Katowice, 15.02.2024

Dr hab. Marek Sikora
Katedra Sieci i Systemów Komputerowych
Politechnika Śląska
ul. Akademicka 16
44-100 Gliwice
Email: marek.sikora@polsl.pl

Recenzja rozprawy doktorskiej

Tytuł rozprawy: Wyjaśnialność i bezpieczeństwo systemów inteligentnych

Autor rozprawy: mgr inż. Katarzyna Filus

Promotor rozprawy: dr hab. inż. Joanna Domańska

Dziedzina: nauki inżynieryjno-techniczne

Dyscyplina: informatyka techniczna i telekomunikacja

1. Temat i cel rozprawy

Tematyka rozprawy obejmuje zagadnienia oceny badania bezpieczeństwa i wyjaśniania decyzji systemów komputerowych, których funkcjonalności bazują na metodach, a w aspekcie inżynierskim na bibliotekach, maszynowego uczenia (ozn. ML). Główny cel pracy to poprawa bezpieczeństwa takich systemów oraz – dla pewnej klasy metod (sieci głębokich o określonej architekturze) – zwiększenie czytelności i zrozumiałości podejmowanych decyzji. Pierwszy cel zrealizowano na kilku poziomach, zarówno poprzez analizę podatności bibliotek zawierających implementacje metod ML oraz opracowanie metod testowana i wykrywania ataków na głębokie sieci neuronowe. Drugi cel osiągnięto poprzez zaproponowanie metody interpretacji działania głębokiej sieci konwolucyjnej dedykowanej do klasyfikacji obrazów.

W pracy nie zdefiniowano tezy głównej ani tez pomocniczych, natomiast jasno zdefiniowano cel i zakres badań oraz szereg celów szczegółowych.

Uzasadnienie wyboru tematu nie budzi żadnych wątpliwości, Autorka bardzo dobrze i trafnie uzasadnia celowość podjęcia badań opisanych w rozprawie. Tematyka rozprawy jest bardzo istotna, objaśnialności i bezpieczeństwo systemów zawierających elementy ML jest aktualnym tematem naukowym, zwłaszcza w kontekście nadchodzących regulacji prawnych dotyczących zaufana do tego typu systemów.

2. Zawartość i charakter rozprawy

Rozprawa składa się z pięciu rozdziałów oraz spisu literatury.

We wstępie Autorka w sposób pogłębiony uzasadnia motywacje podjęcia badań, podając nie tylko argumenty naukowe, ale również prawne, powołując się na uregulowania europejskie i pierwsze propozycje norm związanych z certyfikacją systemów zawierających elementy rozwiązań ML. We wstępie zaprezentowano również listę celów szczegółowych pracy, cele te wiążą się bezpośrednio z unikalnymi propozycjami Autorki związanymi z poprawą bezpieczeństwa i objaśnialności działania systemów inteligentnych. Kolejne rozdziały przedstawiają konkretyzację i opisy tych propozycji.

Rozdział drugi jest rozdziałem wprowadzającym czytelnika w tematykę rozprawy, w rozdziale tym znajdujemy wyjaśnienia przyjętej terminologii oraz rozwinięcie definicji wyjaśnialności i bezpieczeństwa. Na tym tle Autorka przywołuje wybrane metody wyjaśnialności dla modeli uczenia głębokiego. W rozdziale drugim podejmowany jest również bardzo ważny aspekt jakim jest intuicyjności – w szczególność jej pomiar – metod wyjaśnialności. Dalsza część rozdziału poświęcona jest aspektom bezpieczeństwa systemów inteligentnych, w szczególność w rozdziale tym rozróżniono dziedzinowe aspekty bezpieczeństwa związane z naturą systemów ML (np. hazardy wynikające z indukcyjnej natury uczenia czy tzw. ataki adwersarza) oraz aspekty klasyczne, jakie dotyczą każdego systemu komputerowego (np. luki bezpieczeństwa, podatności bibliotek, etc.).

