Face Swapping Video Detection with CNN

Wang Yang, Junfeng Xiong,
Liu Yan, Hao Xin, Wei Tao
Baidu X-Lab

百度安全 有AI更安全

Baidu Security
Advanced AI , Stronger Security

Face Swapping Video Detection with CNN

- Deepfakes video

- A simple and effective CNN
- Face recognition based method

Deepfakes Video

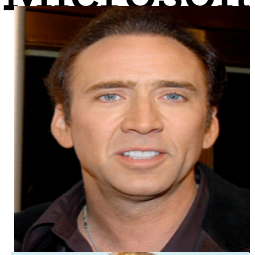


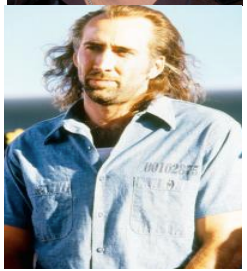
When Face Recognition Systems Meet Deepfakes

Vulnerable face comparison before fake faces

- Microsoft Azure

Real





Fake





azure.microsoft.com





Similarity 86.0%





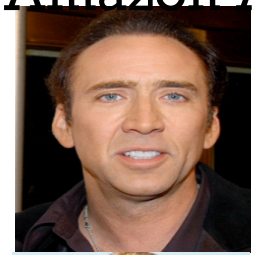
Similarity 70.5%

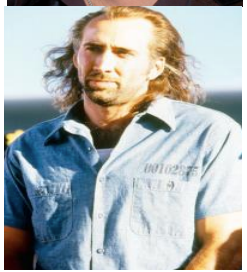
When Face Recognition Systems Meet Deepfakes

