dr hab. Jarosław Bylina
Instytut Informatyki
Uniwersytet Marii Curie-Skłodowskiej
Pl. M. Curie-Skłodowskiej 5
20-031 Lublin
email: jaroslaw.bylina@umcs.pl

Lublin, 19 lutego 2024

Recenzja rozprawy doktorskiej

Tytuł rozprawy:	**Wyjaśnialność i bezpieczeństwo systemów inteligentnych**
Autor rozprawy:	**mgr inż. Katarzyna Filus**
Promotor rozprawy:	**dr hab. inż. Joanna Domańska, prof. IITiS PAN**
Dziedzina:	**nauki inżynieryjno-techniczne**
Dyscyplina:	**informatyka techniczna i telekomunikacja**

1 Tematyka rozprawy i jej aktualność

Pani magister inżynier Katarzyna Filus zajęła się w swojej pracy problematyką bezpieczeństwa — oraz wyjasnialności — systemów inteligentnych. Nie trzeba nikogo przekonywać, jak ważny jest to temat w tej chwili. Obecnie, systemy inteligentne spotykamy na każdym kroku i zaczynają one towarzyszyć nam na codzień, tak w życiu zawodowym, jak i prywatnym. Jednakże, inną sprawą jest zaufanie do takich systemów — ze względu na ich słabą wyjaśnialność i zagrożenia bezpieczeństwa.

Autorka bada wszechstronnie zagadnienia dotyczące bezpieczeństwa i wyjaśnialności akcentując ich silny wzajemny związek i proponując wiele metod mających na celu poprawę tych aspektów oraz potwierdza ich skuteczność porównując ze znanymi rozwiązaniami.

Tematyka rozprawy jest więc bardzo aktualna. Rozwój systemów inteligentnych i coraz szersze ich stosowanie w każdym niemal aspekcie życia wymusza zadbanie o bezpieczeństwo tego rodzaju rozwiązań — także przez poprawę ich wyjaśnialności.

2 Problem naukowy sformułowany w rozprawie

Głównym celem pracy była — cytując Autorkę — „poprawa bezpieczeństwa oraz wyjaśnialności systemów opartych na algorytmach sztucznej inteligencji".

Poprawę tę Autorka zamierzała osiągnąć poprzez wszechstronną analizę problemów z bezpieczeństwem i wyjaśnialnością systemów inteligentnych, a następnie stworzenie różnych narzędzi informatycznych, które by tę poprawę zapewniały.

Problem podejmowany przez Autorkę jest sformułowany bardzo dobrze (nieco lakonicznie jako cel pracy, ale jasno rozwinięty we wstępie), przejrzyście, a co najważniejsze, bardzo trafnie w świetle bieżących zainteresowań informatyki technicznej i telekomunikacji. Rzeczony problem rozwiązany został poprawnie, przy użyciu właściwych do tego celu metod.

3 Zawartość i charakter rozprawy

Rozprawa składa się z pięciu rozdziałów. Pierwszy z nich jest krótkim wstępem, w którym zawarte jest wprowadzenie do pracy, a także jej cele, wraz z uzasadnieniem. Rozdział 2 poświęcony jest analizie tematu — to jest przeglądowi zagadnień związanych ogólnie z problematyką pracy (a więc bezpieczeństwem i jego powiązaniami z wyjaśnialnością). Kolejne dwa rozdziały (3–4) stanowią główną część pracy i poświęcone są autorskim rozwiązaniom poprawiającym bezpieczeństwo (rozdział 3) oraz wyjaśnialność (rozdział 4) systemów inteligentnych. Wraz z tymi rozwiązaniami Autorka pokazuje wyniki potwierdzające ich użyteczność w odpowiednich sytuacjach, a także zamieszcza wnioski dotyczące poszczególnych metod. Na szczególne wyróżnienie zasługuje staranna konstrukcja tych rozdziałów, z przemyślanymi podrozdziałami, które także są odpowiednio podzielone tak, by czytelnikowi ułatwić czytanie i zrozumienie zamiarów oraz wyników Doktorantki. Ostatni rozdział (5) to podsumowanie całej pracy wraz z planami na przyszłość.

