dr hab. inż. Mariusz Borawski, Prof. US
Uniwersytet Szczeciński
Instytut Zarządzania
Wydział Ekonomii, Finansów i Zarządzania

Szczecin, dnia 20.06. 2024

RECENZJA
ROZPRAWY DOKTORSKIEJ MGR TOMASZ FALGOWSKIEGO
PT. „ANALIZA METYLACJI SEKWENCJI CPG W OPARCIU O UCZENIE MASZYNOWE I SIECI NEURONOWE"

Recenzowana rozprawa została napisana pod kierunkiem prof. dr hab. n. med. Tadeusza Sulkowskiego dziedzinie nauk medycznych i nauk o zdrowiu w dyscyplinie nauki medyczne. Objętość pracy wynosi 187 stron.

Recenzję napisano w celu oceny czy rozprawa spełnia warunki określone w art. 187 ustawy z dnia 20 lipca 2018 r. *Prawo o szkolnictwie wyższym i nauce* zwanej dalej Ustawą (Dz. U. 2018, poz. 1668). Wymogiem art. 187 jest, aby:

1. Rozprawa doktorska prezentowała ogólną wiedzę teoretyczną kandydata w dyscyplinie albo dyscyplinach;
2. Wykazywała umiejętność samodzielnego prowadzenia pracy naukowej lub artystycznej;
3. Przedmiotem rozprawy doktorskiej było oryginalne rozwiązanie problemu naukowego.

Recenzję napisano pod kątem cytowanych wymogów ustawy. Obejmuje ona następujące punkty:

1. Wybór tematu pracy.
2. Cel rozprawy.
3. Ocena merytoryczna i formalna pracy na tle jej układu.
4. Ocena źródeł wykorzystanych w pracy.
5. Wnioski końcowe w kontekście wymogów art. 13, ust.1 Ustawy.

1. Wybór tematu pracy

Badacze używający sztucznych sieci neuronowych odnieśli w ostatnim okresie wiele spektakularnych sukcesów. Wynikiem tego było na przykład stworzenie takich narzędzi jak ChatGPT udostępniający szerokiemu gronu odbiorców bardziej naturalną formę komunikacji z komputerem oraz znakomicie ułatwiającą sterowanie przetwarzaniem danych. Wskazuje to, że poszukiwanie nowych zastosowań sieci neuronowych jest bardzo aktualnym i obiecującym kierunkiem badań. Przedstawiona mi do recenzji praca przedstawia zastosowanie sztucznych sieci neuronowych do analizy metylacji sekwencji CPG. W zamyśle Autora zastosowanie sztucznych sieci neuronowych ma podnieść skuteczność tej analizy w stosunku do dotychczas stosowanych metod.

W mojej opinii temat pracy został wybrany właściwie.

2. Cel rozprawy

Celem rozprawy jest ocena czy architektura sieci neuronowej oparta o kombinację sieci neuronowej konwolucyjnej i typu transformers, która została przetrenowana z użyciem syntetycznych danych jest porównywalnym lub lepszym od standardowych metod narzędziem do analizy metylacji DNA. Jest to cel adekwatny do tematu pracy i w mojej opinii został sformułowany właściwie.

3. Ocena merytoryczna i formalna pracy na tle jej układu

Przedstawiona do recenzji rozprawa składa się z trzech rozdziałów (wstęp, materiały i metodyka, wyniki), wniosków i dodatku. Łącznie praca liczy 187 stron.

W tym kształcie stanowi ona spójną całość o logicznej strukturze, zgodną z tytułem, zawierającą jednak pewne drobne niedociągnięcia, które wskażę w dalszej części recenzji. Rozprawę można traktować jako pracę z pogranicza medycyny i informatyki. Jest to bardzo wielką jej zaletą. Autor wychodząc z ułomności i niedostatków używanych obecnie metod analizy destylacji DNA zaproponował rozwiązanie dające znacząco lepsze rezultaty. Ponadto stworzył bibliotekę umożliwiającą wykorzystanie jego rozwiązania przez wszystkie zainteresowane osoby bez konieczności zagłębiania się w jego techniczne aspekty.

