Prof. dr hab. inż. Jacek Rumiński,
Katedra Inżynierii Biomedycznej
Wydział Elektroniki, Telekomunikacji i Informatyki
Politechnika Gdańska

Gdańsk, 10.12.2024 r.

Recenzja

pracy doktorskiej mgr. inż. Łukasza Lepaka

pt. „Task Automation Using Artificial Intelligence Methods".

Promotor: dr hab. inż. Paweł Wawrzyński

Recenzja została przygotowana w odpowiedzi na pismo Rady Dyscypliny Informatyka Techniczna i Telekomunikacja, Politechniki Warszawskiej (z dnia 25 października 2024 r.) z informacją, że na podstawie uchwały Rady nr 739/2024 (podpisanej przez Przewodniczącego Rady, prof. dr. hab. inż. Jarosława Arabasa) zostałem wyznaczony jako recenzent rozprawy doktorskiej Pana mgr. inż. Łukasza Lepaka.

1. Tematyka, cele i tezy rozprawy

Tematyka pracy dotyczy szeroko rozumianej automatyzacji zadań z wykorzystaniem sztucznej inteligencji. W szczególności związana jest z rozwiązywaniem bardzo różnych problemów w obszarze AI, które mogą być zastosowane: w handlu energią elektryczną, detekcji słów kluczowych w nagraniach audio, symultanicznym tłumaczeniu maszynowym oraz dostrajaniu wartości hiperparametrów w trakcie procesu uczenia.

Częścią rozprawy jest seria pięciu artykułów. Całość jest napisana w języku angielskim, dlatego cytowane fragmenty w niniejszej recenzji będą przytaczane w oryginale.

Doktorant sformułował pięć hipotez:

H1: *„Utilizing information relevant to the day-ahead energy market trading process to automate bid creation increases the trading agent's profits and allows the strategy to be used in real-life scenarios."*

H2: *„Finding keywords in call center recordings in a low-resource language is possible using audio similarity matching models."*

H3: *„The short- and long-term influence of the learning algorithm's hyperparameters on the training process can be used to optimize these hyperparameters on the fly without prior knowledge of the solved problem, thus improving the learning performance."*

H4: *„The translation delay in a simultaneous machine translation can be automatically controlled while maintaining high translation quality for sequences of arbitrary length."*

H5: *„Deep, arbitrary transformation of states in a recurrent neural network allows to achieve results comparable with state-of-the-art methods while mitigating training problems and being memory-efficient."*

Hipotezy są interesujące i wskazują ogólne problemy badawcze. Szkoda, że nie zostały uszczegółowione (m.in. poprzez precyzyjnie określone założenia), ponieważ tak sformułowane mogę być oczywiste w świetle aktualnego stanu wiedzy. Przykładowo postawione w H2 stwierdzenie, że coś jest możliwe, jest hipotetycznie zawsze realizowalne w świetle stanu wiedzy i w przedstawionym kontekście (nie jest uszczegółowiona np. czułość czy precyzja).

2. Zawartość rozprawy

Rozprawa napisana jest w języku angielskim i składa się z dziewięciu rozdziałów, spisu literatury oraz załączników, spośród których kluczowym jest zbiór pięciu publikacji.

W rozdziale pierwszym Autor wprowadza do tematyki rozprawy oraz przedstawia i krótko komentuje hipotezy badawcze.

Rozdział drugi prezentuje wybrane i podstawowe aspekty teoretyczne dotyczące uczenia ze wzmocnieniem i sztucznych sieci neuronowych.

Każdy z rozdziałów 3-7 zawiera krótki opis aspektów badawczych i wyników związanych z powiązanymi pięcioma artykułami, związanymi z tematyką: 1) automatyzacją handlu energią z jednodniową prognozą wyprzedzającą, 2) detekcji polskich słów w zapisach audio, 3) strojenia hiperparametrów w trakcie uczenia, 4) symultanicznego tłumaczenia maszynowego oraz 5) głębokiej transformacji stanów w sieciach rekurencyjnych.

Rozdział ósmy przedstawia podsumowanie rozprawy, a ostatni rozdział dziewiąty zawiera opis innych osiągnięć Doktoranta (wystąpienia konferencyjna, projekty).

Autor cytuje 79 pozycji literatury. Większość cytowanych prac, to recenzowane publikacje w czasopismach lub materiałach konferencyjnych. Wykaz zawiera jedynie 7 pozycji, w tym jeden raport techniczny z ostatnich trzech lat (2022-2024).

