Olgierd Unold
Katedra Informatyki Technicznej
Wydział Informatyki i Telekomunikacji
Politechnika Wrocławska

Wrocław, 21.12.2024 r.

RECENZJA
rozprawy doktorskiej mgra inż. Jakuba Łyskawy
„Leveraging the Distribution of Collected Samples in Reinforcement Learning"
Promotor: dr hab. inż. Paweł Wawrzyński

1. Tematyka rozprawy

Tematyka recenzowanej rozprawy doktorskiej ulokowana jest w obszarze uczenia ze wzmocnieniem (ang. reinforcement learning, RL), dodajmy głębokiego RL, które nie bez kozery utożsamiane jest dzisiaj ze sztuczną inteligencją (ang. artificial intelligence, AI). O ile jeszcze nieco ponad 10 lat temu RL traktowane było jako jedno z możliwych podejść uczenia maszynowego, obok uczenia z nadzorem i bez nadzoru, to od roku 2013, od momentu publikacji architektury Deep Q-Network (Mnih et al., 2013) i niezwykle skutecznego połączenia RL z uczeniem głębokim, obserwujemy kolejne modele i zastosowania uczenia ze wzmocnieniem, które idą nieuchronnie, przynajmniej dla niektórych, w kierunku ogólnej czy też silnej sztucznej inteligencji. Dynamiczny rozwój tego obszaru AI skutkuje coraz bardziej wyrafinowanymi, ale też i praktycznymi aplikacjami. Przestają nam wystarczać spektakularne wyniki osiągane w coraz to trudniejszych grach, takich jak Go, DoTA 2, Starcraft 2, wprzęganie RL w mechanizm uczenia wielkich modeli już nie tylko językowych, jak ChatGPT, zaczynamy się pytać o wyjaśnialność tych algorytmów (Milani et al., 2024), ich bezpieczeństwo w rzeczywistych środowiskach (Gu et al., 2024), wreszcie możliwość sterowania obiektami fizycznymi (Tang et al., 2024).

Dysertacja wpisuje się w ten ostatni nurt badań i zgodnie z intencjami jej Autora, szuka potwierdzenia trzech hipotez. Dwie pierwsze z nich wprost stawiają tezę, że istnieje możliwość poprawy miary nazwanej „sample efficiency" (do tej miary wrócimy jeszcze później) w środowiskach sterowania obiektami fizycznymi („physical control environments"), trzecia jeżeli wprost do tego nie nawiązuje, to przecież jest udowadniana przez publikację, w której proponowany algorytm weryfikowany jest w symulowanych środowiskach robotycznych.

Można zatem z całą pewnością stwierdzić, że **problematyka rozprawy jest aktualna, ważna i w pełni uzasadniona**.

2. Kompozycja, redakcja i zawartość rozprawy

Na recenzowaną rozprawę składa się zbiór pięciu opublikowanych artykułów naukowych, opatrzonych wspólnym tytułem „Leveraging the Distribution of Collected Samples in Reinforcement Learning" (Wykorzystanie rozkładu zapamiętanego doświadczenia w uczeniu ze wzmocnieniem). Tworzą one załącznik B do zwartego druku, na który składa się sześć rozdziałów napisanych w języku angielskim, poprzedzonych streszczeniem, również w języku polskim, bibliografia oraz załącznik A, na który składa się lista użytych akronimów.

Rozdział 1 to wprowadzenie w tematykę rozprawy, opis osiągnięć naukowych doktoranta, a właściwie lista trzech postawionych hipotez wraz z syntetycznymi opisami osiągniętych rezultatów oraz lista opisów bibliograficznych publikacji wchodzących w skład recenzowanego cyklu i jednej publikacji spoza cyklu. Autor rozprawy stawia trzy hipotezy. Brzmią one następująco, cytuję w brzmieniu oryginalnym:

- Hypothesis 1. Including the properties of the autocorrelated action noise in the training proces improves the sample efficiency in physical control environments,
- Hypothesis 2. Decreasing the exptected duration of actions increases the sample efficiency in physical control environments while including the expected action duration in the collected data allows a value-based algorithm to efficiently reuse old experience samples.
- Hypothesis 3. The precision of achieving a subgoal in goal-conditioned HRL may be automatically calculated using a distribution of past distances to subgoals.

Zestawmy je z listą publikacji Autora wchodzących w skład cyklu (zachowuję oryginalną numerację) wraz z deklarowanym procentowym wkładem.

[P1] **Łyskawa, Jakub**, and Paweł Wawrzyński. "ACERAC: efficient reinforcement learning in fine time discretization." IEEE Transactions on Neural Networks and Learning Systems 35.2 (2024): 2719-2731, MNiSW=200 pkt, IF=10.4, Wkład=60%.

