Deepfake detection through PRNU and logistic regression analyses

3. July 2020

Catherine de Weever
Sebastian Wilczek

Supervisor:
Prof. dr. ing. Zeno Geradts
Netherlands Forensic Institute

The Problem of Perception

[1] Source: https://www.dijitalx.com/wp-content/uploads/2019/09/deepfake_amyadams_nicholascage.png

How can a forged Deepfake video be differentiated from an authentic one, for forensic purposes?

How can a forged Deepfake video be differentiated from an authentic one, for forensic purposes?

1. What detection methods are already available?
2. Are these detection methods still applicable to modern Deepfakes?
3. If these methods are still applicable, can they be enhanced?
4. If these methods are not applicable anymore, what other approaches could be taken?

Related Work

Video Characteristics

- Retrieving values from video files for comparison
- Koopman, M., Rodriguez, A. M., Geradts, Z.
- PRNU cross-correlation

Data-driven

- Capturing specific artifacts
- Visual features
- Matern, F., Riess, C., Stamminger, M.
- Deep neural network & logistic regression model

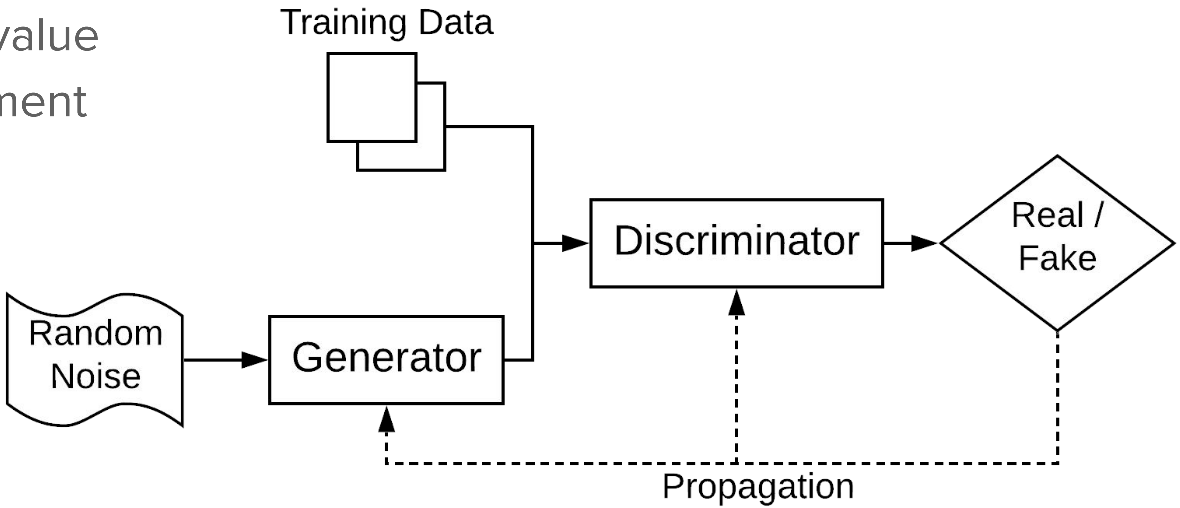
Frequency Domain

- Freq. Domain analysis followed by classifier
- Durall Lopez, R. et al.
- Supposed 100% accuracy

Background Information

GAN

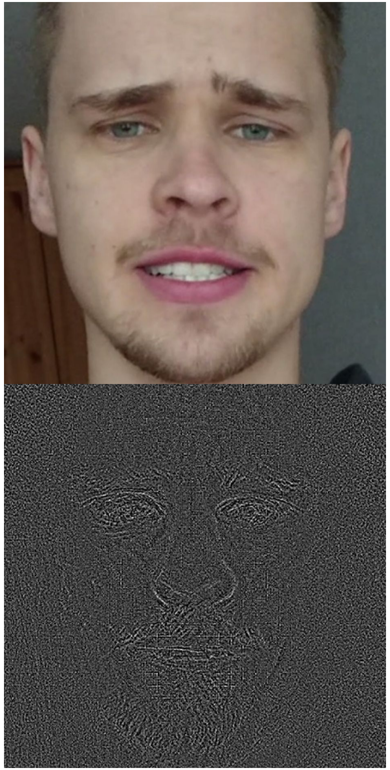
- Generative Adversarial Network
- Random, influenced creation
- Propagation of Loss value
- Continuous improvement



