Research Project Proposal: Explanations for deepfake detection

Samuele Pino
samuele.pino@mail.polimi.it
CSE Track

POLITECNICO
MILANO 1863

HONOURS PROGRAMME
HP-SR
in Information Technology

Abstract

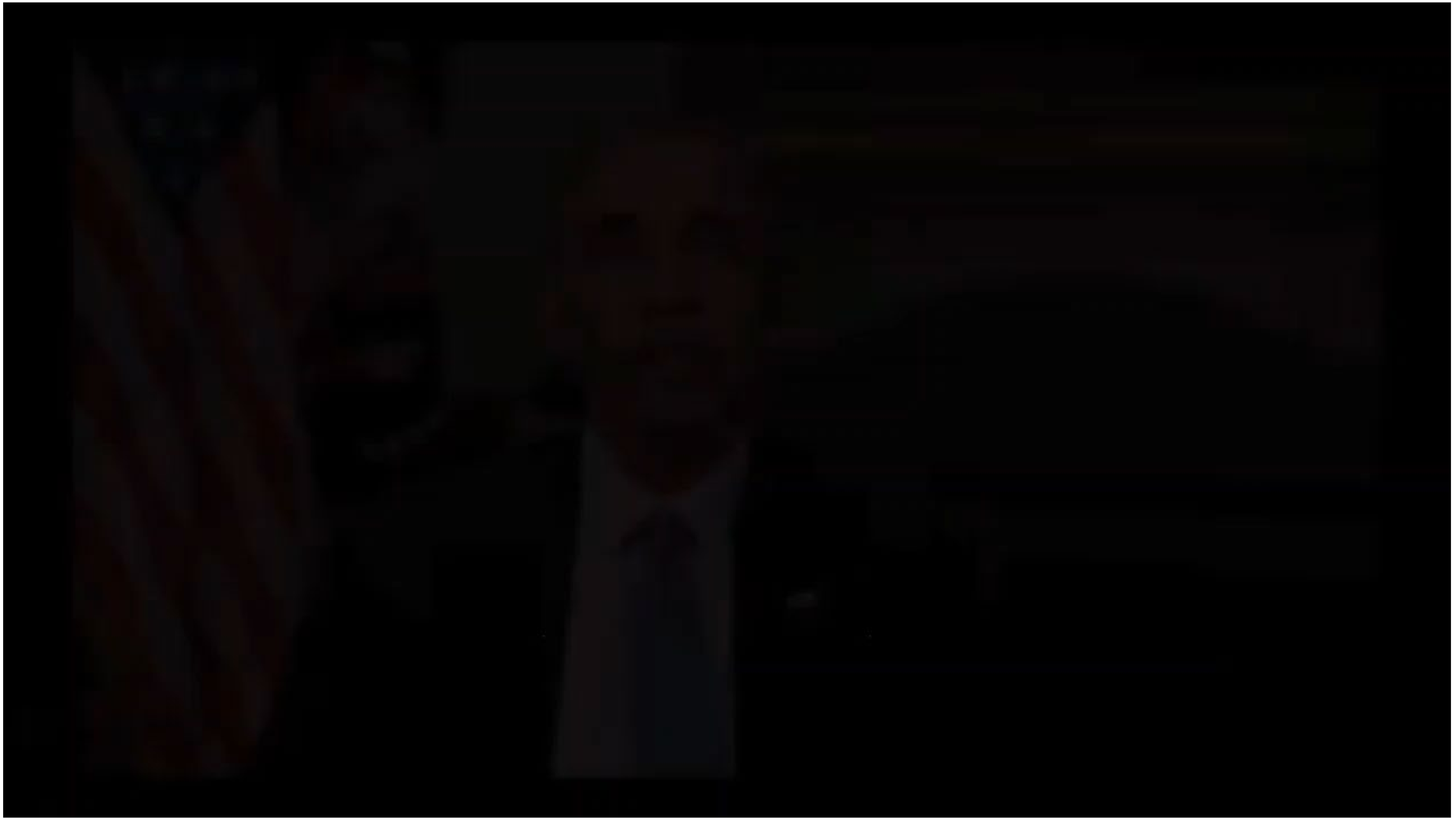
Automatic classificators can already detect if a video is real or manipulated.

We would like to provide human-understandable explanations for these predictions.

Overview

- Video manipulation
- Deepfakes: overview & technical background
- Deepfake detection methods
- Explanation problem & techniques
- The research goal and plan

Video manipulation



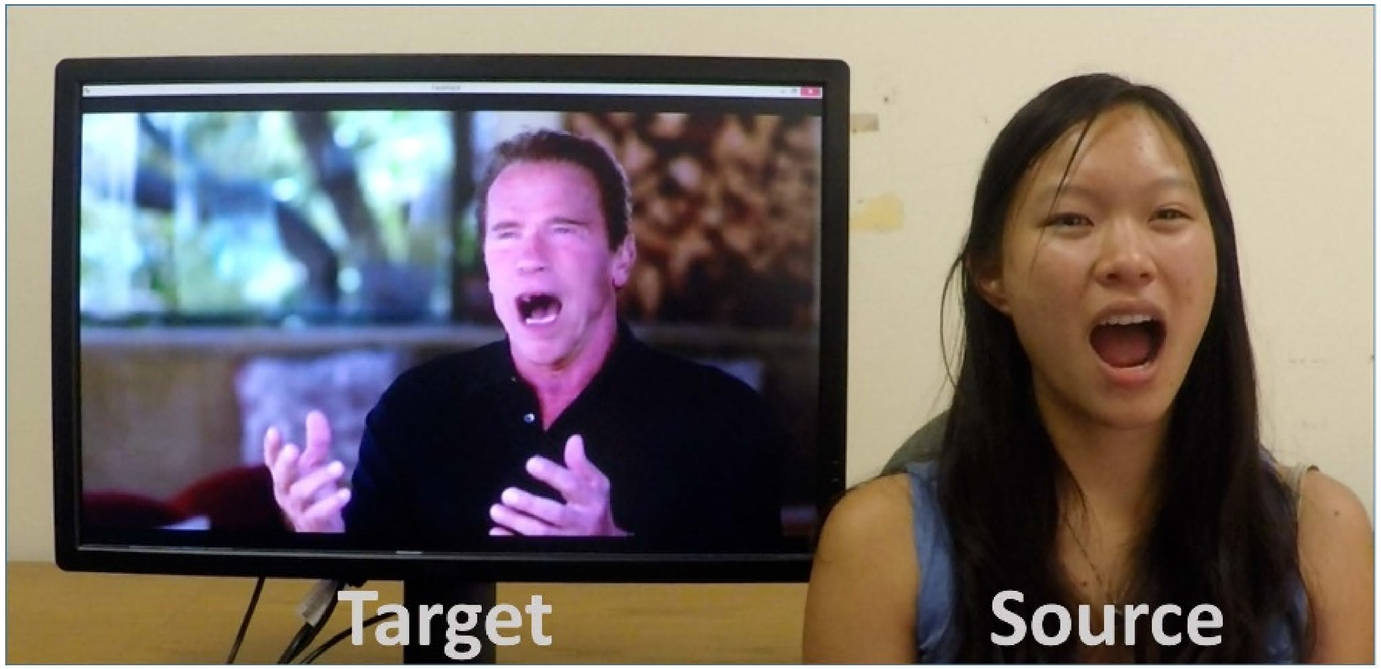
https://www.youtube.com/watch?v=cQ54GDm1eL0

