Deepfake a technologie za ním

ADAM GRYGAR

PB016 Umělá inteligence I

Listopad 2019

1 Úvod

Deepfake je falešné video, kde je pomocí strojového učení nahrazen jakýkoliv obličej nebo hlas komukoliv. V dnešní době je část deepfaků na takové úrovni, že běžný člověk skoro nedokáže poznat rozdíl zda se jedná o falešné video nebo ne. Na Internetu jich existují tisíce. Například lidé nahrazují obličeje slavných osobností obličeji jejich příbuzných.

V deepfake videu lze také nahradit hlas dotyčné osoby.

Technologie vytvářející deepfake jsou jednoduše dostupné. Existují přímo mobilní aplikace, které umožňují výměnu vašeho obličeje za obličej jiné osoby a v reálném čase dané video nahrát.

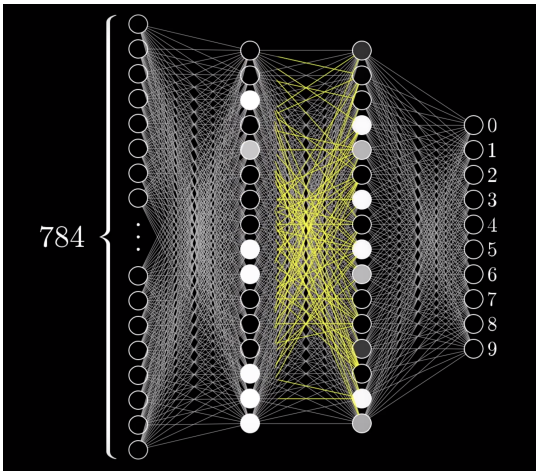
Většina deepfaků na Internetu jsou vytvořena amatérskými tvůrci a rozpoznat je není problém. Problém je pokud je deepfake vytvořen týmem profesionálů, kterých výtvory jsou tak uvěřitelné, že i lidé, kteří se věnují obrazovým médiím, je mají problém rozpoznat. Ještě větší problém nastává, když se vytváří falešné informace, nebo aby hanili známou osobnost nebo politika. [Koř19] [Sve18]

2 Technologie

Jak jsem již v úvodu naznačil, tak deepfake je vytvářen pomocí strojového učení, konkrétně pomocí Generative Adversarial Networks (dále GANs), což jsou dvě neuronové sítě, které se navzájem zdokonalují.

Neuronová síť

Neuronová síť je síť (obvod) neuronů, ze kterých se skládají vrstvy. Neuron je objekt, který drží číslo od 0 do 1. Neurony v sousedních vrstvách jsou všechny propojeny vazbami. Jednoduchá neuronová síť se skládá z tří druhů vrstev. Vstupní, výstupní a mezi nimi se nachází skryté vrstvy. Mějme například neuronovou síť, která má za cíl rozpoznávat ručně napsané číslice. Poté neuron ve vstupní vrstvě bude mít hodnotu jako jas určitého pixelu v obrázku číslice. Výstupní vrstva bude mít 10 neuronů. Podle výstupního neuronu, který bude mít nejvyšší hodnotu, vybereme příslušnou číslici (viz obrázek 2.1). Skryté vrstvy zajišťují více komplexní vztahy mezi vstupními hodnotami. Neuronová síť funguje tak, že velikost aktivace neuronů v jedné vrstvě určuje hodnotu aktivace neuronů v nadcházející.



Obrázek 2.1 [3Bl17]

Základní neuronová síť se dvěma skrytými vrstvami.