Rozprawa jest więc samodzielnym dziełem, w którym Doktorantka opisuje autorskie rozwiązania zauważonych problemów systemów inteligentnych pod kątem ich bezpieczeństwa oraz wyjaśnialności — powołując się też na swoje publikacje w międzynarodowych czasopismach naukowych.

4 Oryginalny wkład Autorki w dyscyplinę

Najważniejsze osiągnięcia autorskie Doktorantki przedstawione w rozprawie to przede wszystkim różnorodne sposoby poprawy bezpieczeństwa i wyjaśnialności systemów opartych na sztucznej inteligencji (wraz z uprzednią analizą, testami i wnioskami):

- analiza i propozycje poprawy jakości bezpieczeństwa popularnej biblioteki *TensorFlow*;
- wykrywanie podatności oprogramowania przy użyciu metryk statycznej analizy kodu wraz z oceną ich przydatności;
- modyfikacja Losowych Sieci Neuronowych w celu dostosowania ich do wydajnego wykrywania ataków sieciowych;
- system automatycznego zbierania danych i etykietowania na potrzeby testowania algorytmów uczenia głębokiego w wizji komputerowej;
- nowy atak adwersarza na końcowe reprezentacje obrazów wewnątrz sieci konwolucyjnej;
- nowa metryka do oceny stopnia szkodliwości ataków adwersarza pod względem podobieństwa etykiet;
- wizualizacja i wyjaśnianie działania ekstraktorów cech;
- nowe metryki do filtrowania systemu lokalizacji opartego na telefonach komórkowych.

5 Analiza źródeł i zastany stan wiedzy

Bibliografia recenzowanej rozprawy doktorskiej obejmuje 279 pozycji — zacytowanych w odpowiednim kontekście. Źródła te dobrze przedstawiają bieżący stan wiedzy na tematy poruszane w pracy. W ogromnej większości są to źródła nowe (ostatnie pochodzą z 2023 roku), co bardzo pożądane w tak aktualnej tematyce. Autorka cytuje także kilka starszych publikacji (z końca XX wieku), co osadza pracę w szerszym historycznie kontekście dyscypliny, pokazując pewną ciągłość nauki.

Wyczerpującym przedstawieniem stanu wiedzy zastanej jest rozdział 2. w którym znacząca część z pozycji bibliografii całej pracy jest wymieniona i opisana pod kątem tematyki rozprawy. Co ważne, Autorka zawiera w pracy także dodatek (A), w którym prezentuje swoje wszystkie prace.

6 Znaczenie wkładu Autorki

Zgodnie z tytułem rozprawy, Autorka prowadzi swoje badania w dwóch aspektach — *bezpieczeństwa* i *wyjaśnialności* systemów inteligentnych — którym poświęca dwa główne rozdziały swojej pracy (odpowiednio 3 i 4).

W kwestii pierwszej, bezpieczeństwa systemów inteligentnych, znaczenie wkładu Doktorantki jest następujące.

- Autorka bada standardowe podatności w bibliteece *TensorFlow*, opisuje i charakteryzuje ich przyczyny i skutki. Zwraca ona uwagę na to, że wykryte luki są całkowicie podobne do innych w językach C/C++, ale mogą być trudniejsza do wykrycia ze względu na specyfikę biblioteki.

- Autorka dokonuje analizy potencjalnej przydatności metryk oferowanych przez analizatory statyczne kodu do wykrywania podatności związanych z zarządzaniem zasobami i ograniczeniami bufora pamięci w kodzie w językach C/C++. Badania wykazują istotność statystyczną większości korelacji między cechami z narzędzi analizy statycznej. Modele uczenia maszynowego osiągnęły dokładność nawet powyżej 90%, a selekcja cech pozwoliła na stworzenie efektywniejszych modeli przy mniejszej liczbie cech oraz określić najsilniejsze wskaźniki podatności na zagrożenia, wskazując na cechy związane z rozmiarem, postacią i złożonością kodu oraz metrykami dotyczącymi utrzymania kodu. Te cechy mogą być sygnałem złych praktyk programistycznych i konieczności przeprowadzenia testów bezpieczeństwa.