PRNU

- Photo Response Non Uniformity
- Fingerprint of digital camera
- Inhomogeneity in silicon of sensor
- Photons translated to electrons slightly differently
- Cross-correlation between patterns

➜ Likelihood of originating from same camera



Logistic Regression

- Classification algorithm
  - predicts the probability of a target variable
- Solves binary classification problems

$$y' = \frac{1}{1 + e^{-(z)}}$$

$$z = b + w_1 x_1 + w_2 x_2 + \ldots + w_N x_N$$

[2]

- Probability score between 0 and 1
  - mapped to a binary category by classification threshold
- Metrics used for evaluating classification model's predictions
  - Accuracy -> best classification threshold is known
  - receiver operating characteristic curve (ROC) curve -> N classification thresholds
    - Area under the ROC curve (AUC) -> aggregate measure of the performance

Experiments

Data Set Retrieval

FaceForensics++

Technical University of Munich

Celeb-DF

University at Albany

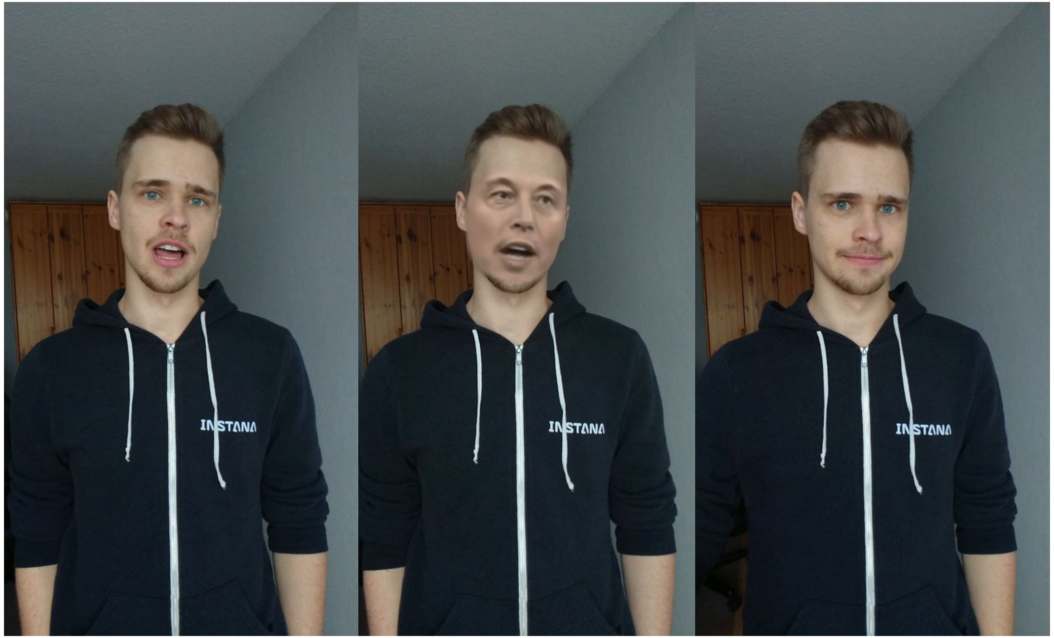
Own creations

University of Amsterdam

Deepfake Creation

- Created using *DeepFaceLab*
- 60.000 - 65.000 Iterations
- *Quick96*
- ca. 10 seconds (220-260 frames)

- NVIDIA Quadro P5000
- Intel Xeon E5-2678 v3
- Sony G8341 Xperia XZ1



Original Fake Check

PRNU Analysis

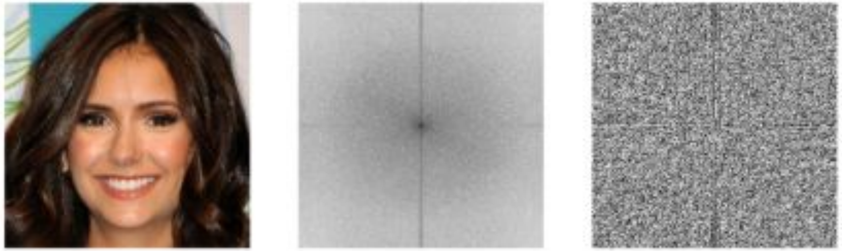
- PRNU Compare
  - developed by NFI
  - extracts patterns
  - computes cross-correlation
- Comparisons
  - Original
  - Deepfake
  - Check (where available)
- Classification
  - High correlation ➔ Original
  - Low correlation ➔ Fake

