Synteettisen median tunnistus

MATINE tutkimusseminaari

FT Ville Hautamäki (UEF)

21.11.2019

Tutkimusryhmä

PhD Ville Hautamäki, senior researcher

PhD Abraham Woubie, postdoc

- Acoustic and visual deep reinforcement learning

MSc Ivan Kukanov, phd student

- Deep multilabel classification theory
- Graduation 2020

MSc Trung Ngo Trong, phd student

- Deep Bayesian semi-supervised learning
- Graduation 2020

MSc Anssi Kanervisto, phd student

- Deep reinforcement learning for computer games
- Graduation 2021

MSc Janne Karttunen, project researcher

- Sim2Real transfer learning
- Deepfake detection

BSc Hannu Sillanpää, project researcher, msc student

- Deepfake detection

BSc Keishi Ishihara, msc student

- Deep reinforcement learning for computer games

Motivaatio

- Syväoppiminen on mahdollistanut videoiden väärentämisen siten, että videoon voidaan muokata toisen henkilön kasvot
- Menetelmät kehittyvät jatkuvasti paremmiksi ja ne ovat saatavilla tavallisille käyttäjille yhä helpommin
- Deepfake -menetelmillä tuotetut videot voivat olla mm. poliittinen, taloudellinen ja oikeudellinen uhka.

THE WALL STREET JOURNAL.

SUBSCRIBE SIGN IN

PRO CYBER NEWS

Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

Scams using artificial intelligence are a new challenge for companies

By Catherine Stupp

Updated Aug. 30, 2019 12:52 pm ET

SHARE AA TEXT

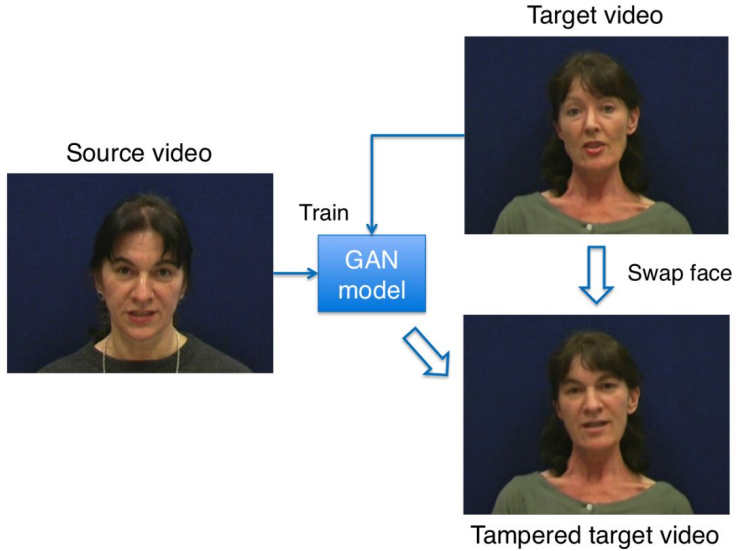
Criminals used artificial intelligence-based software to impersonate a chief executive's voice and demand a fraudulent transfer of €220,000 ($243,000) in March in what cybercrime experts described as an unusual case of artificial intelligence being used in hacking.

Tutkimuksen tavoite

- Kerätä deepfake -materiaalia tilastollisesti merkittävien kokeiden ajamista varten
- Käyttää kerättyä materiaalia kehittämään menetelmä, mikä tunnistaa deepfake -väärennökset oikeista videoista
- On selvää että on olemassa kaksi päätösvirhettä: *deepfake tunnistetaan aidoksi* tai *aito tunnistetaan deepfakeksi*
    - Nämä virheet eivät ole yhtä tärkeitä
    - Opetetaan malli suoraan huomioimaan virheiden tärkeys ero