Vulnerable face comparison before fake faces

- Amazon AWS

Real





Fake





aws.amazon.com





Similarity 95.1%



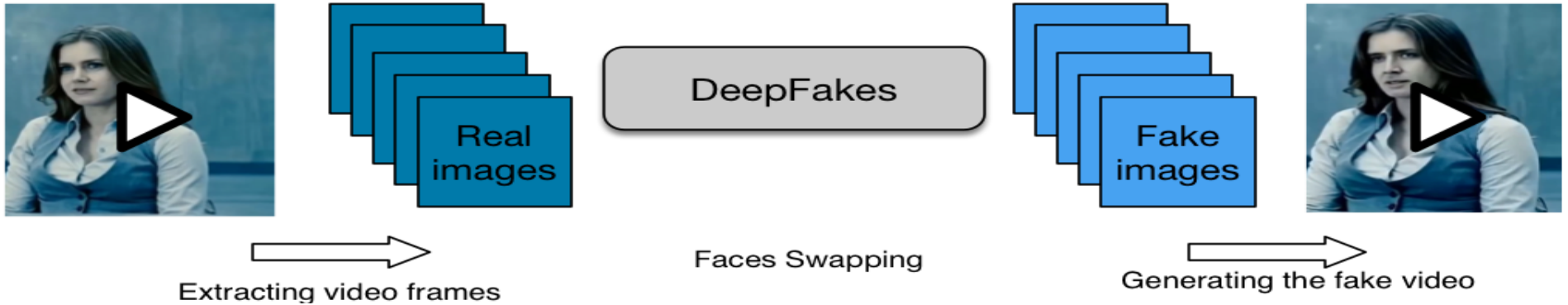


Similarity 87.3%

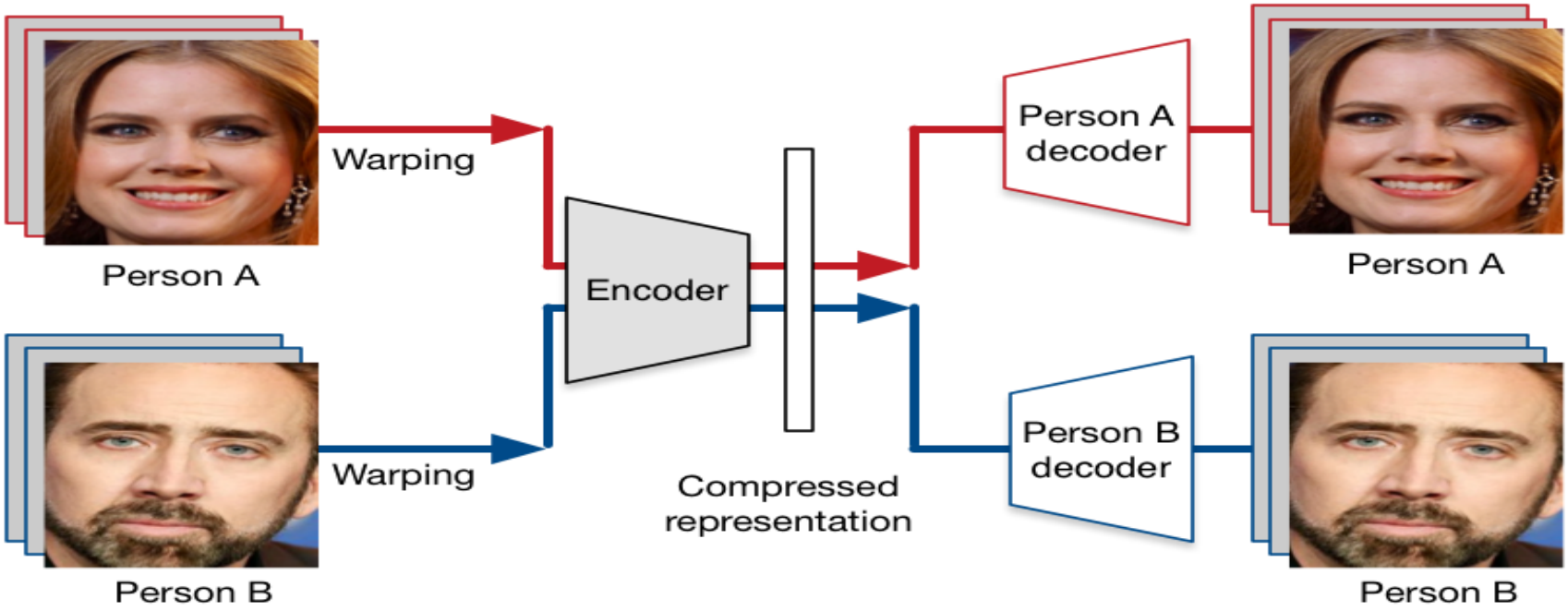
Face Swapping Video Generation

Characteristics

- Swap victim's face in every frame independently
- Not End2End
- Only manipulate central face area
- Autoencoder

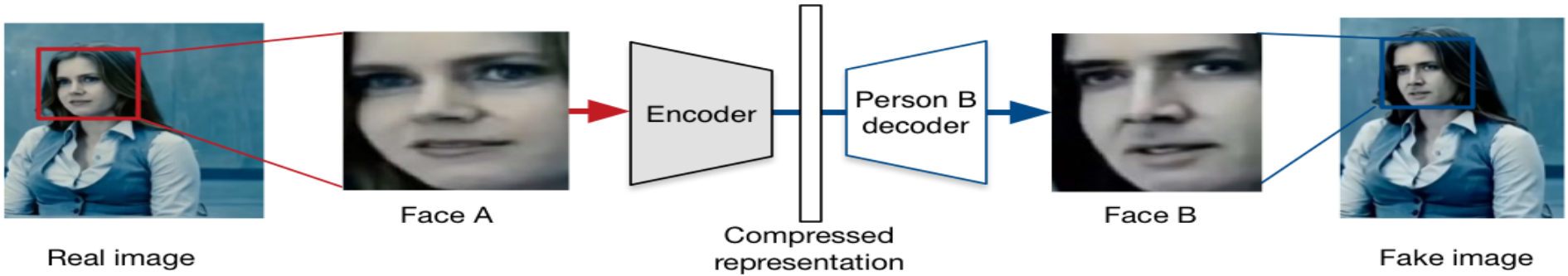


Deepfakes Training Phase



Deepfakes Generation Phase

- Convert
  - Person A Encoder -> Person B Decoder
- Merge back
  - Gaussian Blur/Color Average
  - Poisson Image Editing



Face Swapping Video Detection with CNN

- Deepfakes video
- **A simple and effective CNN**
  - capturing low-level features of the images

- Face recognition based method

A Simple and Effective CNN

Design purpose

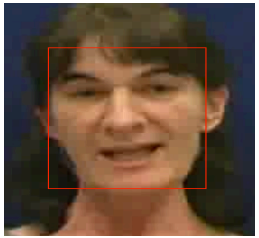
- Input contains marginal(background) information.
- Capture low-level features of the images.