- Doktorantka przedstawia nową metodę inicjalizacji oraz trenowania Losowych Sieci Neuronowych, badając ich skuteczność w detekcji ataków botnetowych. Wykazuje eksperymentalnie, że proponowane zmiany prowadzą do uzyskania lepszych wyników bez negatywnego wpływu na dokładność sieci. Dodatkowo, przeprowadzono badania pokazują potencjał tych metod w poprawie bezpieczeństwa oprogramowania, szczególnie w redukcji fałszywych alarmów.

- Autorka wykazuje skuteczność markerów ArUco do testowania systemów rozpoznawania obiektów w czasie rzeczywistym, co może poprawić jakość i bezpieczeństwo usług opartych na głębokich sieciach neuronowych oraz umożliwia wykorzystanie zebranych danych w celu specjalizacji modeli i poprawy ich dokładności.

- Autorka przedstawia nowy atak adwersarza, który działa niezależnie od klasy i klasyfikatora w ostatniej warstwie konwolucyjnej sieci neuronowej. Wyniki ataku nie są jednoznaczne, ale mogą sugerować większą szkodliwość, co może przydać się w testowaniu sieci.

Natomiast w kwestii wyjaśnialności Autorka wyprowadza następujące wnioski.

- Doktorantka przedstawia własną metodę wizualizacji sieci konwolucyjnych, która dostarcza informacji na temat analizowanej sieci i tego, na czym sieć się skupia —

dzięki czemu można próbować interpretować działanie sieci, w takich aspektach, jak wzorce, stronniczość, ataki adwersarza, inspekcja działania sieci.

- Autorka proponuje wyjaśnialny elastyczny system lokalizacji, który okazuje się być efektywny, skalowalny i względnie prosty do zastosowania.

7 Redakcja rozprawy i prezentacja wyników

Struktura pracy jest uporządkowana i przejrzysta. Podział na poszczególne rozdziały jest logiczny. Wydzielenie w każdym z podrozdziałów rozdziałów 3 i 4 sekcji *Wyniki* oraz *Wnioski* bardzo ułatwia czytanie, zrozumienie i ocenę pracy. Bardzo wyraźnie zaznaczone są dokonania autorskie Doktorantki (wraz z wymienionymi publikacjami, do których odwołania znajdujemy w każdym ze wspomnianych podrozdziałów).

Pracę dobrze się czyta — napisana jest starannie zarówno pod względem językowym, jak i typograficznym (z drobnymi uchybieniami, o których w sekcji poniżej). Wyróżnić należy dobór odpowiedniego narzędzia do składu (LaTeX) — bez niego złożenie pracy wypełnionej wzorami, algorytmami, kodami źródłowymi, rysunkami i tabelami byłoby trudne i nie dałoby zadawalającego wyniku. Bez większych zarzutów należy też odnieść się do wyglądu, opisu i czytelności ilustracji, wykresów oraz tabel (choć i tu znajdą się drobne usterki o których w poniższej sekcji), których duża liczba ułatwia czytanie i zrozumienie rozprawy. Także interpretacja tabel i wykresów zawarta w pracy jest poprawna.

Warto też pozytywnie wyróżnić znajdujące się na końcu spisy (skrótów, rysunków i tabel), które bardzo ułatwiają analizę i nawigację w pracy — a także dodatek ze spisem publikacji Doktorantki.

8 Słabe strony i uwagi krytyczne

Niewiele jest w pracy rzeczy zasługujących na uwagi krytyczne.

- Bardzo pożądana byłaby możliwie ścisła definicja *sztucznej inteligencji* — w rozumieniu Autorki. Niestety, jest tak, że wciąż różne źródła i różne osoby rozumieją to pojęcie różnie, a w takim całościowym dziele jakim jest rozprawa doktorska, czytelnik chciałby pewnie stąpać wśród omawianych tematów.