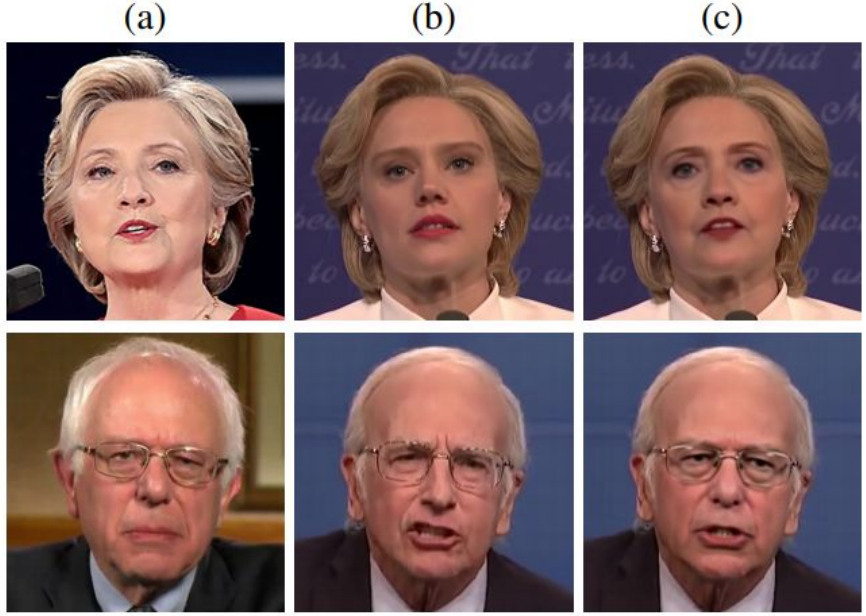
Tutkimus: **Vulnerability assessment and detection of Deepfake videos**, *Pavel Korshunov and Sebastien Marcel* (2019)

- Tutkimuksessa luotiin 620 face-swap videota GANin (*generative adversarial network*) avulla
- Näillä deepfake-videoilla onnistuttiin huijaamaan viimeisimpiä kasvojentunnistusjärjestelmiä (VGG ja Facenet) jopa 95% tapauksista
- Tukivektorikoneella (*support vector machine*) videoista pystyttiin tunnistamaan väärennöksiksi n. 91%



Tutkimus: **Protecting World Leaders Against Deep Fakes**, *Agarwal et al.* (2019)

- Tunnistettiin poliitikoista tehtyjä väärennettyjä videoita kasvojen ilmeiden ja pään liikkeen perusteella
- Jokaista poliitikkoa varten opetettiin oma malli, joka keskimäärin 92% tapauksista tunnisti väärennöksen
- Menetelmä vaatii kuitenkin paljon opetusdataa kohdehenkilöstä ja erikseen häntä varten opetetun mallin



(a) alkuperäinen kuva, (b) imitaattori, (c) väärennös

1. Deepfake-generointi

- Projektin alkupuolella kokeiltu deepfake-generointia
- Datasetin luominen generoimalla sivuutettu
  - Laadukkaan deepfake-mallin luonti yhdelle henkilöparille vie viikon laskenta-aikaa ja mahdollisesti lisäaikaa hienosäädölle
  - Tulos vaihtelee henkilöstä riippuen
  - Helpompaa käyttää valmista materiaalia
  - FaceForensics++: 1000 henkilöä (kevyemmät mallit, enemmän resursseja?)

Faceswap-GAN

1. Deepfake-generointi

DeepFaceLab

Adversarially Learned Inference

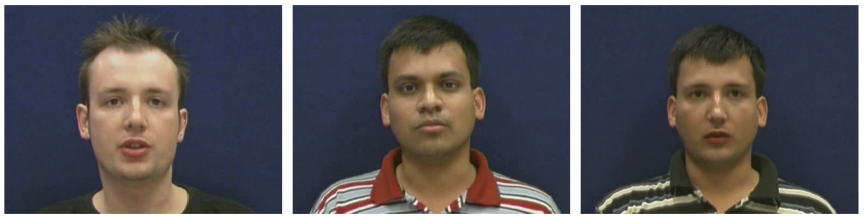
CycleGAN

Faceswap-GAN

2.1 Julkisesti käytössä oleva data

- Koostimme opetusdataksi julkisesti saatavilla olevat FaceForensics++ ja DeepfakeTIMIT datasetit
  - FaceForensics++: 1000 videota (500k kuvaa)
  - DeepfakeTIMIT: 320 videota (35k kuvaa)
- Data jaettu opettamiseen (72%), validointiin (14%) ja testaukseen (14%)

FaceForensics++



(a) Original Donor (b) Original Target (c) Face Swapped

DeepfakeTIMIT

2.2 Kerätty data

- Todenmukaista evaluointia varten keräsimme YouTubessa olevista deepfake-videoista datasetin
- 79 videota, joissa yhteensä 98 kasvonvaihdosta
    - Henkilöiden identiteetti, sijainti videolla ja videoiden leikkaukset merkattu
- 98 aitoa videota kerätty VoxCeleb-datasetistä

