi oboma modelmi slúži ako signál pre vnútornú motiváciu. Predpokladali sme, že destilácia náhodného cieľového modelu poskytuje dostatočný signál iba na začiatku učenia. Preto sme navrhli tri metódy (Pecháč a spol., 2023) samokontrolovaného učenia pre trénovanie cieľového modelu: metódu SND-V založenú na kontrastívnom učení (Chopra a spol., 2005), metódu SND-STD založenú na SpatioTemporal DeepInfomax algoritme (Anand a spol., 2019) a metódu SND-VIC založenú na VICReg algoritme (Bardes a spol., 2022). Tieto metódy vytvorili priestor reprezentácií vhodných pre destiláciu. Spoločnou črtou metód samokontrolovaného učenia (v kontexte reprezentácií v RL) je predpoklad, že stavy, ktoré nasledujú po sebe, by mali mať veľmi podobné reprezentácie, zatiaľ čo stavy, ktoré sa v trajektórii nachádzajú ďalej od seba, by mali mať čo najrozdielnejšie reprezentácie. Tento cieľ je možné dosiahnuť optimalizáciou rôznych chybových funkcií spadajúcich do kategórie samokontrolovaného učenia.

Celkovo sme testovali naše metódy na 6 prostrediach Atari (Montezuma's Revenge, Gravitar, Venture, Private eye, Pitfall, Solaris) a 4 prostrediach Procgen (Coinrun, Caveflyer, Jumper and Climber), ktoré sa považujú za zložité na prehľadávanie prostredia (explorácia), pretože majú veľmi riedku externú odmenu. Z testovaných prostredí iba hra Pitfall predstavovala príliš zložitý problém, nakoľko žiaden zo všetkých testovaných algoritmov vôbec nezafungoval (t.j. nezískal ani jeden bod odmeny). V ostatných 9 prostrediach najlepšie výsledky dosiahli modely založené na SND metódach, pričom v 8 prípadoch to bolo s výrazným náskokom pred existujúcimi algoritmami (v prostredí Venture boli výsledky takmer rovnaké ako pri modeli RND). Pri porovnaní v skóre dosiahli modely SND najvyššie skóre v 5 prostrediach Atari hier a v 3 prípadoch (Montezuma's Revenge, Gravitar, Private Eye) bolo skóre výrazne vyššie ako porovnávané mo-

dely. Dosiahnuté skóre sa používa aj na porovnávanie s modelmi od iných tvorcov. V celosvetovom rebríčku [1] sa naše modely umiestnili zväčšna na 2. mieste. Okrem toho, pokiaľ nám je známe, sme prví, kto úspešne natrénoval agentov v Procgen prostrediach ťažkých na exploráciu.

V analýze sme sa zamerali na preskúmanie priestoru reprezentácií *cieľového modelu* pre RND aj SND algoritmy. Zostrojili sme viacero metód pre jeho analýzu a skúmali sme spojitosť medzi meranými parametrami a výsledkami daného modelu. Hoci ide o nelineárny vysokorozmerný priestor, rozhodli sme sa použiť nástroje lineárnej algebry a aspoň zhruba získať predstavu o niektorých jeho základných vlastnostiach.

Pomocou natrénovaného agenta sme zozbierali 10000 vzoriek pre jednotlivé prostredia. Najskôr sme pomocou QR dekompozície získali vektorovú bázu daného priestoru reprezentácií a zisťovali sme, či sú všetky dimenzie lineárne nezávislé, alebo či je reálna dimenzionalita daného priestoru nižšia. Z tejto analýzy vyšlo, že priestory vytvorené každou z metód majú všetky dimenzie lineárne nezávislé.

V druhom kroku sme vizualizovali pomocou dištančnej matice vzájomnú vzdialenosť vstupných stavov a následne vzájomnú vzdialenosť reprezentácií vytvorených jednotlivými metódami. Tam sme objavili zaujímavé výsledky. Zatiaľ čo RND vytváral v priestore reprezentácií podobnú štruktúru aká bola na vstupe v priestore stavov, SND metódy túto štruktúru do značnej miery potlačili. Posun v priestore stavov vyvolával veľký posun v priestore reprezentácií a bolo jedno "ktorým smerom" sme sa pohli. Inými slovami, malá aj veľká zmena vstupného stavu sa prejavila ako podobná zmena v reprezentácii, čo značne sťažovalo úlohu *učeného modelu*, ktorý sa snažil replikovať tieto reprezentácie. Naopak, pri RND modeli stačilo, ak sa natrénoval *učený model* niekoľko podobných stavov a zvyšné sa už od nich v priestore reprezentácií nelíšili, a preto ich ani nedetekoval ako nové, hoci mohli byť podstatné pre ďalšie napredovanie.