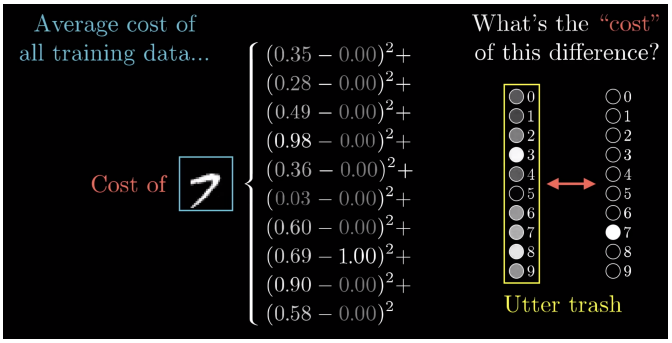
Aktivace neuronu je určena hodnotou neuronů v minulé vrstvě, váhou vazeb a konstantou bias transformovaná na hodnotu mezi 0 a 1, například pomocí funkce sigmoid. Poté může výpočet vypadat takto:

$$\sigma(w_1 \cdot a_1 + w_2 \cdot a_2 + ... + w_n \cdot a_n + b)$$

kde w je hodnota váhy vazby, a je hodnota neuronu v předchozí vrstvě a b je bias. Ve skutečnosti se používají vektory a operaci s nimi. Například stejný vztah mezi neurony v jedné vrstvě s vrstvou předchozí se dá popsat takto:

$$\sigma\left(\begin{bmatrix} w_{0,0} & w_{0,1} & \cdots & w_{0,n} \\ w_{1,0} & w_{1,1} & \cdots & w_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{k,0} & w_{k,1} & \cdots & w_{k,n} \end{bmatrix} \begin{bmatrix} a_0^{(0)} \\ a_1^{(0)} \\ \vdots \\ a_n^{(0)} \end{bmatrix} + \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_0 \end{bmatrix}\right)$$

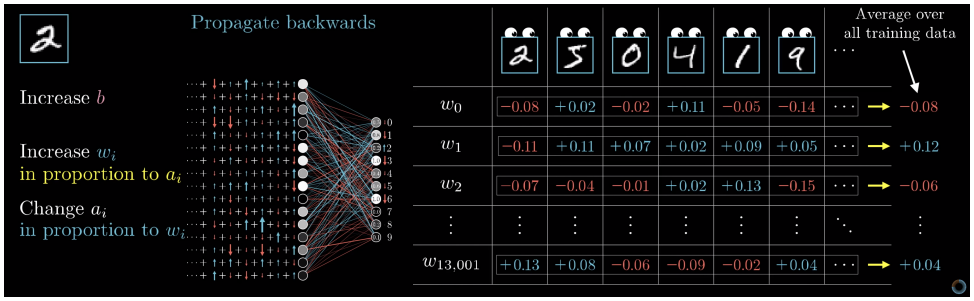
Neuronová síť se snaží minimalizovat cenovou funkci (Cost function), nebo-li najít její lokální minimum. Její trénování probíhá pomocí zpětné propagace (Backpropagation).
Cenová funkce nám říká, jak moc si je síť daným výsledkem jistá (viz obrázek 3.2).



Obrázek 2.2 [3Bl17]

Ilustrace, jak funguje cenová funkce. Levý sloupec je výstup neuronové sítě a napravo je očekávaný výstup.

Zpětná propagace upravuje hodnoty vazeb a biasů v síti tak, aby výsledek cenové funkce byl co nejblíže 0 pro daný vstup. Postupuje postupně od výstupu až ke vstupu. Jsou tři možnosti, jak změnit hodnotu výsledku. První je změna biasu. Druhá je změna hodnot vazeb propojující neurony. A poslední je změna hodnoty neuronu v předchozí vrstvě, pomocí stejných možností. Úprava hodnot vazeb a biasů se z důvodu efektivity dělá po blocích dat. Nejprve se trénovací data náhodně zamíchají a poté se rozdělí do bloků například po 10. Síť až pro průchodu jednoho bloku provede zpětnou propagaci, ne po každém jednom průchodu. [3Bl17]



Obrázek 2.3 [3Bl17]

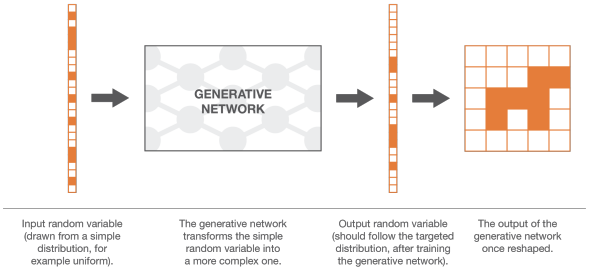
Nalevo zpětná propagace, úprava hodnot vazeb a biasů. Napravo znazornění projití jednoho bloku dat a následné zprůměrované hodnoty výsledku.

