2019 / 2500M-0106
ISSN 1797-3457 (verkkojulkaisu)
ISBN (PDF) 978-951-663-083-3

TIIVISTELMÄRAPORTTI

SYNTEETTISEN MEDIAN TUNNISTUS KÄYTTÄEN GENERATIVE ADVERSARIAL NETWORKS -MALLIA

Ville Hautamäki, Janne Karttunen, Hannu Sillanpää ja Ivan Kukanov
Tietojenkäsittelytieen laitos, Itä-suomen yliopisto
PL 111, 80101 Joensuu
villeh@cs.uef.fi

Manipuloituja videoita on ollut olemassa Lumiere veljesten ajoista lähtien. Tähän asti kuitenkinsellaisten manipuloitujen videoiden tuottaminen, jotka voivat huijata katsojaa, on ollut aikaa vievää. Syvän generative adversarial network (GAN) mallinnuksen tuottaman dramaattisen parannuksen myötä aidon näköisten väärennettyjen videoiden tuottaminen on muuttunut todeksi. Tässä projektissa keskityimme ns. deepfake-videoihin, joissa lähdehenkilön kasvot vaihdetaan kohdehenkilön kanssa. Käsittääksemme turvallisuusnäkökulmasta katsottuna, deepfakeiden tunnistus tulisi nähdä ns. seulontatehtävänä. Esimerkiksi, opetettua mallia voisi suorittaa sosiaalisen median alustoilla, jolloin mallia ajettaisiin suureen määrään videoita päivittäin. Silloin on selvää, että vain pieni osa lähetetyistä videoista on deepfakeja, joten havaitsemisen suorituskykyä on mitattava kustannusherkällä (cost sensitive) tavalla. Tämän projektin tuloksena esittelemme menetelmän, jolla mallin parametreja voidaan estimoida kustannusherkällä tavalla.

1. Johdanto

Vain muutamassa vuodessa suuren yleisön ja tutkimusyhteisön huomio on keskittynyt deepfakejen vaaroihin. Deepfaket yleensä määritellään videoksi missä lähdehenkilön kasvot vaihdetaan kohdehenkilöiden kasvoihin. Tätä kutsutaan myös identiteetinvaihdoksi. Klassisessa biometrisen tunnistuksen kirjallisuudessa deepfake-hyökkäys kuuluu kategoriaan spoofing / representation attacks. Klassisesti, nällä hyökkäyksillä oli tarkoitus huijata automaattista biometrista tunnistusjärjestelmää, kuten kasvojen- tai puhujan tunnistinta. Deepfakeiden tapauksessa hyökkääjään tarkoituksena on pääsääntöisesti huijata ihmisiä koneiden sijasta. On helppo kuvitella tällaisen tekniikan sovelluksia, kuten kyseenalaisten videoiden levittäminen poliitikoista sosiaalisessa mediassa ennen vaaleja. Kiristysvideoiden tehtailu on myös mahdollista, sillä kun teknologia on periaatteessa kaikkien käsillä, voivat kiristyksen kohteena olla myös tavalliset kansalaiset.

Deepfakejen luomiseen on olemassa useita tapoja: Kasvojenvaihto vaihtaa ihmisen kasvot toisen henkilön kanssa videon kuva kerrallaan, huulisynkronointimenetelmät modifioivat videon suun liikkeet vastaamaan vaihdettua puhetta ja nukke-mestari menetelmät siirtävät liikkeet näyttelijästä kohdehenkilöön. Jotta vaihdettu kasvokuva olisi korkealaatuinen, se vaatii tehokasta kasvokuvien generointimallia. Tällaisia malleja ovat esimerkiksi GAN-mallit, kuten StyleGAN, FS-GAN. Ideana on, että uudet kasvot luodaan videon kuva kerrallaan. Tällöin samat lähdehenkilön ilmeet ja kasvojen asento generoituvat uudelle kohdehenkilön kasvokuvalle.

