Deepfake

Marcin Wilkowski

V roku 2019 bol na internete zverejnený videozáznam, na ktorom hovorí zakladateľ Facebooku Mark Zuckerberg. Varuje v ňom pred situáciou, keď má jedna osoba úplnú kontrolu nad údajmi a tajomstvami miliárd ľudí. Skutočným problémom je, že Zuckerberg nikdy nič také nepovedal.

Nahrávka trvá asi tak desať sekúnd. Zuckerberg sedí za stolom v béžovom tričku; je známe, že podobne ako ostatní miliardári zo Silicon Valley sa vyhýba noseniu oblekov a drahého oblečenia. Zuckerberg pri rozprávaní slovným dôrazom a gestami rúk upozorňuje na kľúčové slová: „ukradnuté údaje“, „kontrola“ a „tajomstvá“. Nahrávka však nebola svedectvom o revolúcii v pohľade na súkromie, ktoré Facebook premenil na tovar podliehajúci bezohľadnej komercializácii. Vytvorili ju umelci Bill Posters a Daniel Howe, členovia umeleckého kolektívu Big Dada, pomocou metód strojového učenia. Počítačový program založený na týchto metódach použil pôvodný záznam CBSN z roku 2017, ktorý slúžil ako podklad pre manipulovaný klip. Zuckerbergova tvár bola podrobená automatickému spracovaniu – algoritmus prispôsobil pohyby pier a výraz tváre slovám tohto nepravdivého vyhlásenia.

Big Dada / Public Faces (2019) [prístup 12. 03. 2021],K dispozícii na Bill Posters:http://billposters.ch/projects/big-dada/

↗ Približne trojminútový film s názvom Big Dada/Public Faces (2019), v ktorom sa prostredníctvom technológie deepfake objavujú Marcel Duchamp, Marina Abramovič, Mark Zuckerberg, Kim Kardashian, Morgan Freeman a Donald Trump.

Umelecká akcia kolektívu Big Dada zdôraznila rastúcu hrozbu nahrávok manipulovaných algoritmami strojového učenia.

Manipulácia s filmovými záznamami nie je ničím novým, veď dosiahnutie požadovaných efektov pomocou vhodných záberov alebo vystrihnutím záberov sa praktizovalo dávno predtým, ako sa film ako médium začal zaznamenávať a upravovať digitálne.

Skutočne hlboká manipulácia

„Deepfake“ je pomenovanie, ktoré sa skladá z dvoch častí. Slovo „Deep“ je odvodené od „hlbokého učenia“, ktoré v skratke umožňuje generovať nový obsah na základe súborov použitých ako šablóna, čo umožňuje algoritmu učiť sa. Napríklad milióny fotografií ľudských tvárí možno použiť na trénovanie algoritmu tak, aby bol schopný generovať úplne fiktívne obrázky. Podobne sa milióny textov stávajú základom pre algoritmy generovania prirodzeného jazyka a na základe týchto textov sa algoritmus učí vytvárať nový obsah. Tieto riešenia sa úspešne používajú napríklad pri softvéri na úpravu textu, ktorý navrhuje slová, a v aplikáciách, ktoré generujú krátke informačné texty o počasí, nehodách a výsledkoch obchodovania na burze.

Účinnosť takýchto algoritmov sa meria čo najväčšou prirodzenosťou umelo vytvoreného obsahu, bez ohľadu na to, či ide o text s informáciami o počasí, fotografiu, alebo film zobrazujúci panorámu krajiny, ktorý kombinuje pôvodné záznamy s pokročilým CGI (počítačom generované snímky).

Slovo „Fake“ zasa zdôrazňuje ciele používania takýchto riešení: v tomto prípade je výsledkom používania hlbokého učenia manipulácia a zavádzanie publika. Jeho použitie v hranom filme nie je nesprávne ako prvok fiktívnej tvorby prezentovaného sveta. Ak však ide o dokumentárny film, politický prejav alebo reportáž z podujatia, môže sa stať nástrojom propagandy alebo spôsobiť politické krízy či konflikty.

Deepfake technológia bola použitá napríklad proti ženám, ktorých snímky dostupné online boli vložené do pornografických nahrávok – obeťami neboli len hviezdy a celebrity, ale aj bežní používatelia sociálnych médií. Vytvorenie takéhoto niekoľkominútového klipu môže trvať viac ako 40 hodín, ale je možné ho vytvoriť na bežnom domácom počítači pomocou bezplatného softvéru bez potreby čokoľvek programovať.