Visual Artifact Analysis

[3]

- VA
  - developed by Matern, et al.
  - captures visual artifacts in the eyes, teeth, and facial contours
  - 2 variants
    - VA-MLP ➔ small neural network
    - VA-Logreg ➔ logistic regression model
  - detect and segment ➔ feature extraction ➔ classify
- Non-trained model ➔ default classifiers
- Trained model ➔ created classifiers using the extracted features
- Classification
  - AUC score > 0.5 ➔ deepfake
  - AUC score ≤ 0.5 ➔ original

Frequency Domain Analysis

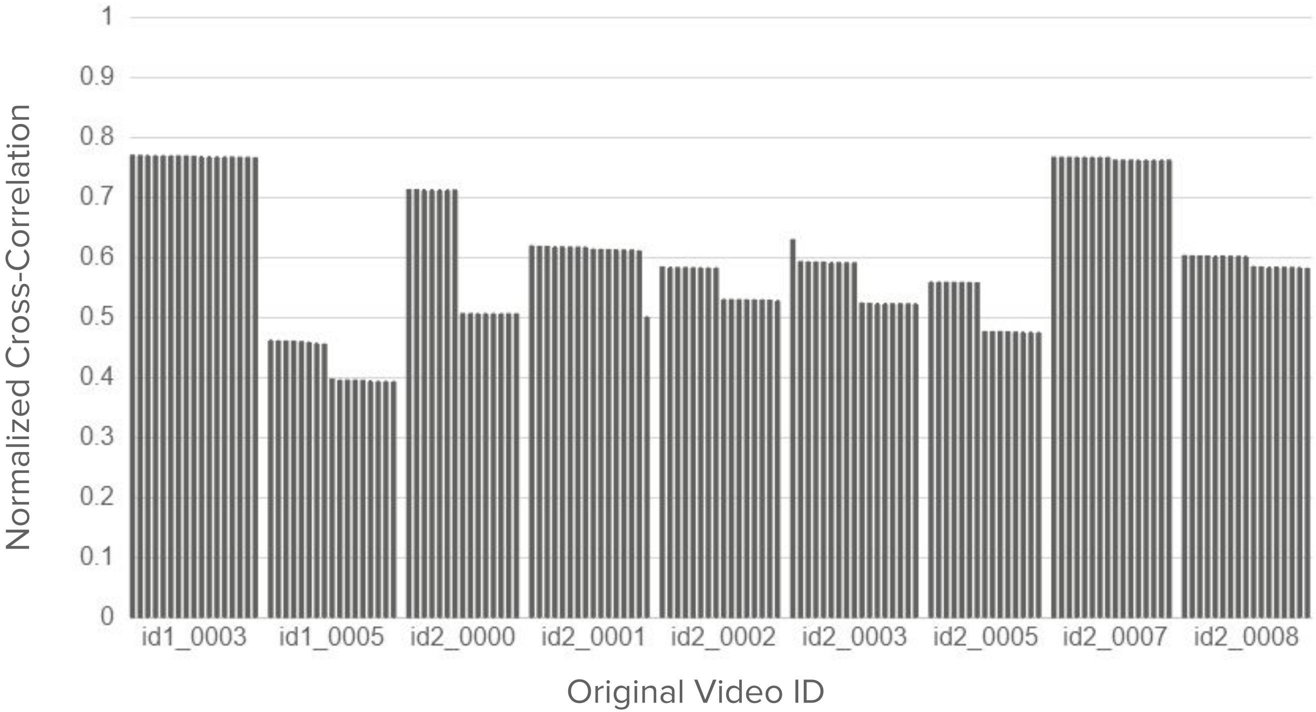


[4]