Prezentując własne osiągnięcia Autor wskazuje m.in. na publikacje w materiałach konferencyjnych 2023 International Joint Conference on Neural Networks (IJCNN), przedstawiając punktację Ministerstwa jako 140 punktów. Taka punktacja widnieje w wykazie publikacji w Politechnice Warszawskiej. Zgodnie z komunikatem Ministra Edukacji i Nauki z dnia 03 listopada 2023 r. „o zmianie i sprostowaniu komunikatu w sprawie wykazu czasopism naukowych i recenzowanych materiałów z konferencji międzynarodowych", w załączonym wykazie dla materiałów konferencji IEEE International Joint Conference on Neural Networks [IJCNN] przypisano 70 punktów. Ta sama punktacja była w we wcześniejszym wykazie z 2023 roku. Wymaga to wyjaśnienia (może były jeszcze inne korekty, o których nie jestem świadom). Nie jest to istotne dla oceny treści merytorycznych rozprawy, ale stanowi uwagę porządkową.

3. Oryginalne osiągnięcia

W pracy podjęto różnorodne problemy naukowe z zakresu sztucznej inteligencji. Do szczególnych, oryginalnych osiągnięć zaliczam:

- O1: Opracowanie i weryfikacja zautomatyzowanej strategii dla metody uczenia ze wzmocnieniem możliwej (w pewnym zakresie) do wykorzystania w handlu energią na Rynku Dnia Następnego (z jednodniową prognozą wyprzedzającą), w której informacje dostępne dla agenta są przetwarzane w krzywe popytu i podaży z wykorzystaniem zbiorów ofert o zmiennym rozmiarze, umożliwiając w ten sposób agentowi racjonalne zachowanie.
- O2: Wykazanie, że detekcja słów kluczowych w języku polskim w niskiej jakości zapisach audio jest częściowo możliwa z zastosowaniem reprezentacji generowanych na bazie modeli przetrenowanych na zbiorze w języku angielskim, szczególnie w przypadku zastosowania zaproponowanej metody post-processingu. Warte podkreślenia są bardzo ciekawe eksperymenty oraz udział Doktoranta w przygotowaniu zbioru danych.
- O3: Udział w opracowaniu algorytmu „Autonomous Stochastic Descent with Momentum v2 (ASDM2)" w wersjach bazujących na z normalizacją (ASDM2/n, na bazie metody ADAM) i bez normalizacji gradientu (ASDM2/b, na bazie techniki momentum), a także udział w stosownej weryfikacji algorytmów.

Ponadto pozytywnie oceniam inne formy osiągnieć związanych z rozwiązywanymi problemami badawczymi:
- I1: Rozwój i testowanie agenta RLST (Reinforcement Learning for on-line Sequence Transformation), zaprojektowanego do transformacji sekwencji o dowolnej długości. W szczególności uzyskanie lepszych wyników dla proponowanego rozwiązania względem metod referencyjnych dla dłuższych sekwencji (> 22 słów, Tatoeba), co opisano w publikacji P4.
- I2: Eksperymentalna weryfikacja modułu DMU (Deep Memory Update), poprzez opracowanie modeli referencyjnych i przeprowadzenie testów opisanych w publikacji P5.

Wyżej wymienione osiągnięcia (I1 i I2) nie zostały przeze mnie włączone do grupy szczególnych i oryginalnych osiągnieć doktoranta, ponieważ nie znalazłem wystarczających dowodów w przedstawionej dokumentacji w zakresie oryginalnego wkładu Doktoranta w autorstwo proponowanych rozwiązań. W przedstawionej dokumentacji widnieje jedynie:

P4: „The PhD candidate assisted in developing, testing, and analyzing the proposed system, reviewed the literature, prepared the manuscript, and presented the method at the conference."
P5: „The PhD candidate developed models to which the proposed solution was compared, and conducted part of the experiments."

Powyższe osiągniecia mieszczą się w obszarze dyscypliny informatyka techniczna i telekomunikacja i wskazują umiejętność samodzielnego prowadzenia pracy naukowej przez Doktoranta. Ponadto, opisane osiągniecia na tle stanu wiedzy demonstrują, że rozprawa doktorska stanowi oryginalne rozwiązanie problemu naukowego.