Rozdział trzeci traktuje o bezpieczeństwie systemów inteligentnych. W pierwszej kolejności podejmowany jest temat zagrożeń bezpieczeństwa systemów inteligentnych jako systemów komputerowych. Autorka, wykorzystując klasyczną metodykę badania podatności oprogramowania, analizuje popularną bibliotekę uczenia głębokiego TensorFlow. Charakteryzuje sześć różnych typów podatność związanych m.in. z zarządzaniem pamięcią. Wyniki prezentowane są w sposób ilościowy (liczba i typ znalezionych podatności), jak również jakościowy. Autorka analizuje przyczyny znalezionych podatność, w szczególności czy przyczyny te mogą wynikać z natury zastosowanych algorytmów maszynowego uczenia. W rozdziale podjęto również tematykę opracowania klasyfikatora jako narzędzia wspomagania decyzji wykrywającego rzeczywiste (elimanacja fałszywych alarmów) luki bezpieczeństwa przez statyczne analizatory kodu. Klasyfikator trenowany jest na podstawie metryk generowanych przez analizatory, stosowane są różne metody klasyfikacji danych tabelarycznych. Autorka przedstawia metodykę trenowania klasyfikatorów oraz omawia ich efektywność, prezentując oddzielne ich czułość i specyficzność. W dalszej części rozdziału zaproponowano zastosowanie tzw. losowej sieci neuronowej do wykrywania zagrożeń bezpieczeństwa (wykrywanie ataków sieciowych, wykrywanie podatności w kodzie). Przedstawiono również dwie modyfikacje wpływające na przebieg trenowania losowej sieci neuronowej: modyfikację sposobu inicjalizacji wag oraz ograniczenie liczby operacji matematycznych podczas trenowania sieci. Autorka przedstawia matematyczne podstawy proponowanych modyfikacji oraz pokazuje ich globalnie pozytywny wpływ na trafność predykcji podczas wykrywana zagrożeń.

Sekcja 3.4 dotyczy tematyki automatyzacji tworzenia zborów danych dla trenowania i testowania sieci głębokich stosowanych w zadaniach rozpoznawania obrazów i autonomicznej jazdy. Autorka przedstawia metodę użycia markerów identyfikujących obiekty, które powinny być rozpoznane przez sieć. Eksperymenty przeprowadzono w środowisku laboratoryjnym. Traktuję ten aspekt rozprawy jako temat istotny z punktu widzenia bezpieczeństwa, ale bardziej rozumianego jako wykrywanie hazardów działania, a nie bezpieczeństwa cybernetycznego. Ponieważ w rozprawie aspekty związane bezpieczeństwem systemów inteligentnych rozumiane jako jakość ich działania (np. unikanie hazardów) nie są w zasadzie adresowane, traktuję tę sekcją jako element poboczny w stosunku do całej zawartości pracy.

Ostatni z tematów podejmowanych w rozdziale trzecim dotyczy ataków (tzw. ataków adwersarza) na sieci głębokie. Autorka argumentuje, że umiejętność testowania odporności modeli głębokich na ataki adwersarza wpływa pozytywnie na bezpieczeństwo działania sieci – na jej zaufanie, dlatego uzasadnione jest tworzenie bardziej wysublimowanych metodycznie ataków i umiejętność ich wykrywania. W tej części rozdziału Autorka przedstawia własną propozycję tego typu ataku (Network Saturation Attack - NetSat). W metodzie tej atakowane są warstwy pośrednie sieci, a więc warstwy, w których przechowywana jest głęboka reprezentacja danych – reprezentacja wyuczonego modelu. Analiza szkód wyrządzonych przez ataki adwersarza powinna zdaniem Autorki uwzględniać nie tylko możliwość binarnej

identyfikacji takiej szkody (wystąpienie lub brak ataku), ale również umożliwiać mierzenie stopnia, w jakim atak adwersarza wpływa na decyzje sieci. W tym celu Autorka przedstawia propozycję metryki niepodobieństwa, która opisuje jak daleko od etykiety zwracanej przez model niezaatakowany jest etykieta zwracana przez model po ataku. Eksperymenty weryfikacyjne wykonano w oparciu o trzy architektury sieci neuronowych.