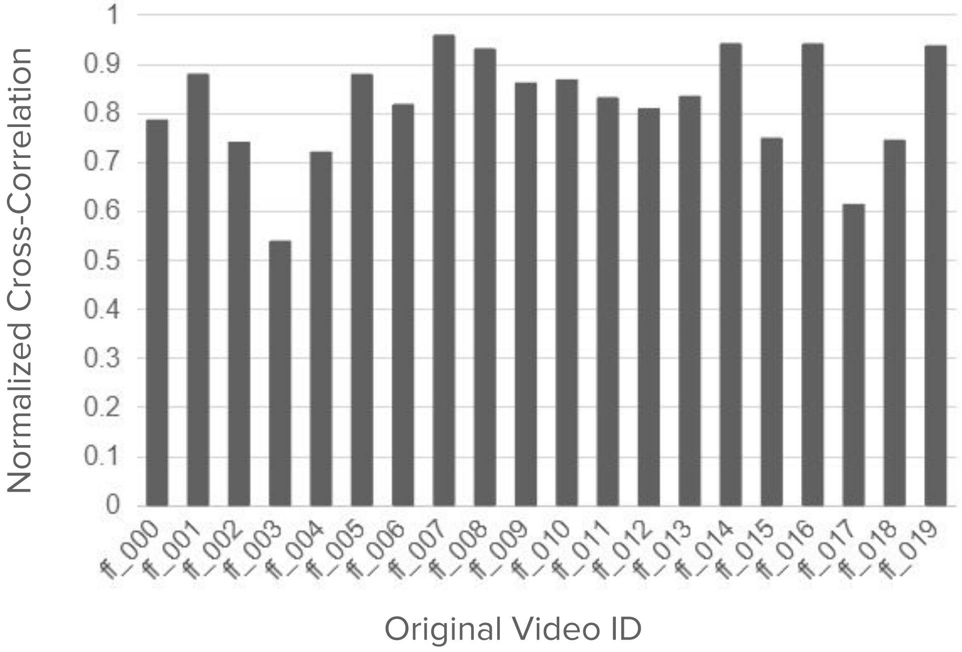
- DeepFakeDetection
  - developed by Durall Lopez et al.
  - frequency domain of an image
  - both real and fake images needed
  - DFT → Azimuthal averaging → Classify
- Data preparation
  - ran face detection
  - square images
- Classification
  - logistic regression model
  - Accuracy (average classification rate) > 0.5 → deepfake