Ďalšou analýzou bol odhad natiahnutia jednotlivých dimenzií priestoru reprezentácií *cieľového modelu* (dalo by sa obrazne povedať, že išlo o odhad tvaru). Pomocou PCA metódy sme odhadli vlastné čísla lineárneho obalu a získali ich distribúciu. Z výsledkov bolo vidieť, že naše modely SND-VIC a SND-V majú oba tendenciu využívať rovnomerne všetky dimenzie, a obaly týchto priestorov majú tvar hyperelipsoidov, ktoré sa blížia k hyperguli. To koreluje s predchádzajúcimi zisteniami, že ľubovolná zmena v stavovom priestore (malá, či veľká) vedie k rovnakej zmene v rámci priestoru reprezentácií. Zdá sa, akoby neexistoval nejaký preferovaný smer, ale reprezentácie sú v ňom homogénne rozložené a každá reprezentácie je od každej približne rovnako vzdialená.

Poslednou analýzou SND modelov je pochopenie ich časového vývoja a schopnosti poskytnúť veľký signál internej odmeny pre predtým nevidené stavy. Pre účely explorácie je najdôležitejšia schopnosť odhaliť stavy v blízkej budúcnosti, ktoré sú veľmi podobné už videným stavom. Zozbierali sme súbor 2700 stavov od nášho najlepšieho agenta pre prostredie Montezuma's Revenge. Počas experimentu sme *cieľové a učené moduly* trénovali iba na vzorkách z minulosti a testovali sme ich citlivosť na vzorkách z budúcnosti, ktoré ešte nevideli. Pri všetkých troch SND metódach je vnútorná motivácia oveľa vyššia pre nevidené stavy a nekonverguje k nule ako pri RND.

Naše zistenia ukazujú, že pri trénovaní cieľového modelu je dôležité zabezpečiť *dekoreláciu reprezentácií a rovnaké využitie všetkých dimenzií príznakov*. Takýto model je pomerne robustný a dostatočne citlivý na novosť vďaka tomu, že jeho reprezentácie reagujú veľkou zmenou aj na malú zmenu v stave. Zároveň samokontrolovaná regularizácia zabraňuje kolapsu motivačného signálu na nulu. Na základe našich výsledkov môžeme konštatovať, že metódy samokontrolovaného učenia sú určite perspektívne pri vytváraní detektorov novosti, ktoré možno úspešne použiť z hľadiska vnútornej motivácie a zlepšiť tak skúmanie prostredia.

Poďakovanie: Tento výskum bol podporený projektmi VEGA 1/0373/23 a KEGA 022UK-4/2023.

[1] https://paperswithcode.com/task/atari-games

Literatúra

Anand, A., Racah, E., Ozair, S., Bengio, Y., Côté, M. a Hjelm, R. D. (2019). Unsupervised state representation learning in Atari. *CoRR, abs/1906.08226*.

Bardes, A., Ponce, J. a LeCun, Y. (2022). VICReg: Variance-invariance-covariance regularization for self-supervised learning. V *International Conference on Learning Representations*.

Burda, Y., Edwards, H., Storkey, A. a Klimov, O. (2018). Exploration by random network distillation. *arXiv:1810.12894*.

Chopra, S., Hadsell, R. a LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. V *International Conference on Pattern Recognition*.

Oudeyer, P.-Y. a Kaplan, F. (2009). What is intrinsic motivation? a typology of computational approaches. *Frontiers in Neurorobotics*, 1:6.

Pecháč, M., Chovanec, M. a Farkaš, I. (2023). Exploration by self-supervised exploitation. *arXiv:2302.11563*.

Ryan, R. a Deci, E. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25(1):54–67.