Rozdział czwarty poświęcony jest zagadnieniom wyjaśnialności. Autorka proponuje mechanizm wizualizacji działania konwolucyjnych sieci neuronowych, polegający na wyznaczaniu punktowych map aktywacji cech, które można wykorzystać m.in. do oceny zdolności sieci do reprezentowania obrazu za pomocą wzorców – metoda Network Activation Mapping (NAM). Przedstawiona propozycja testowana jest na pięciu modelach sieci konwolucyjnych i zestawiana jest z wynikami wyjaśnialności dostarczanymi przez metodę odniesienia, jaką jest metoda mapowania aktywacji sieci GRAD-CAM. Badane są trzy wersje metody NAM: max, mean i var. Każda z tych metod przekłada się na różną interpretację/realizację zadania objaśnialności. Autorka przedstawia również zastosowanie metody NAM do analizy (ilustracji) wpływu ataku NetSat na mapy cech.

Drugie z zagadnień podejmowanych w rozdziale dotyczy specyficznego tematu wyjaśniania decyzji podejmowanych przez system lokalizacji obiektów (np. ludzi w obiektach zamkniętych). Moim zdaniem tematyka tej części rozdziału jest luźniej związana z główną tematyką pracy i dotyczy bardzo specyficznego problemu.

Ostatni rozdział zawiera podsumowanie wyników oraz odniesienie się do celów rozprawy.

3. Analiza źródeł i zastany stan wiedzy

Bibliografia recenzowanej rozprawy doktorskiej składa się ze 279 pozycji. Autorka cytuje je w odpowiednim kontekście. Źródła te dobrze przedstawiają bieżący stan wiedzy w zakresie zagadnień podejmowanych w pracy. W szczególności, Autorka przedstawia propozycje uregulowań prawnych i norm związanych z certyfikacją systemów inteligentnych. Pewien niedosyt pozostawia jednak brak szerszego przeglądu metod objaśnialności, w szczególność pokazania szerszego kontekstu tego zagadnienia w odniesieniu do złożonych modeli ML stosowanych nie tylko w analizie obrazów, ale również w analizie danych tabelarycznych, szeregów czasowych, etc. Zwłaszcza, że istnieje co najmniej kilka metod objaśniania dedykowanych do wszystkich tych typów danych. Przywoływana literatura jest aktualna – duża część cytowanych prac została wydana została po roku 2019.

Autorka powołuje się również na swoje publikacje dotyczące omawianych zagadnień. Pozwala to recenzentowi zorientować się w dorobku publikacyjnym Autorki. Prace publikowane są w czasopismach naukowych i materiałach konferencyjnych. Dorobek publikacyjny Doktorantki oceniam jako bardzo dobry.

4. Oryginalne wyniki i ich znaczenie

Doktorantka podejmuje ważny problem badania bezpieczeństwa i wyjaśnialności systemów inteligentnych. Problemy te są jednymi z najbardziej aktualnych zagadnień nie tylko w środowisku naukowym, ale również – a może nawet bardziej w zawiązku z lawinowym rozwojem generatywnej AI – w środowiskach prawniczych i decydenckich. Wszystkie z przedstawionych do tej pory propozycji norm i uregulowań prawnych dotyczących certyfikacji systemów zawierających elementy ML zawierają silny postulat konieczności zrozumienia działania systemów ML, zwłaszcza systemów bazujących na metodach głębokiego uczenia. Przedstawiana do recenzji praca doktorska dobrze adresuje te zagadnienia dla pewnej grupy metod, jakimi są głębokie sieci neuronowe oraz biblioteki implementujące algorytmy trenowania sieci.

