dr hab. inż. Paweł Dworak, prof. ZUT
Katedra Automatyki i Robotyki
Wydział Elektryczny
Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

Szczecin, dnia 16 lutego 2024 r.

RECENZJA ROZPRAWY DOKTORSKIEJ NA ZLECENIE PRZEWODNICZĄCEGO RADY DYSCYPLINY AUTOMATYKA, ELEKTRONIKA, ELEKTROTECHNIKA I TECHNOLOGIE KOSMICZNE AKADEMII GÓRNICZO-HUTNICZEJ W KRAKOWIE

Tytuł rozprawy: **Applications of reinforcement learning methodologies to autonomous driving**

Autor rozprawy: **mgr inż. Mateusz Orłowski**

I. Cel, zakres i charakter rozprawy

Recenzowana rozprawa dotyczy zastosowania metod uczenia ze wzmocnieniem do realizacji zadania autonomicznej jazdy samochodu. Autor dokonuje szczegółowej analizy problemu i przedstawia metody syntezy algorytmów sterowania wykorzystujących sztuczne sieci neuronowe zdolne do wykonywania tego typu zadań. Autor rozprawy włączył się w nurt badań prowadzonych przez grupy badawcze z kraju i zagranicy i postanowił z jednej strony dokonać gruntownej analizy problemu – sposobów pozyskiwania danych i modelowania przestrzeni samochodu, z drugiej zaproponował metody syntezy układów tj. budowę sieci i algorytmy ich uczenia. Do realizacji tych zadań wykorzystał zaawansowane narzędzia i środowiska programistyczne oraz twórczo je uzupełnił opracowując samodzielnie nowe struktury danych, sieci i techniki ich uczenia.

Swoje badania Autor prowadził z jasno określonymi celami, zdefiniowanymi jako hipoteza badawcza przedstawiona na str. 2 rozprawy.

Metodologia uczenia się przez wzmacnianie ma zastosowanie do rozwiązywania problemów związanych z podejmowaniem decyzji i planowaniem trajektorii pojazdów autonomicznych.

Jej potwierdzenie realizuje się poprzez realizację celów pośrednich sformułowanych jako hipotezy pomocnicze:

1. Metody uczenia ze wzmocnieniem pozwalają na stworzenie wysokopoziomowego interfejsu

sterowania pojazdem.

2. *Wprowadzenie reguł deterministycznych w czasie szkolenia poprawia czas szkolenia i wynikającą z niego politykę.*
3. *Możliwe jest stworzenie z wykorzystaniem metody uczenia ze wzmocnieniem niskopoziomowego systemu sterowania pojazdem z bezpośrednim planowaniem trasy pojazdu.*
4. *Metody uczenia ze wzmocnieniem pozwalają rozwiązać problem koordynacji jednoczesnego ruchu wielu pojazdów.*
5. *Uzależnienie nagrody pojedynczego agenta od celów innych agentów poprawia ogólną średnią wydajność wszystkich agentów.*

Uważam, że tak wyznaczone cele rozprawy są zadaniem ambitnym od strony teoretycznej i implementacyjnej, ciekawym od strony poznawczej i potrzebnym ze względów praktycznych. Są istotne i aktualne na tle obecnego stanu wiedzy, stanowiąc oryginalne zadanie badawcze.

Wyniki analiz ilustrowane są przykładami symulacyjnymi i eksperymentalnymi bezpośrednio weryfikującymi poprawność syntezy i implementacji opracowanych algorytmów, i możliwość ich pracy w rzeczywistym środowisku, co pozwala bezpośrednio wykorzystać jej rezultaty na tym polu. Mają przez to bardzo duże zastosowanie praktyczne, umożliwiają bowiem konstrukcję efektywnych algorytmów autonomicznego sterowania pojazdami.