Mitä asialle voisi sitten tehdä? Pitäisikö lailla kieltää deepfake videoiden tehtailu ja levittäminen? Ongelmana on, että teknologialla on monia legitiimejä käyttökohteita, kuten viihdeteollisuus, mistä hyvänä esimerkkinä ovat YouTube:sta löytyvät deepfake videot. Näissä pyritään tekemään hauskoja videoita vaihtamalla esimerkiksi Sylvester Stallonen kasvot Arnold Schwarzeneggerin kasvojen paikalle. Näin katsoja voi arvioida miten Sylvester Stallone olisi suoriutunut Terminaattorin roolista. On siten vaikea kuvitella, että lainsäädännöllä voisimme ratkaista deepfake-teknologian aiheuttamat sosiaaliset ja kansalliseen turvallisuuteen liittyvät ongelmat.

Kaikki tämä tuo esiintarpeen deepfake-videoiden automaattiseen erotteluun aidoista videoista. Tässä projektissa keskitymme juuri tähän tehtävään, miten ja kuinka hyvin deepfake voidaan erottaa aidosta. Koska aidot videot ovat paljon yleisempiä kuin deepfaket, täytyy tämä ottaa huomioon tunnistamisessa. Myös automaattinen aidon videon merkitseminen deepfakeksi ei ole yhtä vaarallinen virhe kuin se että deepfake merkitään aidoksi. Suoratoistopalvelun moderaattori voi tarkistaa että näyttääkö deepfake merkintä todelliselta ja jos merkintä oli väärä niin tämän väärän päätöksen kustannus oli ainoastaan moderaattorin käyttämä työaika. Mutta deepfaken merkitseminen aidoksi videoksi voi aiheuttaa merkittäviä yhteiskunnallisia kustannuksia jos video pääsee leviämään sosiaalisessa mediassa. Täten kustannusten asymmetria tulisi ottaa huomioon automaattisessa päätöksenteossa. Juuri tähän aiheeseen työmme keskittyi.

Kuva 1. Faceswap-GAN -ohjelmistolla luotuja Deepfake-kuvia. Kuvaparien vasemmalla puolella alkuperäinen kuva ja oikealla väärennös.

Muutamia aineistoja on saatavissa deepfake-tunnistimien kehittämiseen ja testaamiseen. Julkisesti saatavilla olevia DeepfakeTIMIT ja Faceforensics++ (FF++):n deepfake -alajoukkoa on käytetty laajasti tunnistinten opettamiseen ja arviointiin. On uskottavaa että näillä aineistoilla opetettu tunnistin ei välttämättä pysty tunnistamaan deepfakeja jotka on tehty jollain toisella menetelmällä. Toinen menetelmä voi olla myös sama matemaattinen malli mutta hieman eri parametrisoinnilla.

2. Tutkimuksen tavoite ja suunnitelma

Pääasiallinen kysymys on että kuinka hyvin pystymme tunnistamaan deepfakeja yleisesti ja erityisesti sellaisia missä generointimenetelmä ei ole ollut tiedossa tunnistimen opetusvaiheessa. Oletamme että todellisessa käyttötapauksessa juuri uudentyyppisiä deepfakeja tulisimme prosessoimaan.

Toinen kysymys on että voimmeko opettaa mallin operoimaan halutussa kahden virheen (deepfake tunnistuu aidoksi, ja aito tunnistuu deepfakeksi) kompromississa.

