Ochrona przed deepfake’ami staje się coraz pilniejszą potrzebą

8 stycznia 2024

Deepfake stanie się w najbliższych latach jednym z podstawowych narzędzi w rękach cyberprzestępców. Technika łączenia zdjęć, filmów i głosu – często wykorzystywana w celach rozrywkowych – może być także groźnym instrumentem służącym dezinformacji, ale i atakom wyłudzającym dane czy pieniądze. Walka z takimi zagrożeniami staje się przedmiotem prac wielu naukowców. Badacze z Uniwersytetu Waszyngtona w St. Louis opracowali mechanizm obronny AntiFake, który ma zapobiegać nieautoryzowanej syntezie mowy, utrudniając sztucznej inteligencji odczytanie kluczowych cech głosu. Zapotrzebowanie na takie narzędzia będzie rosło wraz z rozwojem rynku generatywnej SI.

„Pojęcie „deepfake” składa się z dwóch członów. Pierwszy z nich, deep, odnosi się do głębokich sieci neuronowych, które stanowią najbardziej rozwinięty aspekt uczenia maszynowego, a fake oznacza tworzenie fałszywego obrazu twarzy lub głosu. W naszym kontekście słowo to oznacza, że cyberprzestępca może wykorzystać głębokie sieci neuronowe, aby podrobić mowę danej osoby” – mówi w wywiadzie dla agencji Newseria Zhiyuan Yu, doktorant na Wydziale Informatyki Uniwersytetu Waszyngtona w St. Louis. „Zastosowanie deepfake’ów do celów złośliwych ataków może przybrać wiele form. Jedną z nich jest wykorzystanie ich do rozpowszechniania fałszywych informacji i mowy nienawiści, również na podstawie głosu osób powszechnie znanych. Przykładowo przestępca użył głosu prezydenta Bidena, aby wypowiedzieć Rosji wojnę. Mieliśmy również do czynienia z sytuacjami, kiedy cyberprzestępca wykorzystał głos użytkownika, aby przejść weryfikację tożsamości w banku. Takie sytuacje stały się częścią naszego życia”.

Eksperci WithSecure przewidują, że w 2024 roku dojdzie do nasilenia problemu dezinformacji z użyciem technologii deepfake. Będzie się tak działo głównie z uwagi na to, że to rok wyborów prezydenckich w USA oraz do Parlamentu Europejskiego. „W obliczu zagrożeń związanych z deepfake'ami proponujemy aktywne przeciwdziałanie im za pomocą technologii AntiFake. Powstałe dotychczas rozwiązania systemowe opierają się na wykrywaniu zagrożeń. Wykorzystują techniki analizy sygnału mowy lub metody wykrywania oparte o fizykę, aby rozróżnić, czy dana próbka jest autentyczną wypowiedzą danej osoby, czy została wygenerowana przez sztuczną inteligencję. Nasza technologia AntiFake opiera się jednak na zupełnie innej metodzie proaktywnej" – mówi Zhiyuan Yu.

Wykorzystana przez twórców metoda bazuje na dodaniu niewielkich zakłóceń do próbek mowy. Oznacza to, że gdy cyberprzestępca użyje ich do syntezy mowy, wygenerowana próbka będzie brzmiała zupełnie inaczej niż użytkownik. Nie uda się więc za jej pomocą oszukać rodziny ani obejść głosowej weryfikacji tożsamości. „Niewiele osób ma dostęp do wykrywaczy deepfake'ów. Wyobraźmy sobie, że dostajemy e-mail, który zawiera próbkę głosu, lub otrzymujemy taką próbkę na portalu społecznościowym. Niewiele osób poszuka wtedy wykrywacza w sieci, aby przekonać się, czy próbka jest autentyczna. Według nas ważne jest przede wszystkim zapobieganie takim atakom. Jeśli przestępcy nie uda się wygenerować próbki głosu, który brzmi jak głos danej osoby, do ataku wcale nie dojdzie" – podkreśla badacz.

