A NOVEL FRAMEWORK FOR ASSESSMENT OF LEARNING-BASED DETECTORS IN REALISTIC CONDITIONS WITH APPLICATION TO DEEPFAKE DETECTION

Yuhang Lu, Ruizhi Luo, Touradj Ebrahimi

École Polytechnique Fédérale de Lausanne (EPFL)
Multimedia Signal Processing Group (MMSPG)

ABSTRACT

Deep convolutional neural networks have shown remarkable results on multiple detection tasks. Despite the significant progress, the performance of such detectors are often assessed in public benchmarks under non-realistic conditions. Specifically, impact of conventional distortions and processing operations such as compression, noise, and enhancement are not sufficiently studied. This paper proposes a rigorous framework to assess performance of learning-based detectors in more realistic situations. An illustrative example is shown under deepfake detection context. Inspired by the assessment results, a data augmentation strategy based on natural image degradation process is designed, which significantly improves the generalization ability of two deepfake detectors.

Index Terms—Assessment framework, Deepfake detection, Data augmentation

1. INTRODUCTION

The performance of the deep convolutional neural networks (DCNN) based detectors are often measured through publicly available benchmarks. However, it has been shown that many DCNN-based methods are vulnerable to real-world perturbation and post-processing operations \[1\2\3\]. In realistic situations, images can face unpredictable distortions from the extrinsic environment, such as noise and poor illumination conditions, or constantly undergo various processing operations to ease transmission. This could create critical challenges for safety and security applications, for instance, in surveillance, in autonomous vehicles, and in digital forensics.

Take deepfake detection as an example. Deepfakes refer to manipulated face contents using deep learning techniques. The advanced deepfake creation algorithms and open source software make it accessible and cheap to the public, posing risks to the security and authenticity of media. Although current deepfake detectors have been pushing the limits of all available benchmarks, they are for most parts, developed under ideal conditions. Malicious agents could easily fool the detector by adding noise imperceptible to human eyes or by applying more aggressive compression ratios to the targeted content. Therefore, it is desired to design a more comprehensive and systematic approach with an in-depth analysis of how the different processing operations impact the detection performance.

In this work, an assessment framework is presented for generic learning-based detection systems, which extensively evaluate the performance of a detector towards different real-world modifications. To generate a broad range of variations on test data in a realistic manner, common post-processing operations such as image transcoding, smoothing, enhancement, resizing, and synthetic noises are employed. The effectiveness of our framework is illustrated under the deepfake detection context, where thorough experiments are conducted and insightful conclusions are drawn. In the end, we have designed a data augmentation strategy based on realistic image degradation modeling process, which significantly improves the generalization ability of a deepfake detector at marginal impact on the performance in ideal conditions.

2. RELATED WORK

Several studies have investigated the vulnerability of CNN-based model to real-world and common image corruptions. Dodge and Karam \[1\] first measured the performance of image classification models on data suffered from noise, contrast variation and image blur. Hendrycks et al. \[3\] proposed a benchmark to evaluate the robustness of image recognition models towards common corruptions. Extensive work \[4\5\] has been carried out in object detection and semantic segmentation and applied to safety-critical applications. Current activities in this area mostly focus on corruptions during data acquisition and apply to only one type of computer vision task. Our proposed assessment framework offers a more general solution and in addition considers the impact of realistic image processing operations in the end-to-end workflow.

Face manipulation detection is a classical problem in computer vision. Currently, the large majority of proposed detectors treat it as a binary classification task. Similar to other computer vision tasks, a number of large-scale datasets, benchmarks, and competitions \[6\7\8\9\] are released and or-
organized to assist the community to resolve this problem. For instance, FaceForensics++ \cite{6} is one of the most popular face forensics dataset. Facebook released one of the biggest deepfake datasets, DFDC \cite{9}, and organized a competition based on the latter. As a results, a large amount of deepfake detection methods \cite{10,11,12,13} have been proposed. One of the first deep learning-based methods for deepfake detection was proposed by Zhou et al. in \cite{10}. Rössler et al. \cite{6} proposed to finetune Xception network on face manipulation dataset and showed outstanding results on the FFpp benchmark. Nguyen et al. \cite{12} leveraged capsule network to detect face manipulation. Besides focusing on spatial domain, recent works \cite{14} attempt to resolve the problem in the frequency domain.

3. PROPOSED ASSESSMENT FRAMEWORK

In this section, we describe a rigorous assessment framework for generic detection and recognition tasks to assess their performance in more realistic conditions. It is hoped that this framework will serve as a broad benchmarking approach to evaluate robustness in realistic image processing and corruptions mechanisms while at the same time providing insights on improving the approach under assessment. In general, our framework contains six categories of processing operations or corruptions with more than ten minor types. Each type consists of over five different severity levels. The details of all operations used in evaluations are described below with the illustration of a typical example in Fig. 1.