Results

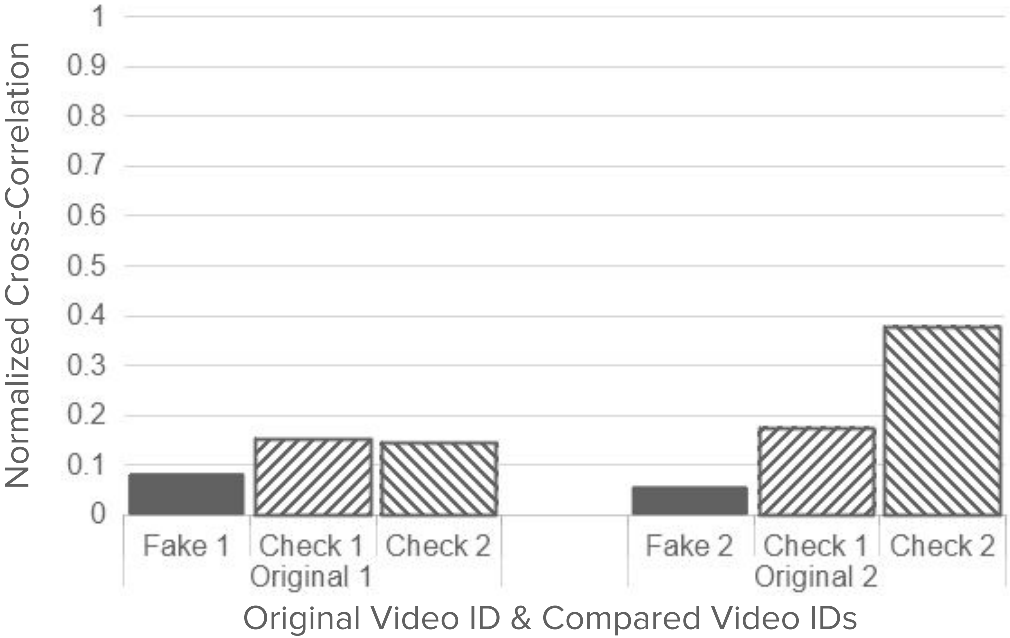
PRNU Analysis - Celeb-DF



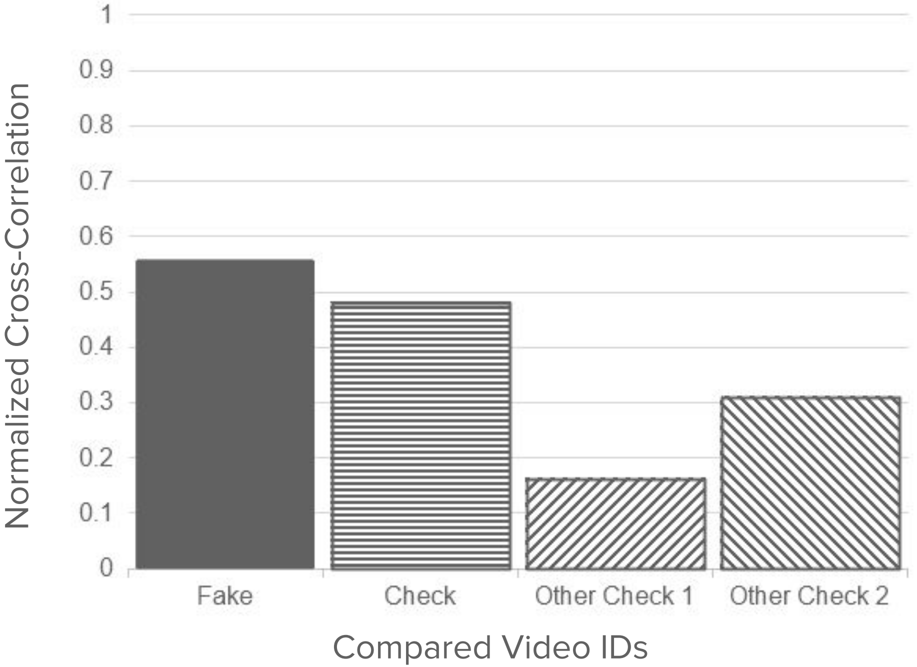
PRNU Analysis - FaceForensics++



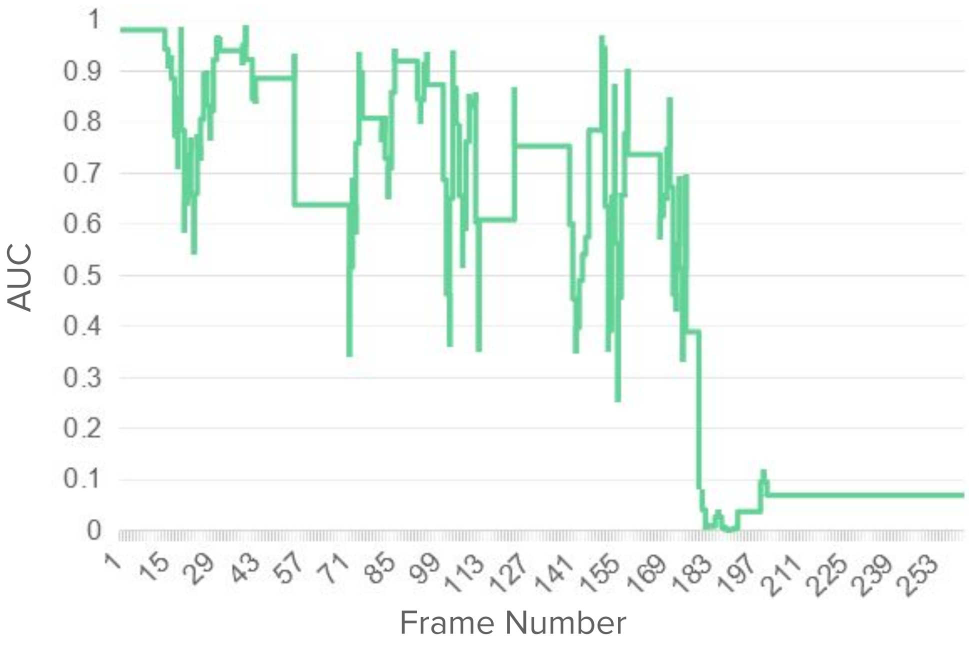
