SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Detekcija uvjerljivog krivotvorenog sadržaja na društvenim mrežama uporabom GAN-ova

Barbara Kos, Matija Pavlović
Voditelj: *prof. dr. sc. Tomislav Hrkać*

Zagreb, siječanj 2024.

Detekcija uvjerljivog krivotvorenog sadržaja na društvenim mrežama uporabom GAN-ova

Sažetak

Ovaj seminarski rad predstavlja metodu detekcije uvjerljivog krivotvorenog sadržaja na društvenim mrežama uporabom GAN-ova. Prvo se detaljno razrađuje sam pojam uvjerljivog krivotvorenog sadržaja, njegova pojava i opasnosti koje proizlaze iz te pojave. Nadalje se definiraju osnovni pojmovi i koncepti GAN-ova, predlaže se implementacija sustava koji obavlja detekciju, razmatraju se prednosti i nedostatci implementiranog sustava u odnosu na druge načine detekcije. Naposljetku, iznose se i prijedlozi budućeg rada na projektu, moguća unaprjeđenja i iznose se zaključci o izvedivosti implementacije predloženog sustava u stvarnosti.

Ključne riječi: d eepfake, detekcija, GAN, strojno učenje, duboko učenje

Abstract

This seminar paper presents a method of detecting persuasive fake content on social networks using GANs. First, the notion of persuasive fake content itself, its phenomenon and the dangers arising from it are discussed in detail. Furthermore, the basic concepts and concepts of GANs are defined, the implementation of a detection system is proposed, the advantages and disadvantages of the implemented system compared to other detection methods are discussed. In conclusion, we discuss future work, improvements and feasability of a real world implementation.

Keywords: d eepfakes, detection, GAN, ML, deep learning

SADRŽAJ

1. Uvod

1.1. Pojava

Prva poznata pojava pojma "deepfake" datira iz prosinca 2017. godine kada je korisnik Reddita osnovao "subreddit" pod nazivom "r/deepfakes".Ovaj podforum uglavnom je sadržavao pornografski sadržaj u kojem su izmijenjena lica kako bi nalikovali poznatim osobama. Ovaj fenomen predstavlja tehnološki napredak, ali istovremeno izaziva zabrinutost zbog potencijalne zloupotrebe.

Takvi sadržaji često su prikazivali poznate i utjecajne osobe u situacijama koje se nikada nisu dogodile. Neki od poznatijih primjera uključuju papu Franju, bivšeg predsjednika SAD-a Donalda Trumpa te druge javne ličnosti. Ovaj trend je brzo stekao popularnost, često zbog senzacionalizma i šoka koji izaziva.

Unatoč negativnom kontekstu u kojem se često spominje stvaranje deepfake sadržaja, važno je napomenuti da postoji i pozitivan aspekt primjene ove tehnologije. Naime, deepfake tehnologija može se koristiti u edukativne svrhe, kao što je stvaranje videa u kojima poznate osobe, poput Davida Beckhama, podižu svijest o globalnim problemima poput malarije na različitim jezicima.

Osim toga, deepfake tehnologija nalazi primjenu u umjetnosti i zabavi, npr. u stvaranju scena u filmovima nakon smrti glumaca ili njihova digitalnog starenja. Također, postoji značajan broj deepfakeova čija je svrha isključivo humoristična, a takvi se sadržaji često viralno šire društvenim mrežama.

Važno je razumjeti da deepfake tehnologija nosi sa sobom i etičke izazove te da njezina primjena zahtijeva odgovornost kako bi se izbjegla potencijalna šteta i zloupotreba.

1.2. Opasnosti

Širenje *deepfake* tehnologije donosi mnoštvo potencijalnih opasnosti koje nadilaze početnu fascinaciju njezinim tehnološkim mogućnostima. Kako *deepfake*ovi postaju sofisticiraniji, rizici povezani s njihovom zlouporabom postaju sve izraženiji.

