DeepFake Detection

Baran Deniz KORKMAZ

September 2020

1 Problem Definition

- Free access to large-scale public databases
- Fast progress of deep learning techniques, especially GANs

have boosted the emergence of deepfake creation.
Such creations can be used with a malevolent intentions, therefore pose a huge threat into privacy.

2 Deepfake Detection

The creation of deepfakes has been rapidly responded by the emergence of detection mechanisms using different kinds of deep and shallow models.

The approaches developed in deepfake detection can be primarily categorized into two classes:

1. Traditional Approach: Contains preliminary detection systems based on analyzing intrinsic and extrinsic features that are existing within media sources.
    - Extrinsic Features: Aims at embedding external unique signals into original images (e.g. digital watermarking). Then the received image can be verified whether it is fake or not by comparing the extracted watermark and the original watermark.
    - Intrinsic Features: Aims at discovering intrinsic and invariant features from the original images. The forgery image should be able to detected by checking the statistical property of the extracted intrinsic feature from the received image because any tampered operation will make the intrinsic feature changed. In summary, this approach aims at finding unusual statistical properties within the images.
2. Modern Approach: Traditional forgery detection techniques do not respond into the requirements for the fake images generated by GANs since their image content are made by deep neural network directly. Therefore,

such images does not contain any abnormalities in statistical properties in the intrinsic features, leading traditional forgery detection approach to fail:

- Highly dependent on the specific training scenario.
- Not robust against unseen conditions.

To overcome this problem, deep neural network architectures are utilized to form the basis for deepfake detection systems.

3 Why Deepfake Detection?

- Deepfakes can be abused to cause political, social, religious, and economic misunderstandings within society. They can be used in a range variety of malicious actions such as misleading public, affecting results in elections, creating chaos in financial markets, blackmailing, violation of privacy and identity, etc.
- Posing a significant threat in a broad range of different aspects, creation of detection systems is inevitable.

4 Deepfake Creation

- Understanding the creation procedures of deepfakes are key to produce an efficient and effective solution.
- The evolution of deep learning techniques have eliminated many manual editing steps in deepfake creation. Modern deepfake generation techniques are mainly based on the following deep learning techniques:
  - Autoencoders
  - GANs
- It is interesting to remark that each fake images may be characterized by a spesific GAN fingerprint just like natural images are identified by a device-based fingerprint (i.e. PRNU - Considered as the fingerprint of digital cameras left in the images by the cameras.)
  CHALLENGE: The recent studies mainly provide good accuracy results for the database(s) used in training (however we still observe poor generalization issues). This is mainly because the GAN-fingerprint information present in fake images generated by GANs (Karras et al.). However, recent studies have been proposed in the literature to remove such GAN fingerprints from the fake images while keeping very realistic appearance. This situation will lead into bigger challenges even for the most advanced manipulation detectors in the future.

- The studies mainly focus on 4 kinds of face manipulations in the genearation of deepfake detection systems:

    1. Entire Face Synthesis: This manipulation creates entire non-existent face images, usually through powerful GANs (e.g. StyleGAN, ProGAN).
    2. Identity Swap: This manipulation consists of replacing the face of one person in a video with the face of another person.
    3. Attribute Manipulation: This manipulation consists of modifying some attributes of the face such as the colour of the hair or the skin, the gender, the age, adding glasses, etc.
    4. Expression Swap (Face Reenactment): This manipulation consists of modifying the facial expression of a person.

- The studies will be presented according to the manipulation they focus on to detect.

5 Challenges

1. Deepfake detection is normally deemed as a binary classification problem. The binary classification approach requires a large database of real and fake images/videos for training.

    - This problem is addressed by the use of GANs in the creation of deepfakes (Koshunov and Marcel - The creators of **Deepfake TIMIT** DB)

2. The release of advanced deepfake creation techniques with the use of autoencoders and GANs, deepfake detection has become a more challenging subject.

3. Training the detection network by deepfakes generated by all kinds of GANs is not only infeasible but also extremely hard. We must also focus pay attention that GANs are constantly evolving and advancing. Most studies do not present a satisfactory generalization. Recent studies aim at finding generative solutions for deepfake detection.

4. Most image detection methods cannot be used for videos because of the strong degradation of the frame data after video compression. Furthermore, videos have temporal characteristics that are varied among sets of frames.

5. The detection in low quality videos is relatively harder, therefore an usual degradation in deepfake detection accuracy in studies has been observed. **Sabir et al. has studied a deepfake detection system that analyzed only the low quality videos which we will see later.**

6 State-of-Art Techniques

6.1 Dang et al.

- **Study:** Dang et al. (2020)
- **Method:** Deep Learning Features
- **Classifiers:** CNN + Attention Mechanism
- **Performance:**
  - ENTIRE FACE SYTHESIS: AUC = 100% EER = 0.1%
  - IDENTITY SWAP: AUC = 99.4% EER = 3.1%
  - ATTRIBUTE MANIPULATION: AUC = 99.9% EER = 1.0%
  - EXPRESSION SWAP: AUC = 99.4% EER = 3.4%
- **Databases:** DFFD(ProGAN, StyleGAN)
- **Description:**
  - Tested in the following facial manipulation types: Entire Face Sythesis, Identity Swap, Attribute Manipulation, Expression Swap
  - Attention mechanisms have been applied to further improve the training process of the detection systems.
  - Use of attention mechanisms and popular CNN models such as XCeption-Net and VGG16 have been proposed.
  - Impressive results show the importance of novel attention mechanisms.