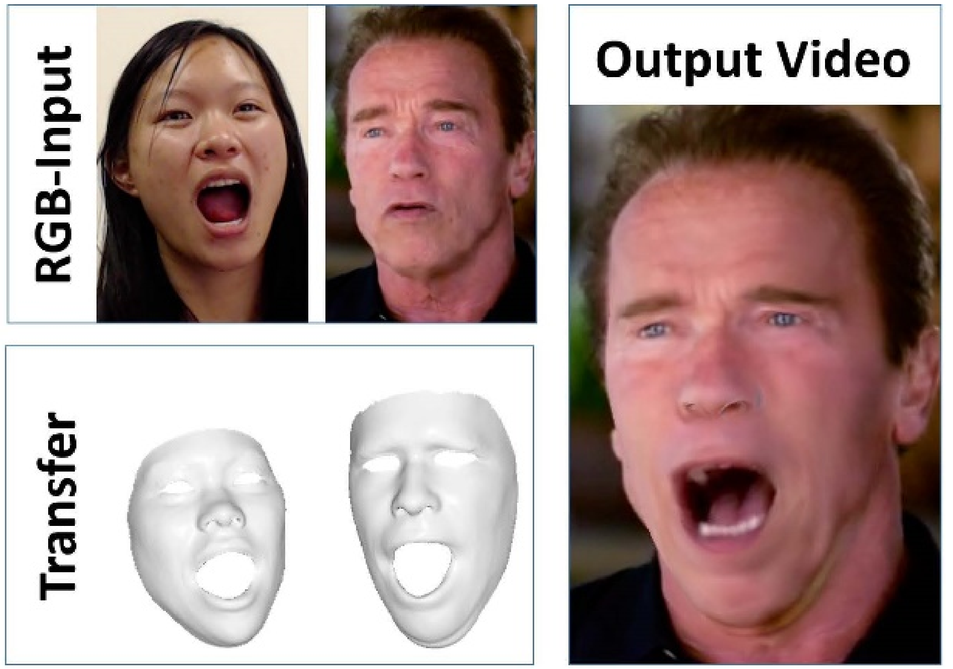
Video manipulation types

- Facial reenactment
- Identity swap

Video manipulation types

- Facial reenactment
- Identity swap

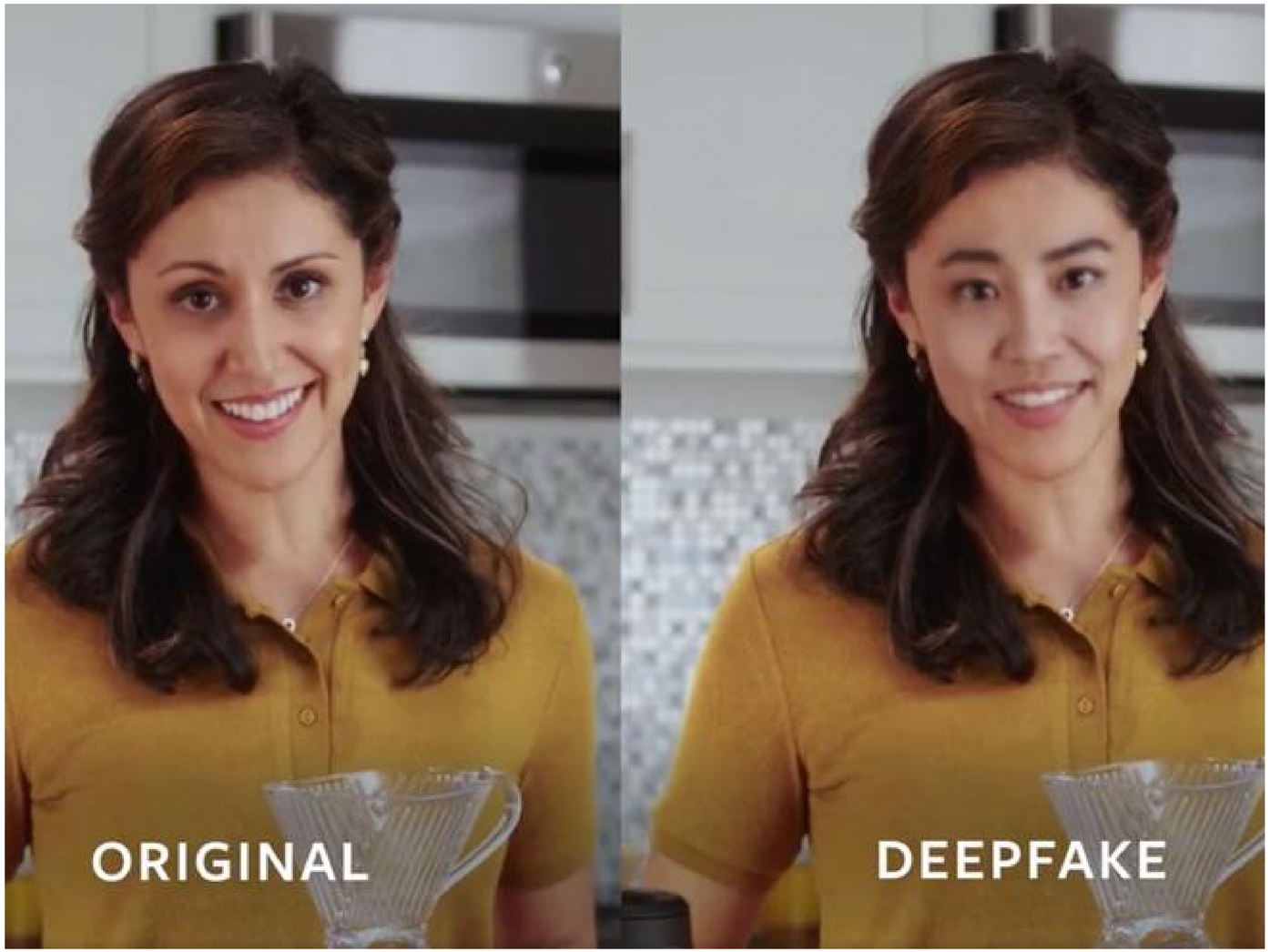




[Thies at al., “Face2Face: Real-time Face Capture and Reenactment of RGB Video”, 2016]

Video manipulation types

- Facial reenactment
- Identity swap



[Deepfake Detection Challenge, 2019]

