Biometria

Projekt 3

Uwierzytelnianie 'in-the-wild'

Piotr Syga

- Termin: 18. czerwca 2024 (grupa wtorkowa), 19. czerwca 2024 (grupy środowe)
- Raport: raport należy wysłać na skrzynkę mailową prowadzącego laboratorium przynajmniej 48 godzin przed oddaniem listy (wtorek, środa 13,środa 15, środa 17)

Celem projektu grupowego jest wykorzystanie wiedzy zdobytej na wykładzie oraz implementowanych wcześniej systemów uwierzytelniania na podstawie fizjologicznej i behawioralnej do zaprojektowania i wdrożenia systemu uwierzytelniania, który radzi sobie z 'zanieczyszczonymi danymi'. Oznacza to, że pobierając próbkę do weryfikacji nie mamy kontroli nad otoczeniem i uwierzytelniająca się osoba może przez brak wiedzy lub celowe działanie dostarczyć próbkę nie spełniającą założonych przez autora systemu parametrów granicznych.
Poniżej znajduje się lista przykładowych utrudnień dla poszczególnych modalności. Każda grupa powinna wybrać przynajmniej 2 (z poniższej listy lub inne, zaakceptowane przez prowadzącego laboratorium).
Dla identyfikacji na podstawie obrazu twarzy może to być:

- wiele twarzy na zdjęciu (w szczególności niektóre twarze mogą znajdować się na obrazach, plakatach etc.), nie mając pewności która z osób powinna być uwierzytelniona,
- detekcja DeepFake, antispoofing, liveness detection, gdyż nie mamy pewności czy jest to żywy użytkownik czy też jego spreparowany obraz (można zainspirować się, np. lekką architekturą detekcyjną Mesonet, skorzystać z FLD-SDK, FaceForensics, warto wykorzystać zbiór CelebDF lub zbiór DFDC),
- zdjęcia o bardzo niskiej rozdzielczości lub niskiej jakości (np. CCTV), można tu wykorzystać zbiór TinFace
- wykorzystane jest zdjęcie z filtrem (np. Snapchatowym czy Instagramowym), mogą to być zarówno elementy tła jak i 'podmienione' oczy, nos, usta czy uszy,

- twarz przy weryfikacji jest zasłonięta przez element garderoby, jak np. okulary, maseczka czy kapelusz (można wykorzystać ROF),
- zamiast jednego z utrudnień można zadbać o zachowanie prywatności użytkowników systemu uwierzytelniania (można się zainspirować Fawkes oraz LowKey)

Dla identyfikacji na podstawie głosu przykładami mogą być: Głos:

- zamiast pojedynczego mówcy mamy do czynienia z dialogami lub rozmową grupy osób (można wykorzystać informacje z wykładu nt. diaryzacji oraz DailyTalk)
- detekcja DeepFake, antispoofing, livness detection (przykładowe rozwiązania cechy Whispera, RawGAT, AASIST)
- nagrania z dużym poziomem hałasu (tzn. amplituda hałasu między 70 a 120% amplitudy sygnału)
- wypowiedzi do uwierzytelniania są w innym języku niż użyte przy wdrażaniu (próbki można stworzyć przy pomocy XTTS, wykorzystać Multilingual Librispeech lub MLAAD)
- zamiast jednego z utrudnień można zadbać o zachowanie prywatności użytkowników systemu uwierzytelniania (można się zainspirować ortogonal, rozdzielenie cech, GAN, SSL)

W przypadku potrzeby wykorzystania większej liczby danych pomocne będzie zestawienie datasetów.
Przykładem innych modalności może być keystroke dynamics (artykuł, zbiór danych). Dla tej modalności utrudnieniem może być identyfikacja na podstawie dowolnego, krótkiego tekstu, uzyskanie części danych 'z symulatora' lub też zachowanie prywatności użytkownika.

Podobnie jak w przypadku wcześniejszych projektów, można skorzystać z gotowego modelu, jednak konieczne jest zadbanie aby 'trudne' próbki nie powodowały istotnego pogorszenia metryk względem 'czystych' próbek. Można założyć, że w przypadku weryfikacji FRR nie powinien wzrosnąć o więcej niż 5 punktów procentowych i jako cel postawić sobie minimalizację FAR. W przypadku identyfikacji True Identification Rate nie powinno spaść o więcej niż 5 punktów procentowych, minimalizujemy liczbę odrzuceń.

W systemie powinno być wdrożonych przynajmniej 100 użytkowników (w tym wszyscy członkowie grupy). Liczba próbek użytych do wdrożenia jest dowolna (wyjątkiem jest identyfikacja na podstawie krótkich wypowiedzi czy krótkich tekstów pisanych). Pobranie próbki do uwierzytelniania nie powinno trwać dłużej niż 30s. Przetestuj skuteczność systemu na przynajmniej 1000 próbkach 'czystych' oraz przynajmniej

5000 'trudnych'. Zadbaj o zbalansowanie zbioru testowego pomiędzy utrudnieniami. Porównaj uzyskane wyniki i omów metody poprawy odporności systemu.

W ramach projektu przygotuj raport, w którym zostaną wskazane:

- dane autorów (imię, nazwisko, nr indeksu, termin laboratorium)
- dokładny opis problemu, przedstawienie źródeł, które były wykorzystywane
- opisana metoda uwierzytelniania, ze szczególnym uwzględnieniem modyfikacji zrobionych w celu uodpornienia systemu
- przedstawiony zbiór danych (oraz podział na odpowiednie podzbiory i metoda subsamplingu) – istotne cechy to liczba osób, liczba próbek dla każdej z osób, parametry próbek takie jak jakość, czas trwania, rozdzielczość; zaznacz wyraźnie na którym etapie był używany który podzbiór danych,
- opisana procedura wdrożenia oraz uwierzytelniania
- opis procedury testowania, prac wykonanych w celu korekty wyników dla zaburzonych próbek, wyniki, porównanie z wynikami próbek 'czystych' oraz z wynikami raportowanymi przez autorów modelu bazowego (lub innych zbliżonych funkcjonalnością modeli) oraz wnioski.