Ako metóda manipulácie je preto technológia deepfake nebezpečná vzhľadom na prirodzenosť dosiahnutých efektov a logiku online komunikácie. Keď sme neustále zaplavovaní informáciami, textami a obrazmi, nemáme čas ani energiu skúmať ich pravosť. V prípadoch, keď je upravovanie obrazov bežnou praxou, nie je možné rozpoznať falšovanie v priebehu niekoľkých sekúnd. Terčom útokov pomocou tejto metódy sa môže stať napríklad medicína, pretože analýza obrazu je už základom na rozpoznávanie niektorých chorôb (napríklad rakoviny pľúc). V roku 2019 umožnil útok vykonaný na výskumné účely v jednom z lekárskych centier v Izraeli vytvorenie deepfake snímok pľúc pomocou počítačovej tomografie. Našťastie nikto neutrpel škodu, pretože „páchatelia“ vlámanie sami odhalili. Počas volebnej kampane v roku 2016 lekár Hillary Clintonovej sprístupnil časť jej zdravotnej dokumentácie na internete v reakcii na obvinenia, že zdravotný stav kandidátky jej nedovoľuje ujať sa úradu prezidenta USA. Ak by sa s lekárskymi záznamami manipulovalo, deepfake by mohol mať vplyv aj na politické procesy.

Deepfake v histórii

Vo vedeckej komunite sa v rámci výskumných projektov zameraných na rozšírenie možností práce s digitálnymi obrazmi vyvíjajú IT metódy na vytváranie hlboko manipulovaného obsahu. Niekedy môžu mať aj originálne vedecké využitie vrátane skúmania minulosti. V archeológii podporuje hĺbková strojová analýza satelitných snímok vyhľadávanie potenciálnych lokalít a rekonštrukciu poškodených predmetov, ako je keramika a mince.

Shira Faigenbaum-Golovin a ďalší vo svojej publikácii z roku 2016 preukázali užitočnosť strojového učenia pri analýze starovekých hebrejských textov napísaných na keramických fragmentoch okolo roku 600 pred n. l. Automatickým porovnávaním písmen z jednotlivých nápisov sa im podarilo dokázať, že texty napísalo šesť rôznych autorov.

V historických vedách sa hlboké učenie úspešne používa aj na rozpoznávanie rukopisov v stredovekých textoch a na dopĺňanie gréckych nápisov. Metódy CBIR (content-based image retrieval) umožňujú prehľadávať obrovské zbierky historického vizuálneho materiálu. Namiesto zadávania kľúčových slov ako v tradičných vyhľadávačoch môže používateľ vyhľadávať podľa obrázka.

Strojové učenie umožňuje aj extrakciu vybraných fráz vo veľkom množstve historických textov, napríklad automatické rozpoznávanie dátumov, miest, osôb a činností opísaných v digitalizovaných vojenských správach a „operačných rozkazoch“. Táto metóda sa nazýva NER (named-entity recognition). Talianski výskumníci použili strojové učenie na rozpoznávanie obrazov zničených historických pamiatok na starých filmových záznamoch, pričom predpokladali, že každá nová informácia o neexistujúcom objekte môže byť cenná a užitočná pri realizácii ich virtuálnej trojrozmernej rekonštrukcie.

Riešenia, ktoré stoja za deepfake manipuláciami, môžu slúžiť ako inšpirácia pre nový pohľad na audiovizuálne historické zdroje. Vlak prichádzajúci do stanice La Ciotat je zobrazený vo *Vlaku prichádzajúcom do stanice*, jednom z prvých nemých filmov, ktoré nakrútili bratia Lumièrovci koncom 19. storočia. Z historických dôvodov sú aj najlepšie digitalizované verzie tohto materiálu nekvalitné, ale pomocou strojového učenia sa záznam prepísal do rozlíšenia 4K, takže jeho detaily sú teraz lepšie viditeľné. V skutočnosti išlo o veľmi podobnú operáciu, aká sa používa pri manipulácii technológiou deepfake: záznam bol doplnený o nové pixely na základe dostupných knižníc umelej inteligencie. Ide teda skôr o reinterpretáciu, než o neutrálnu rekonštrukciu, čo je pre každú historickú štúdiu zásadný rozdiel. Podobne by sa malo pristupovať aj k čiernobielym historickým fotografiám, ktoré sa automaticky vyfarbujú pomocou softvéru na strojové učenie.