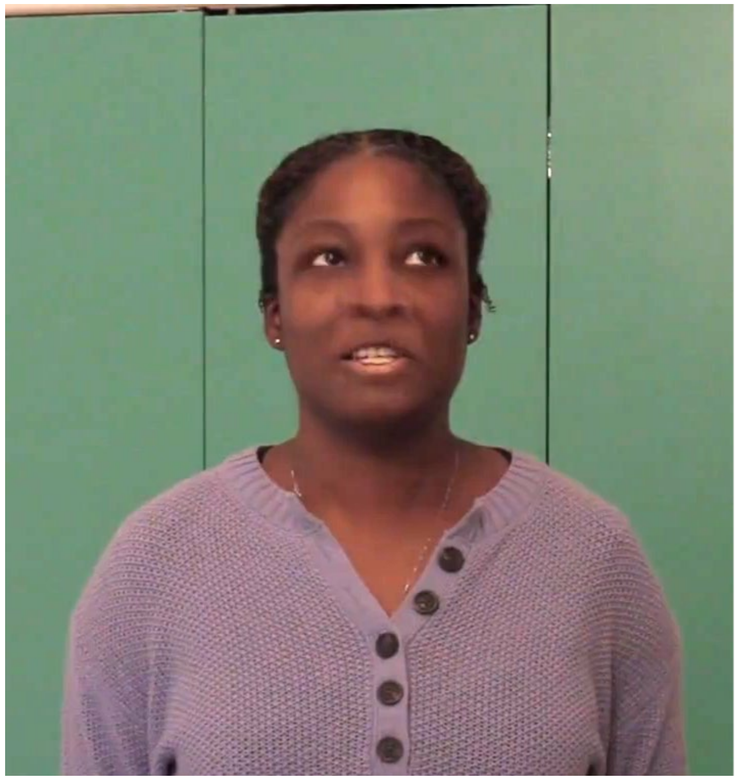
Deepfakes

Deepfakes: overview

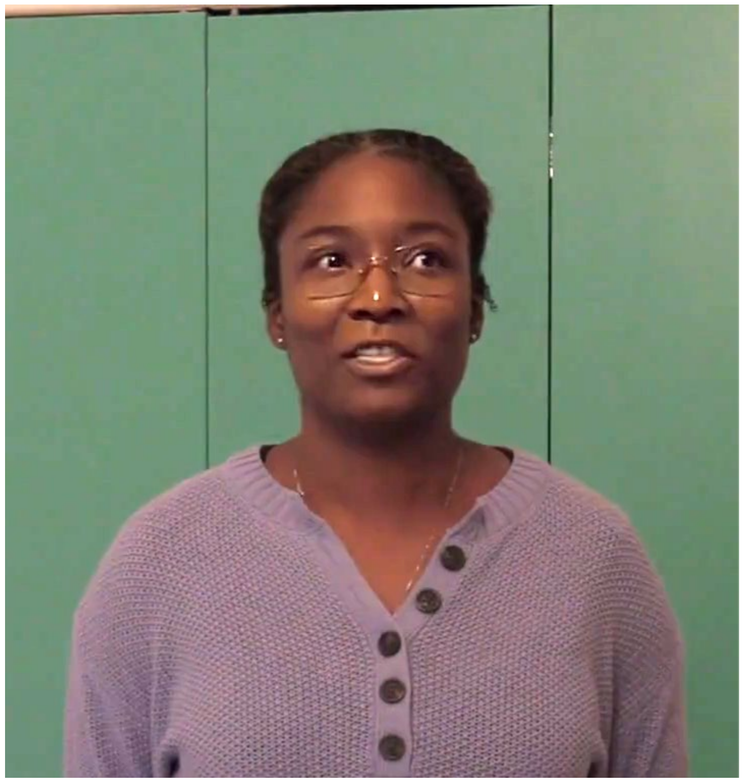
- Replacing the face of a targeted person A by the face of B in a video
- Deep learning technique
- Initially created to generate face-swapped adult contents
- No paper

Deepfakes: overview

REAL



FAKE



FAKE

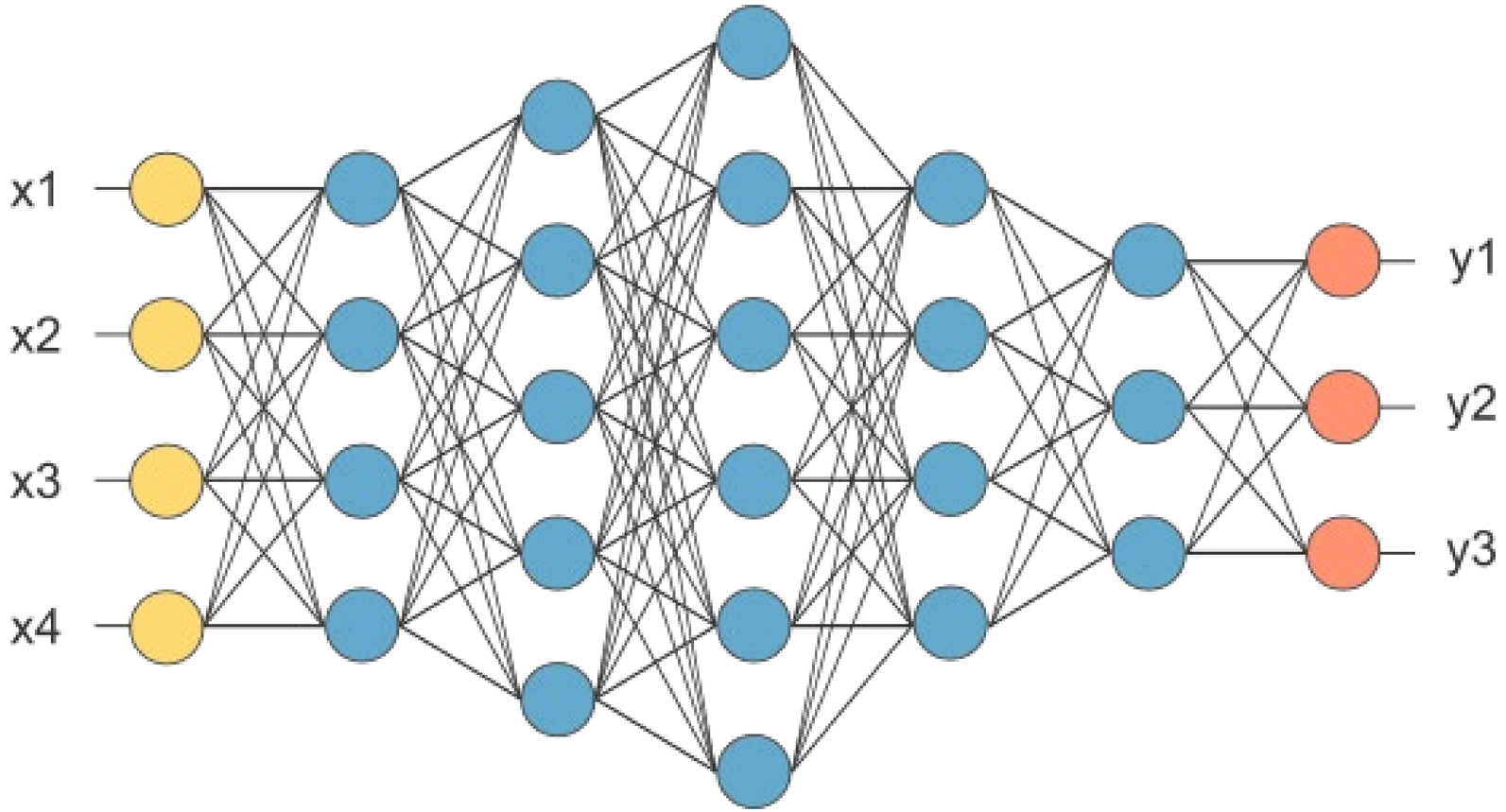
[Deepfake Detection Challenge, 2019]

Deepfakes: overview

- Why is it important to detect them?
- Why “deep”?