Praca dotyczy realizacji problemów w większości do tej pory nie rozwiązanych, których próby realizacji mają dopiero miejsce. Po naszych ulicach nie jeżdżą do tej pory pojazdy autonomiczne, zatem podejmowany w pracy problem jest w oczywisty sposób aktualny. Jego rozwiązanie jest trudne zarówno od strony teoretycznej jak i praktycznej. Wymaga znajomości i umiejętności zastosowania skomplikowanych algorytmów uczenia maszynowego, programowania, dużej wiedzy z zakresu modelowania układów dynamicznych, analizy, syntezy i implementacji algorytmów automatycznego sterowania. Podjęta tematyka wpisuje się bezpośrednio zakres dyscypliny Automatyka, elektronika, elektrotechnika i technologie kosmiczne.

II. Zawartość merytoryczna rozprawy

Rozprawa liczy 131 strony i została podzielona na sześć rozdziałów uzupełnionych spisem literatury. Praca została napisana w języku angielskim.

Pierwszy rozdział, zatytułowany „Introductiton" stanowi wprowadzenie do zasadniczej części rozprawy i przedstawia przegląd literatury z zakresu uczenia ze wzmocnieniem oraz autonomizacji samochodów. W podrozdziałach 2.1, 2.2 przedstawia się ogólną strukturę algorytmu oraz szereg szczegółowych algorytmów i rozwiązań różnych problemów metody uczenia ze wzmocnieniem, tu m.in. sposobów formułowania funkcji kosztu/nagrody. Dalej, w podrozdziale 2.3 Autor przedstawia klasyfikację poziomów autonomiczności systemów

sterowania pojazdami oraz ogólną strukturę architektury ich układów sterowania. Przedstawia się cechy poszczególnych typów sensorów, sposoby lokalizacji i reprezentacji pojazdu i przeszkód w jego otoczeniu oraz sposobów planowania trajektorii jazdy pojazdu. Rozdział ten nie jest może obszerny, ale dość treściwy i potwierdza duża orientację Autora w przedmiocie badań.

Zasadniczą część pracy stanowią rozdziały trzeci, czwarty i piaty.

W rozdziale trzecim przedstawiono sposób realizacji zadania autonomicznej jazdy po autostradzie. Agenta RL uczono do kontrolowania zachowania samochodu z wykorzystaniem wysokopoziomowego interfejsu sterowania, w którym zdefiniowano dopuszczalne manewry i zadaną prędkość ruchu. Celem samochodu, a tym samym agenta w układzie sterowania, było osiągnięcie docelowego miejsca na autostradzie, definiowanego przez nr pasa oraz odległości, w jak najkrótszym czasie. Jako dodatkowe, wydaje się oczywiste, kryterium przyjęto przestrzeganie przepisów ruchu drogowego oraz optymalizację komfortu jazdy.

Opis realizacji tego zadania Autor rozpoczął od sformułowania problemu – wraz z przyjęciem zestawu, w mojej ocenie akceptowalnych, założeń. Wówczas na tle dotychczasowych rozwiązań, przedstawionych krótko w rozdz. 3.3 przedstawiono wykorzystane narzędzia: symulacji jazdy samochodu, środowiska oraz bibliotek uczenia ze wzmocnieniem. Sparametryzowano sposób opisu sterowanego samochodu, drogi, innych uczestników ruchu, dozwolonych manewrów i stanu ich realizacji; opisano sposób generowania trajektorii ruchu i jej realizacji. Rozdział kończy przedstawienie algorytmu optymalizacji, struktury i sposobów uczenia opracowanej sieci neuronowej oraz uzyskanych wyników. Zaprezentowano, w jaki sposób różne strategie wykonania działań agenta wpływają zarówno na funkcjonalność, jak i efektywność treningu.

