Przeciwdziałanie nadużyciom DeepFake – metody i techniki

Franciszek Górski
Katedra Systemów Multimedialnych

Plan wykładu

- Czym jest DeepFake?
- Jakie są rodzaje DeepFake'ów?
- Metody detekcji DeepFake'ów

Ale najpierw mała retrospekcja

Aresztowanie Donalda Trumpa

Papież Franciszek

Ale czy to się wydarzyło naprawdę?

Nie, pokazane treści zostały zsyntezowane i nie są prawdziwe.

Treści te określa się mianem Deepfake'ów.

Czym jest Deepfake?

Czym jest Deepfake?

Deepfake - treść, która powstała w wyniku manipulacji oryginalnej zawartości np. obrazu czy dźwięku, w celu zmanipulowania jej przekazu - *fake (ang. fałszerstwo).*

Deep (ang. głęboki) oznacza, że manipulacja ta została przeprowadzona za pomocą modeli głębokich sieci neuronowych.

Manipulacja ta ma często na celu przekazanie nieprawdziwych informacji, ale może też służyć pozytywnym celom.

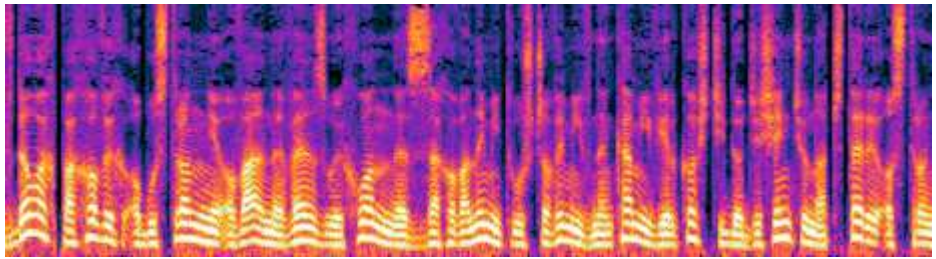
Wzrost popularności Deepfake'ów

- Ludzie od wieków starali się manipulować treścią.
- Wraz z pojawieniem się mediów komunikacyjnych pokusa ta stała się jeszcze większa.
- Ciężko było jednak modyfikować media takie jak obraz czy dźwięk w taki sposób aby manipulacja była dokładna, a co za tym idzie trudna w wykryciu.
- Dużą zmianą okazało się być jednak zastosowanie głębokich sieci neuronowych w tym procesie.

Rodzaje Deepfake'ów?

Rodzaje Deepfake'ów

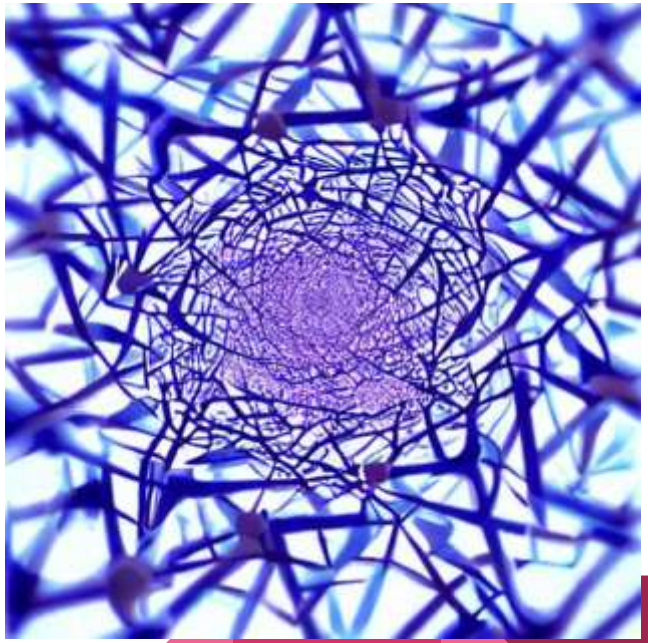
- manipulacja obrazu
- manipulacja głosu
- manipulacja materiału wideo



W jaki sposób tworzone są materiały DeepFake?

Głębokie sieci neuronowe

- Do tworzenia materiałów Deepfake wykorzystywane są tzw. modele generatywne sieci neuronowych
- Są to modele, które projektowane są i trenowane do zadania generowania odpowiednich treści np. obrazów
- Do przykładów można zaliczyć: Generative Adversarial Networks, Diffusion models, Variational autoencoders, Transformers



Głębokie sieci neuronowe

- Celem Deepfake jest prawie zawsze manipulacja wizerunku lub charakterystyki konkretnej osoby

- Osobami tymi są zazwyczaj politycy czy celebryci z 2 powodów:
  - są znani szerokiej publice i ich przekaz jest obserwowany
  - w sieci znajduje się wiele ich zdjęć, nagrań czy wypowiedzi

Głębokie sieci neuronowe

- Materiały zawierające wizerunek *ofiary* zostają wykorzystane w celu trenowania modeli generatywnych

- Po takim treningu modele te są w stanie generować obrazy, nagrania wideo czy same wypowiedzi ofiar z dowolną treścią

