Detecting Facial Manipulation Deepfakes

Evan Kravitz, Huazhe Xu

April 21, 2020

https://www.instagram.com/p/ByaVigGFP2U/

https://www.instagram.com/p/ByaVigGFP2U/

What is a deepfake?

- Synthetic image/video of a person that looks realistic to human viewers, which can be used to perpetrate fraud or spread misinformation
- Deepfakes are a form of social engineering attack
- We have focused our research on detecting facial deepfakes

Social engineering attacks

Face synthesis

StyleGan (2019)

Face swap

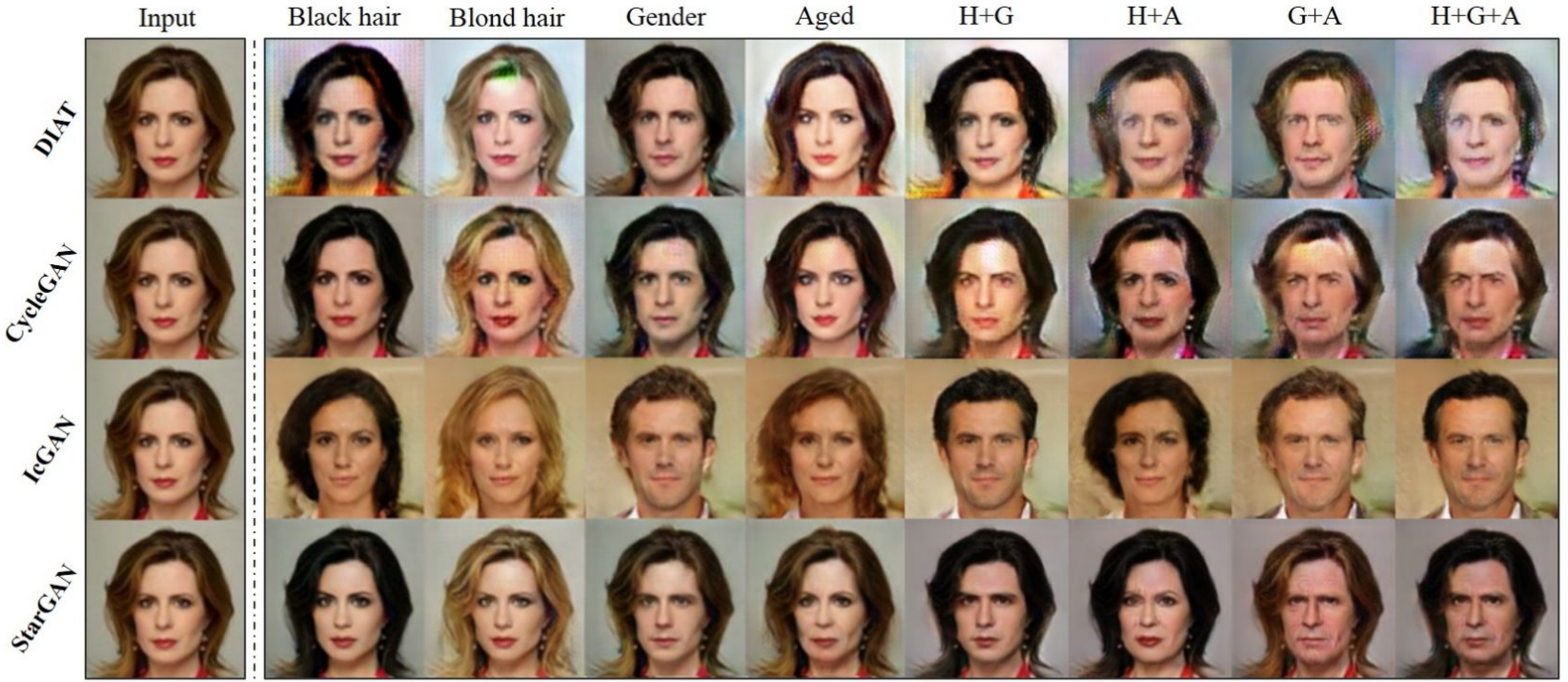
FaceSwap (2016)

Deepfake FaceSwap (2020)

Jennifer Lawrence/Steve Buscemi FaceSwap using the Villain model

Face attribute

StarGAN (2018)