PRNU Analysis - Own Deepfakes



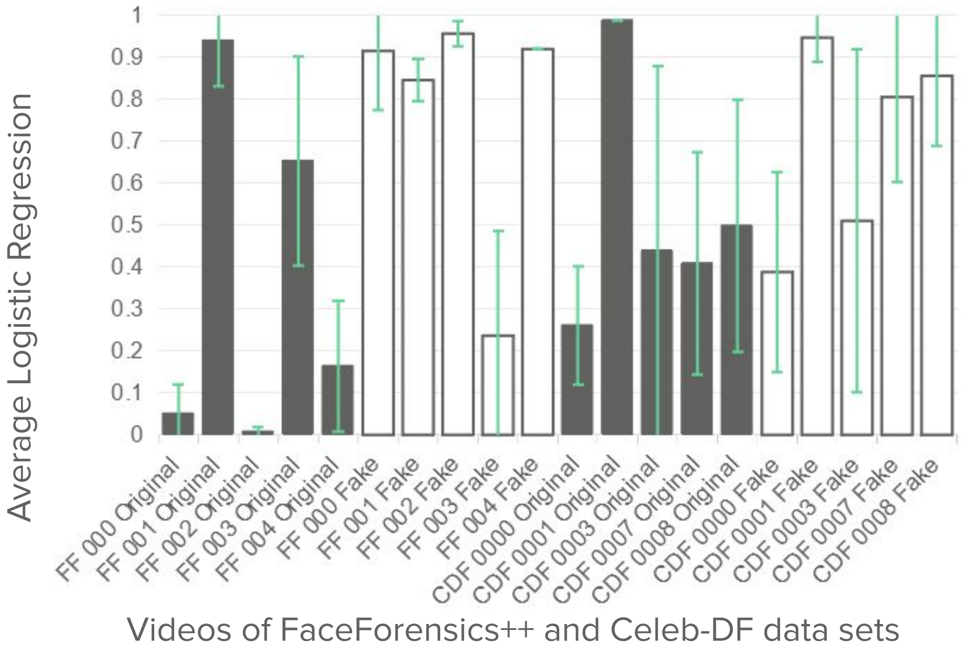
PRNU Analysis - Own Deepfakes, stabilized



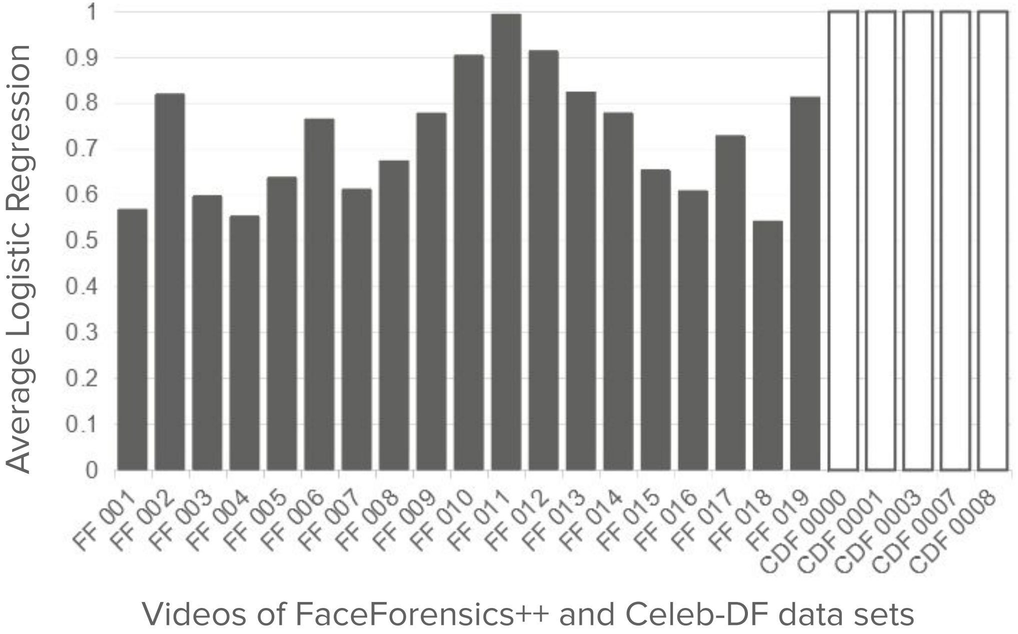
Visual Artifact Analysis - partial Deepfake and original