4. Ocena szczegółowa i pytania

Przedstawiona do oceny rozprawa doktorska bazuje na pięciu publikacjach. Są one związane z dyscypliną informatyka techniczna i telekomunikacja i dotyczą aspektów sztucznej inteligencji. Niemniej są bardzo różnorodne tematycznie, o niejednorodnym wkładzie Doktoranta, wpływając na spójność rozprawy. Każdy z przedstawionych problemów badawczych jest ciekawy, aczkolwiek różnorodny pod względem stosowanych metod i zastosowań.

Przedstawiona dokumentacja w rozprawie doktorskiej pozwala mi przede wszystkim skupić się na hipotezach H1, H2 i H3 jako tych, których podjęcie bezpośrednio związane jest z wykazaniem umiejętności Doktoranta w zakresie samodzielnego prowadzenia pracy naukowej.

Praca badawcza związana z H1 jest bardzo szeroko udokumentowana zarówno w artykule jak i obszernym załączniku. Doktorant jest głównym autorem pracy, a opis wkładu merytorycznego nie pozostawia wątpliwości o jego zasadniczym udziale w prowadzenia tej pracy naukowej. Autor ciekawie rozważył funkcję opisującą działanie agenta (wzór (10), publikacja P1), uwzględniającą reprezentacje sieci neuronowej (g^1 i g^2) dla określonych stanów analizując zmienne kontrolowane i niekontrolowane. W eksperymentach wykorzystał dostępne repozytorium Stable Baselines3. Zaproponowaną strategię zbadał w odniesieniu do innych, potencjalnie prostszych: strategii arbitralnej, bazującej wiedzy dotyczącej historycznie najlepszych godzin w zakresie zakupu i sprzedaży energii oraz strategii godzinowej, w ramach której w każdej godzinie składane są oferty zakupu i sprzedaży.
Kluczowym wkładem autora jest opracowana strategia i zaprojektowanie oraz przeprowadzenie stosownych eksperymentów. Mam jednak klika uwag i pytań związanych zarówno ze strategią jak i eksperymentami.
Po pierwsze, Doktorant porównał wyniki zaproponowanej przez siebie strategii głównie do dwóch innych strategii, również zdefiniowanych przez Autora rozprawy. Szkoda, że nie odniósł się do innych metod proponowanych z literatury. Szkoda również, że Doktorant nie przedstawił wyników weryfikacji strategii najpierw na takich danych, na których można byłoby porównać znaczenie opracowanej metody z innymi, znanymi ze stanu wiedzy (a dopiero później pokazać dla danych krajowych, jako przypadku szczególnego).
Sformułowana strategia bazuje na prostej polityce zakładającej, że działanie w chwili t jest określane bezpośrednio przez wyjście sieci neuronowej (wzór 10) dla określonego stanu s. W związku z tym mam pytania:

P1. Jak dobór parametrów sieci neuronowej wpływał na uzyskiwane wyniki (np., dlaczego 300 neuronów)?
P2. Jaki wpływ na uzyskiwane wyniki miało dodanie szumu? Dlaczego jest to szum gaussowski?

Zakładam, że propozycja określonej strategii ma mieć charakter uniwersalny (przy określonych założeniach) i sprawdzać się nie tylko na danych historycznych (wówczas jest to jedynie wartość poznawcza), ale również w przyszłości. Dlatego, mam kolejne pytanie:

P3. Czy doktorant sprawdzał, jakie wyniki uzyska z zastosowaniem zaproponowanego podejścia dla nowszych danych, tj. po 2019 roku? Warto zauważyć, że od 2020 roku średnio, wartość maksymalna ceny energii w Polsce była notowana w okolicach godziny 19.00, nie 10.00.

Ponadto, nie jest dla mnie jasne, czy Doktorant rozważał wpływ innych agentów, biorących udział w aukcjach. Wydaje się to bardzo ważne.

Wyniki umieszczone w tabeli 7 załącznika do publikacji P1 wskazują, że kluczowy jest dobór algorytmów RL na uzyskiwane wyniki. Stąd kolejne pytanie:

P4: Czy Doktorant przeprowadzał eksperymenty, wskazujące, czy ważniejszy jest dobór strategii czy algorytmu RL?

Doktorant przyjmuje na wejściu 117 obserwacji, z czego 72 wartości związane są ze stanem pogody. Stąd pojawia się pytanie:

P5. Czy Doktorant zbadał wpływ parametrów pogody na uzyskiwane wyniki?