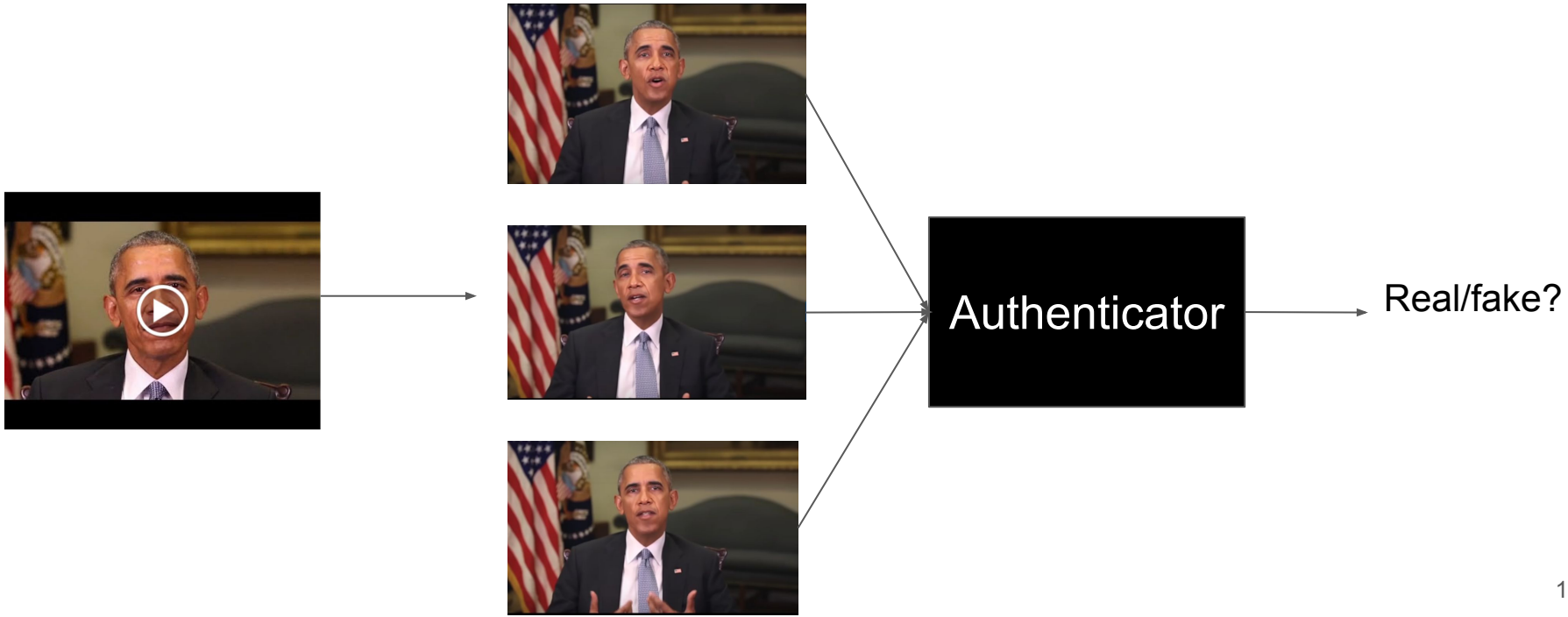


Facial expression

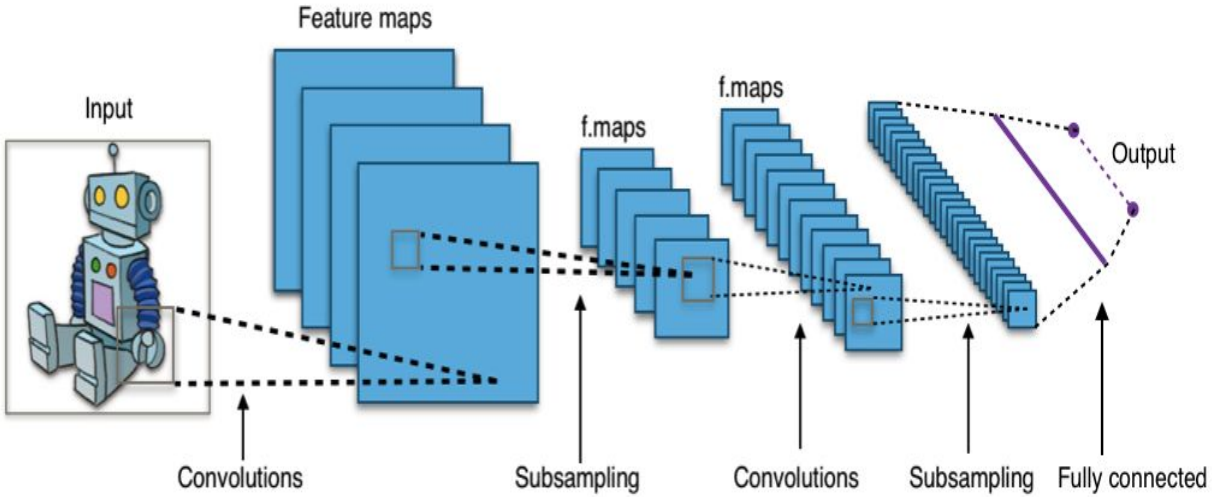
Face2Face (2016)