W pierwszym rozdziale zatytułowanym wstęp Autor przedstawia teoretyczne informacje niezbędne do realizacji pracy. Rozdział można byłoby w mojej opinii podzielić na dwie części: część poświęconą aspektom typowo medycznym i cześć poświęconą aspektom typowo informatycznym. Ze względu na to, że rozdział liczy 56 stron można byłoby

pomyśleć o podzieleniu go na dwa rozdziały: jeden związany z medycyną i drugi związany z informatyką. W części informatycznej rozdziału Autor szeroko omówił problematykę uczenia sieci neuronowych oraz przedstawił sieci konwolucyjne i transformers. O ile omówienie sieci konwolucyjnych można uznać za wystarczające, to sieci transformers można byłoby przedstawić w pracy w moim odczuciu dokładniej. Specyfika sieci została omówiona dość dobrze, ale brakuje między innymi schematu sieci i bardziej szczegółowego omówienia sposobu warstwy kodującej pozycję. Jest to o tyle ważne, że sieć transformers jest częścią rozwiązania zaproponowanego przez Autora rozprawy.

W rozdziale drugim zatytułowanym materiały i metody przedstawiono użyte metody, architekturę sieci neuronowej oraz zaprezentowano uzyskane wyniki. Według mnie opis techniczny wykonanego rozwiązania jest zdecydowanie zbyt uproszczony. Brakuje szerszego opisu sposobu generowania danych symulowanych przeznaczonych do uczenia sieci neuronowej. Sposób przedstawienia architektury sieci neuronowej mógłby być bardziej szczegółowy. Autor powinien również zadbać o więcej szczegółowości w przedstawianiu różnych parametrów związanych z działaniem sieci neuronowej. Te wszystkie mankamenty rekompensuje biblioteka stworzona przez Autora rozprawy, z której można wyczytać wszystkie informacje o budowie i działaniu stworzonej sieci neuronowej. Sprawia to, że badania przeprowadzone przez Autora można bez problemu odtworzyć i powtórzyć, a zainteresowani badacze mogą dalej rozwijać zaproponowane rozwiązanie.

W rozdziale trzecim zatytułowanym wyniki przedstawiono uzyskane przez Autora wyniki. Autor rozprawy wykazał się dużą starannością, aby przedstawić jak najlepiej wyniki swoich badań. Wyniki jednoznacznie wskazują na to, że cel pracy został osiągnięty.

Wnioski poświęcone są analizie wyników. Autor rozprawy na podstawie uzyskanych rezultatów wykazuje, że cel pracy został osiągnięty. Z punktu widzenia informatycznego analiza uzyskanych wyników nie budzi większych zastrzeżeń. Moje zastrzeżenia budzi natomiast rola wniosków w pracy: Czy jest to analiza wyników badań, czy podsumowanie pracy? Materiał powinien być podzielony na dwie części. Cześć poświęcona analizie wyników powinna trafić jako podrozdział do rozdziału zawierającego przedstawienie wyników badań. Na podstawie pozostałego materiału i streszczenia analizy wyników powinno zostać sformułowane podsumowanie pracy.

Praca pod względem edycyjnym wykonana jest w moim odczuciu słabo. Autor powinien więcej uwagi poświęcić na właściwe formatowanie pracy. Kwestie edycyjne zmniejszają czytelność pracy. Czytając ją można się łatwo zgubić – nie wiadomo, który rozdział aktualnie się czyta. Trudno też odnaleźć się w hierarchii rozdziałów, podrozdziałów i

pozostałych części strukturalnych pracy. Z treści pracy nie wynika jednoznacznie wkład Autora. Cel postawiony przez Autora był niezwykle ambitny i został niewątpliwie osiągnięty. Realizacja celu wymagała ogromnego nakładu pracy, co niestety nie znalazło odzwierciedlenia w treści rozprawy. Jednym z głównych filarów współczesnego sukcesu sieci neuronowych są ogromne zbiory treningowe. Przykładowo zbiór treningowy ImagNet zawiera 1281167 obrazów treningowych, 50 000 obrazów używanych w walidacji i 100000 obrazów testowych. Prostszy zbiór treningowy MNIST to „tylko" 60000 obrazów treningowych i 10000 obrazów testowych. W postawionym sobie przez Autora rozprawy zadaniu uzyskanie odpowiednio dużego zbioru danych, aby uzyskać w miarę prosto akceptowalny wynik można uznać za niemożliwe. Doktorant musiał dokładnie zapoznać się z konstrukcją i działaniem różnego rodzaju sieci neuronowych, aby z różnych ich fragmentów złożyć sieć, która mogłaby poradzić sobie z uczeniem się na bardzo małych zbiorach danych. To nadal jest jednak za mało, aby uzyskać sensowny wynik. Mgr T. Falkowski musiał przygotować symulator danych wejściowych, aby wygenerować zbiory treningowe i testowe o odpowiednich wielkościach. Symulator musiał precyzyjnie odzwierciedlać zachowanie się rzeczywistych danych. Jakiekolwiek błędy w działaniu symulacji spowodowałyby problemy z działaniem na danych rzeczywistych. To wszystko sprawia, że zaprojektowanie, przede wszystkim właściwe przetrenowanie sieci neuronowej uważam za spore osiągnięcie doktoranta.