A Simple and Effective CNN

Training

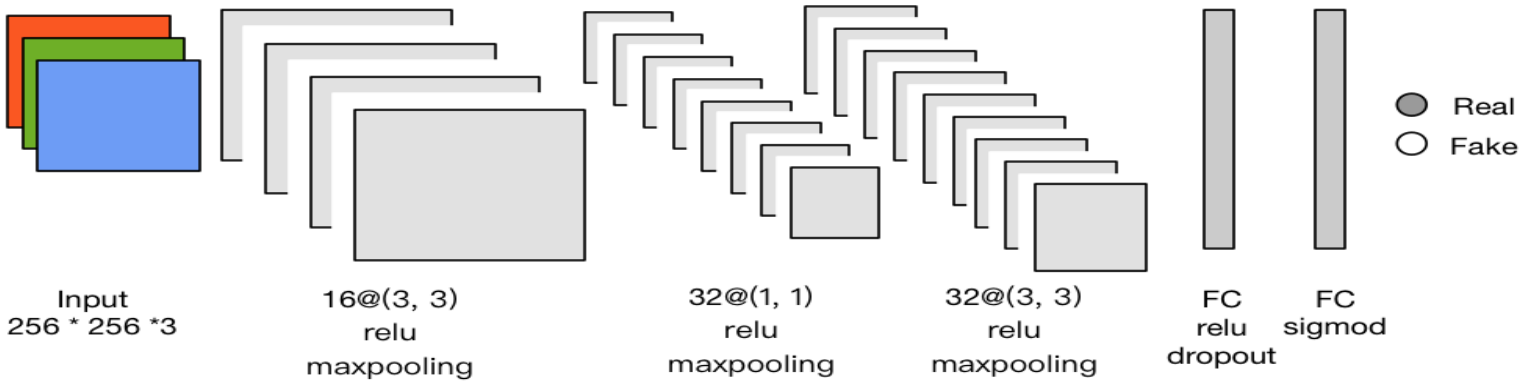
- Dataset from VidTIMIT
  - 67600 fake faces and 66629 real faces
  - low quality and high quality images
- Cropped faces
  - with face landmark detector MTCNN
  - obtain 1.5 scaled bounding box
- Augmented data
  - horizontal flipping
  - randomly zooming
  - shearing transformations



A Simple and Effective CNN

Characteristics

- 3 convolution layers
- Accuracy rate: 99%



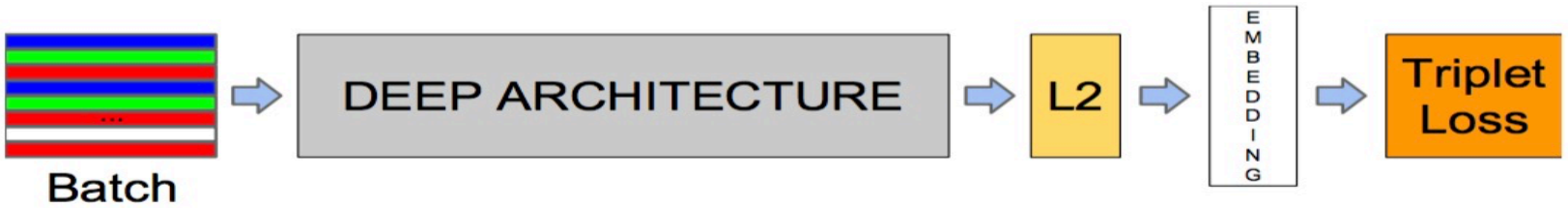
Face Swapping Video Detection with CNN

- Deepfakes video
- A simple and effective CNN
- Face recognition based method
  - capturing high-level features of faces

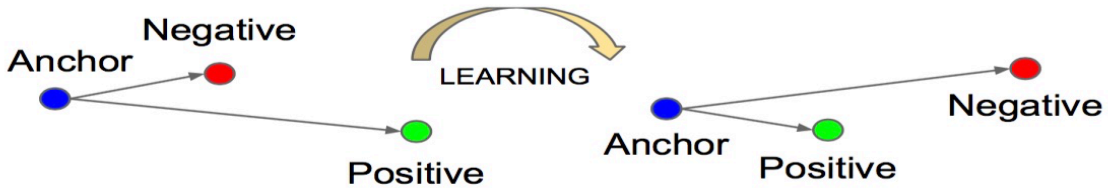
What is FaceNet?