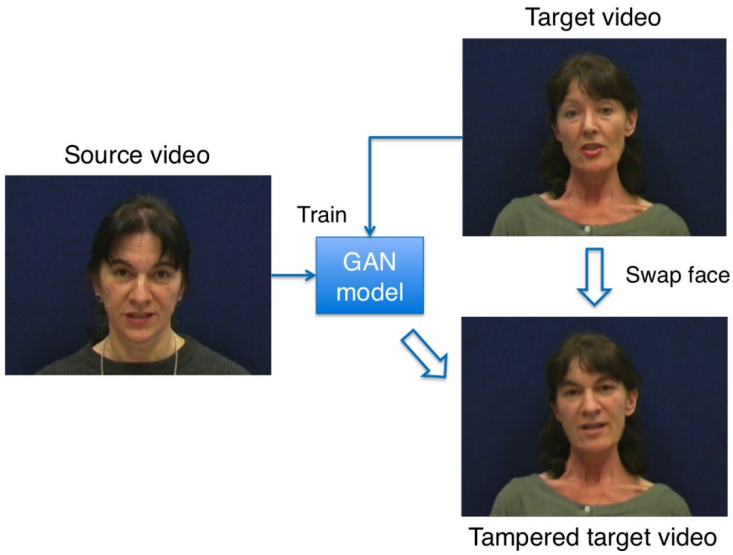
Suunnitelma koostuu kahdesta vaiheesta (kevät ja syksy):

1. Etsiä sopiva aineisto deepfake tunnistimen opettamiseen ja testaamiseen.
    a. Tuotamme itse deepfakeja
    b. Etsimme muiden tekemiä aineistoja
2. Opetamme deepfake tunnistimen kerätyllä aineistolla ja testaamme sitä.
    a. Koska oletettavasti uudet hyökkäykset mitä ei ole nähty opetusvaiheessa tulevat olemaan vaikeita, yritämme ratkaista ongelmaa GAN mallilla. GAN malli koostuu kahdesta osasta: generaattori ja diskriminaattori. Diskriminaattori tuottaa arvion siitä kuinka varmasti kuva on synteettinen. Kokeilemme käyttää tätä.
    b. Kokeilemme erillaisia moderneja malleja tunnistamaan deepfaket.
    c. Suunnittelemme mallin opetuksessa käytettävän kustannusfunktion missä käyttäjä voi säätää sitä kuinka tärkeäksi hän kokee jomman kumman tunnistusvirheen. Testaamme myös malleja tällä valitulla virhemittarilla (siis missä virheiden asymmetria on eksplisiittinen).

3. Aineisto ja menetelmät

Tutkimuksen tavoitteena oli kehittää menetelmiä deepfake-videoiden tunnistamiseen. Tämä tutkimustyö tarvitsee sekä opetusaineistoa että myös testausaineistoa. Opetimme mallejamme yhdistetyillä DeepfakeTIMIT- ja FF ++ -aineistoilla. Keräsimme myös YouTube-sivustolta useita viihdetarkoituksiin tehtyjä deepfake-videoita. Tällä uudella joukolla testaamme sitä, pystyykö opetettu malli tunnistamaan hyökkäyksen mitä se ei opetusvaiheessa ole nähnyt. Molemmat aineistot olemme julkaisseet ryhmämme web-sivuilla [1].

[1] http://cs.uef.fi/deepfake_dataset/



Kuva 2. DeepfakeTIMIT-datasetin videoiden luontiprosessi.

Käytimme tutkimuksessa opetusaineistona FaceForensics++, Deepfake-TIMIT ja VidTIMIT datasettejä. FaceForensics++ koostuu 1000:sta lähdevideosta, joista on luotu väärennettyjä versioita eri menetelmillä. Tässä tutkimuksessa käytimme deepfake-menetelmällä luotuja videoita. DeepfakeTIMIT koostuu 320 väärennetystä videosta, jotka on luotu opetusdatassa myös mukana olevista VidTIMIT-datasetin lähdevideoista.

Koulutettujen mallien arviointia varten keräsimme erillisen datasetin YouTube-videopalvelusta löytyvistä deepfake-videoista. Nämä videot ovat harrastelijoiden luomia ja tarkoitettu viihdekäyttöön, joten monissa niistä on kiinnitetty erityistä huomiota kuvanlaatuun. Videoissa käytetyt menetelmät eivät ole täysin samanlaisia opetusdatan menetelmien kanssa, joten tunnistustarkkuuden odotetaan olevan huonompi, kuten aiemmissa tutkimuksissa on huomattu [2].