Visual Artifact Analysis - Faceforensics++ & Celeb-DF

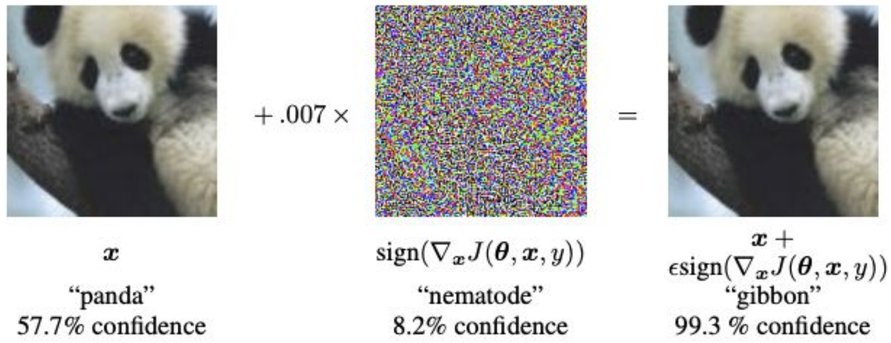


Frequency Domain Analysis - FaceForensics ++ & Celeb-DF



Detection Evasion

- methods using Logistic Regression
  - Misclassification ➜ lower AUC
  - Adversarial Attack using FGSM
    - Add random noise
- PRNU pattern influenced by camera movements
  - Lower cross-correlation scores



[5]

$$adv\_x = x + \epsilon * \text{sign}(\nabla_x J(\theta, x, y))$$

[5]

Media Authentication - The Problem

- GAN can use detection for improvement
- Working towards detection makes Deepfakes better
- Potential solution: Find original instead of fake media
- Requires Provenance Information
  - Chain of origin and modification
  - Problem: Propagation

Media Authentication - Provenance Propagation

- Public key infrastructure
  - Everyone needs public keys
  - Entire file required to read signature ➜ streaming?
  - Files often edited without breaking authenticity
- More sophisticated systems
  - AMP (Microsoft)
  - Manifests with various hashes and metadata
  - Stored with file / in database / using blockchain
  - Watermarking
  - Works for file chunks
  - Pointers to source files ➜ Provenance
  - Other projects (Adobe, The New York Times)

Conclusion

Conclusion

- Varying results for PRNU analysis
  - No cut-off value determinable
  - Drawback: requires comparison media
- Visual artifact analysis unreliable
  - Plenty of invalid frames
  - Results ambiguous for both trained and untrained approach
  - Only needs Deepfakes
- Frequency Domain analysis biased
  - Different data sets return different results
  - Requires both Deepfake and original media
- Detection Evasion
  - Movement for PRNU, FGSM for logistic regression model

Conclusion

- GAN will improve
  - Alternative approaches required
- Provenance systems are promising
- System Design
  - Access, Security, Availability
- Ethical concerns
  - Responsible organizations
  - Perception of unsigned media
  - Privacy of sources

Discussion

- Insufficient data for PRNU analysis
- Little time → Automation
- Data sets not fitting for visual artifact and frequency domain analysis
- Evasion and Media Authentication only discussed in theory

Future Work

- Extension of Measurements
- More custom Deepfakes & Check videos
- Neural network alternatives
- GAN-proof detection
- Design of Media Authentication system
- Ethics of Media Provenance

Thank you for your attention!

Catherine
de Weever

Sebastian
Wilczek

sebastian.wilczek@protonmail.com

https://github.com/sebastianwilczek

References

1. https://www.dijitalx.com/wp-content/uploads/2019/09/deepfake_amyadams_nicholascage.png
2. https://developers.google.com/machine-learning/crash-course/logistic-regression/calculating-a-probability
3. Matern, F., Riess, C., and Stamminger, M. “Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations”. In: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). 2019, pp. 83–92.
4. Durall Lopez, R. et al.Unmasking DeepFakes with simple Features. Available at: https://arxiv.org/abs/1911.00686v3.pdf
5. https://www.tensorflow.org/tutorials/generative/adversarial_fgsm