[P2] Szulc, Marcin, **Jakub Łyskawa**, and Paweł Wawrzyński. "A framework for reinforcement learning with autocorrelated actions." Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 23–27, 2020, Proceedings, Part II 27. Springer International Publishing, 2020, MNiSW=140 pkt, Wkład=30%.

[P3] **Łyskawa, Jakub**, and Paweł Wawrzyński. "Actor-Critic with variable time discretization via sustained actions." International Conference on Neural Information Processing. Singapore: Springer Nature Singapore, 2023, MNiSW=70 pkt, Wkład=75%.

[P4] Bortkiewicz, Michał, **Jakub Łyskawa**, Paweł Wawrzyński, Mateusz Ostaszewski, Artur Grudkowski. Bartłomiej Sobieski, Tomasz Trzcinski "Subgoal Reachability in Goal Conditioned Hierarchical Reinforcement Learning." 16th International Conference on Agents and Artificial Intelligence. SCITEPRESS, 2024, MNiSW=70 pkt, Wkład=20%.

[P5] Bednarski, Bogdan, Łukasz Lepak, **Jakub Łyskawa**, Paweł Pieńczuk, Maciej Rosoł, Ryszard Romaniuk. "Influence of IQT on research in ICT." International Journal of Electronics and Telecommunications 68.2 (2022), MNiSW=70 pkt, IF=0.7, Wkład=16,7%.

Publikacja [P1] datowana jest na rok 2024, ale artykuł o tym samym tytule i zmienionej kolejności autorów był umieszczony w repozytorium arXiv w roku 2021, a wydawnictwo IEEE wskazuje datę publikacji na rok 2022. Wszystkie publikacje mają przypisane punkty ministerialne, dwie z nich również zawierają przypisany współczynnik „impact factor". Cytowania tychże publikacji są (póki co) sporadyczne.

Hipotezę 1. mają dowodzić publikacje [P1] oraz [P2], hipotezę 2. publikacja [P3], natomiast hipotezę 3 publikacja [P4]. W tym zestawieniu nie ma publikacji [P5], co więcej, jest ona artykułem przeglądowym, w której doktorant jest autorem raptem jednostronicowego rozdziału V, a w samej rozprawie wspomina się o niej ledwie w rozdziale 2 opisującym tło literaturowe, nie ma o niej wzmianki też w podsumowaniu rozprawy. Zatem umieszczenie jej w cyklu należy uznać za błąd.

Kolejna uwaga dotyczy konsekwentnie stosowanej w rozprawie, w tym również w hipotezach, ale już nie w samych publikacjach (!?), zbitki słownej „autocorrelated action noise". Publikacje [P1] i [P2] wprowadzają nowy, skądinąd bardzo interesujący, algorytm ACERAC, który jest akronimem jego pełnej nazwy „Actor-Critic with Experience Replay and Autocorrelated aCtions". Stochastyczna zależność pomiędzy kolejnymi akcjami wymuszana jest przez rozszerzenie definicji strategii (zamiennie zwanej polityką) o proces autoregresyjny ξ_t, wymuszający w konsekwencji autokorelację następujących po sobie akcji. Owszem, można traktować ten proces jako rodzaj szumu o rozkładzie Ornsteina-Uhlenbecka, który jest również wykorzystywany przez część algorytmu zwanego krytykiem, kiedy to optymalizuje funkcję wartości-szumu w miejsce optymalizowanej zwykle w tych algorytmach (tj. działających w strukturze aktor-krytyk) funkcji samej wartości. Pojawia się jednak pytanie, na które mam nadzieję uzyskać odpowiedź, jaki był cel zmiany sformułowania istotnej własności algorytmu ACERAC. Czyżby sformułowanie to miało na celu podkreślenie roli szumu towarzyszącego wyborowi akcji na etapie eksploracji oraz szumu autokorelacyjnego zachowującego bliskie następstwo akcji?