3. Deepfake-tunnistin

- Tavoitteena kehittää deepfake-videoiden automaattisen tunnistuksen tarkkuutta
- Vertailtu seuraavia menetelmiä:

  3.1 Convolutional Neural Network (CNN)

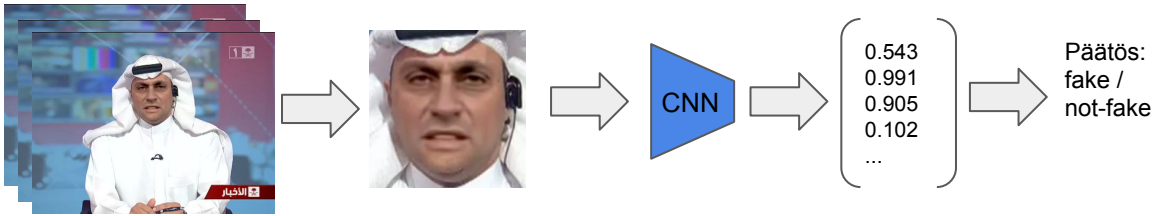
  3.2 Long short-term memory (LSTM)

  3.3 Attentive pooling

  3.4 Maximal Figure-of-Merit (MFoM)

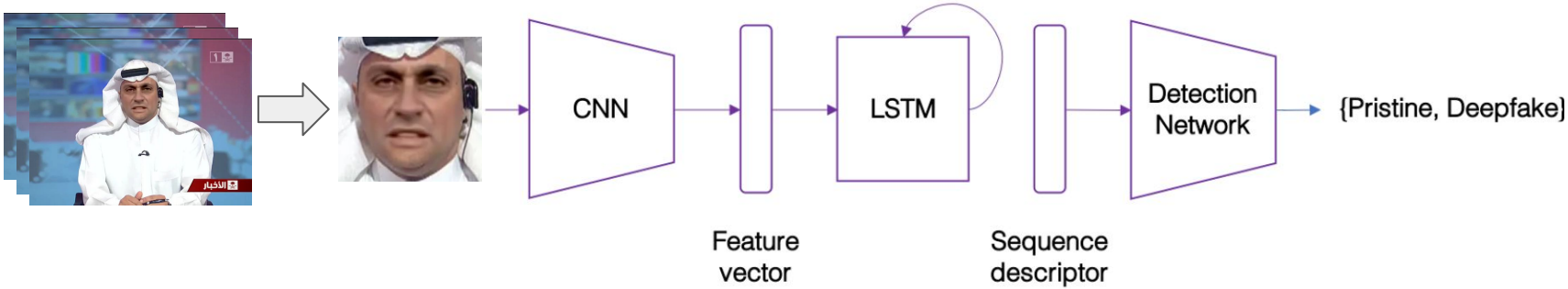
3.1 Convolutional Neural Network (CNN)

- CNN on yleisesti käytetty menetelmä kuviin liittyvässä koneoppimisessa
- Neuroverkko oppii luokittelemaan kuvia näkemänsä opetusmateriaalin perusteella
- Tehtävän helpottamiseksi kuva rajataan kasvojen alueelle kasvojentunnistusmenetelmillä