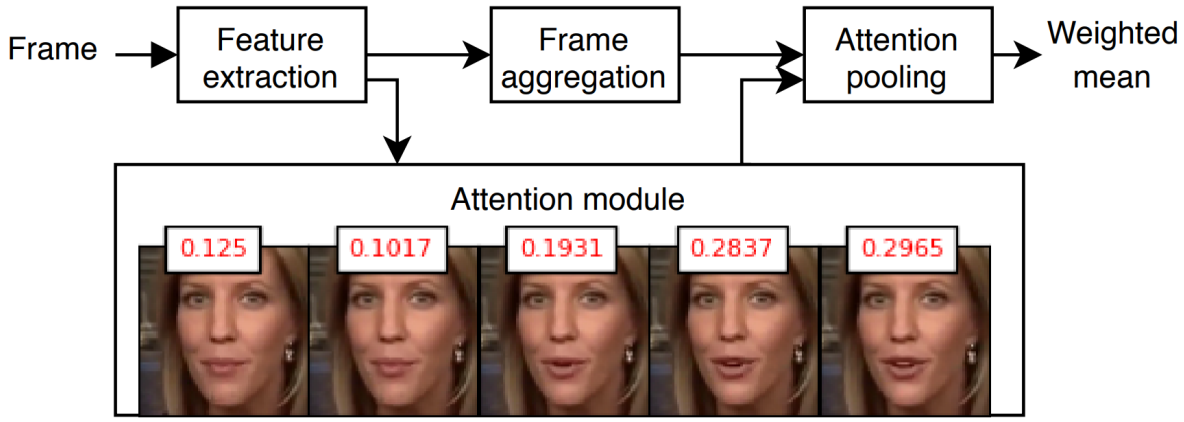
[2] Y. Li, X. Yang, P. Sun, H. Qi, et al., “Celeb-df: A new dataset for deepfake forensics,” arXiv:1909.12962, 2019.

Kuva 3. Esimerkkiotoksia kerätystä aineistosta. Yläreunassa väärennöksiä ja alarivillä aitoja kuvia.

Kerätty aineisto koostuu 79 Deepfake-videosta, joissa on yhteensä 98 dokumentoitua kasvonvaihdosta. Datasettiin on merkattu henkilöiden nimi, sijainti videolla ja videon leikkauskohdat ja metatiedot. Vastapainoksi väärennetyille videoille otimme aineistoon mukaan 98 aitoa videota VoxCeleb2-kokoelmasta.

Vertailukohtamenetelminä käytimme CNN ja LSTM -pohjaisia neuroverkkoja. CNN toteutuksena käytimme MobileNet:iä ja LSTM:n toteutimme tutkimuksen [3] kuvaukseen pohjautuen. Kehitimme deepfake videoille attentive pooling -menetelmän, joka pyrkii painottamaan tuloksessa videon eniten merkitseviä kohtia. Motivaatio tämän menetelmän käytölle on se, että nykyisten deepfake-videoiden luomiseen käytetyt menetelmät käsittelevät jokaista videon kuvaa erikseen ja eristyksissä toisistaan. Tämä voi johtaa siihen että väärennöksen laatu vaihtelee merkittävästi videon eri ajankohdissa. Kuvassa 4 esittelemme menetelmän rakenteen ja esimerkin painotuksesta. On oleellista ymmärtää että malli keksii itse kuville nämä painot, niitä ei näytetä missään vaiheessa opetusta menetelmälle. Opetusvaiheessa menetelmälle kerrotaan mikä video on deepfake ja mikä on aito.

[3] D. Gera and E. J. Delp, “Deepfake video detection using recurrent neural networks,” in AVSS, 2018.



Kuva 4. Attentive pooling diagrammi ja esimerkki yksittäisten kuvien painotuksesta.