Na uwagę zasługuje również fakt, że Doktorant wykonał samodzielnie bibliotekę implementującą autorskie rozwiązanie. Jest to zgodne z najnowszymi trendami w nauce, w których autorzy prac udostępniają opracowane przez siebie metody w sposób jak najbardziej przyjazny osobom, które chciałyby z nich korzystać. Stworzenie biblioteki zdejmuje z potencjalnych użytkowników metody konieczność zrozumienia jak ona działa. Wystarczy do tego użyć funkcji napisanych przez twórcę metody. Obecnie istnieją renomowane czasopisma naukowe takie jak na przykład SoftwareX, które nie publikują wyników badań, ale tylko opisy bibliotek zaliczanych do kategorii wolnego oprogramowania wraz z linkami do repozytoriów. Jeżeli Autor rozprawy nie myśli o skomercjalizowaniu swojego rozwiązania, to dobrze by było jakby pomyślał o publikacji w jednym z takich czasopism. To pozwoliłoby na dotarcie do jeszcze większej grupy ewentualnych osób zainteresowanych zaproponowanym przez Autora rozwiązaniem. Pozwoliłoby też Doktorantowi na dostosowanie swojego repozytorium do ogólnie przyjętych w informatyce zwyczajów. Standardy narzucane przez tego rodzaju czasopisma znakomicie to ułatwiają. Trzeba docenić mgr. T. Falgowskiego, że nie będąc informatykiem nauczył się programować

i to w stopniu na tyle zaawansowanym, że poradził sobie ze stworzeniem zaawansowanej biblioteki.

punktu widzenia informatycznego za nowatorskie można według mnie uznać następujące osiągnięcia Doktoranta:

1. Znalezienie nowego zastosowania dla sieci neuronowych, tj. analiza metylacji sekwencji CPG;
2. Opracowanie metody symulacji danych na potrzeby treningu sieci neuronowych;
3. Zaproponowanie architektury sieci neuronowej rozwiązującej wskazane przez Autora rozprawy zadanie dotyczące analizy metylacji sekwencji CPG

W pracy można zauważyć szereg drobnych uchybień, przykładowo:

1. W całej pracy brak numeracji rozdziałów, podrozdziałów itp.
2. Rozdziały pracy nie rozpoczynają się od nowej strony, przez co łatwo się w pracy pogubić;
3. Stosowanie podpisów pod wzorami czego się w książkach z dziedziny matematyki się nie stosuje (na przykład strony 30, 32, 33);
4. Zbyt dużo tabel i rysunków znajduje się w głównym tekście pracy, co zmniejsza jej czytelność. Dobrze by było przesunąć część z nich do dodatku, np. ze stron 85-105;
5. Zbyt długie akapity. Przykładowo akapit ze stron 74-76 obejmuje ponad trzy strony;
6. Wiele rysunków jest zbyt małych (na przykład rysunki 40, 41, 42 itp.). Należałoby je powiększyć albo zmienić opisy osi, tak aby czcionka była większa;
7. Str. 23, wzór 2. Brak wyjaśnienia co to jest *i*;
8. Str. 23, wzór 2. Niekonsekwencja w oznaczeniach logarytmu, przez co wzór staje się niejasny. Skoro $\log_2$ to logarytm przy podstawie 2 to jak należy rozumieć zapis: $log2\left(\frac{Beta_i}{1-Betai}\right)$? Jako $log_2\left(\frac{Beta_i}{1-Betai}\right)$, czy jako $log\left[2\left(\frac{Beta_i}{1-Betai}\right)\right]$?
9. Str. 23, wzór 2. Brak wyjaśnienia co to jest *Beta*. Czy *Beta* to to samo co β? Jeżeli tak to jest to zapis nieprawidłowy, nie można używać różnych oznaczeń do zapisu tego samego parametru;
10. Str. 23, wzór 2. Czy $Beta_i$ to *Beta* z indeksem *i* czy to jest inny parametr?
11. Str. 30, wzór 3. Nie powinno się używać, w ramach jednego rozdziału tego samego symbolu (β) do oznaczenia różnych parametrów (wzory 1 i 2);
12. Str. 33, wzór 8. Zmienne w i x oraz $\boldsymbol{w}$ i $\boldsymbol{x}$ oznaczają to samo. W matematyce pogrubienie może zmieniać znaczenie co w tym przypadku jest mylące;