3.2 Long short-term memory (LSTM)

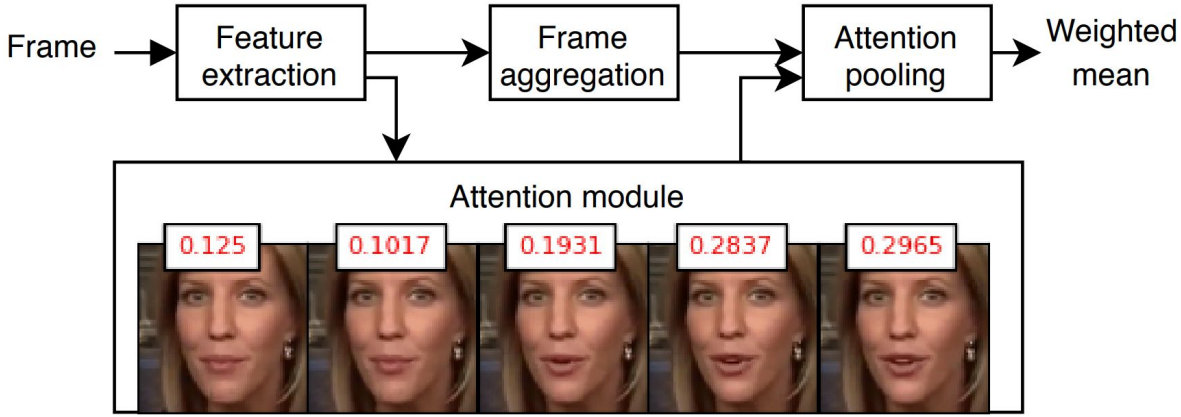
- Vertailukohtana käytettiin aiemmassa tutkimuksessa (Guera & Delp, 2018) käytettyä LSTM-verkkoa (*Long short term memory*)
- CNN-osuus on toteutettu InceptionV3-verkolla ja sen painoarvot ladataan valmiiksi Imagenetillä opetetusta mallista. Painoja ei muuteta treenatessa.
- Verkolle syötetään kerrallaan useita ruutuja, joten teoriassa se voi tunnistaa deepfake-videon aikariippuvaisista ominaisuuksia



Kuva mukaillen: **Deepfake Video Detection Using Recurrent Neural Networks**, *David Guera & Edward J. Delp,* (2018)

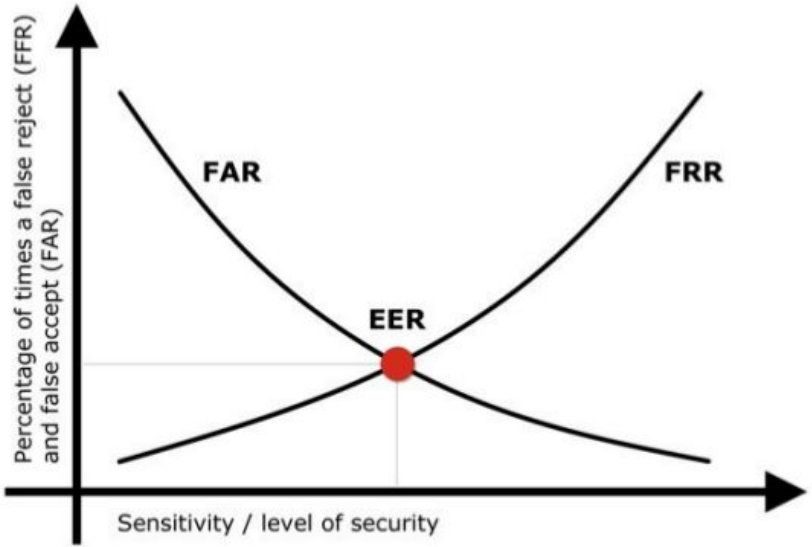
3.3 Attentive pooling

- Yleensä kuvien prediktiot keskiarvostetaan koko videon yli. Kaikki kuvat eivät välttämättä ole yhtä merkittäviä.
- Tutkimme attentive pooling -menetelmää, missä erillinen neuroverkko tuottaa painon jokaiselle kuvalle.



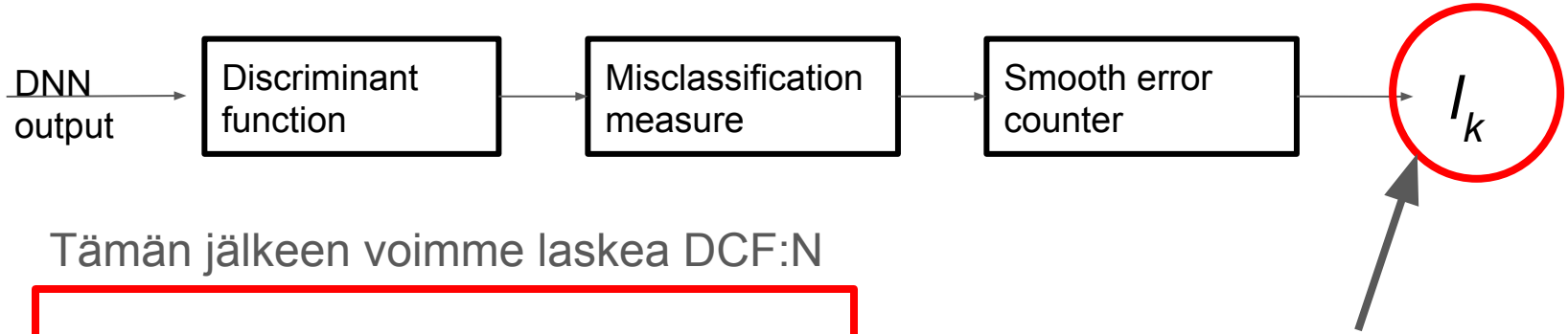
3.4 Maximal Figure-of-Merit (MFoM)