Kompromissi kahden eri virheen välillä

Virhe false alarm (FA), kertoo että aito tunnistettiin deepfakeksi, ja vastaavasti miss (myös tunnetaan nimellä false reject (FR)) kertoo että deepfake tunnistettiin aidoksi. Tunnistusvirhettä voidaan sitten mitata *decision cost function* (DCF):lla:

$$C_{\mathrm{DCF}}(t) = C_{\mathrm{miss}} \cdot P_{\mathrm{tar}} \cdot P_{\mathrm{miss}}(t) + \\ + C_{\mathrm{fa}} \cdot (1 - P_{\mathrm{tar}}) \cdot P_{\mathrm{fa}}(t),$$

missä DCF on kynnysarvon funktion, eli malli tuottaa reaaliluvun, missä suurempi luku tarkoittaa todennäköisemmin kyseessä on deepfake ja pieniluku tarkoittaa että kyseessä on aito video. Virheiden lukumäärät voidaan siten laskea kynnysarvon käytön jälkeen. Parametrit, C_{miss} ja C_{fa} kertovat virheen kalleudesta, tässä voidaan käyttää vaikkapa euroja mittarina ja P_{tar} kertoo arvatun (priori) todennäköisyyden sille että havaittu video on oikeasti deepfake. On uskottavaa että deepfake videot ovat suhteellisen harvinaisia, joten asettamalla haluttu P_{tar} voidaan testata menetelmää tällä halutulla uskomuksella. On huomattavaa että DCF on yliparametrisoitu, eli DCF voidaan aina kirjoittaa yhden parametrin funktiona. Siksipä tässä työssä yksinkertaistimme esitystä ja pidimme kustannus parametrit ykkösinä ja muuttelemme ainostaan P_{tar}:ia.

Toinen yleisesti käytetty kahden luokan luokittelumittari on *equal error rate* (EER), mikä on yksi (teoreettinen) kynnysarvo DCF:llä siten että P_{miss} *ja* P_{fa} ovat yhtä suuret. Tämä on yleensä käyttökelpoinen tapa mitata mallin toimivuutta, koska mittari tuottaa virhearvion prosentteina. Lisäksi jos halutaan tarkastella mallin toimivuutta kaikilla kynnysarvoilla, niin on järkevää piirtää niin kutsuttu *detection error tradeoff* (DET) piirros. Tämä on muunnos ROC piirroksesta, missä viiva on täysin suora ja diagonaalinen jos tunnistustulosten jakaumat ovat normaalit. DET -piirroksen etu ROC -piirrokseen on että menetelmien väliset erot tulevat helpommin esille.

Sovelsimme tässä esityksessä *Maximum Figure-of-Merit* (MFoM)[4] menetelmää DCF ja EER mittareiden suoraan optimointiin. On huomion arvoista, että DCF -mittari vaatii aina kynnystyksen, joten gradientti -pohjaisilla menetelmillä optimointi on mahdotonta. MFoM on kätevä yleinen suunnitteluperiaate missä mikä tahansa mittari voidaan pehmentää ja sitten käyttää neuroverkon parametrien estimoinnissa. Kirjoittamassamme ICASSP 2020 konfferenssiin lähetetyssä artikkelissa on kaikki tarvittavat detaljit esitelty. Julkaisemme myöhemmin teknisen raportin ArXiv palvelussa tästä artikkelista.