Generative Adversarial Networks(GANs)

Máme dvě neuronové sítě. Jedna z nich se stará v našem konkrétním případě o generování snímků videa a říká se ji Generátor (Generator) a druhá Diskriminátor (Discriminator), která se snaží poznat zda se jedná o podvrh vytvořený Generatorem nebo ne.

Generátor je konvoluční neuronová síť (Convolutional neural network). Na vstup přivedeme nějaký náhodný šum, který se postará o vygenerování vždy jiného obrázku. Nechť máme síť, která má za úkol generovat obrázky koček. Poté šum je vektor v n rozměrném prostoru, který podle jeho velikosti a směru udává například, jakou bude mít kočka barvu, kde bude umístěná na

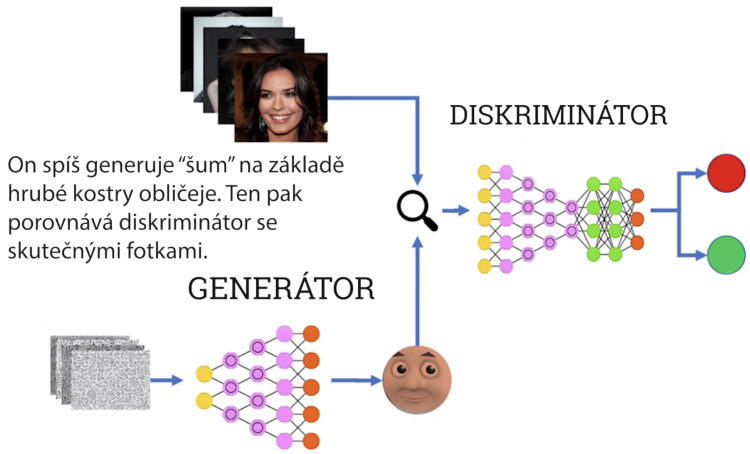
obrázku a podobně. [Roc19] [Com17]



Obrázek 2.4 [Roc19]

Průběh tvorby snímku Generatora.

Úloha Generátora je tvorba čím dál více uvěřitelnějších snímků videa. Úloha Diskriminátora je čím dál lépe rozeznávat zda se jedná o falešný snímek nebo ne. Odměna Generátora je inverzní k výstupu Diskriminátora, takže pokud vyprodukuje snímek, který Diskriminátor pozná, že se jedná o podvod, je Generátor potrestán. Trénují se takto v cyklu dokud Diskriminátor nebude vracet číslo 0.50 nebo-li 50 na 50, čili vůbec netuší zda se jedná o podvod nebo ne. [Met16] [Com17]



Obrázek 2.5 [Kor19]

Průběh funkce GANs.

3 Použití

Pornografie a film

Deepfake se nejdříve začal užívat v pornografii v roce 2017. Obličeje porno hereček a herců byly vyměněny za obličeje hollywoodských hvězd. V drtivé většině případů bylo video vytvořeno bez souhlasu dané osobnosti.
V červnu 2019 vyšla aplikace s názvem DeepNude. Díky této aplikaci, která využívá GANs, mohl kdokoliv odstranit oblečení z obrázku jakékoliv ženy. Aplikace byla stažena z prodeje o pár dní později.[Dic19][Col19]

Deepfake také využívají amatérští tvůrci pro tvorbu zábavných videí, kde například nahradí obličej hlavního herce v určité části filmu nebo rozhovoru obličejem jiné osoby. [Rad18]

Obrázek 3.1 [Gen08] a 3.2 [Fac19]

Nahoře originální snímek z rozhovoru Billa Hadera. Dole snímek z videa,kde je obličej Hadera změněn za obličej Toma Cruise, od uživatele Ctrl Shift Face vytvořený pomocí GANs.