Bardzo podobnie zorganizowano i przedstawia się treści w rozdziale czwartym, w którym analizuje się możliwość realizacji manewrów autonomicznego parkowania. Konstruowany agent miał za zadanie generację trasy pojazdu, aby możliwe było samodzielne zaparkowanie we wcześniej zdefiniowanym miejscu parkingowym, przy czym wyróżnia się trzy rodzaje parkowania: wzdłuż drogi, poprzeczne i skośne. Tak jak w rozdziale poprzednim Autor na początku przedstawił problem i przyjęte założenia oraz opisał sposób parametryzacji i realizacji ruchu pojazdu. W podrozdziale 4.5 przedstawił sposób opisu otoczenia pojazdu i struktury przyjętych sieci neuronowych. Dalej przedstawiono sposób wyliczania funkcji nagrody, etapowy sposób uczenia sieci oraz analizę uzyskanych wyników. Ważną częścią tej analizy jest wydajność obliczeniowa poszczególnych algorytmów, zarówno w fazie uczenia jak i ostatecznej pracy systemu. Pokazano, że opracowane algorytmy można zaimplementować do pracy w rzeczywistym samochodzie i że mogą one pracować z powodzeniem w reżimie systemu czasu rzeczywistego.

W rozdziale piątym uczenie przez wzmacnianie zastosowano do rozwiązania problemu koordynacji ruchu wielu pojazdów w kilku typowych sytuacjach: mijania się pojazdów przy jednoczesnym zwężeniu jezdni, jazdy na tzw. suwak oraz ruchu na typowym skrzyżowaniu

równorzędnym. Do budowy sieci, sposobu parametryzacji ruchu pojazdów, ich względnego położenia oraz opisu otoczenia wykorzystano doświadczenia i sposoby przedstawione w poprzednich zadaniach (rozdziałach). Ponieważ zadanie polegało na koordynacji pracy wielu agentów, kluczowym elementem było tu opracowanie funkcji współdzielonej nagrody, zapewniającej z jednaj strony osiągnięcie celów wszystkich uczestników ruchu, z drugiej szybkość realizacji tych zadań. Wszystkie pojazdy uczestniczące w ruchu były sterowane według tego samego algorytmu, polityki tworzonej w procesie uczenia, a ich agenci byli w stanie opracować skuteczne strategie sterowania we wszystkich badanych scenariuszach. Rozdział kończy przedstawienie i analiza uzyskanych wyników.

W podsumowaniu Autor przedstawił osiągnięcia pracy oraz uwagi dotyczące nierozwiązanych problemów, wymagających dalszej analizy i poszukiwań.

Spis literatury jest dość obszerny, zawiera 171 pozycji, głównie z ostatnich lat. Są one prawidłowo dobrane i zostały zacytowane w treści pracy. 6 z nich stanowią cytowania własnych prac, czterech współautorskich artykułów naukowych oraz dwóch patentów.

Układ pracy generalnie oceniam jako właściwy, choć osobiście czasem dokonałbym lekkich modyfikacji w treści czy kolejności rozdziałów. Na przykład dla sformułowanego problemu, Autor najpierw przyjmuje założenia, a dopiero po nich znajdują się podrozdziały przeglądowe „Prior Art". Dalsza dyskusja rzadko w bezpośredni sposób odnosi się też do tych założeń. Innym przykładem jest formułowanie i dyskusja nad różnymi postaciami funkcji nagrody w rozdziale 5.6 zatytułowanym „Results". Wiem, że sposób uczenia bezpośrednio wpływa na wyniki działania sieci, aczkolwiek funkcja celu definiuje cel do osiągnięcia, a nie sposób uczenia; taka forma, bez wydzielenia do osobnego rozdziału sprawia wrażenie przypadkowości, a nie solidnej analizy wymagań i potwierdzenia spodziewanych wyników. Rozumiem, że może to jednak wynikać z konsekwentnego poszukiwania kolejnych, lepszych rozwiązań.