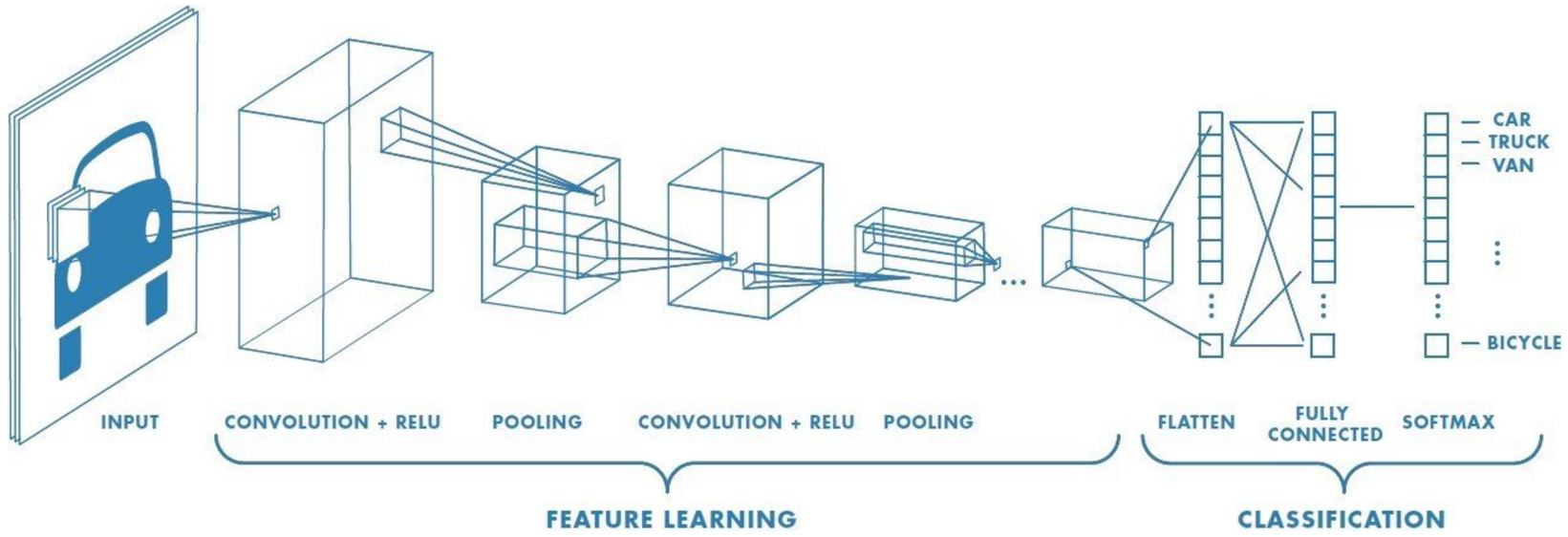
Deepfakes: technical background

- DNN
- CNN
- Auto-encoder
- GAN
- LSTM



Deepfakes: technical background

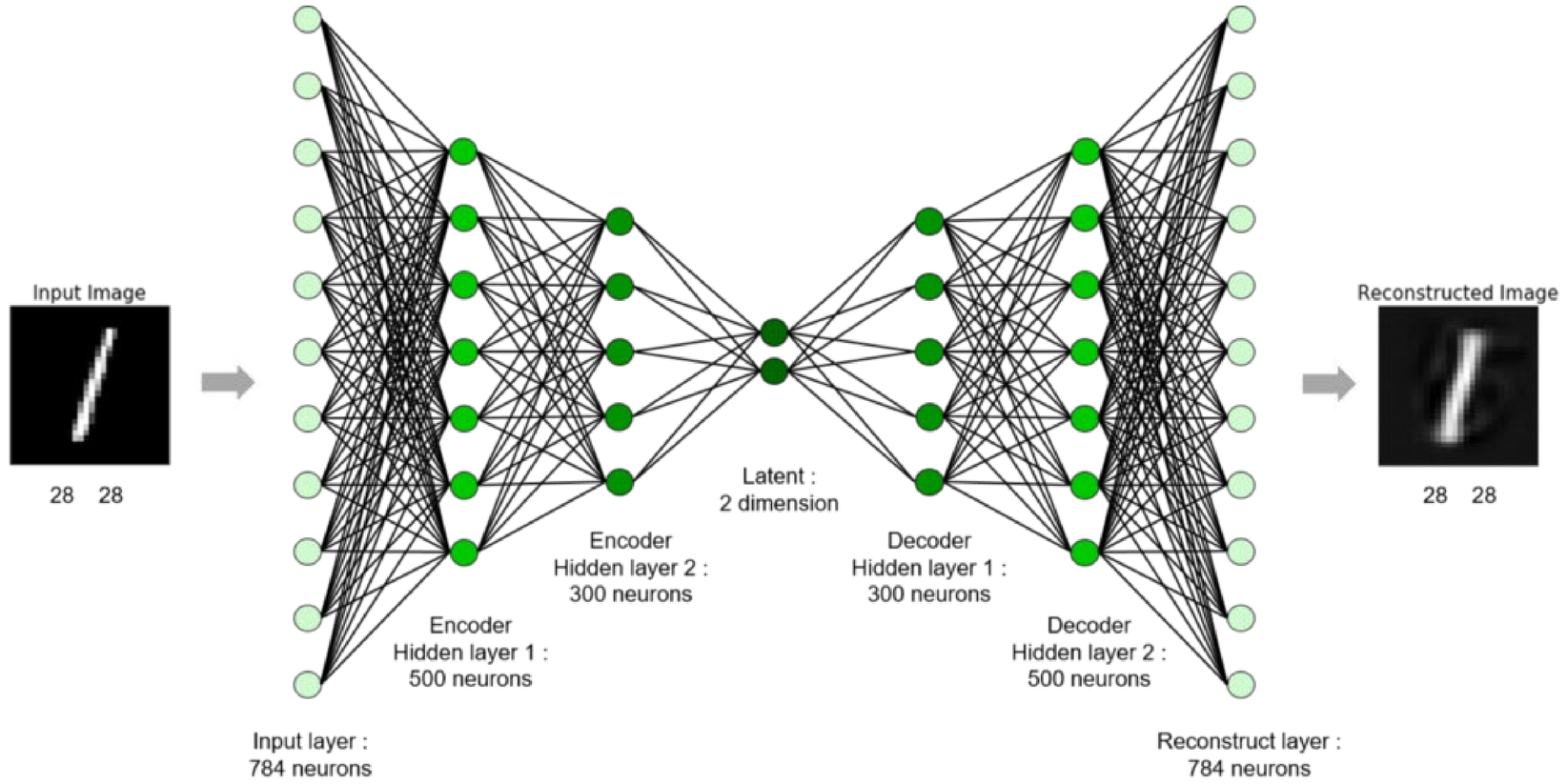
- DNN
- CNN
- Auto-encoder
- GAN
- LSTM



https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Deepfakes: technical background