Protecting against deepfake

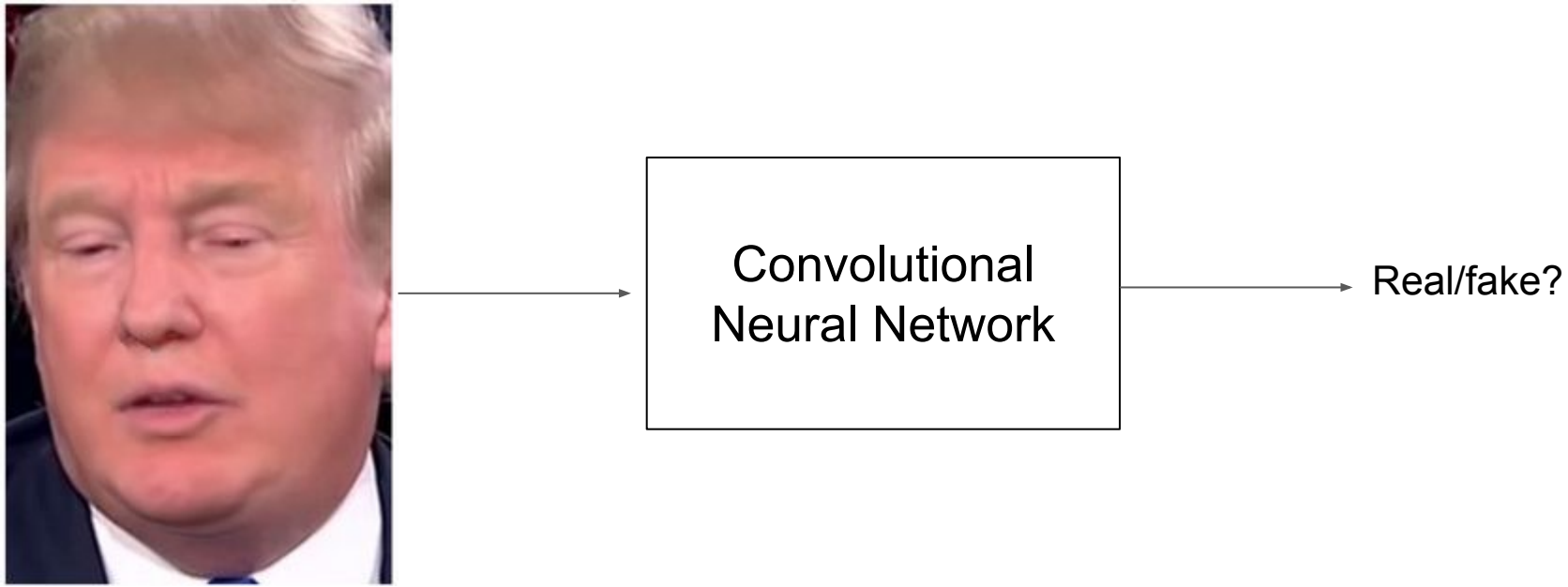
- We need a system for authenticating media

