Model Card - DeepFake Detection Service

Model Details

- Developed by: CERTH-ITI Media Verification Team
- Model date: 03/02/2022
- Model version: 1.0. In this version, an ensemble of five models is deployed.
- Processing pipeline:
  - Download the image/video from the input URL.
  - In case of image:
    1. Use a Face Detector to detect all faces in the image.
    2. Feed each face to the model ensemble to get a DeepFake probability score in the range of $(0, 1)$.
  - In case of video:
    1. Segment the input video into shots.
    2. For each shot, use a Face Detector to detect faces in the shot's frames.
    3. Perform face clustering to discard wrongly detected faces from the detector and organize the remaining faces into groups.
    4. Feed each face to the model ensemble to get a DeepFake probability score in the range of $(0, 1)$.
    5.
- Model input: Video or image URL.
- Model output: The video-level DeepFake probability, and the probability for each detected person in each video shot. Probabilities below 50% and closer to 0% mean **real** and above 50% and closer to 100% mean **fake**.
- Model type:
  - DeepFake prediction: an ensemble model is used based on the average score of five models:
    1. a vanilla EfficientNet-b4,
    2. a Transformer head based on DETR with *fixed* positional embeddings on top of an EfficientNet-b4,
    3. a Transformer head based on DETR with *learned* positional embeddings on top of an EfficientNet-b4,
    4. a Multi-head Transformer based on DETR on top of an EfficientNet-b4,
    5. a vanilla EfficientNet-V2-m. This is the new model addition in this version.
  - Face Detection: we use the facenet-pytorch library.
  - Face Clustering: we employ the method described in this paper, where we extract face features using the pretrained InceptionResnetV1 provided in facenet-pytorch library and used DBSCAN for clustering.
  - Shot segmentation: the feature extraction and similarity calculation described in this paper are used to extract peaks in the graph of distances of the consecutive frames.
- Citation details: (CERTH-ITI Media Verification Team, 2022) MeVer DeepFake Detection service.
- Feedback and Contact: Spiros Baxevanakis (spirosbax@iti.gr), George Kordopatis-Zilos (georgekordopatis@iti.gr), Symeon Papadopoulos (papadop@iti.gr)

Intended Use

- Primary intended use: Detect whether the faces present in the input image or video have been manipulated/generated using Deep Learning methods (DeepFake).
- Primary intended users: Journalists, media verification companies/organizations/groups and researchers working on the problem of DeepFake detection.
- Out-of-scope uses:
  - The service cannot detect audio manipulations.
  - The service does not detect fully synthetic faces such as the ones produced by services like This Person Does Not Exist.
  - The service cannot detect if an image/video has been tampered with using non-facial manipulations or other forgeries (e.g. splicing, copy-move, in-painting).
  - The current version of the service does not process videos longer than 12 minutes containing more than 50 shots due to limits on computational resources. Refer to the *Caveats and Recommendations* section.

Relevant Factors

- Factors affecting service performance include:
  - Manipulations: whether the models have been trained with the presented DeepFake manipulation method or not. Refer to the *Training Data* section for more information.
  - Background faces: if there are many background low-resolution faces in the input image/video, it may affect the final prediction since all detected faces are taken into account equally.
  - Image/Video quality: blurry or low quality faces can lead to erroneous detections (false positives).
  - Adversarial Attacks: alterations in the images/videos to evade detection are detrimental to detection accuracy.

Metrics

- Model performance measures:
  - **Balanced Accuracy (BA)**: defined as the mean of the recall computed on each class. Possible values are in the range 0%-100%. Higher is better.
  - **AUC**: defined as the Area Under Curve (AUC) based on the Receiver Operating Characteristic (ROC) curve with possible values ranging from 0 to 1. Higher is better.
- Metrics decision: these are popular metrics used in the related literature, and they are also appropriate for imbalanced datasets.

Relevant Datasets

- FaceForensics++ (FF++): The dataset is organized in two manipulation categories, *Identity Swap*, implemented using *FaceSwap* and *DeepFakes*, and *Expression Swap*, implemented using *NeuralTextures* and *Face2Face*. FF++ contains 1000 real videos and 4000 fake videos derived by applying the four models on each real video. Evaluation on FF++ provides a performance indicator on different manipulation categories and methods. Compared to more recent datasets (e.g. CelebDF, DFDC.), the DeepFake quality in FF++ is visibly worse.
- CelebDF-V2 (CelebDF): Comprises videos from celebrity interviews that have been manipulated using improved versions of the DeepFake manipulation methods used in the FF++. It consists of 590 real and 5639 fake videos.
- DeepFake Detection Challenge (DFDC): Published by Facebook in the context of a DeepFake Detection Challenge, it

contains 20,000 videos from hundreds of paid actors that have been used to generate 100,000 manipulated videos using improved DeepFake, FaceSwap methods, and three GAN-based manipulations. Due to its size and quality, it is often used both in research and production.

