dr hab. Norbert Jankowski, prof. UMK

Katedra Informatyki Stosowanej, Uniwersytet Mikołaja Kopernika

ul. Grudziądzka 5, 87–100 Toruń
norbert@umk.pl, tel: (0 56) 6113307

Recenzja

rozprawy doktorskiej

„Rigorous evaluation: Towards trustworthy representations in machine learning" mgra inż. Mateusza Gniewkowskiego

W uczeniu maszynowym i sztucznych sieciach neuronowych osiągnęliśmy bardzo duży postęp, szczególnie w ostatnich 30 latach. Potrafimy dziś budować różne modele, które są szeroko przydatne. Jednak pozostaje kwestia wiarygodności tego, co rzeczywiście jest tworzone. Stworzone modele mogą wydawać się wiarygodne, ale nawet wtedy wcale nie muszą takie być. Brak kontroli nad tą wiarygodnością może okazać się nawet bardzo niebezpieczny. Dzisiaj coraz częściej, nawet w poważnych mediach, cytuje się wyniki modeli LLM niemal jak wyrocznie zaniedbując kwestię wiarygodności Dopóki pytamy o rzeczy mało istotne może to nie nieść niebezpieczeństwa, ale w wypadku złożonych aplikacji AI, np. w medycynie, mogłoby to nieść bardzo poważne konsekwencje.

Jak wiadomo powstały już pewne metody związane z badaniem kiedy dany model może być stosowany — szczególnie mam na myśli metody out of distribution (OOD) a także różne próby stworzenia narzędzi wyjaśniania decyzji sieci neuronowych czy modeli uczenia maszynowego, czyli metody XAI.

Jednak w niniejszej rozprawie autor proponuje pójść dalej. Głównym celem stało się prowadzenie badań nad wykorzystaniem różnych metod uczenia, metod OOD i metod XAI. Celem było nie tylko wyznaczenie czy stworzony model może być użyty jako wiarygodny, czy wytłumaczona może być jego klasyfikacja, ale, aby w oparciu o wspomniane metody stworzyć rozszerzony schemat tworzenia finalnego modelu, w którym uczenie jest przedefiniowane tak, by można było w ciekawy sposób kontrolować wiarygodność finalnego modelu przez douczanie/przeuczanie

WPŁYNĘŁO
15-04-2026

RDN-IT/60/2026

kontrolowane i wspomagane przez pośrednie analizy.

Aby osiągnąć zaplanowany cel autor musiał prowadzić badania bazując na dużej ilości różnych koncepcji uczenia maszynowego i sztucznych sieci neuronowych. Dlatego w rozprawie jest duży rozdział poświęcony zdefiniowaniu uczenia modeli w różnych kontekstach, miarom błędów uczenia, różnym metodom uczenia, które czasem bazują na danych tabelarycznych, czasem tekstowych czy obrazowych. Nie ominięto ważnych i najnowszych algorytmów, jak i szeregu innych z których autor korzysta sprawnie w kolejnych rozdziałach. Rozdział ten należy uznać za zdecydowanie dobrze opracowany, ciekawy i wystarczająco wprowadzający w dalszą część pracy. W rozdziale raz, czy dwa zdarzył się mały błąd we wzorze, gdzie widać dwa minusy obok siebie zamiast jednego (np. str 33). Jednak to wyjątki i poprawność całości oceniam bardzo dobrze.

W rozdziale trzecim mamy zaproponowane bardzo ciekawe badania i propozycje zupełnie nowego projektowania modeli na podstawie specjalnego scenariusza uczeń i kontroli.

Główną propozycją jest, by zastąpić uczenie poprzez: uczenie reprezentacji danych wybranym modelem uczenia, a stworzenie właściwego modelu (na przykład klasyfikatora) bazuje na wyuczonej reprezentacji danych. Po czym następuje, jak napisał autor *audit–measure–improve*, czyli najważniejsza propozycja, aby wcale nie kończyć już uczenia całości po nauczeniu klasyfikatora, lecz badać jakość modelu i w oparciu o tą informację przeprowadzać szczególnie zaplanowane przeuczanie.

Jest to bardzo ciekawa propozycja ponieważ zamiast zostawiać użytkownika z modelem o potencjalnie wątpliwej jakości autor proponuje, aby w sposób subtelnie kontrolowany prowadzić korektę modelu, a nie uczenie od nowa, co, jak wiemy, jest najbardziej standardową procedurą, gdy nie jesteśmy zadowoleni z szacowanego poziomu poprawności, jednak prowadzi po prostu do innych problemów, a nie rozwiązuje je.

Jednak tutaj autor pokazuje, że nawet jeśli jesteśmy zadowoleni z poziomu poprawności uczenia, to wcale nie musi oznaczać, że jest to wiarygodne z powodu możliwości zaistnienia korelacji pozornych, które mogły poprowadzić proces uczenia w złym kierunku. Dlatego właśnie zaproponowane scenariusze analizy i douczenia prowadzonego w bardzo szczególny sposób mogą wprowadzać zmiany we właściwym kierunku.

Badania zostały przeprowadzone na kilku zbiorach danych dotyczących detekcji w analizie tek-

stów http/url.

Na początku pokazane zostało jak pozornie tej samej jakości systemy (mające zbliżoną poprawność) mogą różnić się w wynika tego, że reprezentacji danych uczyły się na zupełnie różne sposoby. Jak już zostało wspomniane autor rozkłada całość uczenia na uczenie reprezentacji/struktury danych z pomocą BoW lub RoBERTa i samą część klasyfikacyjną. Autor pokazał wyraźną różnicę pomiędzy użyciem BoW a RoBERTa, dzięki czemu łatwo się przekonać o pozorności korelacji w przypadku zastosowania tego pierwszego. W efekcie pokazane zostało, jak wyraźna może być różnica w reprezentacji, której nauczyły się modele obu scenariuszy.