Jeżeli już jesteśmy przy hipotezach, to pewne wątpliwości budzi optymalizowana metryka „sample efficiency", która pojawia się podczas formułowania hipotezy 1. i 2. Nie ma jej w pracy [P2], gdzie się mówi tylko o efektywności uczenia, pojawia się w pracy [P1], która jest rozszerzeniem [P2] i to głównie w zakresie zebranego i analizowanego materiału eksperymentalnego, ale już nie innych metryk. Dla precyzji dodajmy, że [P1] redefiniuje też nieco funkcję wartości-szumu. Zatem w [P1] pojawia się jej właściwa interpretacja, cytuję „the learning should be efficient in terms of the amount of experience needed to optimize the agent's behawior". Za tak mierzoną efektywność odpowiedzialny jest bufor pamiętający powtarzane doświadczenia, czyli w oryginalnym sformułowaniu experience replay (ER). Może się mylę, jeżeli tak, to liczę, że doktorant wyprowadzi mnie z błędu, ale w obydwu wspominanych pracach mierzy się co najwyżej wielkość średniego zwrotu (ang. return) (choć nota bene w legendach do wykresów mowa jest o średniej sumie nagród (ang. reward)). Nie ma analizy ani wielkości bufora ER, ani metody jego próbkowania, na efektywność uczenia (jakkolwiek mierzonej). Zresztą, dobroczynny wpływ bufora ER na szybkość uczenia algorytmu aktor-krytyk udowodnił już promotor w swojej nietrywialnej propozycji opublikowanej w 2009 roku (Wawrzyński, 2009).

Przyjrzyjmy się jeszcze kolejnym rozdziałom rozprawy. Rozdział 2. w podrozdziale 1. wprowadza skrótowo czytelnika w zagadnienie samego uczenia ze wzmocnieniem, ze szczególnym uwzględnieniem schematu uczenia aktor-krytyk z buforem ER, który to model jest oryginalną propozycję promotora z roku 2009 i stanowi punkt wyjścia dla prac [P1, P2]. Podrozdział 2. klasyfikuje metody hierarchicznego uczenia ze wzmocnieniem (ang. hierarchical RL, HRL), poświęcając oddzielną sekcję na metody HRL uwarunkowane celem (ang. goal-conditioned). W obrębie tych metod porusza się praca [P3]. Ostatni podrozdział jest krótkim przeglądem ostatnich osiągnięć literaturowych w obszarze kwantowego RL.

Rozdział 3 dotyczy autorskich już metod wymuszania podobieństwa (w tytule rozdziału Autor używa słowa „similarity", choć może być ono nieprecyzyjne) następujących po sobie akcji. Pierwszy podrozdział dedykowany jest metodzie autokorelacji akcji i wzmiankowanych już wcześniej kilkakrotnie publikacji [P1] oraz [P2]. Zgodnie z deklaracją samych autorów, praca [P1] jest rozszerzeniem pracy [P2] i to przede wszystkim ilościowym, bo obejmującym nowy materiał badawczy, a w zakresie modyfikacji samego algorytmu, zmianie uległa funkcja wartości szumu, uwzględniająca również stan autoskorelowanego szumu. Model ACERAC, czy to w wersji pierwszej czy drugiej, jest bardzo interesującą propozycją dostosowania metod uczenia ze wzmocnieniem do rzeczywistych problemów spotykanych w sterowaniu fizycznymi obiektami, jakimi są roboty. Po pierwsze, próba zastosowania jakiegokolwiek

modelu RL do świata rzeczywistego napotyka niełatwą barierę konieczności działania w ciągłym czasie. A to wymusza odpowiednio gęstą jego dyskretyzację. Po drugie, obiekty fizyczne mają swoje ograniczenia wynikające z zastosowanej w nich mechaniki. Nie da się, a przynajmniej nie powinno, wymuszać na obiekcie diametralnie różnych akcji w następujących po sobie chwilach czasu. Propozycja strategii z autoskorelowanymi akcjami (tak o niej pisze promotor w swojej książce „Uczące się systemy decyzyjne" z roku 2021), zawarta i rozwinięta w publikacjach [P2] i [P1], niewątpliwie jest udaną próbą implementacji warunku ograniczania odległości pomiędzy uruchamianymi akcjami w kolejnych stanach (choć oczywiście ciągłość funkcji π pozwala poruszać się czasie pseudociągłym). Wprowadzenie szumu autoregresyjnego o rozkładzie Ornsteina-Uhlenbecka do definicji funkcji polityki w modelu ACER (czyli modelu RL pracującego w reżimie aktor-krytyk i dodatkowo zapamiętującego doświadczenia w buforze ER) oraz zastąpienie funkcji wartości-akcji funkcją wartości szumu okazało się oryginalną i co więcej umocowaną formalnie propozycją. Przeprowadzone eksperymenty za pomocą silnika fizyki pyBullet dla czterech różnych zadań wykazały wyższą efektywność uczenia (choć nie wiem, czy „sample efficiency") w porównaniu do bazowej metody ACER oraz dwóch sztandarowych metod aktor-krytyk, czyli Proximal Policy Optimization (PPO) i Soft Actor Critic (SAC), a w przypadku publikacji [P1] jeszcze Continuous Deep Advantage Updating (CDAU). Co może być interesujące, ta ostatnia metoda też stosuje autokorelacje akcji. Praca [P1] rozszerza jeszcze materiał badawczy o badanie wpływu różnych poziomów dyskretyzacji czasu na działanie modeli. Dla zdecydowanej większości przebadanych konfiguracji, ACERAC wykazał wyższe tempo uczenia i większą średnią sumę zwrotów.