- WildDeepFake (WDF): In contrast to the above datasets where the manipulations were applied by the dataset creators, this contains real-world DeepFakes collected from various video-sharing websites as well as their corresponding real versions. It consists of 3800 real and 3500 fake videos. Due to its real-world nature, it is considered a challenging dataset.

Evaluation Data

- Datasets: FF++, CelebDF, WDF
- Preprocessing: The WDF dataset is already preprocessed via the procedure described in the original paper. For each video in the FF++ and CelebDF datasets, we follow the same processing scheme used in the service. All face images are resized to 300 × 300 and normalized by the ImageNet mean and standard deviation.
- Postprocessing: we use the *Aggregations Strategy* described in the *Model Details* for all evaluation datasets.

Training Data

- Models $1-4$ were trained on the DFDC dataset while model 5 was trained on the WDF dataset.
- We expect that the models will demonstrate good performance on facial manipulations included in the *DFDC* and *WDF* datasets, i.e. Identity Swap manipulations based on DeepFake, FaceSwap, and GAN-based algorithms, and various real-world DeepFake manipulations included in WDF.

Caveats and Recommendations

- General performance: the performance of DeepFake detectors highly depends on the manipulations they have seen during training. For example, if a detector is trained using only one kind of DeepFake manipulation, it would perform very poorly in most other manipulation types. The generalization to novel manipulations is an open research issue that almost all approaches suffer from, including our service. Our training data contain various manipulations, yet we cannot guarantee good performance on unseen manipulations.
- Video quality: it is also recommended that the input media be of the best quality possible since factors like quality and compression significantly affect detection accuracy.
- Video length: to ensure high-quality predictions and avoid computational overload, it is not recommended to submit long videos with many shots (cf. *Out-of-scope uses*).
- Adversarial attacks: an adversarial attacker might affect detection accuracy using methods such as a Projected Gradient Descent (PGP) attack. Even though these attacks might not be visible to the naked eye, they can fool a DeepFake detector into assessing that a DeepFake video is real.
- Facebook videos: The service does not guarantee successful processing of Facebook videos due to the Facebook policies that restrict video downloading.

Quantitative Analyses

Manipulation	BA	AUC
FaceSwap	78.40%	0.8674
DeepFakes	86.20%	0.9468
NeuralTextures	57.65%	0.6276
Face2Face	59.02%	0.6402

Table 1: BA and AUC for each manipulation in FF++.

Dataset	BA	AUC
FaceForensics++	70.31%	0.7705
CelebDF	82.75%	0.9259
WildDeepFake	84.94%	0.9373

Table 2: BA and AUC for the service on three datasets.

Dataset	norm-1	norm-2	norm-inf
FaceForensics++	70.31%	64.04%	50.53%
CelebDF	82.75%	76.01%	50.00%
WildDeepFake	84.94%	63.04%	50.00%

Table 3: BA scores on three datasets adversarially manipulated with the PGP attack (hyperparameters: $eps = 0.2$).

Performance Intuition

- *BA* is the average of the accuracy per class. Since our datasets are imbalanced, it would be misleading to only report the overall accuracy. For example, in a dataset where 90% of the data are DeepFakes, a naive classifier that outputs only Fakes regardless of input would get 90% accuracy.
- *Area Under the Curve (AUC)* takes into account the Miss Rate or, in other words, how often the model wrongly thinks a DeepFake is Real, as well as the True Positive Rate, meaning how often the model correctly classifies DeepFakes. Thus the AUC is an overall metric describing these two rates, and in a classification system, such as ours, higher is better. However, it does not consider the 0.5 decision threshold, which is essential in practice; therefore, we consider it as an auxiliary metric.
- Tables 1, 2 make clear that our system performs much better on the CelebDF and WDF datasets rather than FF++. This is likely due to our training data lacking *Expression Swap* examples, which is one of the two manipulation categories of that dataset (cf. *Relevant Datasets* and *Training Data* sections).
- In Table 1, we observe worse performance in NeuralTextures and Face2Face manipulations (both of type *Expression Swap*), hence we recommend the current version of the service to be used for the detection of Identify Swap rather than Expression Swap.
- In Table 3, the white-box PGP attack is used on all samples from the presented datasets in order to evaluate the service robustness to adversarial attacks. All attacks try to fool the detector into assessing that the input media are real. The norm-1 attack does not have any noticeable effect on the performance in comparison to the original performance from Table 2. However, the norm-2 attack considerably affects the detection accuracy, even though the models still retain decent performance. The strongest norm-inf attack highlights the susceptibility of our model to such attacks as it can no longer distinguish between real and DeepFake videos. Thankfully, traces of an adversarial attack are visible with the naked eye only in images with the norm-inf attack.