A Comparative Study on Discriminative and One–Class Learning Models for Deepfake Detection

Marija Ivanovska[1], **Vitomir Štruc**[1]

[1]*Faculty of Electrical Engineering, University of Ljubljana*
E-pošta: {marija.ivanovska, vitomir.struc}@fe.uni-lj.si

Abstract

Recently, deepfakes or manipulated face images, where a donor's face is swapped with the face of a target person, have gained enormous popularity among the general public. With recent advancements in artificial intelligence and generative modeling such images can nowadays be easily generated and used to spread misinformation and harm individuals, businesses or society. As the tools for generating deepfakes are rapidly improving, it is critical for deepfake detection models to be able to recognize advanced, sophisticated data manipulations, including those that have not been seen during training. In this paper, we explore the use of one–class learning models as an alternative to discriminative methods for the detection of deepfakes. We conduct a comparative study with three popular deepfake datasets and investigate the performance of selected (discriminative and one-class) detection models in matched- and cross-dataset experiments. Our results show that disciminative models significantly outperform one-class models when training and testing data come from the same dataset, but degrade considerably when the characteristics of the testing data deviate from the training setting. In such cases, one-class models tend to generalize much better.

1 Introduction

Detecting artificially synthesized data is a longstanding problem in the artificial intelligence field with wide ranging implications considering the potential of such data for malicious applications. With the recent wide adoption of generative neural networks, the emergence of deepfakes has drawn even greater attention to this problem. Deepfakes are a special type of AI–generated visual data, where the face of a source individual is seamlessly replaced by the face of a target individual, while retaining the initial facial expressions and head poses. Such manipulated data can then be used to harm the reputation of the target individual or to spread misinformation.

With the advancements made in artificial intelligence and image generation, deepfakes are becoming even more convincing and thus difficult to recognize by an average, untrained human. Nevertheless, the generation of manipulated face images usually leaves visual inconsistencies, which adequate machine learning technique should be able to detect. Some studies on this topic, for example, propose explicit modelling of the expected modification traces/artifacts [6, 7, 8] to detect face manipulations. Others rely on data–driven models, that learn to discriminate between normal and synthesized faces without any a priori knowledge [9, 10, 13]. Either way, the detection task is typically formulated as a binary classification problem, which requires a large, diverse and well balanced training set. The collection of such a set is not only challenging, but also represents a never-ending task, as new deepfake generation methods emerge on regularly in the literature.

Motivated by these limitations, we investigate in this paper the ability of existing one–class learning models to detect deepfakes, by learning only from real face images. Additionally, we analyze the advantages and disadvantages of such one-class learning techniques in comparison to the more commonly used discriminative models.

2 Related work

Since the first appearance of deepfakes, the research community has shown great interest in preventing the abuse of this technology. As a countermeasure, numerous detection models have been proposed in the literature [5]. Many of these models have been trained to identify specific traces of visual data manipulations. Although such modifications are often subtle for humans, they can be successfully detected by analyzing different spatial and temporal artifacts that are created during the manipulation process. Agarwal *et al.* [6], for example, exploit the idiosyncratic behavioral patterns of synthesized faces, Yang *et al.* [7] expose deepfakes as face videos with incoherent head poses, while Li *et al.* [8] observe that unlike real videos, deepfakes lack reasonable eye blinking.

Recently, various detection models have been proposed that do not rely on any specific temporal or spatial artifacts, but are instead given examples of real and fake faces to autonomously learn distinguishing data characteristics. Afchar *et al.* [9], for example, use a two–stage convolutional neural network for the detection of both, swapped faces and face reenactment, while Guera *et al.* [10] develop a temporal model based on recurrent neural networks (RNN). Classification networks such as Xception–Net and VGG16 have also achieved high detection

Supported in parts by the ARRS reserach project J2-9433 (B) "Detection of inconsistencies in complex visual data using deep learning".

accuracy in a study carried out by Dang *et al.* in [13].

Although discriminative methods show impressive results on benchmark datasets, in practice they often suffer from robustness and generalization issues and perform poorly, when employed in real–world scenarios. Next to being affected by domain shifts, such methods are also less likely to discover fake images containing unseen inconsistencies. To avoid fine–tuning of existing models to (previously) unseen deepfakes, Khalid *et al.* [11] proposed a reformulation of the deepfake detection task into a one–class anomaly detection problem. Using real face images only, the authors trained a one–class Variational Autoencoder (VAE) that successfully detected deepfakes as deviations from real data. Ortiz *et al.* [12] also followed this one–class learning paradigm and trained VGGFace2 and ResNet50 models that were then used together with an attribution based confidence (ABC) metric to segregate deepfakes from real faces. Despite the promising results of such one–class learning approaches, their potential is not yet explored to a full extent.