Ponadto mam następujące pytanie:

P6: Czy inne zmienne ekstremalne (np. wojna, katastrofy) mogą być uwzględnione w niepewności określonej przez zaproponowaną strategię lub co trzeba byłoby zrobić, aby to uwzględnić?

Hipoteza H1 jest zdefiniowana bardzo ogólnie i dla takiej jej postaci Doktorant dostarczył argumentów za jej potwierdzeniem.

Kolejna praca, związana z hipotezą H2, skupia się na wykorzystaniu dwóch modeli sieci neuronowych do uczenia reprezentacji danych w celu detekcji słów kluczowych w niskiej jakości zapisach audio wraz z ich weryfikacją na zbiorach danych w języku angielskim i polskim. Wytrenowany model dla zbioru języka angielskiego osiąga gorsze wyniki niż większość modeli znanych z literatury (tabela 3, publikacja P2). Autorzy w artykule stwierdzili, że „ (...) *we deemed the differences not significant enough to warrant the trade-off these models introduce in terms of complexity and training time.*". Nie mogę się jednak zgodzić z tym wnioskiem, przy braku odpowiednio szczegółowych założeń w zakresie pracy badawczej związanej z P2 i hipotezą H2. Niemniej model ten jest wykorzystywany jako baza w szeregu kolejnych eksperymentów i dlatego mogę przyjąć, że sam w sobie stanowi założenie kolejnych eksperymentów.
W dalszych eksperymentach Doktorant przeprowadza ocenę detekcji słów kluczowych dla języka angielskiego i polskiego. Są to bardzo interesujące analizy. Szczególnie ciekawe są analizy przekrojowe, tj. w sytuacji, gdy model był trenowany na mieszanych zbiorach danych. W przypadku eksperymentów dla pojedynczego języka wyraźnie widać różnicę jakości detekcji słów kluczowych pomiędzy językiem polskim, a angielskim. Dla słów kluczowych w języku polskim, z wykorzystaniem syntetycznego zbioru ($KW_{PL,mix}$) wartość F-score jest ponad czterokrotnie niższa niż dla eksperymentów ze zbiorem z języka angielskiego (F-score= 0,22

vs. 0,94 w tabeli 4, P2). Dla rzeczywistego zbioru ewaluacyjnego ($KW_{PL,real}$) najlepsza wartość F-score=0,02. To oczywiście nieakceptowalne wyniki. W świetle oceny pracy Doktoranta pojawiają się następujące pytania:

P7. Proszę o wyjaśnienie, dlaczego w opisach eksperymentów i tabelach widnieje notacja wskazująca na 22 słowa kluczowe (np. tabela 5: „3: Evaluation is performed on Polish mixtures (KWPL,mix) and real recordings (KWPL,real) with 22 keywords."), podczas gdy w tekście czytamy: „The dataset contains ten examples for each of the 22 target keywords, although only 19 keywords are actually present in the evaluation recordings."?

P8. Jaką miarę F-score zastosowano w pracy?

Klasyczna definicja to F-score = 2* (precision * recall)/(precision + recall), podczas gdy w pracy P2 jawnie wskazano, że „The F-score is defined as F= precision + recall/2 , (...)"? Taka niestandardowa definicja powoduje, że niska wartość F-score wskazuje na jeszcze mniejszą jakość, niż w przypadku klasycznej definicji. Uważam, że taki dobór miary F-score nie jest właściwy. Co więcej, w tabeli 7 (P2) Autor wskazał wartości „micro" dla precyzji i czułości jako np. 0,4% i 18,5% i powiązana wartość F-score: 0,01. Nie wiem jak ta wartość (i inne z tabeli 7) zostały wyliczone. Klasyczna interpretacja trybu „micro" oznacza uzyskanie wartości globalnych precyzji i czułości. Zatem F-score liczone według wzoru z P2 powinno wynosić: (0,004+0,185)/2=0,189/2=0,0945; natomiast liczone według klasycznej definicji F-score= 2*(0,004*0,185)/(0,004+0,185)=0,00148/0,189=0,00783. Proszę o wyjaśnienie jak wyznaczono wartość F-score.