Za najbardziej wartościowe wyniki uzyskane przez Doktorantkę uważam:

- Badanie podatności biblioteki TensorFlow oraz metodę wykrywania (predykcji) podatności oprogramowania za pośrednictwem metryk generowanych przez statyczne analizatory kodu. Wyniki uzyskane przez Autorkę powinny znaleźć duże zainteresowanie nie tylko w środowisku badaczy, ale również inżynierów korzystających z biblioteki TensorFlow. Ponadto uważam, że pomysł predykcji podatność jest ciekawym rozwiązaniem użytecznym nie tylko dla badania bezpieczeństwa systemów inteligentnych.
- Zastosowanie sieci neuronowej (konkretnie losowej sieci neuronowej wraz z przedstawionymi w pracy modyfikacjami mechanizmów jej trenowania) do wyrywania ataków sieciowych oraz do wykrywania podatności oprogramowania na zagrożenia bezpieczeństwa.
- Zaproponowanie metryki niepodobieństwa umożliwiającej głębszą analizę szkód wyrządzonych przez ataki adwersarza.
- Zaproponowanie metody wizualizacji i wyjaśniania działania głębokich konwolucyjnych ekstraktorów cech. W szczególności, podanie interpretacji wyjaśnialności (dokładniej, jaki aspekt wyjaśnialności jest adresowany) dla trzech konkretyzacji metody NAM (max, mean, var).

5. Redakcja rozprawy i prezentacja wyników

Rozprawa zredagowana jest w sposób dobry. Układ pracy jest nieco nieczytelny, być może lepsza byłaby taka organizacja pracy, w której kwestie bezpieczeństwa systemów inteligentnych i zastosowania metod ML do poprawy bezpieczeństwa systemów inteligentnych zostałyby rozdzielone (osobne rozdziały).

Sposób prezentacji wyników mógłby być zdecydowane lepszy – czytelniejszy. W szczególności, teksty umieszczane na rysunkach są nieczytelne, wartości umieszczane w macierzach pomyłek wymagają dokładniejszego opisu – raz jest w nich zamieszczana po

prostu liczba pomyłek, innym razem są to wartości liczbowe, których znaczenia należy szukać w tekście – brak jednoznacznego podpisu pod niektórymi rysunkami.

Graficzna ilustracja wyników działania klasyfikatorów – histogramy dla wielu wyników – jest również nieczytelna w tym sensie, że trudno w łatwy sposób zidentyfikować różnice pomiędzy klasyfikatorami. Częściowo wynika to z przyjętej metodyki porównań, do czego odniosę się jeszcze w dalszej części recenzji.

Pracę dobrze się czyta – napisana jest starannie pod względem stylistycznym i typograficznym.

6. Słabe strony i uwagi krytyczne/dyskusyjne

Tematyka pracy jest bardzo obszerna, tytuł nieco błędnie sugeruje, że rozważane będą aspekty bezpieczeństwa i wyjaśnialności szerokiego spectrum metod ML działających na różnych typach danych. Wydaje się, że nieznaczne zawężenie tytułu lepiej odzwierciedlałoby zawartość pracy.

Autorka omawia dosyć szerokie spectrum zagadnień bezpieczeństwa i kilka aspektów wyjaśnialności. Zdaniem recenzenta część z tych zagadnień mogłaby zostać pominięta (np. sekcja 3.4 oraz sekcja 4.2). W zamian można by się dokładniej przyjrzeć różnym aspektom bezpieczeństwa systemów inteligentnych, poprzez wydzielenie oddzielnego rozdziału dotyczącego identyfikacji i zapobiegania cyberatakom (rozumianym tutaj bardzo szeroko, również jako ataki na poprawność działania) na modele ML oraz zastosowania metod ML do identyfikacji ataków (również na modele ML).