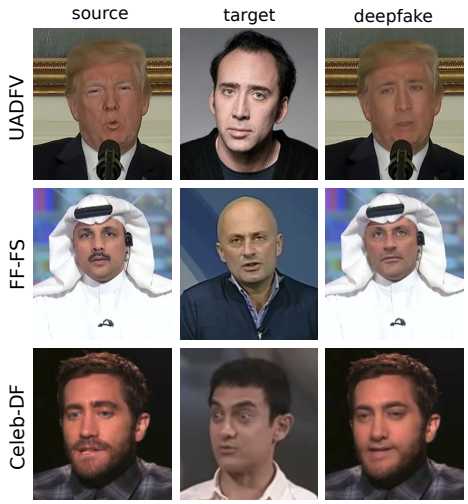

Figure 1: Examples of deepfakes from the three datasets used in this paper: UADFV [2] (top row), FF–FS [4] (middle row) and Celeb–DF [1] (bottom row). Deepfakes have been generated by swapping a source face (left most column) with the face of a chosen target person (middle column).

3 Datasets

We use three popular benchmarks for our investigations, i.e.: UADFV [2], FaceForensics++ [4] and Celeb-DF [1]. Example images from the datasets are shown in Figure 1.

UADFV is a recent deepfake dataset and contains 49 real YouTube videos of public figures and 49 deepfake videos [2]. Deepfakes were generated with the FakeAPP approach, by swapping the original faces with the face of Nicolas Cage. The length of the videos ranges between 3 and 15 seconds. All videos have a frame rate of 30 fps.

FaceForensics++ (FF) is a dataset consisting of 1004 YouTube videos with over 1.8 milion frames containing face images of random people [4]. Each video sequence represents a unique identity. Up to this date, 5 different methods have been used for the generation of the fake videos: deepfakes, Face2Face, FaceSwap, NeuralTextures and FaceShifter with FaceShifter [3] being the latest state–of–the–art method also known by its ability to successfully handle both, non-occluded as well as occluded face images. In this work we only use the 1004 fake images generated by FaceShifter (denoted FF–FS).

Celeb–DF is a recent dataset, whose real videos are collected exclusively from YouTube and feature 59 celebrities of diverse genders, ages, and ethic groups [1]. The dataset has 890 real videos, each around 13 seconds long, with a frame rate of 30 fps. This real data has been used as a source for generation of 5639 high-quality deepfake videos. Fake videos have been synthesized using state-of-the-art synthesis algorithms. The amount of notable visual artifacts has been further minimized by various post-processing methods, making Celeb-DF one of the most challenging benchmarks to date.

4 Experiments

4.1 Methods

In our comparative study we train and evaluate two distinct models: (i) ResNet50 [14], a commonly used discriminative network with residual blocks, and (ii) GANomaly [15], a recent state–of–the–art representative of the one–class learning methods.

ResNet50. Residual networks or ResNets were introduced in an attempt to further improve the performance of existing convolutional neural networks (CNNs) and mitigate the vanishing gradient problem [14] and represent state-of-the-art CNN models typically used for various recognition tasks. In this study, we use ResNet50 - a ResNet variant with 48 convolutional layers and an output layer with a sigmoid activation function at the top, for classification of input samples into fake or real. The model is trained with a binary cross–entropy loss.

GANomaly [15] is a recent state-of–the–art model for detection of anomalies in a one–class learning setting. As its name implies, it is a generative adversarial network (GAN), consisting of a generative part that maps input samples into a latent space and then reconstructs them back to the image space. It is hypothesized, that once such networks learn the identity mapping of normal, non–anomalous samples, they fail to accurately reconstruct anomalies, since they were not seen during the training phase. In our case, we train GANomaly on real face data only, using the objective proposed by the authors of the model [15].

4.2 Data Preprocessing

All datasets described in Section 3 consist of videos, where faces are not zoomed in, so each video frame has a significant amount of redundant information, i.e. non–face image areas. Since faces represent our regions-of-interest (ROIs), we preprocess all videos frames by by subjecting them to a face detection procedure. For this purpose we use the pre–trained MTCNN face detector, proposed by Zhang *et al.* in [16]. The algorithm returns coordinates of the bounding boxes containing a face, which are then used to crop out rectangular face areas. False positives

are manually removed from the final set of images. At the end, all face images are resized to 224×224 pixels for the training of the ResNet50 model and to 256×256 pixels for the training of the GANomaly model.