- Wiarygodność (jakość) tych materiałów zależy od jakości wytrenowania tych modeli, a także od dokładności wydanego im polecenia w postaci np. *promptu*

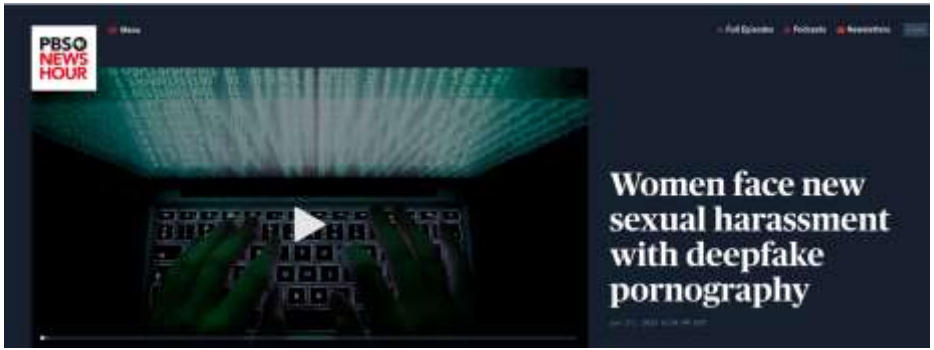
Cele tworzenia materiałów Deepfake

Dezinformacja

- Wykorzystując wizerunek znanej osoby szerzone są nieprawdziwe informacje, mające wprowadzić w błąd opinię publiczną np.:
  - Prezydent Żeleński wzywa Ukraińców do kapitulacji
  - Prezydent Obama obraża Donalda Trumpa
  - Aresztowanie prezydenta Trumpa w Nowym Jorku

Oczernienie/poniżenie kogoś

- Wykorzystanie czyjegoś wizerunku, zazwyczaj samej twarzy, w celu umieszczenia jej w materiałach niekorzystnych dla *ofiary*, najczęściej w pornografi



Podszycie się

- Wykorzystanie czyjegoś wizerunku lub głosu w celu podszycia się pod daną osobę i uzyskania dostępu do jej zasobów np. konta bankowego, telefonu komórkowego

- Wykorzystywane w tym celu jest klonowanie twarzy i jej ruchów oraz klonowanie głosu

TheStreet

TECH RETAIL INVESTING MARKETS PERSONAL FINANCE

HOME › TECHNOLOGY

Cybersecurity expert says the next generation of identity theft is here: 'Identity hijacking'

Scammers last week used deepfake technology to steal $25 million from a multinational Hong Kong firm.

Ian Krietzberg · Feb 5, 2024 4:06 PM EST

Deepfake technology – a synthetic representation of a person's likeness – is not new technology. Think Mark Hamill's de-aged return as a young Luke Skywalker in an episode of 'The Mandalorian' in 2019.

Artificial intelligence is likewise nothing new.

But when it launched at the end of 2022, ChatGPT made AI technology cheaply accessible to the masses, simultaneously setting off a race between nearly all of the mega-cap tech corporations (and a bunch of startups, too) to ship more powerful models.

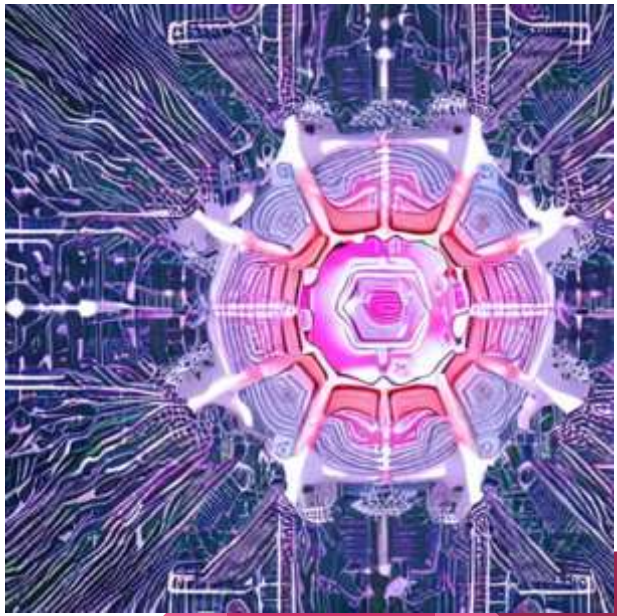
Sposoby detekcji materiałów Deepfake

Jak wykrywać materiały Deepfake?

- Wraz z postępem w rozwoju AI metody tworzenia Deepfake'ów pozwalają na stworzenie coraz dokładniejszych manipulacji

- Manipulacje te stają się coraz trudniejsze do rozpoznania przez *samego* człowieka

- W takim wypadku stosowane są inne modele sieci neuronowych pozwalających na detekcję zmanipulowanego materiału



Rodzaje detekcji

Sposoby wykorzystywane do detekcji zależą od tego jakiego medium dotyczy Deepfake.