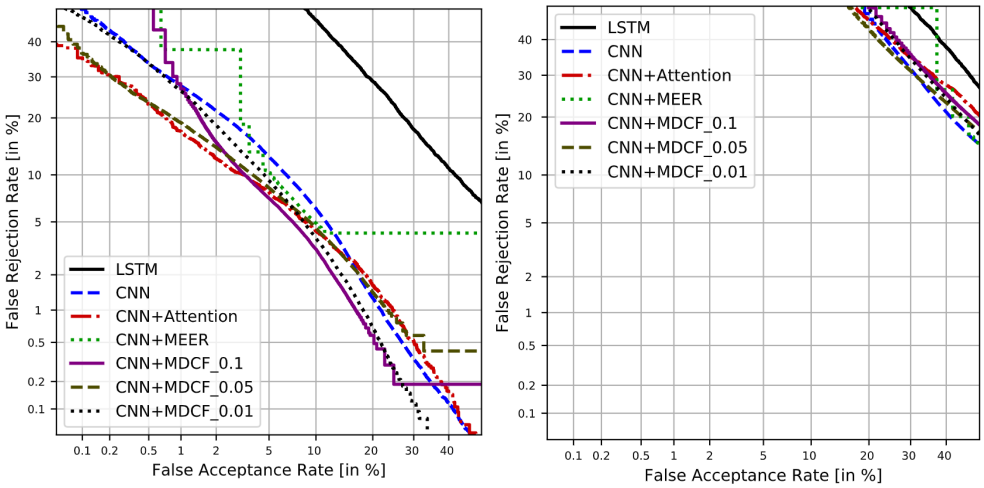
4. Tulokset ja pohdinta

Method	EER	minDCF with P_{tar}			Eval
		0.1	0.05	0.01	EER
LSTM [14]	24.1	0.88	0.92	0.96	38.90
CNN	8.07	0.37	0.43	0.60	30.80
CNN+Attention	6.55	**0.26**	**0.32**	**0.43**	32.90
CNN+MEER	7.16	0.44	0.50	1.00	32.32
CNN+MDCF_0.1	**6.03**	0.33	0.46	0.88	32.49
CNN+MDCF_0.05	6.67	0.28	**0.32**	0.46	**30.56**
CNN+MDCF_0.01	6.72	0.35	0.43	0.56	32.09

Taulukko 1. Testattujen tunnistusmenetelmien suorituskyky opetusdatan testiosuudella sekä kerätyllä arviointimateriaalilla.

Tunnistusmenetelmien arvioinnin tulokset on listattu taulukossa 1. Emme päässeet LSTM-menetelmää käyttäen yhtä hyviin tuloksiin kuin mallia aiemmin testanneessa tutkimuksessa [5], mikä johtunee eri aineistojen käytöstä tai eroista ohjelmistojen toteutuksissa. CNN:llä pääsimme 8,07 % EER:ään ja tulos parantui 6,55 %:iin attentive pooling -menetelmän ansiosta. CNN-menetelmän tulokset parantuivat myös MFoM-menetelmillä. Taulukon oikeanpuoleisessa sarakkeessa on tunnistusmenetelmien tulokset kerätyllä arviointimateriaalilla. Odotetusti tulokset huonontuivat paljon, sillä väärennösten laatu on parempi kyseisessä aineistossa. Tuloksia lähemmin tarkastelemalla huomattiin, että matalampilaatuiset deepfaket tunnistettiin oikein, mutta suurin osa aineiston videoista oli parempilaatuisia kuin opetusmateriaalin aineisto.

[4] S. Gao, W. Wu, C.-H. Lee, and T.-S. Chua, “A maximal figure-of-merit (MFoM)-learning approach to robust classifier design for text categorization,” ACM Trans. on Inf. Syst., vol. 24, 2006.
I. Kukanov, V. Hautamäki and K.A. Lee," Maximal Figure-of-Merit Embedding for Multi-label Audio Classification", ICASSP 2018.
[5] D. Gera and E. J. Delp, “Deepfake video detection using recurrent neural networks,” in AVSS, 2018.



Kuva 4. Detection error tradeoff (DET) -kuvaajat tunnistusmenetelmien tarkkuudesta. Vasemmalla opetusdatan testiosuus ja oikealla arviointiaineiston tulokset.

5. Loppupäätelmät

Ottaen huomioon että vuoden alussa lähdimme tilanteesta missä meillä ei ollut aineistoja, eikä menetelmiä testattavaksi saimme aikaan uskottavat testit kirjallisuudessa esitetyille menetelmille (LSTM ja MobileNet). Myös kehitimme omia menetelmiä attentive pooling ja MFoM deepfakeiden tunnistukseen. Testien valossa näyttää siltä että kummatkin menetelmät toimivat hyvin, seuraavaksi kokeilemmekin näiden menetelmien yhdistelmää. Keräsimme myös oman testiaineiston YouTubesta.Tällä aineistolla itse kehittämämme menetelmät toimivat parhaiten, mutta tulos on edelleen kehno. Tämä vastaa kirjallisuudessa esitettyjä tuloksia ja näyttää, että puhtaasti ohjattuun oppimiseen perustuvat mallit eivät ole hyviä deepfakeiden tunnistukseen. Vastaavaa viestiä on kuultu myös DARPA:n jo päättyneestä MediaFor -ohjelmasta (tulokset julkaistiin ICML 2019 konferenssissa).