Przeprowadzone eksperymenty w trybie mieszanym wskazują, że np. model (Siamese) wytrenowany na zbiorze w języku angielskim daje lepsze wyniki ewaluacji na zbiorze w języku polskim, niż ten samo model wytrenowany albo na zbiorze z językiem polskim, albo na połączonych zbiorach w obu językach. Jest to bardzo zastanawiające. We wszystkich przypadkach wskazywane wartości F-score są bardzo niskie. Najlepsze wyniki uzyskano, gdy zastosowano przetrenowany model Google (pozycja [6] literatury w P2) do generacji reprezentacji danych, wykorzystanych w detekcji polskich słów kluczowych ($KW_{PL,mix}$). Uzyskana wartość F-score to 0,69. Stąd pojawia się kolejne pytanie do Doktoranta:

P9. Czy wyniki uzyskane dla przetrenowanego modelu Google (tabela 6, praca P2), wskazują, że wcześniej rozważane modele są bezużyteczne? Czy nie należało najpierw przeprowadzić eksperymentu z dostępnymi modelami ze stanu wiedzy, a dopiero potem podejmować decyzję o ewentualnym projektowaniu/wyborze modelu bazowego do generacji reprezentacji?

Podsumowując, wyniki przeprowadzonych eksperymentów pokazane w publikacji P2, pozwalają w pewnym sensie potwierdzić hipotezę H2: *„Finding keywords in call center recordings in a low-resource language is possible using audio similarity matching models"*. Jest to jednak możliwe tylko dlatego, że hipoteza została zdefiniowana zbyt nieprecyzyjnie i jest właściwie, w zakresie stanu wiedzy, oczywista.

Trzecia hipoteza podjęta została w publikacji P3. Doktorant jest trzecim autorem tej pracy, ale w rozprawie wskazał swój udział jako: „developed, tested, and analyzed different versions of

the proposed algorithm". Na tej podstawie zakładam, że również poprzez adresowanie H3 w P3 Doktorant zdobywał kompetencje do formułowania problemów badawczych, stawiania hipotez, dobierania i stosowania właściwej metodologii a także analizowania wyników prowadzonych badań. Zaproponowany algorytm (i jego wersje) są rozwinięciem pracy promotora z 2017 roku (na co jawnie wskazano w rozprawie), w którym optymalizowane są wartości wag algorytmów bazowych typu ADAM i SGD z techniką momentum. Wprowadzone modyfikacje są interesujące i wykazano eksperymentalnie ich przydatność, co oceniam pozytywnie. Mam jednak dwa pytania:

P10: Czy początkowa wartość kroku β (krok 3 algorytmu) może być rzeczywiście dowolna i jak wpływa na przebieg optymalizacji?
P11: W algorytmie ADAM, momentum definiowane jest jako: $m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$, lub po uwzględnieniu współczynnika szybkości uczenia (kroku): $\alpha \cdot m_t = (\alpha\beta_1) m_{t-1} + (\alpha(1-\beta_1)) g_t$. Biorąc pod uwagę notację w równaniu 2 w publikacji P3 oraz przebieg proponowanego algorytmu, czy w przeprowadzonych eksperymentach obserwowana była relacja pomiędzy λ_t i β_t?

Powyższe uwagi i pytania wskazują pewne problemy dotyczące wybranych aspektów rozprawy. Będę wdzięczny za ustosunkowanie się do nich w przypadku publicznej obrony.

5. Konkluzja recenzji

Podsumowując, stwierdzam, że przedstawione w rozprawie hipotezy zostały zweryfikowane w stopniu, aby stwierdzić, że powiązane problemy badawcze zostały rozwiązane, przynajmniej na podstawowym poziomie. Metody badawcze zostały w większości właściwie dobrane i zastosowane, a eksperymenty interesująco przeprowadzone.

Biorąc pod uwagę przedstawione wyżej analizy i uzasadnienia stwierdzam, że rozprawa doktorska prezentuje ogólną wiedzę teoretyczną Doktoranta w dyscyplinie informatyka techniczna i telekomunikacja. Doktorant wykazuje umiejętność samodzielnego prowadzenia pracy naukowej a rozprawa doktorska zawiera oryginalne rozwiązanie problemu naukowego.

Dlatego stwierdzam, że rozprawa mgr. inż. Łukasza Lepaka spełnia wymagania sformułowane w Ustawie z dnia 20 lipca 2018 roku – Prawo o szkolnictwie wyższym i nauce, Dz. U. 2018, poz. 1668, z późniejszymi zmianami. Wnioskuję do Rady Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej o dopuszczenie mgr. inż. Łukasza Lepaka do dalszych etapów postępowania w sprawie nadania stopnia naukowego doktora.

prof. dr hab. inż. Jacek Rumiński, prof. PG
Katedra Inżynierii Biomedycznej, ETI, PG