- Tunnistusvirheet FA (aito tunnistettiin deepfake:ksi) ja FR / miss (deepfake tunnistettiin aidoksi), ovat kompromississa keskenään.
- Virheet eivät myöskään ole yhtä tärkeitä!
- Valitsemalla sopiva painotus, valitsemme yhden pisteen ROC / DET piirroksesta.
- Merkittävin näistä on EER (FAR = FRR).
- NIST DCF on mittari missä voimme säätää painotusta.



Maximal Figure-of-Merit (MFoM) auttaa optimoinnissa

- Koostuu kolmesta osasta:



- Tämän jälkeen voimme laskea DCF:N

$$\hat{P}_{\text{miss}} \triangleq \frac{\sum_{k=1}^{M} FN_k}{P} = \frac{\sum_{k=1}^{M} \sum_{\mathbf{X} \in \mathbb{T}} l_k \cdot y_k}{P},$$

$$\hat{P}_{\text{fa}} \triangleq \frac{\sum_{k=1}^{M} FP_k}{N} = \frac{\sum_{k=1}^{M} \sum_{\mathbf{X} \in \mathbb{T}} (1 - l_k) \cdot \overline{y}_k}{N},$$

MFoM:n hyödyntäminen verkon opettamisessa

- Opetetaan verkko käyttäen ristientropiaa (cross-entropy, CE)
- Valitse haluttu virheiden painotus
- Sijoita “smooth error counter” osaksi verkon arkkitehtuuria (lineaarinen kerros)
  - parametrit voidaan opettaa backpropilla
- Jatka verkon opettamista MFoM:llä ja valitulla painotuksella.

