Streszczenie rozprawy doktorskiej „On Generalization and Robustness of Audio DeepFake Detection"

Piotr Kawa

W niniejszej rozprawie przedstawiono rozwiązania mające na celu poprawę metod detekcji wygenerowanej komputerowo mowy tzw. *dźwiękowych manipulacji DeepFake*. Zaproponowane usprawnienia dotyczą problemów dziedzinowych: poprawy generalizacji detektorów, zwiększenia wydajności algorytmów, poprawy odporności detektorów na ataki oraz eksploracji metod reprezentacji danych w oparciu o głębokie sieci neuronowe.

Rozprawa doktorska oparta jest na następujących publikacjach:

1. Kawa, P., Plata, M., Syga, P. *Attack Agnostic Dataset: Towards Generalization and Stabilization of Audio DeepFake Detection*, Interspeech 2022,
2. Kawa, P., Plata, M., Syga, P. *SpecRNet: Towards Faster and More Accessible Audio DeepFake Detection*, TrustCom 2022,
3. Kawa, P., Plata, M., Syga, P. *Defense Against Adversarial Attacks on Audio DeepFake Detection*, Interspeech 2023,
4. Kawa, P., Plata, M., Czuba, M., Szymański, P., Syga, P. *Improved DeepFake Detection Using Whisper Features*, Interspeech 2023.

Generalizacja detektorów DeepFake jest powszechnym i wciąż nierozwiązanym problem niższej skuteczności modeli na danych testowych, których rozkład różni się od danych treningowych. Jednym z powodów jest nadmierne dopasowanie (ang. overfitting) — modele zamiast nauczyć się ogólnych wzorców charakteryzujących próbki DeepFake, uczą się artefaktów poszczególnych generatorów. Ponadto, aktualnie wykorzystywane zbiory treningowe, często nie zawierają próbek stworzonych z użyciem najnowszych generatorów. W celu poprawy generalizacji i stabilności metod detekcji, określanych przez efektywność na danych spoza zbioru treningowego, zaproponowano podejście *Attack Agnostic Dataset*. Metoda ta polega na rozłącznym podziale ataków pomiędzy zbiory treningowe, walidacyjne i testowe. Wyniki omawiane w rozprawie oparte są na trzech zbiorach: WaveFake, FakeAVCeleb oraz ASVspoof 2019 LA. Metody podziału ataków pomiędzy podzbiory mają na celu symulowanie różnych scenariuszy np. trening na zbiorze opartym tylko na podobnych metodach generowania DeepFake. Analiza wyników poszczególnych podziałów pozwala wybrać modele o najlepszej generalizacji na danych spoza zbioru treningowego. Metoda *Attack Agnostic Dataset* może być w łatwy sposób rozszerzona o kolejne zbiory i podziały. W ramach prac zaproponowano nową reprezentację danych opartą na połączeniu spektrogramu melowego oraz parametrów liniowo-cepstralnych w wyniku czego uzyskano wyniki EER o 5% niższe niż w przypadku pojedynczych reprezentacji.

Globalna natura problemu manipulacji DeepFake, której wynikiem jest duża ilość wygenerowanej treści wymaga istnienia nisko-kosztowych metod detekcji zapewniających wysoki poziom efektywności. Pozwala to zarówno na użycie wyżej wymienionych metod w celu analizy treści umieszczanej np. na portalach społecznościowych, jak również, zwiększa dostępność tych rozwiązań pozwalając zwykłym obywatelom na samodzielną weryfikację treści. W tym celu stworzono sieć neuronową *SpecRNet*. Jest to rekurencyjny model przetwarzający dwuwymiarowe reprezentacje sygnału dźwiękowego. Architektura rozwiązania zainspirowana została popularnym modelem anti-spoofingowym RawNet2 przetwarzającym sygnał dźwiękowy przy użyciu 60 razy większej liczby parametrów. Architektura *SpecRNet* pozwala na 40% obniżenie czasu przetwarzania próbek w porównaniu do modelu LCNN uważanego ówcześnie za jeden z najszybszych i najskuteczniejszych metod detekcji. *SpecRNet* odznacza się porównywalnymi wynikami, co zostało dodatkowo

potwierdzone na podstawie zaproponowanych przez nas testów dotyczących m.in. scenariuszy małej ilości danych treningowych czy krótkich próbek dźwiękowych.

W celu zmniejszenia szansy wykrycia wygenerowanej próbki, adwersarz jest w stanie wykorzystać różne rodzaje modyfikacji by ukryć artefakty charakteryzujące sztuczne dane. Jedną z najbardziej efektywnych metod są ataki adwersarialne, które poprzez niezauważalne modyfikacje danych wejściowych są w stanie w skuteczny sposób zaburzyć działanie sieci neuronowych. W rozprawie przedstawiono pierwszą analizę wpływu wyżej wymienionych ataków na detektory DeepFake. W ramach badań przeprowadzono analizę wpływu dwóch typów scenariuszy — white-box oraz transferability różniących się zakresem wiedzy adwersarza na temat atakowanego systemu. Wykorzystując trzy różne rodzaje ataków (FGSM, PGD oraz FAB) zwiększono EER od wyjściowej wartości 0.0221, do 0.9905 w przypadku scenariusza white-box oraz do 0.4867 w przypadku scenariusza transferability. Następnie zaproponowano nową metodę adaptatywnego treningu adwersarialnego mającego na celu zwiększenie odporności sieci na ataki. Wykorzystanie zaproponowanej metody pozwala na obniżenie EER do 0.0982 w scenariuszu white-box oraz 0.1091 w scenariuszu transferability.

Reprezentacja sygnału dźwiękowego przetwarzanego przez klasyfikatory DeepFake ma istotny wpływ na ich efektywność. Coraz większa liczba proponowanych rozwiązań oparta jest na danych uzyskiwanych z sieci neuronowych trenowanych w ramach metod uczenia self-supervised (SSL). Reprezentacje te pozwalają na osiągnięcie lepszych wyników w porównaniu do standardowych algorytmów przetwarzania sygnałów. W ramach rozprawy jako alternatywę dla rozwiązań SSL zaproponowano wykorzystanie modelu rozpoznawania mowy Whisper. Architektura ta została wytrenowana na największym zbiorze przetwarzania mowy składającym się z ponad 860.000 godzin nagrań. Przeprowadzono analizę wykorzystania modelu zarówno jako samodzielny ekstraktor cech, jak również jako integralną część detektorów używaną w procesie dostrajania. Zaproponowana reprezentacja danych składająca się z danych z modelu Whisper połączonych z parametrami mel-cepstralnymi pozwala na poprawę wyjściowego wyniku na zbiorze In-The-Wild o 26%.

18.06.24