4.3 Training and Evaluation Protocol

In our experiments, we follow the $80/20$ split rule, where 80% of the samples are used in the training phase, while the remaining 20% are used for testing. Both, training and testing samples are randomly selected such that they equally represent real and fake data. To ensure a fair comparison, the training and testing splits remain unchanged throughout the execution of the experiments. ResNet50 is trained on both classes, while GANomaly is trained using real data only. Hyperparameters of each model are set to their default values, reported in the respective papers. Following common practice, each trained model is evaluated with the AUC (Area Under the ROC curve) metric. In ResNet50 experiments, the ROC curve is generated with the probabilities of the output layer. In GANomaly experiments the L_2 reconstruction error of the testing samples is used instead.

Experiments are run in two stages: (i) a *matched dataset* stage, and (ii) a *cross-dataset* stage. In the first (matched dataset) stage, both training and testing samples come from the same dataset, assuming that the image quality and the deepfake generation technique have remained the same. In practice, such expectation is unlikely to be met. Therefore, the second (cross dataset) stage consists of cross–dataset experiments, where the testing data comes from a different dataset.

5 Results

We summarize the results from both experimental stages (carried out in accordance with the experimental protocol described above) In Table 1.

Matched-Dataset Results. As can be seen from the presented first-stage results, ResNet50 significantly outperforms GANomaly on all three benchmarks by correctly predicting the ground truth label of over 90% of the data when imagery with similar characteristics is used for training and testing. Although less accurate, GANomaly still shows a relatively high detection rate despite being trained without any fake samples. We can observe that GANomaly is most successful in the detection of the lower–quality UADFV deepfakes and least successful in detection of the higher-quality Celeb–DF deepfakes. In the latter case, the model achieves an AUC score of 0.746 – down by close to 0.2 when compared to ResNet50. To better understand the differences between the two models, we perform a visual analysis of the deepfake samples that were successfully detected by ResNet50 and missed by GANomaly. Representative examples from each dataset are shown in Figure 2. As can be seen, GANomaly is less likely to recognize a deepfake, where visual artifacts are less apparent or might also represent naturally occurring skin spots. It also fails to detect deepfakes generated from images taken in non–standard lighting conditions. A probable reason for such failures is the high generalization capacity of generative one–class methods, causing a successful reconstruction of some image inconsistencies. Because of the discriminative nature of ResNet50, the model is able to detect such deepfakes despite the somewhat more challenging image characteristics.

Table 1: Deepfake detection performance for ResNet50 and GANomaly in terms of AUC scores. Experiments with matched datasets show that ResNet50 significantly outperforms GANomaly on all three benchmarks. Cross–dataset experiments, on the other hand, suggest that in contrast to GANomaly, ResNet50 fails to detect deepfakes coming from datasets not seen during training pointing to better generalization capabilities of one-class models.

Stage	Training dataset	Testing dataset	ResNet50 (discriminative)	GANomaly (one-class)
Matched-dataset	UADFV	UADFV	0.947	0.880
	FF–FS	FF–FS	0.966	0.868
	Celeb–DF	Celeb–DF	0.938	0.746
Cross-dataset	UADFV	FF–FS	0.622	0.768
	UADFV	Celeb–DF	0.531	0.704
	FF–FS	UADFV	0.763	0.794
	FF–FS	Celeb–DF	0.561	0.608
	Celeb–DF	UADFV	0.574	0.772
	Celeb–DF	FF–FS	0.527	0.680

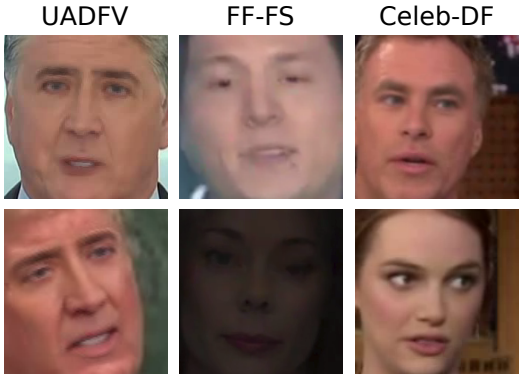

Figure 2: Examples of deepfakes successfully detected by ResNet50 and missed by GANomaly when training and testing samples come form the same dataset.

Cross-Dataset Results. In contrast to the matched-dataset experiments, the results of the cross-dataset experiments show that GANomaly is much more accurate in the classification task, where the model is tested on samples from a dataset different from the one used for training. ResNet50, on the other hand, shows a significant drop in efficiency, in some cases performing similarly to a random classifier. In this case, one of the main problems is simply the domain shift, causing the model to miss both, real and fake samples. Another issue we could observe is the inability of ResNet50 to detect deepfake artifacts that have not been introduced during the training phase. Some deepfake samples that were successfully detected by GANomaly and missed by ResNet50 in this experimental stage are shown in Figure 3. As can be seen, ResNet50 failed to recognize very low–quality UADFV deepfakes with badly aligned face masks and incorrectly adjusted colors. Such deepfakes can not be found in the datasets, utilized for the training process. In contrast to

GANomaly, ResNet50 was also unable to detect unique artifacts such as errors in the reconstruction of the teeth, fake faces with asymmetrical or crossed eyes and sharp illumination transitions. Moreover, training ResNet50 on UADFV and FF-FS, both highly biased towards the Caucasian race, results in very low classification probabilities of real and fake Asian faces.