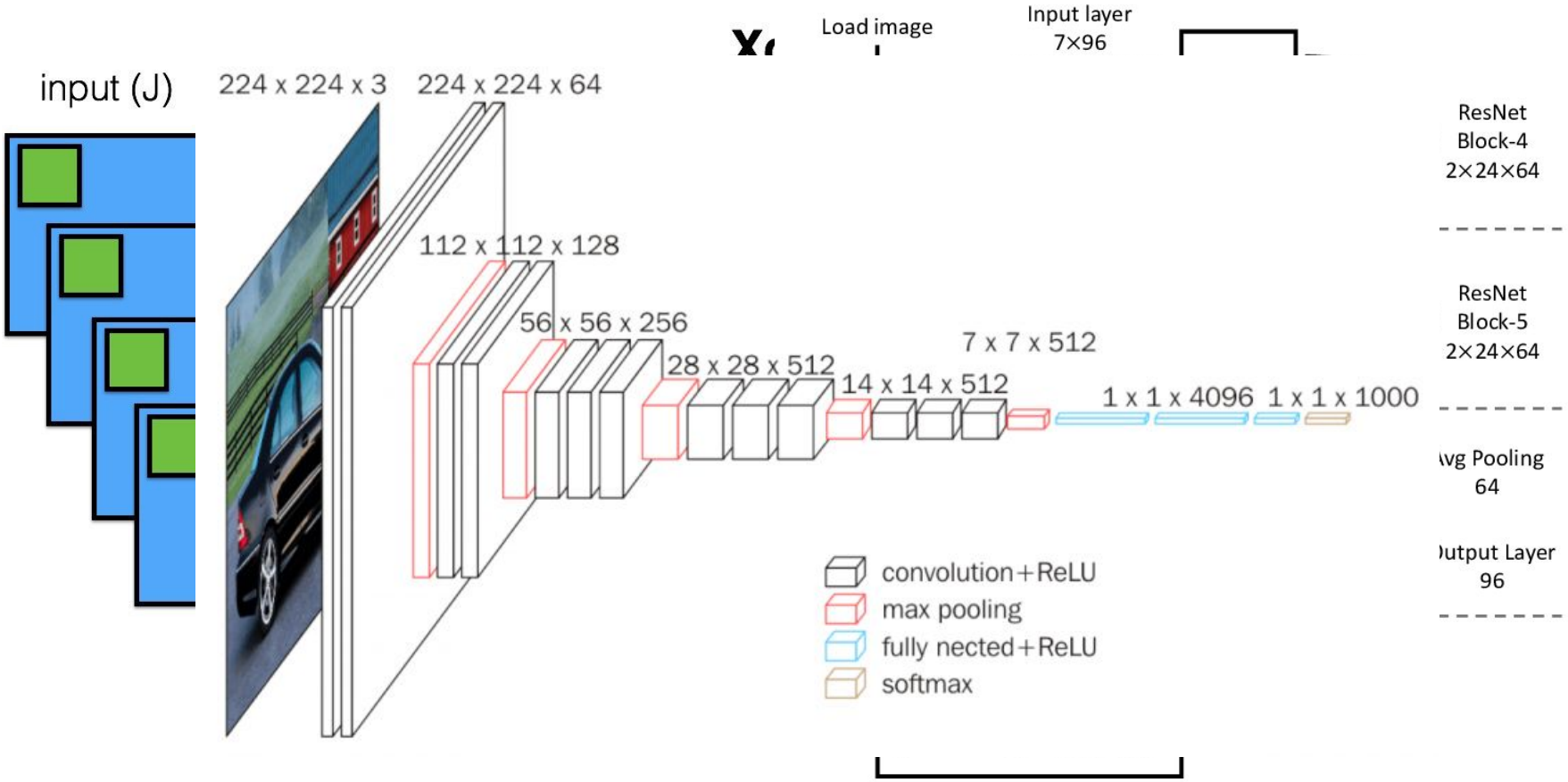


Convolutional Neural Network (CNN)



Convolutional Neural Network (CNN) cont.




Load image
Input layer
7×96
input (J)
224 x 224 x 3
224 x 224 x 64
112 x 112 x 128
56 x 56 x 256
28 x 28 x 512
14 x 14 x 512
7 x 7 x 512
1 x 1 x 4096
1 x 1 x 1000
convolution+ReLU
max pooling
fully nected+ReLU
softmax
ResNet
Block-4
2×24×64
ResNet
Block-5
2×24×64
vg Pooling
64
utput Layer
96

Optical flow

Amerini et al., 2019

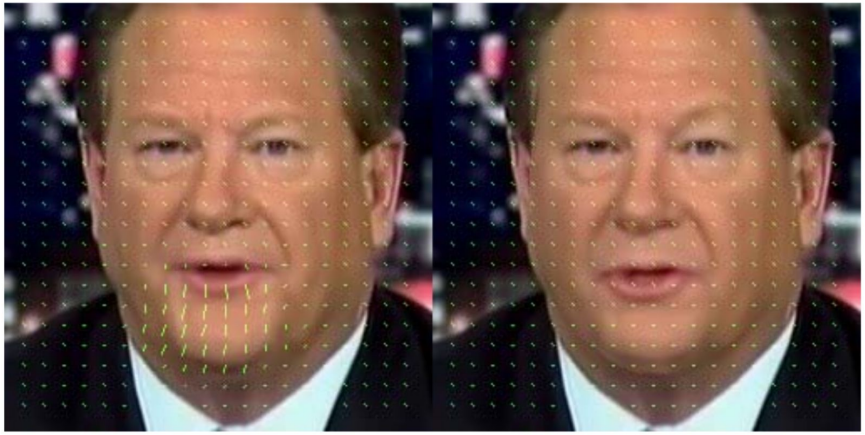
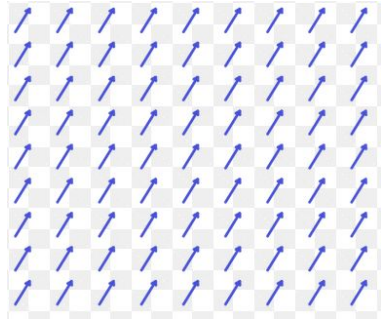
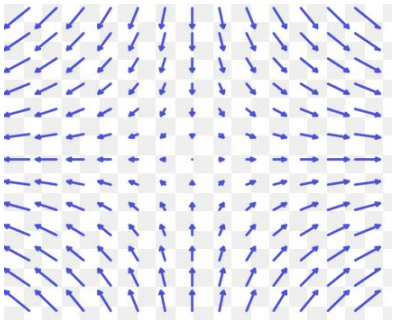


Figure 2. Optical flow for original (left) and deepfake (right) videos.

CNN’s with optical flow



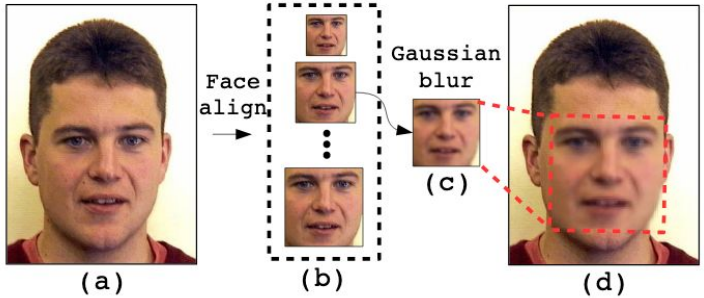


Convolutional Neural Network

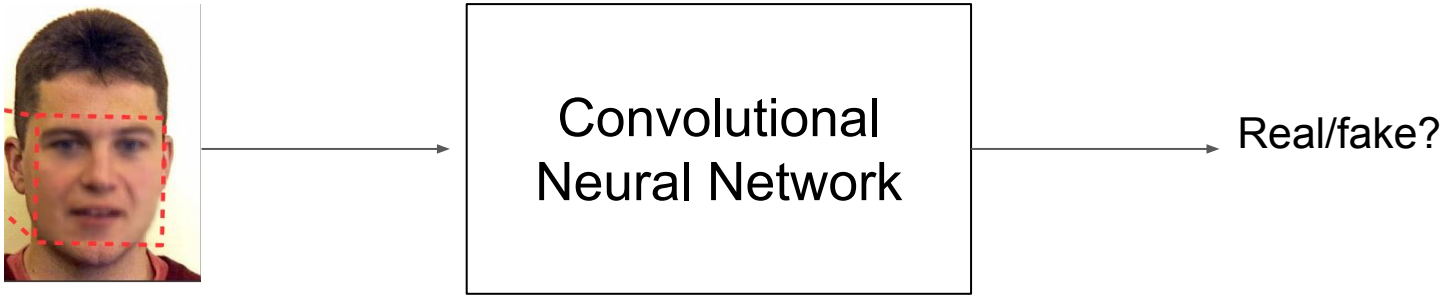
Real/fake?