Jedna od glavnih prijetnji je dezinformacija koja proizlazi iz *deepfake*ova. *Deepfake* tehnologija stvara uvjerljiv lažan sadržaj izmišljanjem realističnih scenarija koji uključuju javne ili utjecajne osobe. To može dovesti do ozbiljnih posljedica kao što su širenje lažnih vijesti, manipuliranje javnim mišljenjem ili čak uplitanje u društvene i političke procese. Mogućnost prikazivanja pojedinaca kako govore ili rade stvari koje nikada nisu učinili ugrožava integritet informacija i dovodi u pitanje autentičnost i vjerodostojnost medija i vlasti.

Moguća opasnost je i potencijalni utjecaj na osobni i profesionalni ugled pojedinca. *Deepfake*ovi se mogu zloupotrijebiti za kreiranje obmanjujućih prikaza, što rezultira ozbiljnim društvenim posljedicama i oštećenju ugleda. Posebno su izložene riziku slavne osobe, političari i drugi javni pojedinci, budući da ih uvjerljivo manipulirani video materijali mogu prikazati u nestvarnim kontekstima koji mogu biti i skandalozni.

Opasnosti *deepfake* tehnologije proširuju se na pravnu domenu, izazivajući zabrinutost zbog mogućnosti manipulacije dokazima i potencijalnim narušavanjem integriteta kaznenopravnog sustava. *Deepfake*ovi mogu biti korišteni za lažiranje dokaza u pravnim procesima, što bi dovelo do sumnje u autentičnost video-dokaza i ozbiljno ugrozilo proces traženja istine unutar pravnog okvira.

Manipulacija emocijama i mišljenjima javnosti predstavlja značajnu opasnost koju donosi *deepfake* tehnologija. Stvaranje lažnih dojmova putem ovih manipulativnih sadržaja potiče ljude na donošenje odluka ili podržavanje ideja temeljenih na potpuno izmišljenim informacijama. Takva vrsta manipulacije može imati ozbiljne društvene posljedice, uključujući rast podrške pogrešnim političkim ili društvenim pokretima. Uz to, *deepfake*ovi mogu dovesti do organiziranog javnog djelovanja na temelju tih lažnih ideja i dojmova. U situacijama kada se koriste za manipuliranje masama, mogu potaknuti neželjene reakcije ili čak potaknuti nasilne događaje. Ova opasnost naglašava potrebu za aktivnim pristupom u detekciji i suzbijanju *deepfake*ova kako bi se očuvala društvena stabilnost.

2. Razrada

2.1. Metode stvaranja

Stvaranje *deepfake* sadržaja podrazumijeva primjenu sofisticiranih tehnika temeljenih na naprednom strojnom učenju, među kojima se ističu generativne kontradiktorne mreže (GAN). GAN-ovi igraju ključnu ulogu u ovom procesu, potičući suparničku dinamiku između generatora i diskriminatora, čime se postižu uvjerljive replikacije lica i pokreta te unaprjeđenje u stvaranju sintetičkog sadržaja.

Jedna od značajnih metoda unutar spektra stvaranja *deepfake* sadržaja je zamjena lica (engl. *face-swap*). Ova tehnika uključuje besprijekorno prenošenje karakterističnih crta lica jedne osobe na drugu u videozapisima ili slikama. Često implementiran putem GAN-ova, ovaj proces obuhvaća detekciju i poravnanje karakterističnih točaka lica kako bi se postigla precizna i neprimjetna zamjena lica.

Druga značajna metoda je *puppet-master* tehnika. Idući korak dalje od zamjene lica, ova tehnika manipulira pokretima cijelog tijela kako bi kontrolirala ponašanje ciljane osobe u videozapisu. Korištenjem algoritama dubokog učenja, ove tehnike analiziraju i repliciraju obrasce kretanja, obuhvaćajući karakteristične točke lica, procjenu položaja tijela i animaciju.

Ove metode, bilo da se temelje na GAN-ovima ili drugim naprednim arhitekturama dubokog učenja, naglašavaju prilagodljivost i složenost stvaranja *deepfake* sadržaja.