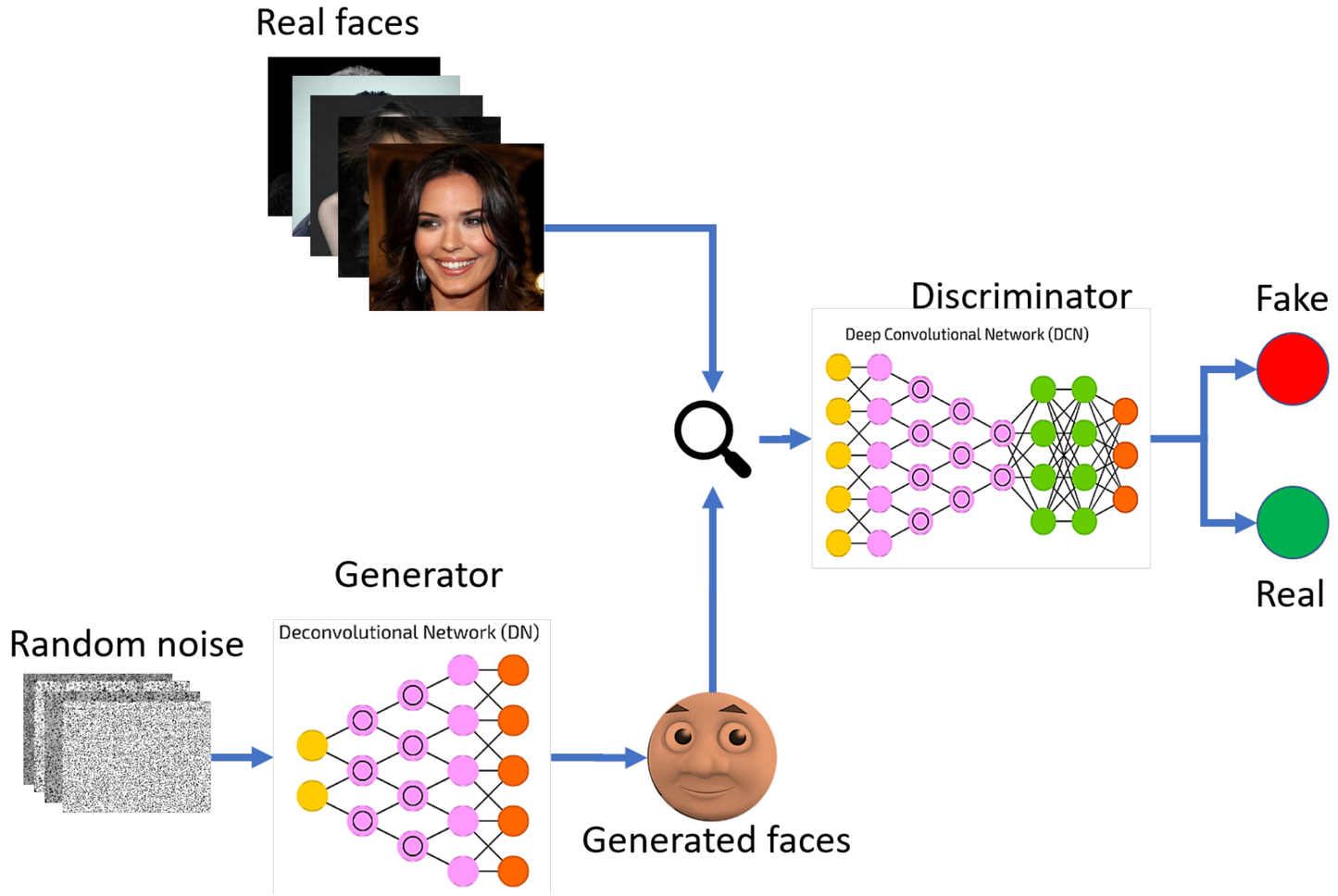
- DNN
- CNN
- Auto-encoder
- GAN
- LSTM



https://mc.ai/auto-encoder-in-biology/

Deepfakes: technical background

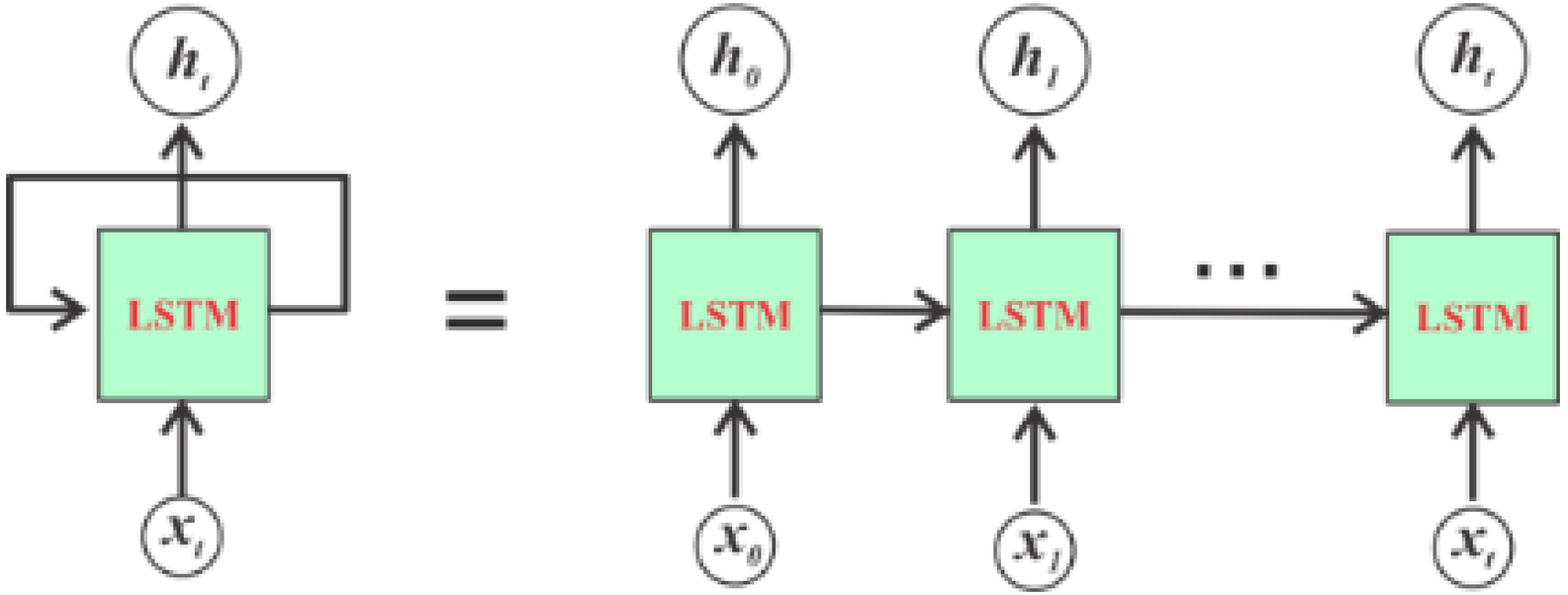
- DNN
- CNN
- Auto-encoder
- **GAN**
- LSTM