Politika

Dezinformace se v politice šíří běžně. Ať už jde o letáky s nepravdivou informací, až po úmyslně napsané celé lživé články s upravenými fotkami, který není postaven na faktech, nebo jsou fakta záměrně překroucena. Cílem dezinformací je ovlivnit smýšlení lidí k dané problematice. Například v prezidentských volbách v roce 2018 se jich šířilo mnoho pomocí řetězových emailů a sociálních sítí. Nejčastěji líčící Miloše Zemana jako kandidáta proti přijímaní uprchlíků a Michala Horáčka a Jiřího Drahoše jako proupchlické kandidáty. Bylo potvrzeno mnoha institucemi, že se jednalo o lež. Problémem je, že k lidem, které této lži uvěřili, se nedostala zpráva, že se jedná o lež, nebo ji ignorovali. Pomocí dezinformací lze ovlivnit výsledky voleb, jak už se již stalo například ve Velké Británii při referendu o Brexitu. [Čes] [Gol18] [SW19]

Deepfake videa jsou ve vytváření dezinformací velmi mocný nástroj. S relativně malým týmem odborníků lze od základu vytvořit video nebo nějaké pozměnit tak, aby hanilo politika nebo politickou stranu. Lze například vytvořit proslov politika, který se nikdy nestal, a pomocí falešných účtů na sociálních sítích proslov rozšířit mezi mnoho lidí a popřípadě změnit výsledek voleb. [Ven19]

Obrázek 3.3 [BBC19]

Nalevo snímek videa vytvořeného pomocí deepfake. Napravo osoba jejíž hlas a pohyb obličeje je použita pro vygenerování videa nalevo.

4 Obrana proti deepfake

Na odhalení deepfake videí pracuje mnoho organizací. Jednou z nich je DARPA se svým programem MediFor.
MediFor přivádí dohromady nejlepší světové výzkumníky s cílem vyvinout technologii na detekci manipulací s videem i fotografiemi. [Tur18]
Siwei Lyu je profesor na State University of New York at Albany. spolu se svým týmem vygenerovali okolo 50 falešných videí a zkusili celou řadu tradičních rozpoznávacích metod. Metody někdy fungovali, někdy ne. Poté při studiu videí přišel s poznatkem, že lidé ve falešném videu skoro vůbec nemrkají a pokud ano, tak mrkání vypadá nepřirozeně. Je to hlavně způsobené nevyvážeností trénovacích dat, kde snímky obličej je většinou zachycen pouze v případě, když má otevřené oči. Ostatní pracovníci zapojení ve výzkumu deepfake se také zaměřují na nepřirozené pohyby hlavy, divné barvy očí a podobné maličkosti. Tyto maličkosti bude čím dál více těžší rozeznat, protože deepfake videa se za poslední rok značně zlepšila a pořád zlepšují. [Kni18] [Ven19]

5 Závěr

Lidé jsou dezinformacemi zasypáváni každý den. Bez dostatečné mediální vzdělanosti jim mnoho lidí věří, jak u nás, tak i ve světě. Proto je deepfake hrozba, na kterou zatím nedokážeme najít účinnou odpověď. Mnoho lidí neví, že nějaká podobná technologie existuje a už vůbec, jak ji rozpoznat, když se za velmi krátkou dobu tato videa posunula na úroveň, že i experti mají problém je rozeznat. Dá se s jistotou říct, že deepfake bude figurovat v mnoha politických bitvách a zcela jistě ovlivní mnoho voličů. Velká zkouška čeká obyvatele Spojených států Amerických s nadcházejícími prezidentskými

volbami v roce 2020.

Technologie GANs, která za tvorbou deepfake stojí, má velký potenciál pro budoucí vývoj oblasti strojového učení a ne jen pro tvorbu deepfake videí.