2.2. Artefakti

Artefakti su anomalije i nepravilnosti koje se pojavljuju u uvjerljivom krivotvorenom sadržaju. Najčešće se manifestiraju kao nepravilnosti u osvjetljenju i sjenama, nedostatak ili pogrešni detalji na području usta, ruku i očiju, razmazivanja rubova, promjena teksture površina(često na licima). U nekim slučajevima dolazi i do razlike u boji očiju. Sve ove pojave pomažu pri otkrivanju uvjerljivih krivotvorenog sadržaja,

golim okom ili pak nekom drugom metodom detekcije. Ukoliko razmatramo uvjerljivo krivotvorene videouratke česte su i inkonzistencije u pokretima, ne poklapanja zvučnih i vizualnih elemenata itd. Ukoliko artefakti nisu uočljivi na prvi pogled, a ipak su prisutni, promatraču mogu stvoriti osjećaj nelagode zbog razlike u očekivanom i percipiranom ponašanju, mimici i pokretima.

2.3. GAN-ovi

GAN-ovi su predstavljeni 2014. godine od strane Ian Goodfellowa i njegovih kolega. Osnovna ideja iza GAN sustava jest postojanje dva glavna dijela mreže: generator i diskriminator. Tijekom treninga, generator i diskriminator se natječu jedan protiv drugoga. Generator pokušava poboljšati svoje sposobnosti generiranja tako da vara diskriminator, dok diskriminator nastoji postati sve bolji u razlikovanju pravih podataka od lažnih. Ovaj suparnički proces dovodi do poboljšanja kvalitete generiranih podataka tijekom vremena.

2.3.1. Generator

Generator GAN-a je ključni dio sustava čija je zadaća stvaranje novih uzoraka ili podataka koji bi trebali biti što je moguće sličniji stvarnim primjerima iz skupa podataka na kojem je mreža trenirana. Na ulaz generatora dovodi se nasumični šum, a zatim se on izmjenjuje u izlaz koji nalikuje podatcima iz skupa za treniranje. Uvođenjem nasumičnog šuma te uzorkovanjem iz različitih točaka ciljne distribucije postižemo raznolikost generiranih podataka. Generator dakle kreira sadržaj suparničkim pristupom, pokušavajući prevariti diskriminator, poboljšava svoj izlaz iz iteracije u iteraciju.

2.3.2. Diskriminator

Diskriminator u okviru GAN-a se jednostavno može opisati kao klasifikator, čija je osnovna zadaća razlikovanje stvarnih podataka od onih koje generira sam generator. Proces treniranja diskriminatora uključuje podatke iz dva izvora, čime se postiže njegova sposobnost prepoznavanja stvarnih i generiranih podataka. Stvarni podaci, poput autentičnih slika ljudi, služe kao pozitivni primjeri tijekom učenja. Ovi primjeri omogućuju diskriminatoru da nauči karakteristike stvarnih podataka i stvori referentnu točku za usporedbu tijekom identifikacije lažnih podataka. S druge strane, negativni primjeri dolaze iz lažnih instanci podataka koje proizvodi generator. Diskriminator

tijekom treniranja razvija sposobnost razlikovanja autentičnih, stvarnih i generiranih, lažnih podataka. Ovaj proces suparništva potiče dinamički odnos između generatora i diskriminatora, čime se postiže postupna konvergencija prema visokokvalitetnim generiranim podacima.

2.3.3. Primjena GAN-ova za detekciju deepfake-ova

Ideja za primjenu dolazi iz činjenice da se upravo GAN-ovi često koriste kako bi se kreirali deepfakeovi i oni sami interno pokušavaju detektirati deepfakeove pomoću svog diskiminatora. U nastavku ovog rada ekstraktirati ćemo diskriminator već istreniranog GAN modela, te pomoću njega pokušati označiti uvjerljive lažne sadržaje.

3. Rezultati i rasprava

Rekreirali smo taj i taj rad, ostvarili te i te rezultate, možemo ih koristiti za to i to

4. Zaključak

Budući rad na ovom projektu uključuje izgradnju te treniranje vlastitog GAN modela, treniranje istog te ekstrakciju tako istreniranog diskriminatora. Ovako dobiveni diskriminator može biti iskorišten u raznim programskim rješenjima koja bi u stvarnom vremenu pregledavale sadržaje na društvenim mrežama, detektirale krivotvoreni sadržaj i upozoravale korisnike na pojavu deepfakeova. Smatramo da bi navedena primjena učinila društvene mreže sigurnijima, reducirala širenje dezinformacija i smanjila potencijal za .

5. Privitci