Niedosyt zarówno w sensie odniesień do literatury przedmiotu, jak również w zakresie wykonanych badań, pozostawia dosyć wąskie potraktowanie tematyki objaśnialności, sprawdzające się de facto do zaproponowania jednej metody dedykowanej dla określonej grupy metod oraz określonego typu danych. Akurat tę cześć pracy oceniam bardzo wysoko, jednak nie da się nie zauważyć, że zdecydowana większość zagadnień rozprawy związana jest z tematyką bezpieczeństwa.

Pewne zastrzeżenia budzi metodyka prowadzonych eksperymentów w zakresie oceny eksperymentalnej propozycji przedstawianych przez Autorkę. Po pierwsze stosowana jest metodyka train and test, która nie daje pespektywy dotyczącej stabilności działania metod. Po drugie Doktorantka rzadko – badając efektywność swoich rozwiązań – sięga po szersze spektrum metod odniesienia, a więc metod stosowanych standardowo w danej grupie zagadnień.

W przypadku niektórych zagadnień brak również głębszej analizy teoretycznej różnic – wad i zalet – pomiędzy rozwiązaniami proponowanymi przez Autorkę a rozwiązaniami istniejącymi. Uwaga ta dotyczy w szczególności metryk dedykowanych do pomiaru szkodliwości ataku na sieć głęboką (str. 124).

Nie zawsze w jasny i zrozumiały sposób opisano sposób przygotowania i charakterystykę danych treningowych, np. w sekcji 4.2 nie do końca jest dla mnie jasne, jaką informację przechowuje zmienna decyzyjna.

Sposób prezentowania i podsumowania wyników badań eksperymentalnych mógłby być zdecydowanie lepszy – histogramy złożone z wielu serii danych są dosyć trudne w interpretacji.

Pytania szczegółowe:

1. Czy zaproponowane modyfikacje dla losowej sieci neuronowej znajdują zastosowanie jedynie w obszarze zainteresowania rozprawy (bezpieczeństwie), czy też propozycje te mają, zdaniem Autorki, pozytywny wpływ na trenowanie sieci w innego rodzaju zadaniach? Dlaczego nie zweryfikowano przedstawionych propozycji na większej liczbie danych benchmarkowych?
2. Dlaczego wybrano metodę train and test jako metodę eksperymentalnej oceny jakości klasyfikatorów występujących w rozprawie?
3. Dlaczego badania weryfikacyjne wykonywano na tak małej liczbie zbiorów?
4. Czy metodę NAM można zastosować do modeli działających na innych typach danych (np. szeregach czasowych)?
5. Czy opracowane metody – ich implementacje – są ogólnodostępne?
6. Czy zbiór danych będący wynikiem prac opisanych w rozdziale 4.2 jest lub będzie dostępny dla szerszego grona badaczy?

7. Podsumowanie i wniosek końcowy

Po analizie rozprawy mogę stwierdzić, że zamieszczone w niej rezultaty badań uzyskano w sposób rzetelny, a wyniki stanowią nowy wkład w dyscyplinę informatyka techniczna i telekomunikacja – w szczególności wnoszą wkład do metodyki oceny bezpieczeństwa systemów inteligentnych. Rozprawa potwierdza zdolność Doktorantki do dalszej pracy naukowej. Uwagi krytyczne i dyskusyjne nie umniejszają mojej jednoznacznie pozytywnej oceny rozprawy, a ich celem jest m.in. chęć podjęcia dyskusji podczas obrony pracy.

Stwierdzam, że recenzowana rozprawa pt. „Wyjaśnialność i bezpieczeństwo systemów inteligentnych" przygotowana przez mgr. inż. Katarzynę Filus spełnia wymagania i warunki określone w ustawie z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (jednolity tekst Dz. U. z 2023 r. z późn. zm.) i wnoszę o przyjęcie ww. rozprawy doktorskiej, dopuszczenie jej do publicznej obrony i dalszych etapów postępowania doktorskiego.

Marcin Szlenk