Niewątpliwie, model zaprezentowany w publikacjach [P1, P2] warty jest uwagi. Naturalnie rodzą się pytania, czy zmiana sposobu pobierania próbek z bufora ER albo wprowadzenie, jak ma to miejsce w nowszych architekturach aktor-krytyk, funkcji przewagi (ang. advantage) nie poprawiłoby jeszcze efektywności i/lub stabilności działania modelu?

Tutaj pojawia się jednak jeszcze inne pytanie, dotyczące rzeczywistego wkładu doktoranta. Co prawda, dla dwuautorskiej publikacji [P1] zadeklarowany jest większościowy jego udział, bo na poziomie 60%, ale w opisie tzw. wkładu autorskiego czytamy, że doktorant był odpowiedzialny za implementację i testy, a wraz z drugim autorem za opracowanie metody i publikację. Jeżeli zestawimy ten wkład z opisem towarzyszącym publikacji wcześniejszej, czyli [P2], w której mamy 3. współautorów, doktorant nie jest już autorem pierwszym, jego wkład procentowy jest na poziomie 30%, a opis mówi o odpowiedzialnościach głównie technicznych (przegląd literatury, przygotowanie eksperymentów, przegląd kodu i testy), to

nasuwają się wątpliwości, na ile najnowsza wersja modelu ACERAC jest jego dziełem (biorąc po uwagę niewielkie różnice modelowe, wspomniane już wcześniej).

Drugi podrozdział rozdziału 3. dotyczy propozycji nowej strategii „podtrzymujących akcji" (ang. sustaining actions). W tym przypadku nie ma już żadnych wątpliwości co do poziomu udziału Autora rozprawy w opracowaniu tej propozycji. Strategia to został opisana w dwuautorskiej publikacji [P3], w której doktorant jest pierwszym autorem, jego procentowy wkład jest na poziomie 70%, a z opisu wynika, że był odpowiedzialny za przygotowanie metody, implementację, testy i prezentację na konferencji. Zaproponowana strategia, zaimplementowana w postaci modelu Sustained-actions Actor-Critic with Experience Replay (SusACER) szuka balansu pomiędzy gęstą dyskretyzacją, która w większości przypadków prowadzi do lepszego uczenia, a rzadką dyskretyzacją, która ułatwia naukę. Model osadzony jest znowu ramach wspominanego już wielokrotnie ACERa, ale – podobnie trochę jak to miało miejsce w modelu ACERAC – rozszerza definicję polityki π o prawdopodobieństwo podtrzymania akcji w kolejnym kroku. W pracy można znaleźć ciekawą dyskusję, popartą matematycznymi wyprowadzeniami, w jaki sposób w takim modelu zachowywać oczekiwany poziom eksploracji. Model rozpoczyna uczenie od dłuższych akcji (tj. trwających przez większą liczbę kroków), by następnie redukować stopniowo czas podtrzymywania akcji do kroku pojedynczego. Eksperymentalne wyniki obejmowały porównanie własnej propozycji z modelami ACER, SAC i PPO dla czterech zadań ze środowiska pyBullet oraz wpływ dwóch parametrów opisujących strategię podtrzymywania akcji na działanie modelu. Wykazano przydatność modelu.

Zaproponowany model stanowi oryginalną i inspirującą propozycję dynamicznej dyskretyzacji czasu. Warto by sprawdzić działanie tej modyfikacji w nowszych niż ACER architekturach.

Rozdział 4. rozprawy wprowadza do publikacji [P4], która jest propozycją modyfikacji algorytmu hierarchicznego RL sterowanego (pod)celami. W publikacji zostały przebadane dwa uznane modele z tego obszaru, czyli HAC oraz HitS, które wyposażono w stosunkowo prosty algorytm adaptacyjnego dystansu do wymaganego podcelu (adaptive subgoal required distance, ASRD). Autorzy publikacji pokazali, że dla prawie wszystkich badanych środowisk, zmodyfikowane algorytmy o moduł ASRD uzyskiwały szybciej odpowiednią precyzję sterowania symulowanymi obiektami fizycznymi.