13. Str. 49, rysunek 49. Brak opisu osi;
14. Str. 50, wzór 14. W Polsce separatorem miejsc dziesiętnych jest przecinek, a nie kropka;
15. Str. 51, wzór 15. Zapis wzoru jest niezgodny z regułami sztuki. Kiedy $L\left(Y,\hat{f}(X)\right)$ równa się $\left(Y-\hat{f}(X)\right)^2$, a kiedy $\left[Y-\hat{f}(X)\right]$? Powinny być we wzorze słowa: jeżeli lub dla. W tym jednak przypadku nie ma powodu, aby wzór zapisywać w tej formie. Powinno się stworzyć dwa odrębne wzory;
16. Str. 51, wzór 15. Czy we wzorze na pewno chodzi o błąd absolutny? $\left[Y-\hat{f}(X)\right]$ to nie jest zapis błędu absolutnego. Błąd absolutny zapisuje się w ten sposób: $\left|Y-\hat{f}(X)\right|$;
17. Str. 52, wzór 17. W pracy powinien być jednolity system oznaczeń. Podstawę logarytmu należy zapisywać w jeden sposób. Tymczasem we wzorze 7 jest on zapisany jako e, a we wzorze 17 jako exp;
18. Str. 52, wzór 19. softmax(x)i czyta się jako softmax(x) razy i. We wzorze raczej chodzi o zapis $\text{softmax}_i(x)$;
19. Str. 52, wzór 19. Znaki znad i spod Σ „uciekły". Przez co zapis wygląda, jakby zero było podnoszone do potęgi K-1;
20. Str. 60. Błąd odnośnika. Nie można znaleźć odwołania;
21. Str. 61. Niekonsekwentne stosowanie odstępu między liniami w wypunktowaniach. Na stronie 34 jest większy niż na stronie 61;
22. Str. 65. Cel pracy powinien znajdować się na początku pracy, aby czytając część teoretyczną wiadomo było po co dany materiał jest wprowadzany.
23. Str. 68. Połowa strony jest pusta. Prawdopodobnie wynika to z błędnego działania automatycznego odnośnika;
24. Str. 78, rysunek 31. Rysunek powinien być obrócony o 180 stopni względem obecnej pozycji. Taką formę orientacji stosuje się zwyczajowo względem tego typu rysunków dla wygody czytelnika.

Pragnę podkreślić, że w mojej opinii wymienione uchybienia w żaden sposób nie zmniejszają wartości merytorycznej recenzowanej pracy.

Z formalnego punktu widzenia praca jest w miarę poprawna, napisana dobrym językiem. Autor dowiódł umiejętności konstruowania tekstu naukowego.

4. Ocena źródeł wykorzystanych w pracy

Bibliografia obejmuje 373 źródła. Wśród pozycji literatury przedmiotu ok. 32% stanowią publikacje z ostatnich 5 lat (2019-2024), co świadczy o tym, że Doktorant jest na bieżąco z najnowszymi doniesieniami w zakresie objętym pracą. Wśród cytowanych źródeł znakomita większość (99%) stanowią źródła anglojęzyczne. Jeżeli istnieje taka potrzeba Autor nie ma obiekcji, aby powoływać się na źródła bardzo stare (50 lat i więcej). Jednak takich pozycji jest względnie niewiele. Wykaz źródeł z zakresu sieci neuronowych został przygotowany starannie, adekwatnie do postawionego sobie przez Doktoranta celu.

5. Wnioski końcowe w kontekście wymogów art. 187 Ustawy

Podsumowując stwierdzam, że w recenzowanej rozprawie Doktorant:

- wykazał się posiadaniem ogólnej wiedzy teoretycznej w odniesieniu do tematyki jakiej dotyczy rozprawa,
- podjął oryginalny problem, ważny i aktualny zarówno pod względem naukowym jak i praktycznym, dla którego zaproponował autorskie rozwiązanie,
- dowiódł, że posiadł umiejętność samodzielnego prowadzenia pracy naukowej.

Wyczerpuje to wymagania art. 187, Ustawy, zatem stwierdzam, że recenzowana praca w pełni spełnia wymagania stawiane rozprawom doktorskim i może być podstawą do ubiegania się o nadanie stopnia naukowego doktora w dziedzinie nauk medycznych i nauk o zdrowiu w dyscyplinie nauki medyczne. W związku z powyższym wnioskuję o dopuszczenie mgr. Tomasza Falgowskiego do dalszych etapów postępowania w przewodzie doktorskim.

Mariusz Borowski