**Compression:** JPEG compression is included in the proposed framework with multiple compression factors. As deep learning-based compression technique are becoming increasingly popular, the technique developed by Ballé et al. \cite{15} is also included in this framework.

**Smoothing:** Image blurring, also known as smoothing, is a widely employed operation to reduce noise which simultaneously results in a reduction of details. Three frequently used filters with various kernel sizes are considered in our framework, including Gaussian, Median, and Average filters.

**Noise:** The acquisition of images can be easily affected by noise. Our framework applies Additive White Gaussian noise (AWGN) with 5 levels of variance. To better reflect the realistic situations, a synthetic Poissonian-Gaussian noise is also considered, the parameters of which are learned from real-world noisy images.

**Enhancement:** Image enhancement is generally a very frequently used technique of adjusting images for better display or further image analysis. We change the contrast and brightness of images by separately applying linear adjustment and Gamma correction.

**Resizing:** Low-resolution data can significantly reduce the performance of modern deep learning-based detectors \cite{16,17}. This is often the case when the detector is employed in an outdoor environment, where captured data could have limited resolution. In this framework, we mainly measure the impact of image downscaling.

**Combinations:** A mixture of two or three operations above is also considered, such as combining JPEG compression and Gaussian noise, making the test data better reflect more complex real-world scenarios.

To validate the assessment framework, one detector should be trained on its original target datasets, such as FFpp for deepfake detection tasks. Processing operations and corruptions are not applied on training data. Furthermore, several parameters were used for processing operations to better understand their impact on the detection.

4. ILLUSTRATIVE EXAMPLE FOR DEEPFAKE DETECTION

The proposed framework is assessed in a deepfake detection scenario, where the robustness of two deepfake detectors are evaluated in presence of realistic processing operations.

4.1. Assessment Datasets

To show the effectiveness of our assessment framework, it is applied to two widely used Deepfake detection databases. FaceForensics++ \cite{6} is used as the main database for our assessment framework. It contains 1000 pristine and 4000 manipulated contents generated by four different approaches. Video contents are also compressed with two quality settings.
using the H.264, denoted as C23 and C40. Consequently, video contents corresponding to three quality levels are used in the training, while only the highest quality contents are used in the final assessment. **Celeb-DF** [7] is a high-quality face forensics dataset. Additional experiments were conducted on the latter to validate the effectiveness of our framework. The test data was selected as recommended in [7] while the validation and training data was split in 20% versus 80% respectively.

### 4.2. Detection Methods

In this section, two deep learning-based deepfake detection approaches are evaluated to demonstrate the usefulness of the proposed assessment framework. Both have been reported to have outstanding performance on public benchmarks.

**Capsule-Forensics** [12] combines traditional CNN and Capsule networks, which requires fewer parameters and at the same time achieves high detection accuracy. The model is trained from scratch on the above-mentioned two datasets.

**XceptionNet** was first introduced in [18] and became a popular CNN architecture used in multiple computer vision tasks. Kößler et al. [6] further adopted it for deepfake detection. It has been shown to outperform several alternative detectors in FFpp benchmark on both uncompressed and compressed contents. We leverage the model pretrained on ImageNet and finetune on the above-mentioned two datasets respectively.

### 4.3. Experiments and Results

#### 4.3.1. Implementation Details

For training, the Capsule-Forensics model was trained using the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a learning rate of $5 \times 10^{-4}$ for 25 epochs. The XceptionNet was finetuned for 10 epochs with learning rate of $1 \times 10^{-3}$.

For both datasets, 100 frames were randomly extracted from each video for training purpose and 32 frames were extracted for validation and testing. Extracted frames were preprocessed and cropped around face regions. Processing operations and corruptions from our assessment framework were used only to test data for further evaluation.

During the evaluation, the Accuracy (ACC), the Area Under Receiver Operating Characteristic Curve (AUC), and F1-score were used as performance metrics in all experiments.

#### 4.3.2. Assessment Results

The two deepfake detectors were trained on the original unaltered training sets of both FFpp and Celeb-DF. Table 1 shows the evaluation results using our assessment framework. Due to the page limit, only AUC scores and a subset of operations and intensity values are presented in this paper.