CNN's with self-labeled data (Li et al., 2019)

1. Generate "negative" examples that contain deepfake generation artifacts

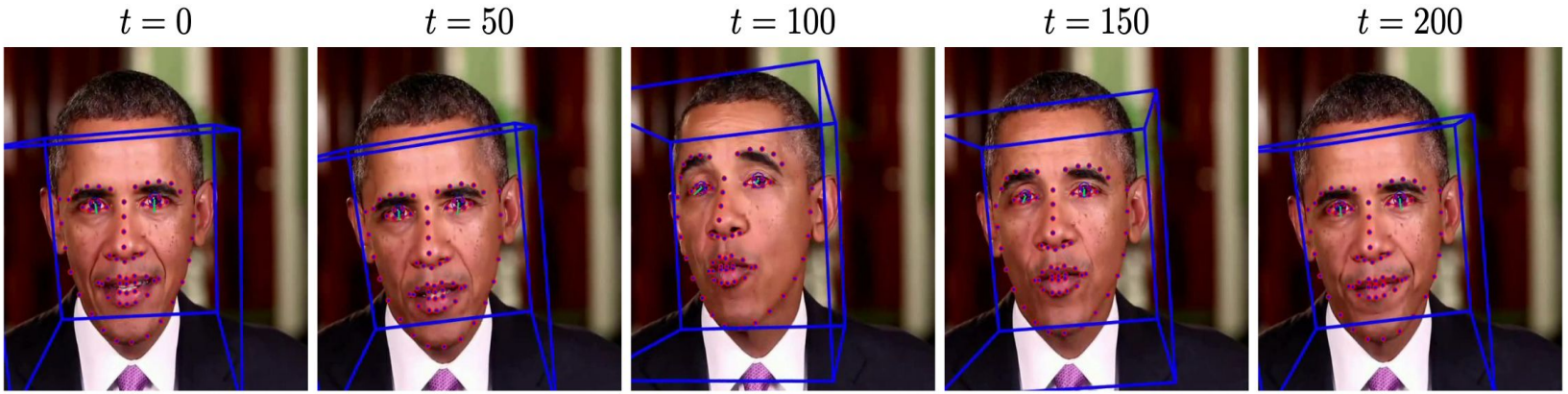


2. Use "negative" examples to train a CNN



Forensic deepfake detection

- Forensic approach
  - Generate correlations between facial features in a video to determine “signature motion” (Agarwal et al., 2019)



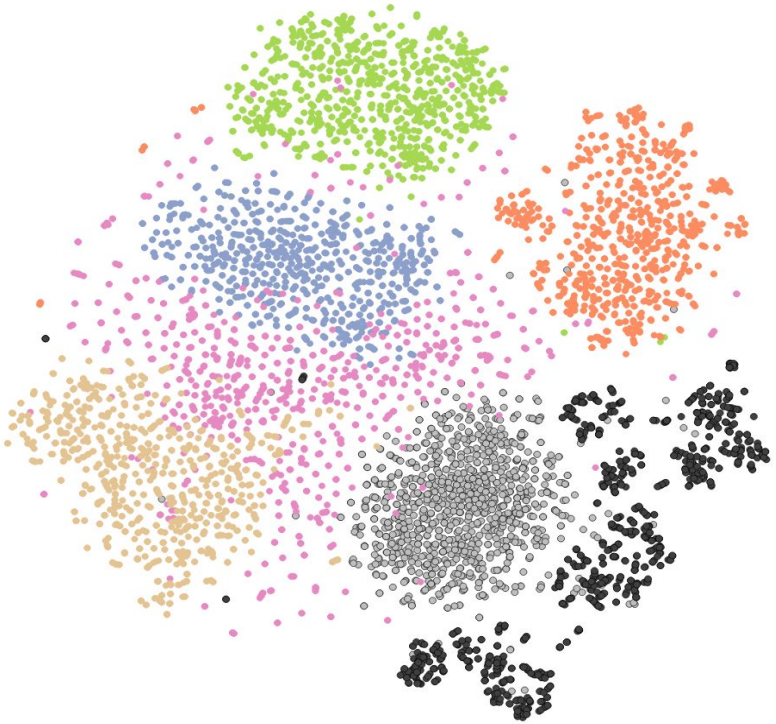
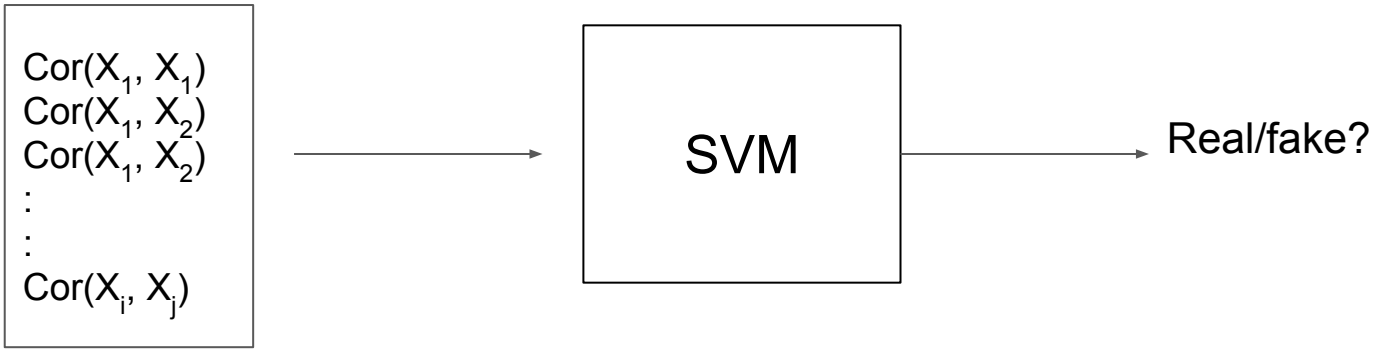