https://www.spindox.it/en/blog/generative-adversarial-neural-networks/

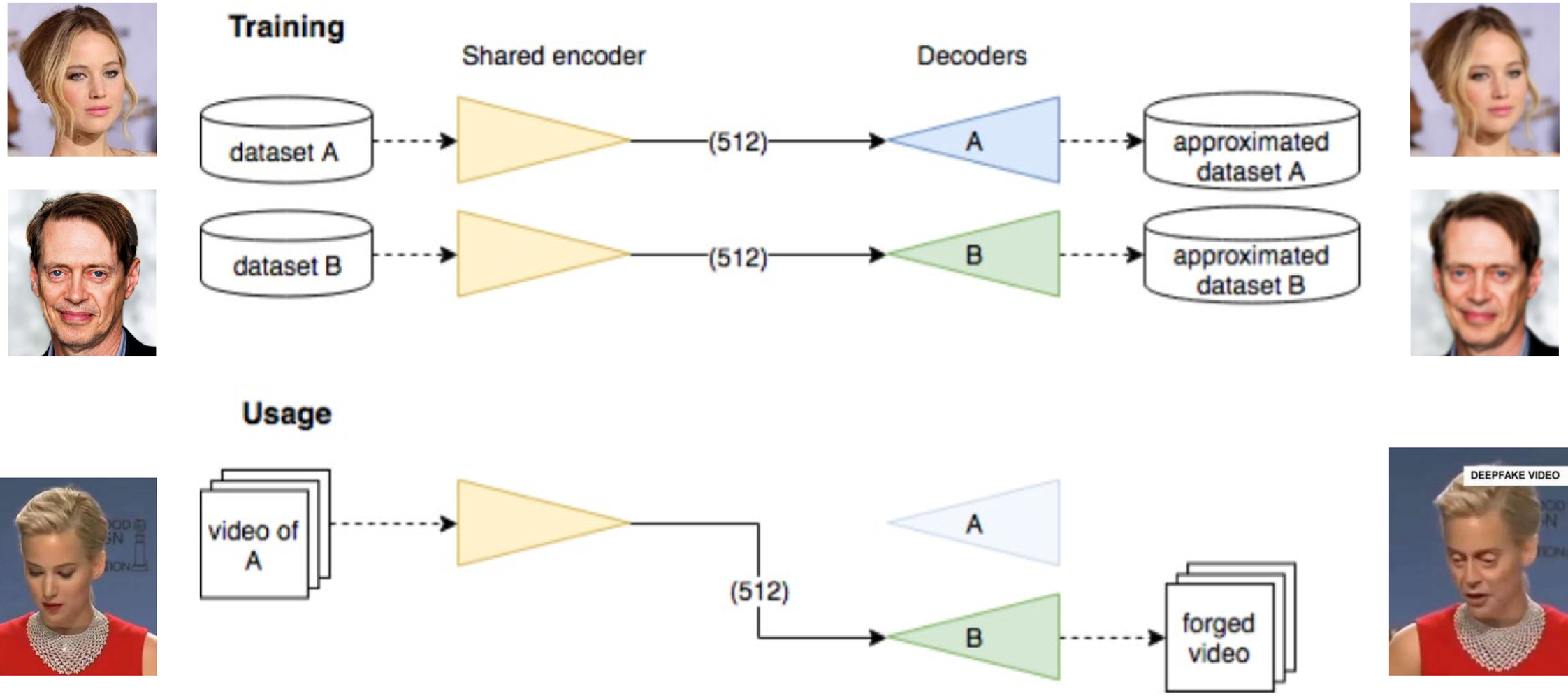
Deepfakes: technical background

- DNN
- CNN
- Auto-encoder
- GAN
- LSTM



https://www.codeproject.com/Articles/5165357/In-depth-LSTM-Implementation-using-CNTK-on-NET-pla

Deepfakes: technical background



[Afchar et al., “MesoNet: a Compact Facial Video Forgery Detection Network”, 2018]

Deepfake detection methods

Detection methods

Use the flaws of the generation pipeline

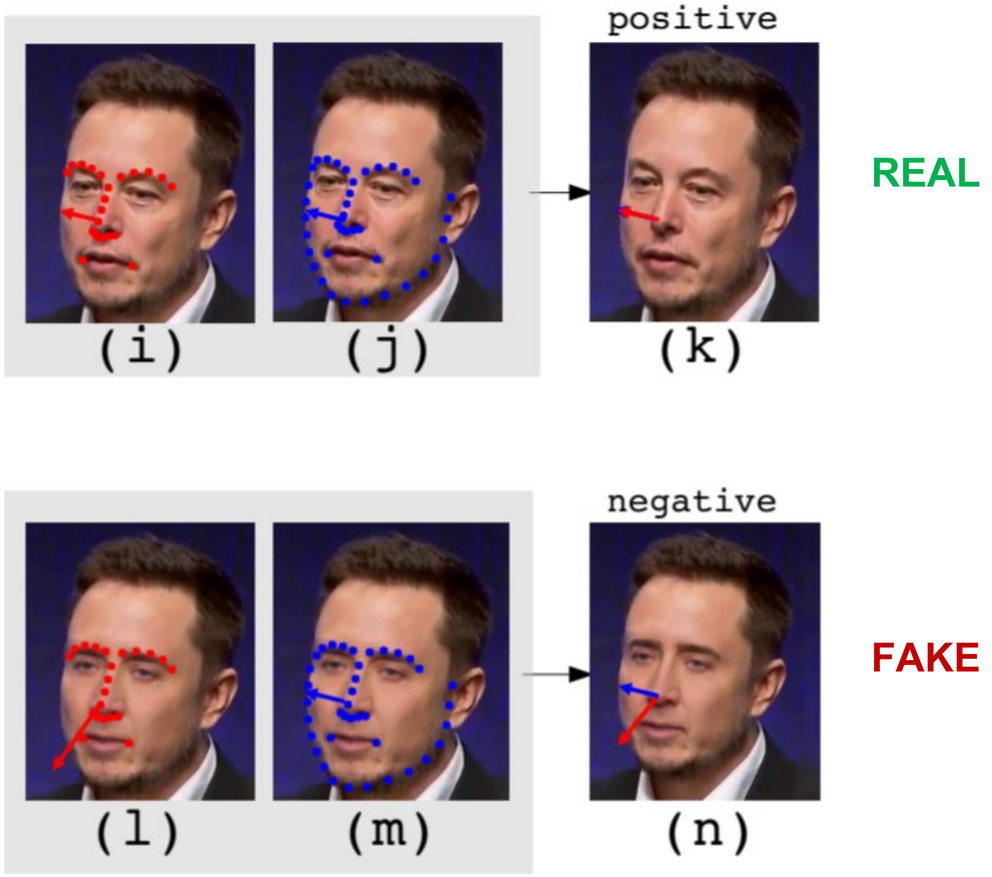
- Specifically chosen features
- Purely learned features
- Temporal inconsistency

[Deepfake Detection Challenge, 2019]

Detection methods

Specifically chosen features

- Affine transformations [Li and Lyu]
- Head-face poses [Yang et al.]
- Visual artifacts [Matern et al.]
- Face/head actions [Agarwal et al.]

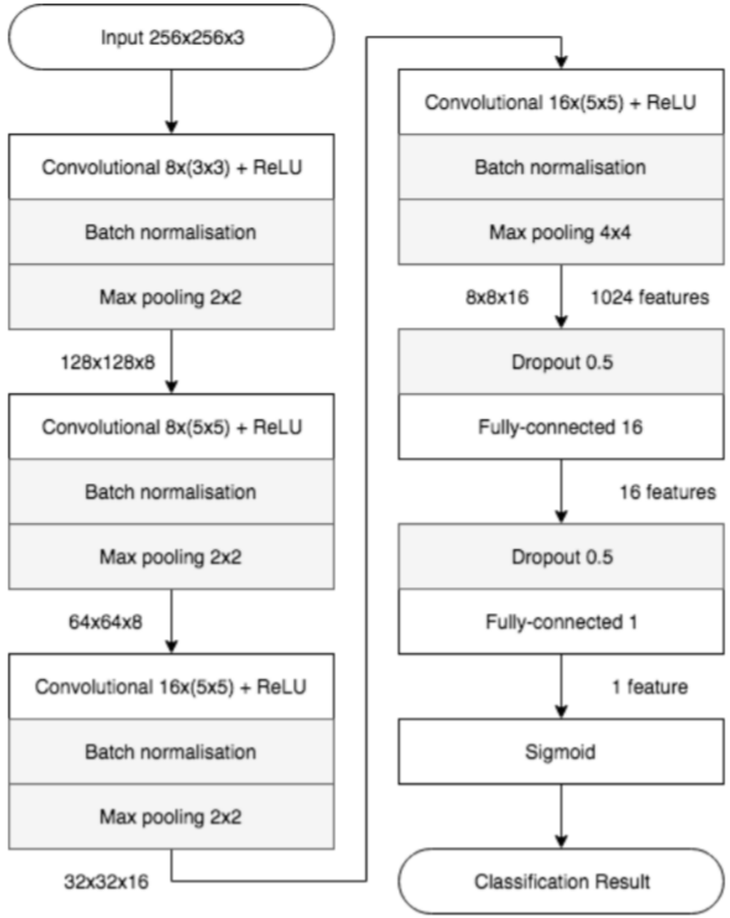


[Tang et al., ‘‘Exposing deep fakes using inconsistent head poses’’, 2019]

Detection methods

Purely learned features

- CNN (Mesonet) [Afchar et al.]
- CNN (XceptionNet) [Rössler et al.]
- CNN + CapsNet [Nguyen et al.]