Odkazy

[3Bl17] 3Blue1Brown. *Neural networks*. 2017. URL: https://www.youtube.com/playlist?list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi.

[BBC19] BBC. *How the Obama / Jordan Peele DEEPFAKE actually works — Ian Hislop's Fake News - BBC*. 2019. URL: https://www.youtube.com/watch?v=g5wLaJYBAm4.

[Col19] Samantha Cole. *This Horrifying App Undresses a Photo of Any Woman With a Single Click*. 2019. URL: https://www.vice.com/en_us/article/kzm59x/deepnude-app-creates-fake-nudes-of-any-woman.

[Com17] Computerphile. *Generative Adversarial Networks (GANs) - Computerphile*. 2017. URL: https://www.youtube.com/watch?v=Sw9r8CL98N0.

[Čes] Ministerstvo vnitra České republiky. *Definice dezinformací a propagandy*. URL: https://www.mvcr.cz/cthh/clanek/definice-dezinformaci-a-propagandy.aspx.

[Dic19] EJ Dickson. *Deepfake Porn Is Still a Threat, Particularly for K-Pop Stars*. 2019. URL:

https://www.rollingstone.com/culture/culture-news/deepfakes-nonconsensual-porn-study-kpop-895605/.

[Fac19] Ctrl Shift Face. *Bill Hader channels Tom Cruise [DeepFake]*. 2019. URL: https://www.youtube.com/watch?v=VWrhRBb-1Ig.

[Gen08] GeninSider. *Letterman w/ Bill Hader 10/02/2008*. 2008. URL: https://www.youtube.com/watch?v=zS1Aee2X3Yc.

[Gol18] Ondřej Golis. *E-mailové lži. Předvolební dezinformace psali a šířili lékař, advokát, důchodce nebo farmář*. 2018. URL: https://www.irozhlas.cz/zpravy-domov/dezinformace-e-mail-prezident-volby-drahos-zeman_1802090600_ogo.

[Kni18] Will Knight. *The Defense Department has produced the first tools for catching deepfakes*. 2018. URL: https://www.technologyreview.com/s/611726/the-defense-department-has-produced-the-first-tools-for-catching-deepfakes/.

[Kor19] Patrik Kořenář. *HROZBA JMÉNEM DEEPFAKE*. 2019. URL: https://www.youtube.com/watch?v=RP9bKAHaaAo.

[Met16] Alec Radford Luke Metz. *UNSUPERVISED REPRESENTATION LEARNING WITH DEEP CONVOLUTIONAL GENERATIVE ADVERSARIAL NETWORKS*. 2016. URL: https://arxiv.org/pdf/1511.06434.pdf.

[Rad18] Petrana Radulovic. *Harrison Ford is the star of Solo: A Star Wars Story thanks to deepfake technology*. 2018. URL: https://www.polygon.com/2018/10/17/17989214/harrison-ford-solo-movie-deepfake-technology.

[Roc19] Joseph Rocca. *Understanding Generative Adversarial Networks (GANs)*. 2019. URL: https://towardsdatascience.com/understanding-generative-adversarial-networks-gans-cd6e4651a29.

[Sve18] PhD Sven Charleer. *Family fun with deepfakes. Or how I got my wife onto the Tonight Show*. 2018. URL: https://towardsdatascience.com/family-fun-with-deepfakes-or-how-i-got-my-wife-onto-the-tonight-show-a4454775c011.

[SW19] Marianna Spring a Lucy Webster. *European elections: How disinformation spread in Facebook groups*. 2019. URL: https://www.bbc.com/news/blogs-trending-48356351.

[Tur18] Dr. Matt Turek. *Media Forensics (MediFor)*. 2018. URL: https://www.darpa.mil/program/media-forensics.

[Ven19] Siddharth Venkataramakrishnan. *Definice dezinformací a propagandy*. 2019. URL: https://www.ft.com/content/4bf4277c-f527-11e9-a79c-bc9acae3b654.