Figure 3. Shown is a 2-D visualization of the 190-D features for Hillary Clinton (brown), Barack Obama (light gray with a black border), Bernie Sanders (green), Donald Trump (orange), Elizabeth Warren (blue), random people [23] (pink), and lip-sync deep fake of Barack Obama (dark gray with a black border).

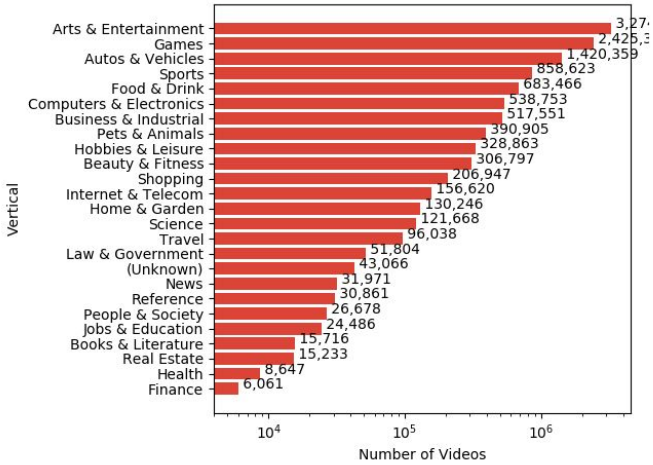
Forensic deepfake detection cont.



Our contribution

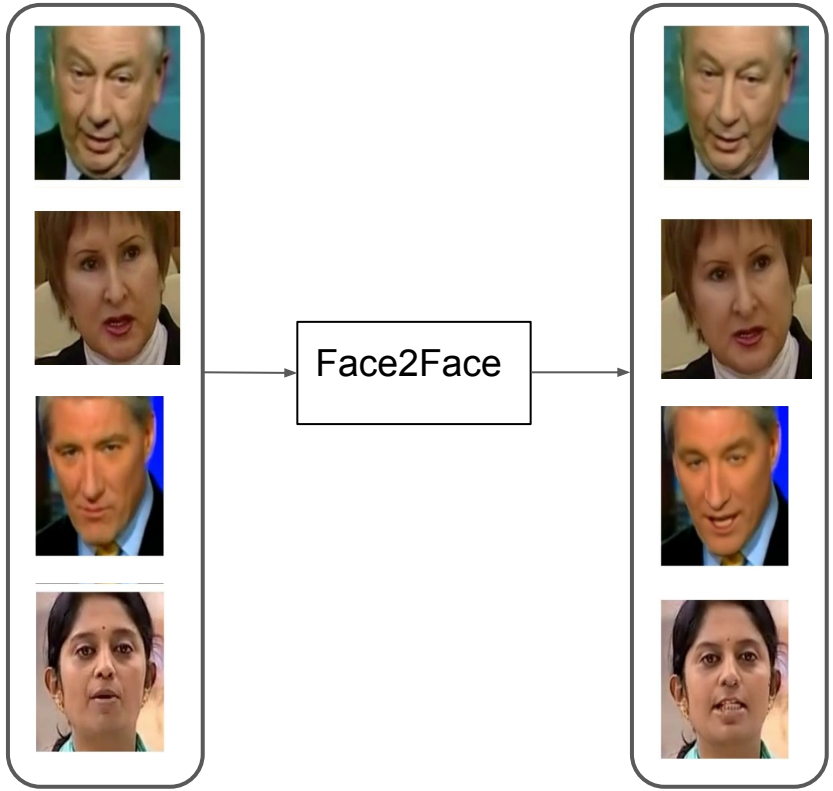
- We aim to improve upon existing neural network and forensic feature models.
  - ✓ Feature augmentation and enhancement
  - ✓ Better classification model