III. Ogólna ocena rozprawy

Recenzowana rozprawa bardzo dobrze wpisuje się w wyniki prac innych grup badawczych zajmujących się algorytmami autonomicznej jazdy pojazdów. Autor umiejętnie wykorzystuje nowoczesne narzędzia - znane i opracowywane osobiście - uzyskując nowe, zaawansowane układy sterowania pojazdami, potwierdzając ich konkurencyjność i zasadność wykorzystania. Przedstawia kolejne elementy opracowywanych algorytmów oraz dowodzi poprawności analizy i skuteczności opracowanych systemów. Mają one szansę w bezpośredni sposób być wykorzystane do budowy pojazdów o zaawansowanym poziomie autonomiczności lub wprost autonomicznych

W mojej opinii cele pracy zostały osiągnięte, a tezy pracy udowodnione, metody uczenia ze wzmocnieniem można z powodzeniem wykorzystać do budowy funkcji autonomicznej jazdy samochodem. Do głównych osiągnięć Autora podczas realizacji pracy zaliczam:

- pokazanie sposobu budowy interfejsu wysokiego poziomu, zawierającego m.in. dozwolone zachowania pojazdu, na potrzeby budowy sieci neuronowej sterującej pojazdem i uczenia jej przez wzmacnianie;
- wykazanie, że wykorzystanie tych deterministycznych zasad wraz z opracowanym mechanizmem dostępnych działań i generowanie trajektorii, skutkuje wyższą efektywnością uczenia sieci i pracy systemu;
- pokazanie, że metodę uczenia przez wzmacnianie można z powodzeniem wykorzystać do budowy niskopoziomowego układu sterowania ruchem pojazdu; układu który podczas uczenia nie wykorzystywał zdefiniowanych reguł postępowania i w sposób bezpośredni generował ścieżkę ruchu pojazdu;
- wykazanie, że możliwe jest takie sformułowanie funkcji nagrody, aby uczenie przez wzmocnienie zastosować do budowy wieloagentowego układu sterowania, w którym niezależne od siebie pojazdy podejmują autonomiczne decyzje i skutecznie realizują wyznaczone cele – indywidualne i zespołowe;
- weryfikację opracowanego z wykorzystaniem uczenia ze wzmocnieniem systemu parkowania z systemem sterowania samochodu testowego, co potwierdziło możliwość pracy tych układów sterowania w reżimie czasu rzeczywistego;
- opracowanie sposobów reprezentacji pojazdów i elementów świata zewnętrznego w systemach służących symulacji ruchu pojazdów i na potrzeby syntezy układów sterowania, w tym wykorzystujących metody nauki ze wzmocnieniem.

Ponadto stworzony symulator ruchu drogowego jest bardzo zaawansowanym narzędziem możliwym do stosowania na potrzeby syntezy układów sterowania autonomicznej jazdy, w tym sieci neuronowych z zastosowaniem metod uczenia ze wzmocnieniem. Narzędzie pozwala porównywać różne modele obserwacyjne środowiska oraz ustalać różne architektury i sposoby uczenia sieci, które mogą podlegać dalszym modyfikacjom i dostosowaniom do nowych wymagań.

Na uwagę zasługuje też duże wyczucie Autora w zakresie problemów praktycznej realizacji analizowanych układów sterowania, umiejętność właściwej interpretacji danych i wyciąganie z nich właściwych wniosków, ich znaczenia dla fizycznej pracy układu.

Liczba i waga prezentowanych przez Autora wyników skłania mnie do wyrażenia pozytywnej opinii merytorycznej. Oczywiście jak w przypadku każdej pracy, w której prezentuje się nowe technologie do realizacji równie nowych zadań, wcześniej nie rozwiązanych (lub przynajmniej nie w pełni rozwiązanych) pojawia się mnóstwo pytań natury szczegółowej, skąd i dlaczego takie, a nie inne założenie, dlaczego taka, a nie inna wartość jakiegoś parametru, dlaczego nie przetestowano jeszcze takiego czy innego przypadku?. I mnie takie się nasuwają. Zdaję sobie

jednak sprawę, że nie sposób wszystko przetestować i przedstawić tego wyniki w skończonym czasie. Stąd wielu takich pytań nie zadam. Podczas lektury rozprawy nasunęło mi się jednak kilka pytań i uwag które chciałbym zasygnalizować:

1. Autor w zasadzie nie odnosi się w pracy i nie przedstawia bliżej znanych już rozwiązań analizowanych problemów, np. dobrze znanych i będących powszechnie dostępnych asystentów parkowania. Ciekawiłoby mnie choćby, w jakim stopniu uzyskane wyniki są (mogą) być lepsze od tych osiąganych obecnie innymi metodami. W wynikach symulacji przedstawionych w tablicy 3.6 procent kolizji wynosił od 0,4 do 1,2%, w kolejnych rozwiązaniach był jeszcze wyższy, nie mówiąc już o danych prezentowanych w tabeli 4.7. Jak to się ma do skuteczności innych rozwiązań? Czy takie badania są jakoś ustandaryzowane, aby porównanie algorytmów było rzetelne?
2. Nie jest dla mnie do końca jasne jakie dokładnie elementy (narzędzia, biblioteki) zostały przez Autora wykonane w ramach pracy, a jakie wykorzystane i podlegały jedynie konfiguracji. W jakim stopniu sieci neuronowe, które stanowiły zasadniczy element pracy, zostały zbudowane, a w jakim skonfigurowane? Autor dysponował bowiem symulatorem TrafficAI, który chyba nie bez powodu ma w nazwie AI. Pomijam tu fakt, że Autor miał po prostu dostęp do tego narzędzia i nie miał w tym zakresie specjalnego wyboru, w zasadzie nie podejmował decyzji, z którego narzędzia skorzysta. Widać to wyraźnie w rozdziale 3.4.1, gdzie pomimo deklaracji „simulation tool must be selected", nie przedstawia się innych konkurencyjnych narzędzi.
3. W celu poprawy komfortu podróży Autor wprowadza ograniczenie ujemnego przyspieszenia ustalając jego zakres, np. w tabeli 3.3 jako <-0.054, 0>. Czy na komfort jazdy nie wpływają również duże dodatnie wartości przyspieszenia? Pytam, bo w zasadzie nie zmieniło by to struktury danych, można by po prostu powyższy przedział przyjąć np. jako <-0.054, 0.054>.
4. Niektóre reguły wskazywane jako deterministyczne, w zasadzie takimi nie są. Co dokładnie oznacza bowiem reguła R4 ze strony 47 „do not change the lane when it is unsafe"?
5. Zasady bezpiecznej jazdy były przyjmowane i symulowane w rozdz. 3 dla Ego, ale czy były również stosowane dla pozostałych uczestników? Czy byli oni (raczej nie) lub czy mogliby być niezależnymi uczestnikami ruchu jak w ostatniej części pracy?
6. Coś czego mi w pracy brakuje, to dokładniejsze wyjaśnienie struktury wykorzystywanych sieci. W zasadzie w każdej z nich znajdują się elementy nie opisywane bezpośrednio w treści pracy, bo w treści Autor skupił się na strukturze danych opisujących samochód i jego otoczenie. Autor nie wyjaśnia jednak dlaczego czasem wybiera jedną, a czasem dwie warstwy ukryte, a struktura sieci stanowi niezwykle ważny element pracy. Nie rozumiem potrzeby pętli z MLP w prezentowanej na rys. 4.7 sieci GNN. Co w tej samej sieci (i kolejnych) reprezentuje blok „Value estimation"? Co ten fragment sieci wylicza?
7. W tabeli 5.3 pierwsze trzy wskaźniki, dla każdego z przypadków, nie sumują się do 100, dlaczego?

IV. Uwagi szczegółowe

Rozprawa napisana jest dość staranie, lecz w czasie lektury zauważa się sporo drobnych błędów edycyjnych i językowych. Pozwolę wskazać kilka przykładów:

- dlaczego rysunek 2.1 znajduje się na stronie 11, pięć stron dalej niż odnośnik w treści pracy?;
- value function – na str. 6 jest opisywana jako pochodna funkcji nagrody, podczas gdy już na stronie 7 skumulowaną nagrodą;
- nazwa funkcji „rewards-to-go" pojawia się w pracy jedynie w algorytmie 2 (no i spisie oznaczeń), ale nigdzie indziej w tekście pracy;
- nad znakiem sumy w równaniu 2.3 brakuje n;
- oznaczenia w równaniu 2.1 i poprzedzającym go zdaniu nie są spójne, czym jest A_t?;
- w sąsiadujących zdaniach znajdujemy różnice w pisowni, np.: „action-value" i „action value", „execute the follow-lane" i „execute follow-lane";
- brak wyjaśnienia symbolu v_f w równaniu 3.2?;
- stosowanie zamiennie indeksów górnych i dolnych, np. dla obiektów o w algorytmie 1;
- pomimo deklaracji, że opisywany algorytm PPO-Clip będzie wariantem bez dywergencji Kullbacka-Leiblera, jest ona dalej uwzględniana w równaniu 3.7, a w każdym eksperymencie jako parametr podaje się parametr beta, równy zero;
- w podpisie rysunku 3.12 wskazuje się elementy sieci zaznaczone na niebiesko, które w tej strukturze i na tym rysunku nie występują. Z kolei w kolejnych schematach brakuje tej legendy, a Autor odsyła czytelnika do legendy rysunku 3.12;
- rysunek 4.1 nie ma odnośnika w treści pracy;
- skrót VCS występuje jedynie w tablicach 4.2, 4.3 i 5.1, a nigdzie w treści pracy i spisie oznaczeń;
- na rysunku 4.6 i w tekście pracy używa się słów „point", podczas gdy w strukturze danych z tabeli 4.2 są to „nodes".

Mankamentem pracy jest też styl wypowiedzi wynikający w dużej mierze z faktu, że praca napisana została w języku angielskim i przez informatyka. Poszczególne zdania formalnie budowane są poprawnie, ale są często – zdania i akapity – trudne w odbiorze, na tyle, że po ich przeczytaniu czytelnik ma trudność w zrozumieniu ich sensu. Językowo wynika to m.in. z konstruowaniu zdań wielokrotnie złożonych, czy takiej ich konstrukcji jakby były bezpośrednimi tłumaczeniami z języka polskiego. Zdarzają się równoważniki zdań. Pomimo poprawności konstrukcji „of the ... of the" po zastosowaniu czterech takich zagnieżdżeń, zdanie

pomimo formalnej poprawności czyta się dość trudno. Autor jako informatyk, często wyjaśnia proces specyficznymi pojęciami z zakresu informatyki i stosowanej teorii (metody), np. „The agent policy, based on observation of the ego's relative position to the goal and obstacles, will derive action, which includes movement and turn angle." To jest zdanie z rozdziału 4.2.1, w którym przedstawiany jest problem, a nie jego rozwiązanie – agent to nie to samo co samochód, skręca i parkuje samochód, a nie agent, a zadaniem jest tu po prostu opracowanie strategii parkowania na podstawie danych z obserwacji otoczenia, w tym statycznych przeszkód i wzajemnych prędkości pomiędzy innymi użytkownikami drogi. Zdarzało się, że pewne fragmenty pracy czytałem po trzy, cztery razy aby zrozumieć ich sens. Rozumiem jednak, że uwagi powyższe nie muszą być podzielane przez innych czytelników i wynikają z osobistych przyzwyczajeń i stylu wypowiedzi.

V. Wnioski końcowe

Dyplomant w recenzowanej Rozprawie jednoznacznie dowiódł swoich umiejętności systematycznej analizy problemu technicznego, dużej wiedzy z zakresu teorii sterowania i robotyki, umiejętności analizy, syntezy i weryfikacji wyników opracowywanych układów sterowania. Wykazał wiedzę i umiejętności właściwe dyscyplinie naukowej *Automatyka, elektronika, elektrotechnika i technologie kosmiczne* oraz potwierdził predyspozycje do prowadzenia badań naukowych. Stwierdzam, że przedstawiona mi do recenzji praca doktorska mgra inż. Mateusza Orłowskiego, pt. „Applications of reinforcement learning methodologies to autonomous driving" merytorycznie spełnia wymagania Ustawy z dnia 20 lipca 2018 Prawo o szkolnictwie wyższym i nauce, stawiane rozprawom doktorskim, i wnioskuję o jej dopuszczenie do publicznej obrony.