Autor zaproponował, aby jako jednego z narzędzi audytu użyć jakości klasteryzacji — aby mierzyć na ile jakość klasteryzacji pokrywa się z planowaną klasyfikacją na bazie uzyskanej reprezentacji danych.

Z kolei jako miar OOD autor używał podejścia bazującego na kNN i mierze Mahalanobisa.

Ale punktem kulminacyjnym propozycji było zaproponowanie bardzo szczególnego sposobu przeuczania, które bazuje na opracowaniu celu uczenia jako kompozycji: uczenia kontrastowego, *anchor distillation* i dodatkowemu trzeciemu członowi (*drift penalization*) ze stratyfikowanym losowaniem. Razem dało to bardzo ciekawy efekt. W wyniku takiego uczenia autorowi udało się utrzymać poprawność uczenia na poziomie mocno zbliżonym do poprzedniego testu. Za to jakość klasteryzacji po takim uczeniu podniosła się znacząco, a nawet bardzo znacząco, co mówi, że udało się uzyskać lepszą reprezentację danych. To z kolei może przełożyć się pozytywnie na wiarygodność tak stworzonego modelu.

Znaczące poprawy uzyskano także w analizie OOD.

Autor zaproponował także inne ciekawe badanie powyższego sposobu uczenia poprzez rzutowanie punktów danych do nisko wymiarowej przestrzeni (przez MDS, t-SNE i UMAP) i analizę jakości klasteryzacji. Wyniki również potwierdzają, że uzyskana reprezentacja jest atrakcyjniejsza i wskaźnik AMI jakości klasteryzacji znacząco się poprawiły niezależnie od wcześniejszego typu rzutowania do nisko wymiarowej przestrzeni (dla wszystkich typów rzutowania).

Doktorant przeprowadzał także udane badania związane z konstrukcją symulowanych ataków (zaproponowano kilka wariantów) na modele sieci tekstowych.

Tutaj również udało się zaproponować nowatorskie metody, które okazały się efektywne w testach. Dzięki tym metodom można lepiej, czyli dokładniej, oszacować rzeczywistą wiarygodność

stworzonych modeli.

Rozdział 4 mógł prezentować nieco szerzej obejmowany materiał.

W dalszej części rozprawy doktorant badał również możliwości konstruowania ataków na modele sieci neuronowych analizujących obrazy.

Tutaj zaproponowano bardzo ciekawą i znaczącą modyfikację funkcji błędu dla przeuczania sieci tak, aby sieć neuronowa lepiej radziła sobie w obliczu puli przypadków, które mogłyby posłużyć jako obszary ataku na modele sieci. Zmodyfikowana funkcja błędu stała się wręcz 5-elementowa i oprócz standardowej funkcji błędów ma mechanizmy *obronne*.

W efekcie uzyskano znaczącą poprawę na obszarach zagrożenia, natomiast utrzymano znacznie zbliżony poziom poprawności ogólnej.

Na stronie 103 część oznaczeń we wzorach nie została opisana. Myślę, że tę część rozdziału można było wręcz znacząco rozbudować, ponieważ jest bardzo ciekawa, a ułatwiłoby to rozumienie wszelkich detali.

Podsumowanie

Jako główny dorobek doktoranta należy uznać rozprawę doktorską i artykuły naukowe z czasopism i konferencji. Autor doktoratu wykazał się bardzo dużą wiedzą z zakresu uczenia maszynowego i sztucznych sieci neuronowych. Operuje on płynnie wieloma metodami, z których potrafi budować znacznie bardziej złożone schematy uczenia i analizować je.

Zaproponowane sposoby tworzenie wiarygodniejszych modeli AI są bardzo ciekawe i nowatorskie.

Dlatego można powiedzieć, że jest to istotny wkład w rozwój informatyki.

Doktorant wykazał się bardzo dobrą wiedzą z informatyki. Umie się nią posługiwać a także widać dużą samodzielność.

Na pochwałę zasługuje jakość merytoryczna pracy, ale i jakość techniczna, dbałość o plan pracy, wzory, rysunki czy tabele.

PODPIS ZAUFANY
NORBERT
JANKOWSKI
15.04.2026 07:20:25 GMT+0200
Dokument podpisany elektronicznie podpisem zaufanym

Przedstawiona bibliografia jest zdecydowanie właściwa i odpowiada prezentowanemu materiałowi. Praca została zaplanowana z dużą dbałością. Jest jasna, posiada liczne wykresy i tabele ułatwiające analizę badawczą, a także charakteryzuje się pożądaną poprawnością naukowo–techniczną.

Pan Mateusz Gniewkowski jest autorem 18 publikacji, w tym 12 z obecnej listy punktowanych czasopism i konferencji.

Jednocześnie **wnoszę o wyróżnienie** pracy mgra Gniewkowskiego za bardzo dobrą pracę doktorską a także za znaczące publikacje.

Kończąc, oceniam pozytywnie dorobek doktoranta.

Uważam, że zaprezentowana rozprawa spełnia warunki dotyczące prac doktorskich i stawiam wniosek o dopuszczenie jej autora do dalszych etapów przewodu doktorskiego.

Toruń, 3.04.2026

Norbert Jankowski

ZA ZGODNOŚĆ
Z ORYGINAŁEM
Główny Specjalista
mgr Anna Paula Sobok
15.04.2026
podpis zweryfikowany (raport)