Dataset



Original labeled data

Altered labeled data



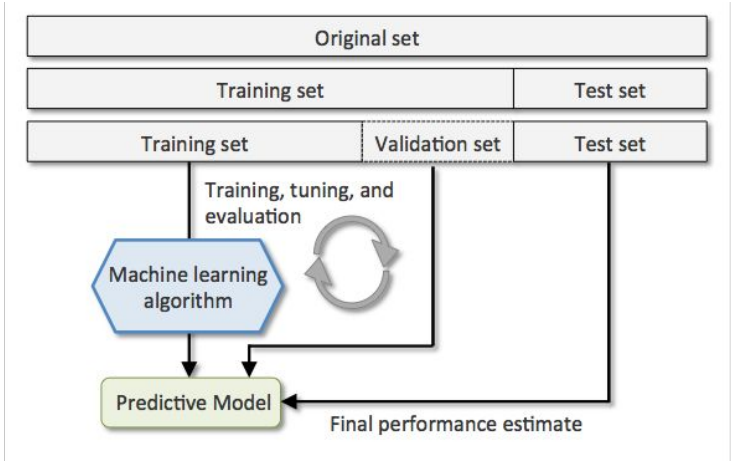
Entire YouTube 8M Dataset

Cropped faces from video frames

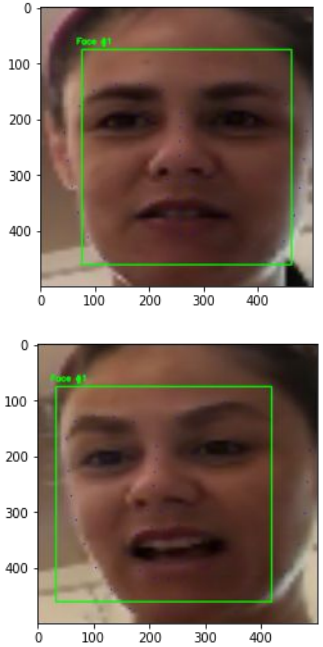
Face2Face manipulated video frames

Dataset cont.

- 704 videos for training (368,135 images)
- 150 videos for validation (75,526 images)
- 50 videos for testing (77,745 images)



Forensic analysis of facial landmarks



68 *(x,y)* coordinates = 136 features

PCA

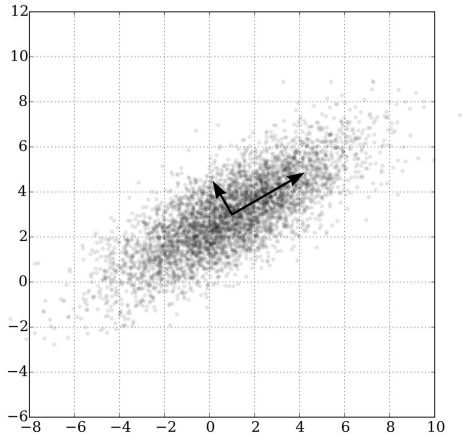
50 features

Classifier

Prediction

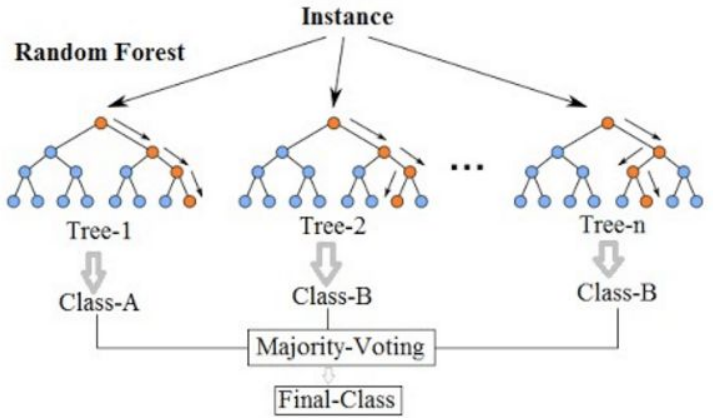
Principal Component Analysis (PCA)

- Popular technique for dimensionality reduction
- Transform feature space into orthogonal basis features, only capture most prominent features
- Fewer features → less variance, less overfitting



Method: Random forest classifier

- Pros:
  - Works with few features
  - Lower variance compared to regular decision tree
  - Explainable model
  - Low cost

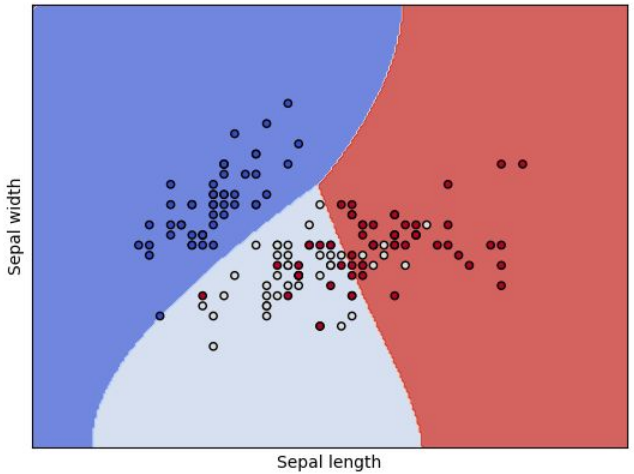


- Cons:
  - Hard to tune

https://towardsdatascience.com/random-forest-classification-and-its-implementation-d5d840dbead0