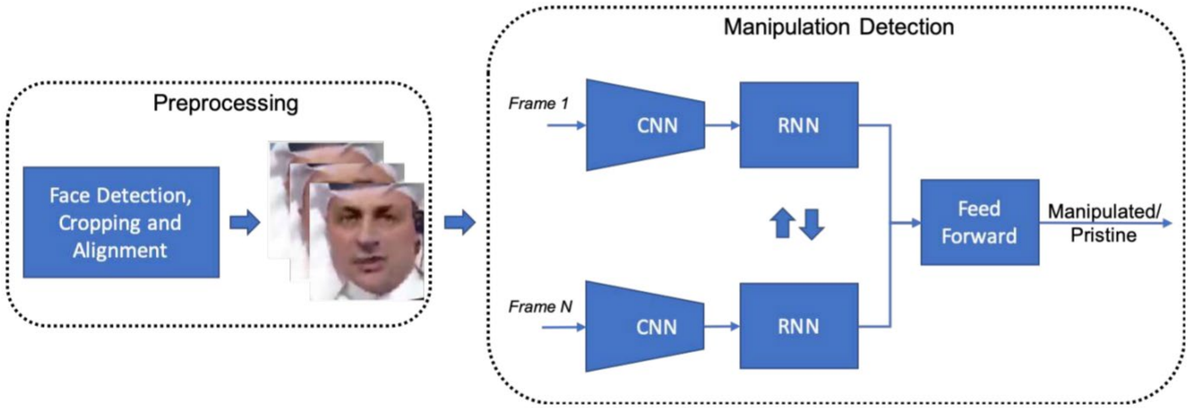


[Afchar et al., “MesoNet: a Compact Facial Video Forgery Detection Network”, 2018]

Detection methods

Temporal inconsistency

- CNN + LSTM
  - Lip sync [Korshunov et al.]
  - Eye blink [Li et al.]
  - Frame consistency [Sabir et al.]



[Sabir et al., ‘‘Recurrent Convolutional Strategies for Face Manipulation Detection in Videos’’, 2019]

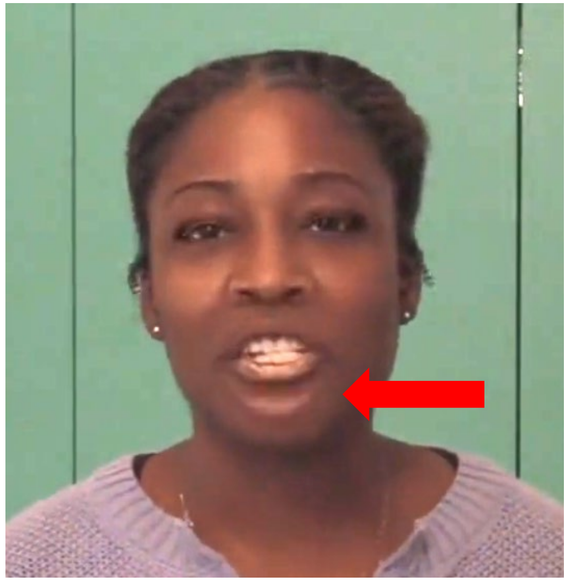
Detection methods

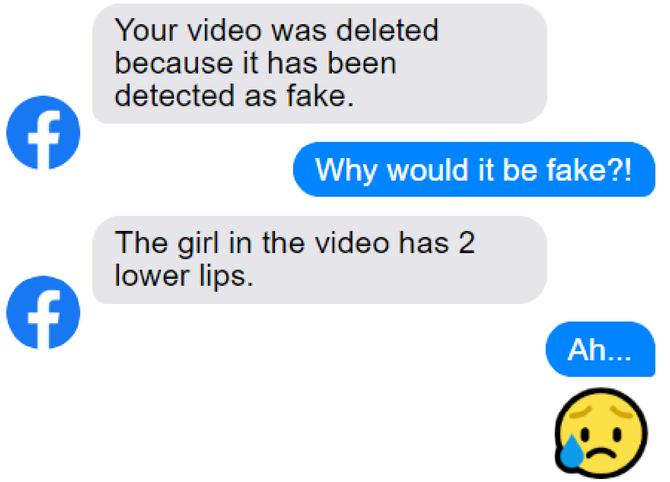
Can we trust these techniques in decision making processes?

Explanation problem

Explanation problem

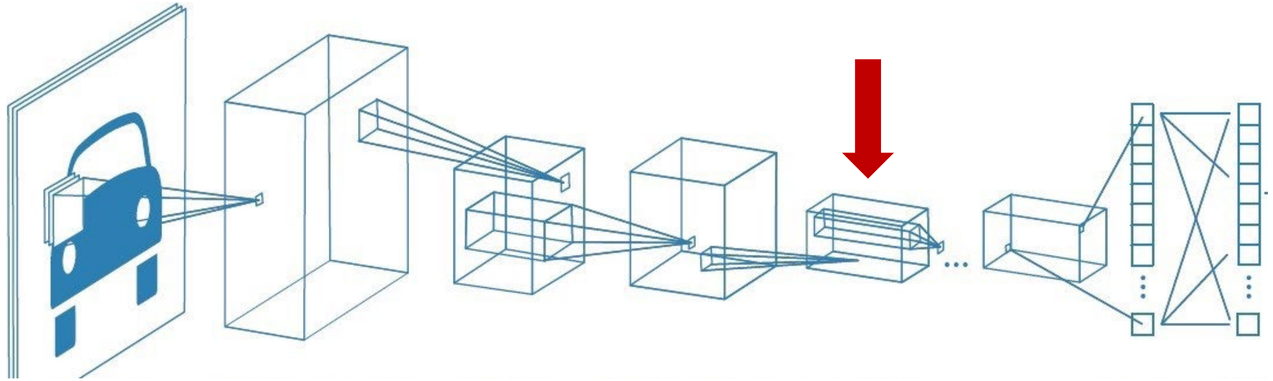
- We want a «correct prediction for the correct reason»
- Complexity-interpretability trade off
- For images: attention maps or natural language
- Why do we need it:
  - law enforcement
  - journalists
  - dispute resolution in social media