Voisiko lainsäätäjä tehdä jotain deepfakeiden tuottamien ongelmien minimoimiseksi? Kuten johdannossa totesimme, kieltäminen ei liene mahdollista tahi järkevää. Mahdollisesti lainsäätäjä voisi miettiä voisiko deepfaken käyttämisestä rikollisessa tarkoituksessa säätää kovempi rangaistus. Hieman samaan tapaan kuin väkivaltarikoksissa jos tekijä on motivoitunut uskontoon kohdistuneen vihan tai rasismin perusteella, voidaan rangaistusta koventaa. Lainsäädännöllä voitaisiin myös velvoittaa streaming palvelut käyttämään deepfakeiden tunnistus ohjelmistoja.

Tutkimuksen seuraavat askeleet ovat selvät. Kansainvälisellä tasolla mielenkiinto deepfakeiden tunnistukseen on suuressa kasvussa. DARPA aloittaa kesällä 2020 SemaFor ohjelman, ja Facebook organisoi keväällä 2020 Deepfake Challenge kilpailun, missä on jaossa yhteensä 1M USD palkintorahoja. Kummassakin hankkeessa on tavoitteena merkittävästi parantaa deepfakeiden tunnistuksen tarkkuutta. Kuten tuloksistamme huomaa, nykyisillä menetelmillä ennalta tuntemattoman hyökkäyksen tunnistaminen on vaikeaa. On ilmiselvää että sekä DARPA:n että Facebookin hankkeessa juuri tähän kansainvälisellä tasolla tullaan

keskittymään.

Otamme itsekin keväällä 2020 askeleet tähän suuntaan. Suunnittelemme syväoppimiseen perustuvaa tilastollista mallia, missä pyritään mallintamaan pääsääntöisesti aitoja videoita. Tämä erona nykyisiin menetelmiin missä mallinnetaan aitojen ja deepfakeiden eroa. Lisäämällä aineistoon pieni määrä tunnettuja hyökkäyksiä pyrimme sitten hyvän mallin aidoille videoille. Näin malli näkisi deepfaket selvästi erillaisina aitoihin nähden. Olemme käyttäneet jo tällaista mallia bioinformatiikan kontekstissa, joten toiveena on että malli toimisi hyvin myös deepfakeille.

6. Tutkimuksen tuottamat tieteelliset julkaisut ja muut mahdolliset raportit

Tutkimuksesta on lähetetty raportti tarkastettavaksi ICASSP 2020 konferenssiin. Tavoitteena on lähettää tammi-helmikuussa 2020 laajempi artikkeli lehteen projektin tuloksista. Myös Hannu Sillanpään gradu tullaan julkaisemaan keväällä 2020.

Projektin aikana myös tuotettiin seuraavat julkaisut:

Trung Ngo Trong, Roger Kramer, Juha Mehtonen, Gerardo Gonzalez, Ville Hautamäki, Merja Heinäniemi, "Semi-Supervised generative Autoencoder for single cell data". Journal of Computational Biology, 2019.

Trung Ngo Trong, Kristiina Jokinen, Ville Hautamäki, "Enabling Spoken Dialogue Systems for Low-Resourced Languages -- End-to-End Dialect Recognition for North Sami" In: D'Haro L., Banchs R., Li H. (eds) 9th International Workshop on Spoken Dialogue System Technology. Lecture Notes in Electrical Engineering, vol 579. Springer, Singapore, 2019.

Bilal Soomro, Anssi Kanervisto, Trung Ngo Trong and Ville Hautamäki, "Towards Debugging Deep Neural Networks by Generating Speech Utterances", Interspeech 2019.