### 5. IMPROVED TRAINING STRATEGY BY DATA AUGMENTATION

To reduce the strong negative impact of real-world processing and corruptions on model performance observed above, we propose a simple yet efficient data augmentation approach that leads to a robustness improvement.
Previous experimental results suggest that properly mixing low quality training data, slightly improves the generalization ability of detectors. Some researchers have explored using Gaussian noise corrupted data \cite{19, 20} or stylized data \cite{4} for training to improve model performance under corruption. However, according to our experiments with deepfake detectors, Gaussian noise-based augmentation only benefits limited type of realistic corruption and meanwhile deteriorates the performance on original unaltered data. Based on the observation of data distortions in realistic conditions, a carefully designed augmentation chain was conceived. It mainly includes 4 realistic data degradation types with various intensity levels. Each process will be randomly applied to certain proportion of data. Figure 3 illustrates the performance improvement of two models on four types of distortions and processing after using our augmentation scheme.

Enhancement: At the beginning of the augmentation chain, there is 50% probability that either the brightness or the contrast of the training data will be non-linearly modified by a factor randomly sampled from $[0.5, 1.5]$.

Smoothing: For each batch of training data, an image blurring technique will be applied with probability of 50%. Either Gaussian blur or Average blur filter is used with a kernel size varying from $[3, 15]$.

Additive Gaussian Noise: Gaussian noise is added in the augmentation chain with probability of 30%. The standard deviation value is randomly sampled from $[0, 50]$.

**JPEG Compression**: In the end, JPEG compression is applied with a probability of 70%. Each time the quality factor is randomly sampled from $[10, 95]$.

**6. CONCLUSION**

Many detectors are designed to be as high performing as possible on specific benchmarks. But this often results in sacrificing generalization ability to more realistic situations. The proposed assessment framework is capable of assessing detectors in more realistic conditions and provides valuable insights on designing more robust techniques. A carefully conceived augmentation chain based on a natural data degradation process is proposed and significantly improves the model robustness against various distortions.
7. REFERENCES

[1] Samuel F. Dodge and Lina Karam, “Understanding how image quality affects deep neural networks,” 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEx), pp. 1–6, 2016.

[2] Yiren Zhou, Sibo Song, and Ngai-Man Cheung, “On classification of distorted images with deep convolutional neural networks,” 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1213–1217, 2017.

[3] Dan Hendrycks and Thomas Dietterich, “Benchmarking neural network robustness to common corruptions and perturbations,” Proceedings of the International Conference on Learning Representations, 2019.

[4] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel, “Benchmarking robustness in object detection: Autonomous driving when winter is coming;” arXiv preprint arXiv:1907.07484, 2019.

[5] Christoph Kamann and Carsten Rother, “Benchmarking the robustness of semantic segmentation models,” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun 2020.

[6] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner, “FaceForensics++: Learning to detect manipulated facial images,” in International Conference on Computer Vision (ICCV), 2019.

[7] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu, “Celeb-df: A large-scale challenging dataset for deepfake forensics,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[8] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy, “Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection,” 2020.

[9] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuow Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer, “The deepfake detection challenge dataset,” 2020.

[10] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis, “Two-Stream Neural Networks for Tampered Face Detection,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), July 2017, pp. 1831–1839, ISSN: 2160-7516.

[11] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen, “MesoNet: a Compact Facial Video Forgery Detection Network,” in 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Dec. 2018, pp. 1–7, ISSN: 2157-4774.

[12] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen, “Use of a Capsule Network to Detect Fake Images and Videos,” arXiv:1910.12467 [cs], Oct. 2019, arXiv: 1910.12467.

[13] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo, “Face x-ray for more general face forgery detection,” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5000–5009, 2020.

[14] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu, “Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain,” 2021.

[15] Johannes Ballé, David C. Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston, “Variational image compression with a scale hyperprior,” ArXiv, vol. abs/1802.01436, 2018.

[16] Tomasz Marciniak, Agata Chmielewska, Radoslaw Weychan, Marianna Parzych, and Adam Dabrowski, “Influence of low resolution of images on reliability of face detection and recognition,” Multimedia Tools and Applications, vol. 74, 06 2013.

[17] Pei Li, Loreto Prieto, Domingo Mery, and Patrick J. Flynn, “On low-resolution face recognition in the wild: Comparisons and new techniques,” IEEE Transactions on Information Forensics and Security, vol. 14, pp. 2000–2012, 2019.

[18] François Chollet, “Xception: Deep learning with depthwise separable convolutions,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800–1807, 2017.

[19] Nic Ford, Justin Gilmer, Nicholas Carlini, and Ekin Dogus Cubuk, “Adversarial examples are a natural consequence of test error in noise,” in ICML, 2019.

[20] Evgenia Rusak, Lukas Schott, Roland S. Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge, and Wieland Brendel, “Increasing the robustness of dnns against image corruptions by playing the game of noise,” ArXiv, vol. abs/2001.06057, 2020.