Tulokset

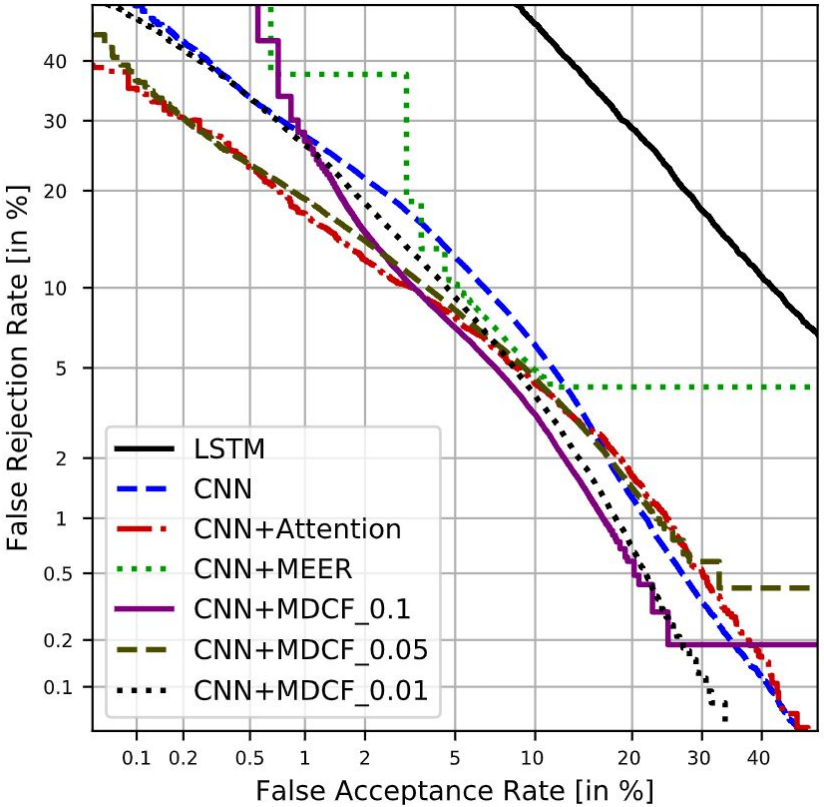
Test set

YouTube set

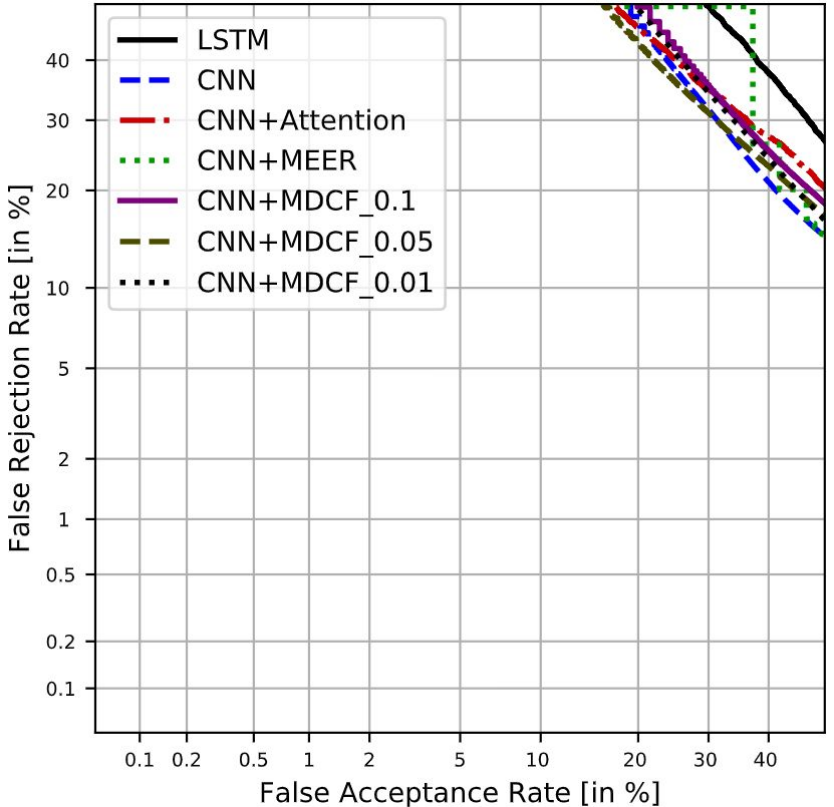
Method	EER	minDCF with P_{tar} 0.1	0.05	0.01	Eval EER
LSTM [14]	24.1	0.88	0.92	0.96	38.90
CNN	8.07	0.37	0.43	0.60	30.80
CNN+Attention	6.55	**0.26**	**0.32**	**0.43**	32.90
CNN+MEER	7.16	0.44	0.50	1.00	32.32
CNN+MDCF_0.1	**6.03**	0.33	0.46	0.88	32.49
CNN+MDCF_0.05	6.67	0.28	**0.32**	0.46	**30.56**
CNN+MDCF_0.01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

Tulokset

Test set



YouTube set



Loppupäätelmät

- Suorituskyvyssä merkittävä ero opetusaineiston ja YouTube aineiston välillä!
- Attentive pooling -menetelmä onnistuu painottamaan videon eri kuvia.
- Onnistuimme myös optimoimaan verkon parametreja haluttuun kustannusten painotukseen.
- GAN diskriminaattorin käyttäminen ei kuitenkaan toiminut (EER 47,5 %).
- Vuoden MATINE rahoituksella saimme projektin hyvälle alulle.

Seuraavaksi DARPA Semantic Detection

- Facebook järjestää deepfake-videoiden tunnistus -kilpailun. Osallistumme siihen Prof. Yamagichi:n ryhmän kanssa.
- DARPA:lla on käynnistymässä SemaFor -projekti, missä on tarkoitus tunnistaa deepfake eri medioilla (video, ääni, kuva, teksti) ja yrittää selvittää automaattisesti miksi kyseessä on deepfake ja kuka sen teki.
- Ennalta tuntemattoman deepfaken tunnistaminen on hankalaa, kuten huomasimme omalla YouTube-setillä.
    - Kokeilemme semi-supervised malleja tähän tehtävään.
- Tuotamme prototyyppisovelluksen, missä käyttäjä voi uploadata oman videotiedostonsa ja sovellus kertoo oliko video deepfake ja kenen kasvot oli vaihdettu.