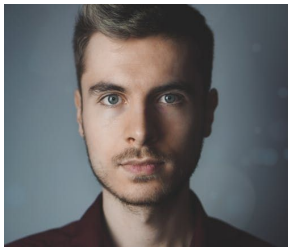
Method: Support vector machine

- Pros:
  - Supports non-linear decision boundaries
- Cons:
  - Hard to tune kernel and hyperparameters



https://pythonmachinelearning.pro/classification-with-support-vector-machines/

Method: Neural Network with facial landmarks



Facial Landmark detector

FC neural Net

Output

Loss: Cross Entropy loss

Features

PCA for dimension reduction

Pros:

- Lightweight --- single GPU training
- Large batchsize

Cons:

- Data hungry
- Need extensive tuning

Metrics

Accuracy: (True Positive + True Negative) / total samples

Precision: True Positives / All the predicted positives

Recall: True Positives / All the actual positives

Results: in-distribution samples (small scale)

- Near perfect performance for random forest
- What does this imply? We can perfectly detect fake/real across the web if we have label for part of a clip.
- 10K training images

	SVM	Random Forest	NN
Accuracy	80.00%	98.10%	85.12%

Table 1: Accuracy for different models

	Random Forest	NN
Precision	98.52%	92.81%
Recall	98.72%	85.01%

Table 2: Precision and Recall for top 2 models

Results: out-of-distribution training and testing

- Both methods drops significantly
- Neural Net performs slightly better (the training accuracy for NN is 90% and for random forest 99.9%)
- Training data is too little!
- 14K training images

	SVM	Random Forest	NN
Accuracy	N/A	70.50%	73.78%

Table 1: Accuracy of Random and NN model

	Random Forest	NN
Precision	77.15%	79.23%
Recall	58.82%	63.44%

Table 2: Precision and Recall for top 2 models

Public Benchmark Results w/ ~5 times our current training data

- Larger net
- More data
- Utilize video property

Method	Info	Deepfakes	Face2Face
Sentinel		**0.964**	**0.905**
Xception	P	0.964	0.869
Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Nießner: FaceForensics++: Learning to Detect Manipulated Facial Images. ICCV 2			
Inception Resnet V1		0.936	0.839
Nika Dogonadze, Jana Obernosterer: Deep Face Forgery Detection. Advanced Deep Learning for Computer Vision Course at TUM			
SCAN	C	0.909	0.825
EfficientNet-b4		0.955	0.796
Face Defence		0.945	0.766
swap_classify		0.909	0.759
XceptionNet Full Image	P	0.745	0.759
Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Nießner: FaceForensics++: Learning to Detect Manipulated Facial Images. ICCV 2			
Bayar and Stamm		0.845	0.737
Belhassen Bayar and Matthew C. Stamm: A deep learning approach to universal image manipulation detection using a new convolutional layer. ACM Workshop on Informatic			
Steganalysis Features		0.736	0.737
Jessica Fridrich and Jan Kodovsky: Rich Models for Steganalysis of Digital Images. IEEE Transactions on Information Forensics and Security			
GAEL-Net		0.718	0.686
Recasting		0.855	0.679
Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva: Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detectio Hiding and Multimedia Security			
Forged face detection : fCNN		0.791	0.642
Rahmouni		0.855	0.642
Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen: Distinguishing computer graphics from natural images using convolution neural networks. IEEE Worl Security,			

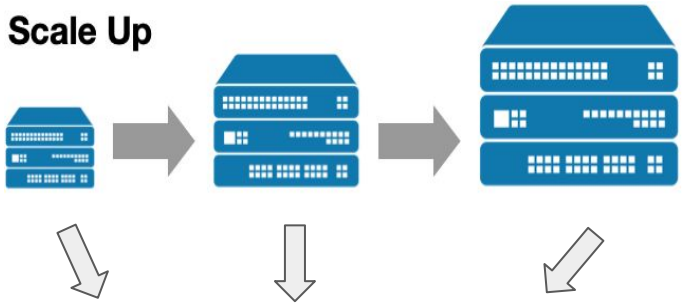
http://kaldir.vc.in.tum.de/faceforensics_benchmark/index.php?sortby=dface2face

Visualized Examples

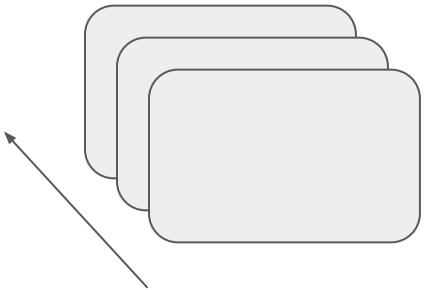
Original Image

Altered Image

Next steps

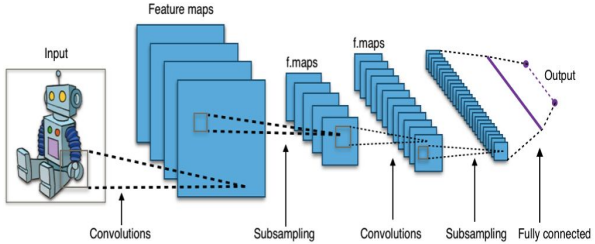


Scale up & Analysis



Temporal Features

Compare with public Benchmark



CNN + Forensic Features

Thank you!