6.2 Li et al.

- **Study:** Li et al. (2019)
- **Method:** Face Warping Features
- **Classifiers:** CNN
- **Performance:**
  - IDENTITY SWAP:
    - AUC = 97.7% - UADFV
    - AUC = 99.9% - DeepfakeTIMIT (LQ)
    - AUC = 99.7% - DeepfakeTIMIT(HQ)
    - AUC = 93.0% - FF++/DFD
    - AUC = 75.5% - DFDC Preview
    - AUC = 64.6% - Celeb-DF

- **Databases:** UADFV, DeepfakeTIMIT (LQ/HQ), FF++/DFD, DFDC Preview, Celeb-DF
- **Description:**
  - Tested in the following facial manipulation types: Identity Swap
  - Li and Lyu suggested that some DeepFake algorithms can only create images with limited resolution, which need to be further warped to match the original faces in the source video.
  - Such transforms leave distinctive artifacts in the resulting fake video.
  - The authors proposed a detection system based on CNNs to detect the presence of such artifacts from the detected face regions and the surrounding areas.
  - The proposed detection approach outperforms the state of the art for UADFV and DeepfakeTIMIT DBs.

6.3 Rössler et al.

- **Study:** Rössler et al. (2019)
- **Method:** Mesoscopic Features, Steganalysis Features, Deep Learning Features
- **Classifiers:** CNN
- **Performance:**
  - IDENTITY SWAP:
    1. * Acc. = 94.0% FF++ (Deepfake,LQ)
       * Acc. = 98.0% FF++ (Deepfake,HQ)
       * Acc. = 100.0% FF++ (Deepfake,RAW)
    2. * Acc. = 93.0% FF++ (Deepfake,LQ)
       * Acc. = 97.0% FF++ (Deepfake,HQ)
       * Acc. = 99.0% FF++ (Deepfake,RAW)
  - EXPRESSION SWAP:
    1. * Acc. = 91.0% FF++ (Deepfake,LQ)
       * Acc. = 98.0% FF++ (Deepfake,HQ)
       * Acc. = 100.0% FF++ (Deepfake,RAW)
    2. * Acc. = 81.0% FF++ (Deepfake,LQ)
       * Acc. = 93.0% FF++ (Deepfake,HQ)
       * Acc. = 99.0% FF++ (Deepfake,RAW)
- **Databases:** FaceForensics++
- **Description:**

- Tested in the following facial manipulation types: Identity Swap, Expression Swap
- Five different detection systems were evaluated:
  1. a CNN-based system trained through handcrafted steganalysis features
  2. a CNN-based system whose convolution layers are specifically designed to suppress the high-level content of the image
  3. a CNN-based system with a global pooling layer that computes four statistics (mean, variance, min, max)
  4. the CNN MesoInception-4 detection system
  5. the CNN-based system XCeptionNet pretrained using ImageNet database and re-trained for the face manipulation detection task (provided best results)
- The detection systems were also evaluated using different video quality levels
- The accuracy of all detection systems decreased as the video quality gets lower, **remarking how challenging this task is in real scenarios**

6.4 Sabir et al.

- **Study:** Sabir et al. (2019)
- **Method:** Image + Temporal Features
- **Classifiers:** CNN + RNN (LSTM)
- **Performance:**
  - IDENTITY SWAP:
    - AUC = 96.9% - FF++ (Deepfake,LQ)
    - AUC = 96.3% - FF++ (FaceSwap,LQ)
  - EXPRESSION SWAP:
    - Acc. = 94.3% - FF++ (Face2Face,LQ)
- **Databases:** FaceForensics++
- **Description:**
  - Tested in the following facial manipulation types: Identity Swap, Expression Swap
  - Detection of fake videos based on using the temporal information present in the stream.
  - The intuition is to exploit temporal discrepancies across frames.
  - **Only the low-quality videos were considered in the analysis.**

6.5 Tolosana et al.

- **Study:** Tolosana et al.
- **Method:** Facial Regions Features
- **Classifiers:** CNN
- **Performance:**
  - IDENTITY SWAP:
    - AUC = 100.0% - UADFV
    - AUC = 99.4% - FF++ (FaceSwap, HQ)
    - AUC = 91.0% - DFDC Preview
    - AUC = 83.6% - Celeb-DF
- **Databases:** UADFV, FF++, DFDC Preview, Celeb-DF
- **Description:**
  - Tested in the following facial manipulation types: Identity Swap
  - The discriminative power of each facial region for the detection of fake videos were studied.
  - Detection system is based on XCeptionNet.
  - Lower accuracy levels on 2nd generation databases are observed compared to 1st generation databases.
  - **A separate fake detection system was spesifically trained for each database.**