Lektura publikacji skłania do kilku uwag. Po pierwsze, dlaczego autorzy nie porównywali się również do nie-hierarchicznych modeli (płaskich), zwłaszcza, że rozwiązywali znane problemy tzw. rzadkich nagród (ang. sparse rewards)? Idąc za tą myślą, warto by też wykonać

porównanie z metodami stosującymi nowsze metody od „hindsight experience replay" (Andrychowicz et al., 2017). Pewne wątpliwości budzi sposób dowodzenia hipotezy H1, który w ogólności polegał na świadomym odejściu od ustawionych jako optymalne parametrów metod. Zdaje się, że w ten sposób można dowieść dowolnej tezy. Sam tekst artykułu pozostawia też wiele do życzenia. Znajdują się w nim wcale niesporadyczne błędy w referencjach literaturowych, również błędne odnośniki literaturowe i niepełne opisy bibliograficzne.

Zachodzi też pytanie, na ile ta publikacja jest powiązana tematycznie z tytułem cyklu publikacji? Wreszcie, podobnie jak w przypadku publikacji [P1] i [P2], co jest rzeczywistym wkładem Autora rozprawy, jeżeli procentowy wkład został oceniony na 20% (publikacja ma 7 autorów), a w opisie czytamy, że doktorant przeglądał kod, brał udział w przygotowaniu wniosków i samej publikacji.

Rozdział 5 rozprawy zawiera półstronicowe wnioski, w kolejnym rozdziale mam wymienione pozostałe osiągnięcia doktoranta, takie jak prezentacja konferencyjna, zrealizowane projekty (badawcze, przemysłowe?), wzmianka o nagrodzie dydaktycznej i działalności popularno-naukowej. Całość kończy nienumerowana bibliografia odwołująca się do aktualnego piśmiennictwa, w której znaleźć można dwa niepełne opisy bibliograficzne. Rozprawę zamykają dwa dodatki.

Sama kompozycja i struktura rozprawy nie budzą większych zastrzeżeń, tym bardziej, że rozprawa nie jest zawartym dziełem, ale głównie zbiorem powiązanych publikacji poprzedzonych niezbędnym wstępem.

3. Oryginalne osiągnięcia Autora

Do najważniejszych i oryginalnych wyników przedstawionych w recenzowanej rozprawie należy zaliczyć:

- Opracowanie i przebadanie algorytmu SusACER.
- Współudział w opracowaniu i przebadaniu algorytmów ACERAC i ASRD.

4. Uwagi krytyczne i dyskusyjne

Większość uwag krytycznych oraz dyskusyjnych zostało wplecionych w tekst recenzji zawartości rozprawy. Spodziewam się, że Autor odniesie się do nich podczas obrony. Wydaje się, że opublikowanie rozprawy w formie tematycznie powiązanych publikacji mogło nie być najlepszym pomysłem, aczkolwiek z pewnością wymagało najmniejszego nakładu pracy. Dwie publikacje z recenzowanych pięciu są – przynajmniej wg opinii piszącego te słowa –

poza głównym tematem cyklu, aczkolwiek trzeba przyznać, że praca [P4] broni postawionej hipotezy nr 3 (której zakres tematyczny chyba jednak wychodzi poza tytuł dzieła). Pojawiły się też wątpliwości co do wkładu merytorycznego doktoranta w niektóre prezentowane rozwiązania. Zatem być może najbezpieczniejszym rozwiązaniem byłoby jednak przedstawienie do oceny jednoautorskiej monografii.

Wszystkie wymienione wyżej krytyczne uwagi, spostrzeżenia i pytania nie podważają ostatecznie pozytywnej oceny pracy.

5. Wniosek końcowy

Recenzowany cykl opublikowanych i tematycznie powiązanych artykułów naukowych stanowi oryginalne rozwiązanie postawionego problemu naukowego, istotnie rozszerzając wiedzę o modelach uczenia ze wzmocnieniem działających w środowiskach robotycznych. Doktorant wykazał się właściwą wiedzą teoretyczną i praktyczną w dziedzinie nauk technicznych, w dyscyplinie naukowej informatyka techniczna i telekomunikacja, dowiódł też opanowania umiejętności samodzielnego prowadzenia pracy naukowej.

Biorąc pod uwagę powyższe fakty, łącznie z zawartymi w recenzji uwagami stwierdzam, że rozprawa spełnia wymagania Ustawy z dnia 20 lipca 2018 r. – Prawo o szkolnictwie wyższym i nauce (Dz. U. z 2024 r. poz. 1571) **i wnioskuję o dopuszczenie mgra inż. Jakuba Łyskawy do dalszych etapów postępowania o nadanie stopnia doktora.**

Olgierd Unold