- Podobny problem widzę w używanym pojęciu *wyjaśnialności* w kontekście, w którym porusza się Autorka. O ile pojęcie *bezpieczeństwa* jest w informatyce okrzepłe i łatwo je przenieść z innych tematów na sztuczną inteligencję, to wyjaśnialność jest rzeczą specyficzną domagającą się ścisłego określenia. Nie można tu jednak zarzucić Autorce braku obszernego omówienia pojęcia — chodzi jedynie o uściślenie definicyjne.

- W pracy mogłoby być więcej odwołań do kodów źródłowych, które Autorka stworzyła na jej potrzeby — brakuje trochę wskazania repozytoriów (poza jednym przypadkiem w podrozdziale 3.4) z kodem źródłowym, których obecność przyczyniłaby

się do zapewnienia powtarzalności doświadczeń i pełniejszego zrozumienia dokonań Doktorantki.

- W podrozdziale 3.3 brakuje mi dokładniejszej analizy, jak zmiany dokonane w Losowych Sieciach Neuronowych wpływają na dokładność i szybkość ich pracy.

- W podrozdziale 3.4 wątpliwości wzbudza sposób 'łatania' w pozyskanych obrazach miejsc po markerze (znaczniku). Brakuje tu moim zdaniem analizy wpływu przedstawionej metody (i być może metod alternatywnych) na dokładność rozpoznawania obiektów.

- W pracy Autorka zawiera ogromną ilość wyników (i to jest bardzo dobra rzecz), ale z tego względu cierpi nieco ich prezentacja — na niektórych wykresach/diagramach Doktorantka stara się zawrzeć maksymalnie dużo informacji i przez to stają się słabo czytelne (np. na s. 59, 79–80, 82, 84–92, 161), niektóre z nich sprawiają wrażenie grafik rastrowych zamiast wektorowych, a w tabelach przydałoby się wyróżnienie (np. przez pogrubienie, czy podkreślenie, jak to jest np. na s. 81) wartości zwracających uwagę (to jest najmniejszych, największych, odbiegających od spodziewanej normy itp.) w morzu cyfr (np. na s. 59, 61).

- Bardzo drobnymi problemami są błędy typograficzne, jak używanie łącznika w miejscu myślnika/pauzy (właściwie wszędzie) lub minusa (s. 115).

- Również bez większego wpływu na zrozumienie pracy, ale nieco gryzące w oczy jest mieszanie polskich i angielskich nazw we wzorach matematycznych (jak np. *Entropia* i *InfoGain* na s. 75).

- Także we wzorach matematycznych można zauważyć (lekko utrudniające czytanie) odbiegnięcie od zwykłego sposobu zapisu standardowych funkcji matematycznych (które powinny być zapisane pismem prostym, nie kursywą, jak *log* na s. 75).

Powyższe uwagi krytyczne w ogóle nie wpływają na ocenę merytorycznej wartości pracy; co więcej, wymienione uchybienia są całkiem zrozumiałe przy tego rodzaju i wielkości pracy, a uwagi są subiektywnym zdaniem recenzenta.

9 Podsumowanie i wniosek końcowy

Po analizie rozprawy Doktorantki, mogę stwierdzić, że jest ona przygotowana rzetelnie i wnosi znaczący wkład w dyscyplinę *informatyka techniczna i telekomunikacja*. Potwierdza ona też zdolność Kandydatki do prowadzenia dalszej pracy naukowej samodzielnie. Świadczy ona o ugruntowanej ogólnej wiedzy Autorki w zakresie nauk inżynieryjno-technicznych i szczegółowej wiedzy odpowiadającej zakresowi badań.

Stwierdzam, że recenzowana rozprawa pt. „Wyjaśnialność i bezpieczeństwo systemów inteligentnych" spełnia warunki określone w odpowiedniej Ustawie, a w związku z tym, wnioskuję o dopuszczenie rozprawy doktorskiej Pani mgr

inż. Katarzyny Filus do publicznej obrony i dalszych etapów postępowania doktorskiego.

Ponadto, z uwagi na wysoką wartość merytoryczną pracy — potwierdzoną publikacjami w recenzowanych czasopismach naukowych mających Impact Factor — oraz jej szeroki zakres wnioskuję o stosowne wyróżnienie recenzowanej rozprawy doktorskiej.

Jarosław Bylina