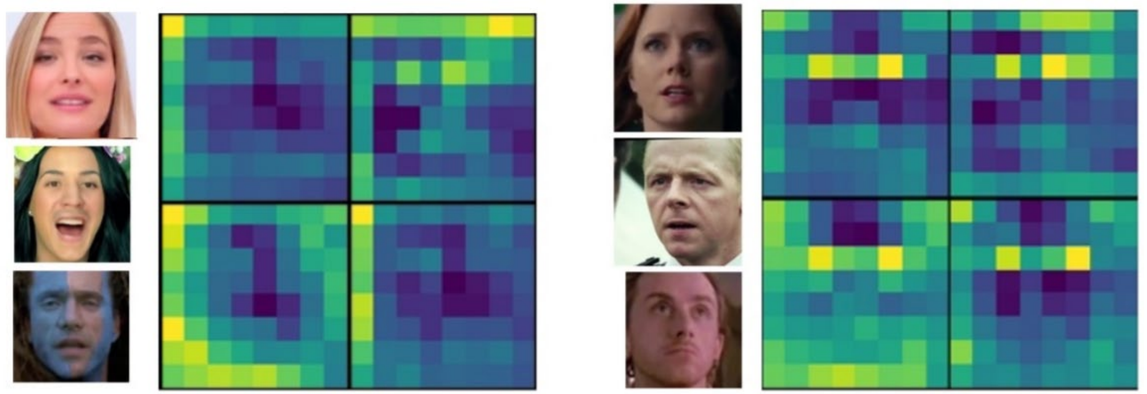




Explanation techniques

- Model specific
- By design
- Black box





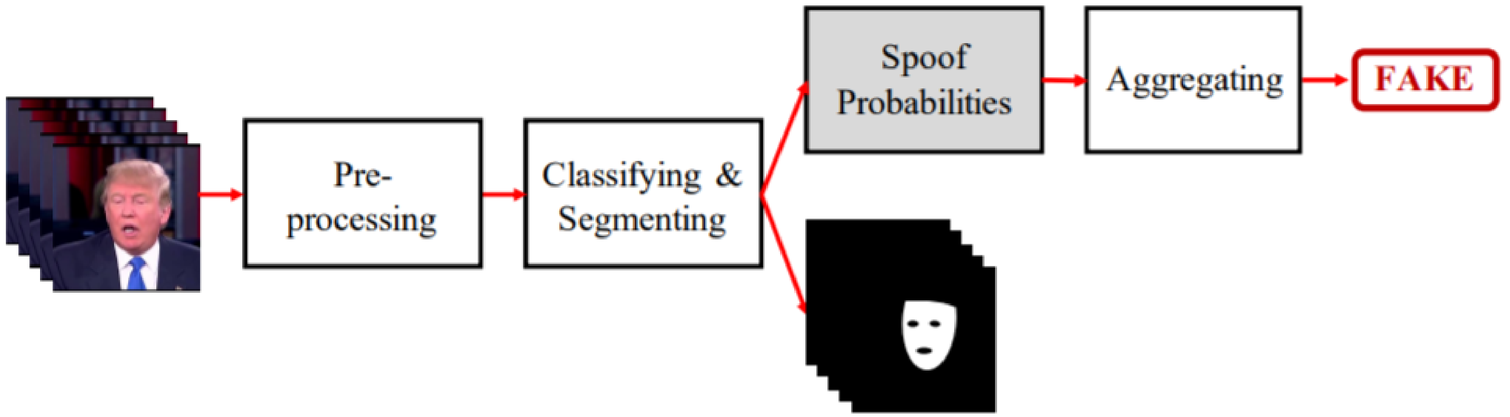
mean layer output of 100 *deepfake* faces

mean layer output of 100 real faces

[Afchar et al., “MesoNet: a Compact Facial Video Forgery Detection Network”, 2018]

Explanation techniques

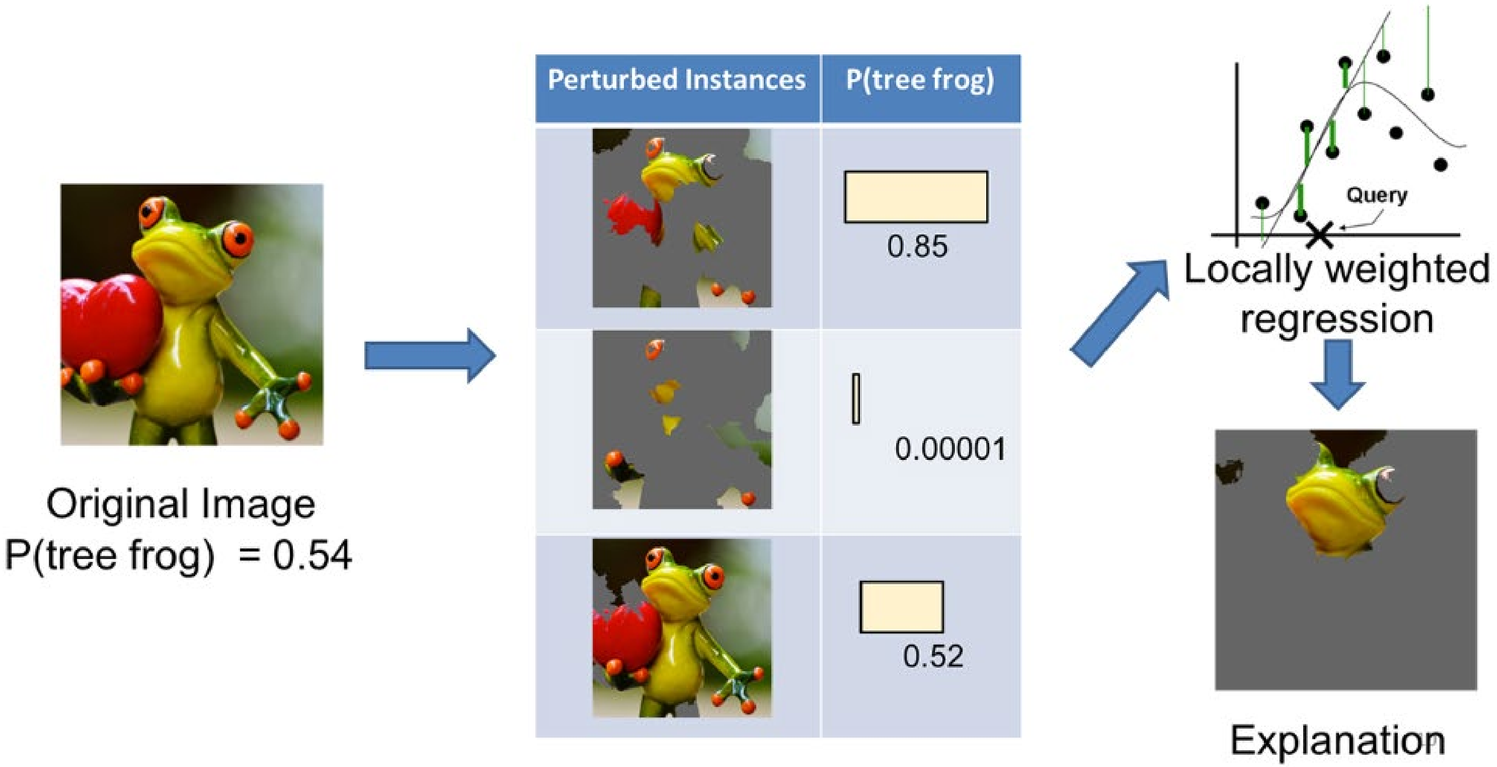
- Model specific
- By design
- Black box



[Nguyen et al, “Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos”, 2019]

Explanation techniques

- Model specific
- By design
- Black box

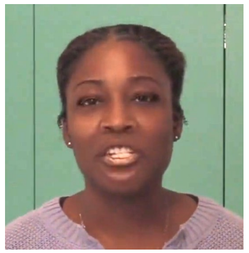


LIME - https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/

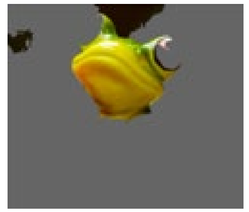
Our research

Research goal

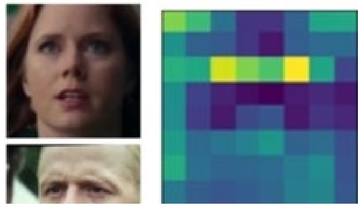
- Investigate explainability of deepfakes
- Do similar models use different features?
- Black box vs. model-aware techniques
- Explanation for video inputs
- Using this knowledge to improve models



Fake lips!



vs.



Research plan

- Implementation of the baseline detectors
- Implementation of known explanation algorithms
- Investigation of extensions and improvements
- Evaluation design
- Results collection and analysis

Questions?

Thank you