Od strony technologicznej rozwiązanie bazuje na dodaniu zakłóceń do próbek dźwięku. Projektując system, twórcy musieli się zmierzyć z kilkoma problemami projektowymi. Pierwszym było to, jakiego rodzaju zakłócenia powinny zostać nałożone na dźwięk tak, by były nieuchwytne dla ucha, a jednak skuteczne. „To prawda, że im więcej zakłóceń, tym lepsza ochrona, jednak gdy oryginalna próbka będzie się składać z samych zakłóceń, użytkownik nie może z niej w ogóle korzystać" – mówi Zhiyuan

Yu.

Ważne było też to, by stworzyć rozwiązanie uniwersalne, które sprawdzi się w różnego rodzaju syntezatorach mowy. „Zastosowaliśmy model uczenia zespołowego, na potrzeby którego zebraliśmy różne modele i stwierdziliśmy, że enkoder mowy jest bardzo ważnym elementem syntezatora. Przekształca on próbki mowy na wektory, które odzwierciedlają cechy mowy danej osoby. Jest on najważniejszym elementem syntezatora stanowiącym o tożsamości osoby mówiącej w wygenerowanej mowie. Chcieliśmy zmodyfikować osadzone fragmenty z zastosowaniem enkodera. Użyliśmy czterech różnych enkoderów, aby zapewnić uniwersalne zastosowanie do różnych próbek głosu" – wyjaśnia naukowiec z Uniwersytetu Waszyngtona w St. Louis.

Kod źródłowy technologii opracowanej na Wydziale Informatyki Uniwersytetu Waszyngtona w St. Louis jest ogólnodostępny. W testach prowadzonych na najpopularniejszych syntezatorach AntiFake wykazał 95-proc. skuteczność. Poza tym narzędzie to było testowane na 24 uczestnikach, aby potwierdzić, że jest dostępne dla różnych populacji. Jak wyjaśnia ekspert, głównym ograniczeniem tego rozwiązania jest to, że cyberprzestępcy mogą próbować obejść zabezpieczenie, usuwając zakłócenia z próbki. Istnieje też ryzyko, że pojawią się nieco inaczej skonstruowane syntezatory, co do których tak pomyślana technologia nie będzie już skuteczna. Zdaniem ekspertów będą jednak w dalszym ciągu toczyć się prace nad wyeliminowaniem ryzyka ataków za pomocą deepfake'ów. Obecnie AntiFake potrafi chronić krótkie nagrania, ale nic nie stoi na przeszkodzie, aby rozwijać go do ochrony dłuższych nagrań, a nawet muzyki.

„Jesteśmy przekonani, że niedługo pojawi się wiele innych opracowań, które będą miały na celu zapobieganie takim problemom. Nasza grupa pracuje właśnie nad ochroną przed nieuprawnionym generowaniem piosenek i utworów muzycznych, ponieważ dostajemy wiadomości od producentów i ludzi pracujących w branży muzycznej z zapytaniami, czy nasz mechanizm AntiFake można wykorzystać na ich potrzeby. Ciągle

więc pracujemy nad rozszerzeniem ochrony na inne dziedziny” – podkreśla Zhiyuan Yu.

Future Market Insights szacował, że wartość rynku generatywnej sztucznej inteligencji zamknie się w 2023 roku w kwocie 10,9 mld dol. W ciągu najbliższych 10 lat jego wzrost będzie spektakularny, bo utrzyma się na poziomie średnio 31,3 proc. rocznie. Oznacza to, że w 2033 roku przychody rynku przekroczą 167,4 mld dol. Autorzy raportu zwracają jednak uwagę na to, że treści generowane przez sztuczną inteligencję mogą być wykorzystywane w złośliwy sposób, na przykład w fałszywych filmach, dezinformacji lub naruszeniach własności intelektualnej.

Źródło: Newseria.pl