Characteristics

- SOTA CNN for face recognition
- Model structure



- Triple Loss

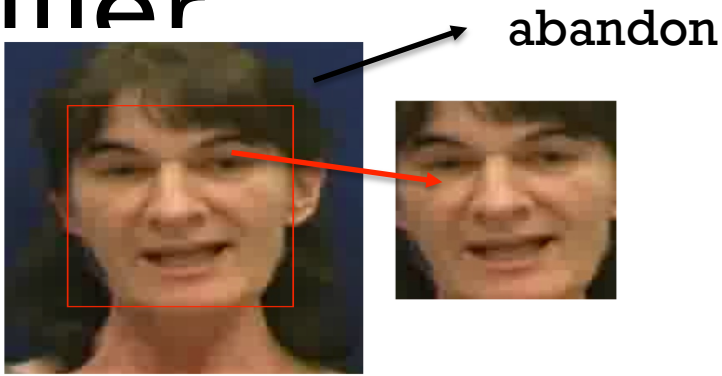


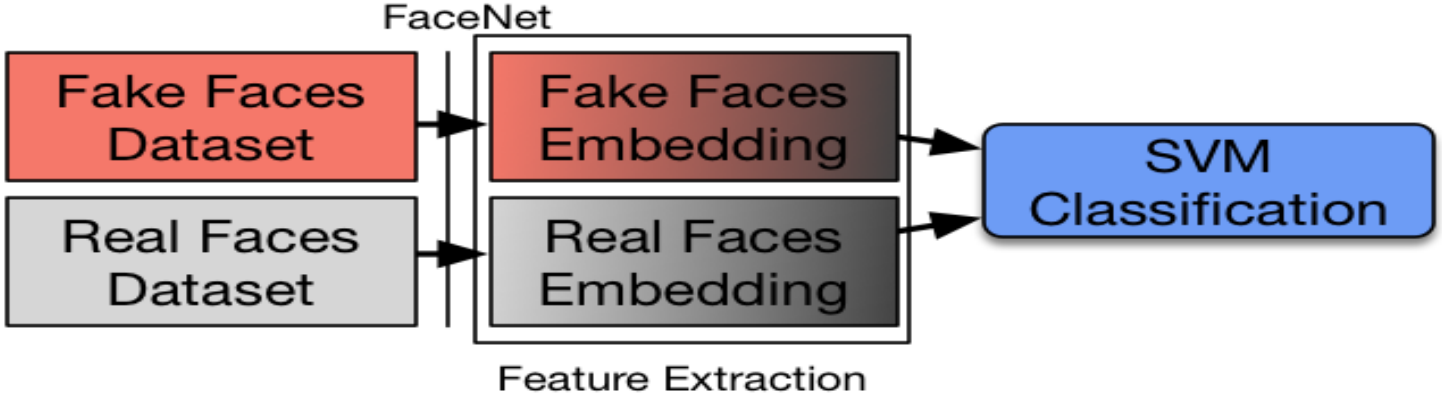
FaceNet: A unified embedding for face recognition and clustering

A FaceNet based SVM Classifier

Training

- Central face area
  - No margin/background
  - Only face area
- Dataset from VidTIMIT

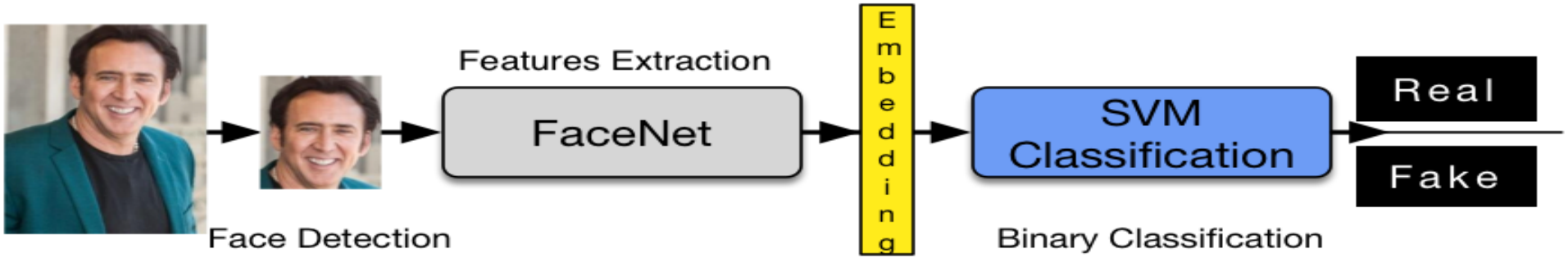




A FaceNet based SVM Classifier

Characteristics

- FaceNet used for extracting face features
- SVM for binary classification
- Accuracy rate: 94%



A Simple and Effective CNN

Accuracy rate: 99%



A FaceNet based SVM Classifier

Accuracy rate: 94%



Summary

- CNN for image classification
  - A simple architecture can work well.
  - catching low-level features: contours, edges…
- A FaceNet based SVM classifier
  - using FR to catch features of fake faces
  - using SVM for binary classification
  - 64% accuracy rate for the misclassification set from the simple CNN based classifier

Thank You!
Q&A