Andy Myers, *[Skenovanie v rozlíšení 4K] Príchod vlaku do stanice La Ciotat*, Bratia Lumièrovci, 1896, 2020 [prístup 09. apríla 2021]. Dostupné na YouTube: https://www.youtube.com/watch?v=1FAJ9fJQRZA

Bratia Lumièrovci zaznamenali príchod vlaku na stanicu La Ciotat v roku 1896. Vo februári 2020 bola zverejnená obnovená, digitálne spracovaná verzia záznamu vo vylepšenej kvalite v 4K rozlíšení.

Zodpovedný prístup k umelej inteligencii si vyžaduje aj otázky o etike vykonávaných činností. V pripravovanom filme *Finding Jack* (réžia Anton Ernst a Tati Golykh) sa objavia nové zábery s podobou Jamesa Deana zrekonštruovanou zo starých nahrávok. Prekračujú takéto projekty prijateľnú hranicu?

V projekte, ktorý realizuje Nadácia USC Shoah Foundation v Los Angeles v Kalifornii, sa svedkovia holokaustu stali modelom, na základe ktorého vznikli interaktívne hologramy: mechanizmy umelej inteligencie sa používajú nielen na oživenie virtuálnej postavy, vytvorenej obrazom z 52 kamier, ale aj na prípravu odpovedí na otázky, ktoré kladú účastníci rozhovoru.

Tento dynamický, zdanlivo prirodzený dialóg, ktorý tvorí jadro projektu, je založený na zozname tisícov otázok a odpovedí, ktoré boli špeciálne vypracované v súvislosti so životopismi svedkov a realitou šoa. Hoci autori zdôrazňujú, že hologram vychádza z autentických spomienok skutočného svedka, ktorý súhlasil s účasťou na

experimente, možno sa pýtať, či tu nie je priestor na určité zneužitie, najmä ak mechanizmus odpovede navrhuje výpovede, ktoré výrazne presahujú tie pôvodné.

USC Shoah Foundation, Nimrod „Zigi" Ariav, New Dimensions in Testimony, 2016 [prístup 09. apríla 2021]. Dostupné na stránke Nadácie USC Shoah: https://sfi.usc.edu/gallery/nimrod-zigi-ariav-new-dimensions-testimony- nov-3-2016

← Svedok holokaustu Nimrod „Zigi" Ariav nahráva svoje svedectvo v hebrejčine, Los Angeles, Kalifornia, apríl 2016.

Boj proti hrozbe deepfake spočíva v zlepšovaní metód jeho rozpoznávania. Je možné analyzovať drobné chyby a nezrovnalosti v usporiadaní pozadia, pohyboch postáv a výraze tváre a na programátorskej úrovni je teraz možné skúmať aj zmeny, ku ktorým dochádza medzi snímkami záznamu nasledujúcimi po sebe, a zistiť nezrovnalosti, ktoré naznačujú manipuláciu. Falzifikáty možno odhaliť aj pomocou štatistických metód, napríklad pomocou Benfordovho zákona. Tento zákon určuje pravdepodobnosť výskytu určitých prvých číslic v sérii čísel a používa sa už desaťročia napríklad pri odhaľovaní daňových podvodov. Ukázalo sa, že tieto zákonitosti možno nájsť aj v údajoch sociálnych médií a v štruktúrach počítačových súborov. Narušenie Benfordovej distribúcie môže naznačovať, že grafický súbor alebo videozáznam bol zmanipulovaný, aj keď sa jeho vizuálna vrstva zdá byť autentická.

Možno sa všetky tieto metódy a nástroje čoskoro stanú súčasťou metodológie historického výskumu a podporia prácu s digitálnymi zdrojmi, ktoré nemajú pôvodnú analógovú podobu a dokumentujú dôležité udalosti nedávnej histórie. Technickej analýze tohto typu zdroja by však určite malo predchádzať jeho kritické zhodnotenie: kontrola pôvodu materiálu, kontextu, v ktorom bol vytvorený, a jeho autorstva – takýto

prístup byť užitočný v prípade akéhokoľvek materiálu šíreného nezávisle prostredníctvom sociálnych médií.

Preklad: Ústav pamäti národa (ÚPN)