6.6 Afchar et al.

- **Study:** Afchar et al. (2018)
- **Method:** Mesoscopic Features
- **Classifiers:** CNN
- **Performance:**
  - IDENTITY SWAP:
    1. ∗ Acc. = 98.4% Own
    2. ∗ AUC = 84.3% UADFV
    3. ∗ AUC = 87.8% DeepfakeTIMIT (LQ)
       ∗ AUC = 68.4% DeepfakeTIMIT (HQ)
    4. ∗ Acc. = 90.0% FF++ (Deepfake,LQ)
       ∗ Acc. = 94.0% FF++ (Deepfake,HQ)
       ∗ Acc. = 98.0% FF++ (Deepfake,RAW)

5. * Acc. = 83.0% FF++ (FaceSwap,LQ)
   * Acc. = 93.0% FF++ (FaceSwap,HQ)
   * Acc. = 96.0% FF++ (FaceSwap,RAW)
6. * AUC = 75.3% DFDC Preview
7. * AUC = 54.8% Celeb-DF

- EXPRESSION SWAP:
  1. * Acc. = 83.2% FF++ (Face2Face,LQ)
     * Acc. = 93.4% FF++ (Face2Face,HQ)
     * Acc. = 96.8% FF++ (Face2Face,RAW)
  2. * Acc. = 75.0% FF++ (NeuralTextures,LQ)
     * Acc. = 85.0% FF++ (NeuralTextures,HQ)
     * Acc. = 95.0% FF++ (NeuralTextures,RAW)

- **Databases:** Own (Private), UADFV, DeepfakeTIMIT (LQ/HQ/RAW), FF++, DFDC Preview, Celeb-DF
- **Description:**
  - Tested in the following facial manipulation types: Identity Swap, Expression Swap
  - Proposed two different networks composed of few layers to focus on mesoscopic features of the images:
    1. a CNN network comprised of 4 convolutional layers followed by a fully-connected layer
    2. a modification of Meso-4 consisted of a variant of the Inception module, named MesoInception-4
  - The pre-trained model was tested against unseen databases after tested on the private database and the results have proved their approach to be **robust** in some cases such as with FaceForensics++.

6.7 Hsu et al.

- **Study:** Hsu et al. (2020)
- **Method:** Two-Phase Deep Learning
- **Classifiers:**
  1. Stage 1: CFFN, Siamese Network
  2. Stage 2: CNN
- **Databases:** Celeb-A
- **Description:**

– Experimental results show the superior performance of proposed method against its competing methods that are:
    1. Hsu et al. (2018)
    2. Farid et al. (2009)
    3. Huaxiao Mo et al. (2018)
    4. Marra et al. (2018)

7 Available Data Sets

Publicly available databases can be classified according to the manipulation technique used in the creation of their content:

1. Entire Face Synthesis:

Database	Real Images	Fake Images
100K-Generated Images (Karras et al. (2019))	-	100K (StyleGAN)
100K-Faces	-	100K (StyleGAN)
DFFD (2020)	-	100K (StyleGAN) 200K (ProGAN)
iFakeFaceDB (2020)	-	250K (StyleGAN) 80K (ProGAN)

2. Identity Swap:

1st Generation

Database	Real Images	Fake Images
UADFV (2018)	49 (YouTube)	49 (FakeApp)
DeepfakeTIMIT (2018)	-	620 (FaceSwap-GAN)
FaceForensics++ (2019)	1000 (YouTube)	1000 (FaceSwap) 1000 (DeepFake)

2nd Generation

Database	Real Images	Fake Images
DeepFakeDetection (2019)	363 (Actors)	3068 (DeepFake)
Celeb-DF (2019)	890 (YouTube)	5639 (DeepFake)
DFDC Preview (2019)	1131 (Actors)	4119 (Unknown)

3. Attribute Manipulation:
    - Diverse Fake Face Dataset (DFFD) - 18,416 - Face App, 79,960 - StarGAN (Fake Images)
4. Expression Swap:
    - FaceForensics++

Other Public Datasets

The following data sets include real&fake samples together or can be used as a base for deepfake creation. Creation of deepfakes using FaceSWAP-GAN is a good example conducted by Kashunov and Marcel [46].

1. VidTIMIT: Comprised of video and corresponding audio recordings of 43 people, reciting short sentences.

2. Idiap Research Institute replay-attack data set.

3. Deepfake Face Swapping data set created by Afchar et al. [1-83].

4. The facial reenactment FaceForensics data set created by Face2Face method [1-99].

5. The fully computer-generated image data set by Rahmouni et al. [1-101]

6. DARPA MediFor GAN Image/Video Challenge data set [102].

8 References

1. Deep Learning for Deepfakes Creation and Detection: A Survey; Nguyen et al.; 2020

2. DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection; Tolosana et al.; 2020