W ostatnim roku pojawiło się wiele narzędzi opartych o modele uczenia maszynowego zwane sieciami neuronowymi, których zadaniem jest tworzenie zdjęć i filmów obiektów, które nie istnieją w świecie rzeczywistym. Te narzędzia wykorzystywane są często do generowania filmów ze znanymi osobami (potocznie zwanych deepfake: https://en.wikipedia.org/wiki/Deepfake), w których osoby te mówią coś, czego w rzeczywistości nie powiedziały. Kilka przykładów takich filmów możemy znaleźć tutaj:

1. https://www.youtube.com/watch?v=AmUC4m6w1wo
2. https://www.youtube.com/watch?v=AH7Beg5urxY
3. https://www.youtube.com/watch?v=ehD3C60i6lw

Jednak możliwości zastosowania tych narzędzi są dużo szersze, gdyż potencjalnie pozwalają na generowanie dowolnego obrazu czy dźwięku (na przykład generowania głosu znanej osoby na podstawie tekstu pisanego). W połączeniu z narzędziami do generowania tekstu (Jak na przykład model GPT-3: https://en.wikipedia.org/wiki/GPT-3) mogą one służyć do generowania fake newsów na niespotykaną dotąd skalę i o niespotykanej dotąd jakości.

**Jak zatem możemy się bronić przed takimi narzędziami?**

Jednym z podejść jest stosowanie innych sieci neuronowych do wykrywania tych fałszywych treści (na przykład w tym konkursie – https://www.kaggle.com/c/deepfake-detection-challenge). Na ten moment bliższa analiza obrazu jest w stanie ujawnić pewne cechy, które odróżniają te obrazy od obrazów prawdziwych, więc teoretycznie możemy nauczyć model, który zastępuje w tym człowieka. Z tym podejściem jest jednak pewien problem, który postaram się pokrótce objaśnić. Najpierw jednak musimy zrozumieć, jak uczone są takie algorytmy. Skupimy się tutaj na algorytmach do generowania obrazu, gdyż obrazują one dobrze, gdzie może leżeć problem.

Obecnie najlepszymi i najszerzej stosowanymi modelami generujących obrazy są architektury sieci nuronowych typu GAN (https://en.wikipedia.org/wiki/Generative_adversarial_network). Są one uczone w bardzo ciekawy sposób. Otóż stawiamy obok siebie 2 sieci neuronowe: Pierwsza z nich (tzw. Generator) ma za zadanie generować obrazy fałszywe (np, twarze), a druga (tzw, Dyskryminator) dostaje w losowej kolejności obrazy prawdziwe i fałszywe i musi nauczyć je odróżniać od siebie. W tym samym czasie zadaniem Generatora jest nauczyć się jak oszukiwać Dyskryminator tak, by ten rozpoznawał fałszywe obrazy jako prawdziwe.

Na początku uczenia modelu Dyskryminator ma łatwe zadanie, jednak po jakimś czasie Generator staje się tak dobry, że Dyskryminator nie jest w stanie odróżnić obrazu

Materiał pobrany ze strony Instytutu Sobieskiego, sobieski.robocza.ovh, autor: Piotr Tempczyk.

prawdziwego od fałszywego. Nie oznacza to od razu, że wygenerowane obrazy są w stanie oszukać człowieka. Zależy to w dużej mierze od jakości Dyskryminatora: Im lepszy Dyskryminator, tym większa szansa, że Generator nauczy się tworzyć obrazy zbliżone do prawdziwych. Jeśli założymy, że do stworzenia Generatora użyjemy bardzo dobrej sieci, to jedynym ograniczeniem staje się jakość Dyskryminatora. I tutaj dochodzimy do sedna problemu: **każde narzędzie oparte o sieci neuronowe, które zostanie stworzone do wykrywania deepfake może jednocześnie posłużyć do treningu GAN-ów i sprawić, że będą one w stanie oszukać te algorytmy po jakimś czasie.**

Wydaje się więc, że zwalczanie ognia ogniem w tym przypadku może nie dać dobrych rezultatów i zawczasu powinno się też pomyśleć o innych metodach (np. kryptograficznych) do weryfikacji czy dane wideo jest prawdziwe, gdyż w pewnym momencie narzędzia te dojdą do poziomu, gdzie nikt nie będzie w stanie odróżnić prawdy od fałszu, używając tylko algorytmów do rozpoznawania obrazu.

Narodowy Instytut Wolności
Centrum Rozwoju Społeczeństwa Obywatelskiego
NIW
Program Rozwoju Organizacji Obywatelskich na lata 2018–2030
PROO

Komentarz opublikowany w ramach projektu "30 idei dla Polski na trzecią dekadę XXI wieku", który jest współfinansowany ze środków Narodowego Instytutu Wolności - Centrum Rozwoju Społeczeństwa Obywatelskiego ze środków PROO na lata 2018-2030.