Wyróżniamy 3 obszary (media) działania detekcji:

- statyczny obraz
- głos
- nagranie wideo (zazwyczaj) z głosem

Detekcja zmanipulowanego obrazu

- Jeden z trudniejszych przypadków do wykrycia
- Dzisiejsze techniki pozwalają praktycznie na idealne generowanie obrazów w tym twarzy
- Szansy na detekcję należy się dopatrywać w sytuacji umieszczenia czyiś twarzy w innym kontekście i wykrycia drobnych niedociągłości w obrazie

Detekcja zmanipulowanego głosu

- Również bardzo trudne zadanie ze względu na coraz większą doskonałość syntezatorów mowy

- Detekcja syntezowanego głosu może się głównie skupiać na wykrywaniu nienaturalności wypowiedzi, jej sztucznego brzmienia

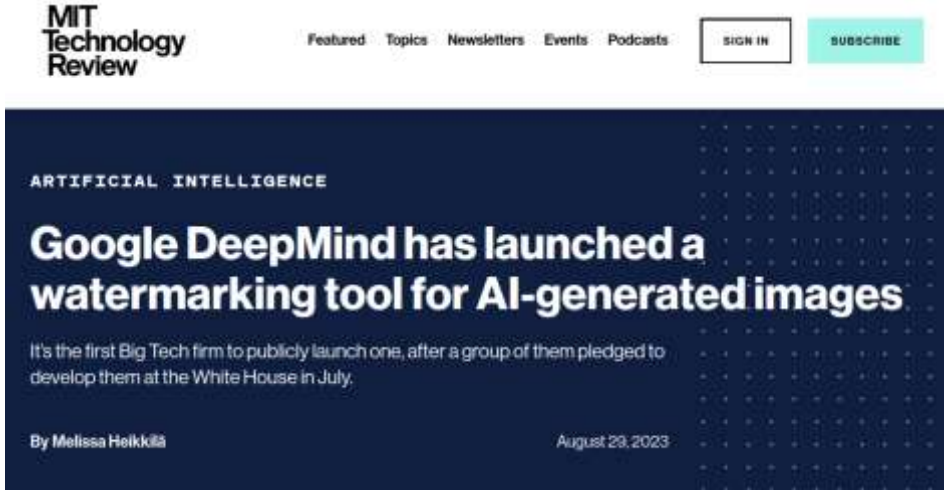
Detekcja zmanipulowanego nagrania audio-wizualnego

- Chyba najpowszechniejszy przypadek stosowania Deepfake'ów
- Ze wszystkich 3 mediów pozwala na najwięcej podejść do detekcji takich jak:
  - Wykrywanie braku synchronizacji między ruchami postaci, zwłaszcza ust a mówionym tekstem
  - Badanie *żywotności* wygenerowanej treści, czy postać nie wydaje się zbyt sztuczna
  - Oparcie detekcji na podstawie wykrywania drobnych artefaktów w materiale

Znakowanie wodne a Deepfake

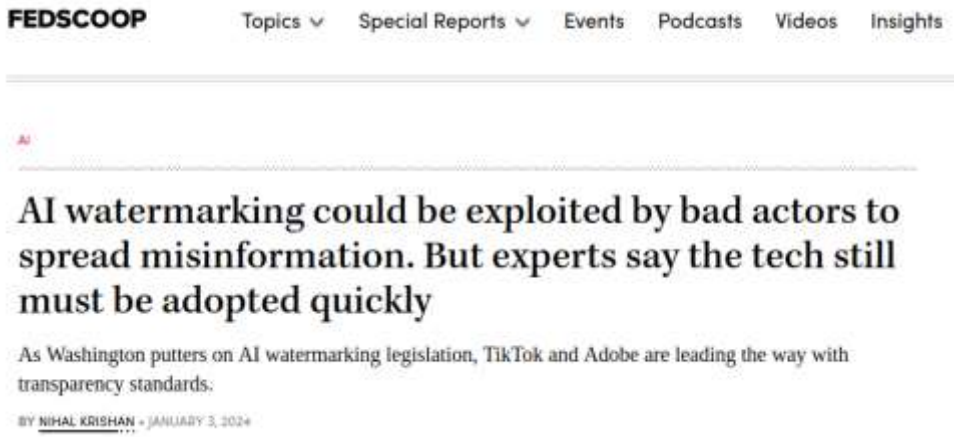
Czy znakowanie wodne?

- W debacie publicznej pojawiają się informacje o stosowaniu znaków wodnych (*ang. watermarks*) jako sposobu na detekcję czy przeciwdziałanie Deepfake'om
- Pomysł ten polega na umieszczenia znaku wodnego w obrazie lub materiału audio i na tej podstawie rozróżnianie treści prawdziwych od fałszywych



Czy znakowanie wodne?

- Nie jest to jednak rozwiązanie idealne
- Znaki wodne wciąż mogą być usuwane
- Na dodatek rozwiązanie to może i tak nie mieć wpływu na tworzenie treści na podstawie już istniejącego materiału



Pytania?

Dziękuję za uwagę!