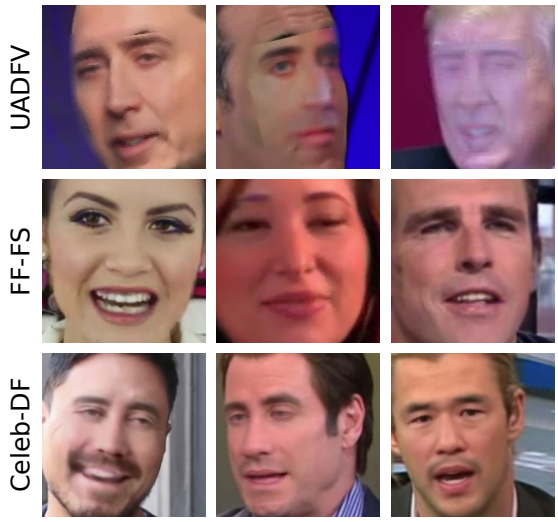

Figure 3: Examples of deepfakes successfully detected by GANomaly and missed by ResNet50 during the cross–dataset experiments, where testing samples come from a different dataset than the training samples.

6 Conclusion

In this paper we presented an experimental study comparing the deepfake detection performance of a discriminatively learned ResNet50 model and the one–class learning GANomaly approach. Using three common deepfake datasets, we conducted a series of matched- and cross–dataset experiments, where testing samples were either taken from the same dataset as the training data or were coming from one of the other two datasets. The matched–dataset experiments showed a clear advantage of the discriminative ResNet50 model. However, in the cross–dataset experiments ResNet50 was severely affected by the domain shift, also failing to recognize deepfakes with visual artifacts, that have not been introduced during the training. Unlike the discriminative model, GANomaly was shown to be less prone to distribution shifts and changes in the technique, used for the generation of deepfakes. Although GANomaly was not able to produce state–of–the–art detection results, our experiments suggest that one–class learning models have considerable potential for deepfake detection due to the fact that they can be trained without any fake samples and, thus, do not rely on prior assumptions about the appearance of the deepfake images.

References

[1] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu: Celeb-DF: A Large-Scale Challenging Dataset for deepfake Forensics, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3204-3213

[2] Y. Li, M.C. Chang, and S. Lyu: In Ictu Oculi: Exposing Ai Generated Fake Face Videos by Detecting Eye Blinking, IEEE International Workshop on Information Forensics and Security (WIFS), 2018

[3] L. Lingzhi, B. Jianmin, Y. Hao, C. Dong, W. Fang: Advancing High Fidelity Identity Swapping for Forgery Detection, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020

[4] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner: FaceForensics++: Learning to Detect Manipulated Facial Images, International Conference on Computer Vision (ICCV),2019

[5] M. Masood, M. Nawaz, K. Mahmood Malik, A. Javed, A. Irtaza: Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward, https://arxiv.org/abs/2103.00484, 2021

[6] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li: Protecting world leaders against deep fakes. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019

[7] X. Yang, Y. Li, S. Lyu: Exposing deep fakes using inconsistent head poses, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019

[8] Y. Li, M. Chang, Siwei Lyu: In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking, IEEE International Workshop on Information Forensics and Security (WIFS), 2018

[9] D. Afchar, V. Nozick, J. Yamagishi, I.Echizen: Mesonet: a compact facial video forgery detection network, IEEE International Workshop on Information Forensics and Security (WIFS), 2018

[10] D. Guera, E. J Delp: deepfake video detection using recurrent neural networks, IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2018 images and videos, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019

[11] H. Khalid, S. S. Woo: OC-FakeDect: Classifying deepfakes Using One-class Variational Autoencoder, IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020

[12] S. Fernandes, S. Raj, R. Ewetz, J. S. Pannu, S. K. Jha, E. Ortiz, I. Vintila, M. Salter: Detecting deepfake Videos using Attribution- Based Confidence Metric, IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020

[13] H. Dang, F. Liu, J. Stehouwer, X. Liu, A. Jain: On the Detection of Digital Face Manipulation, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020

[14] K. He, X. Zhang, S. Ren, J. Sun: Deep Residual Learning for Image Recognition, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016

[15] S. Akcay, A. Atapour-Abarghouei, T. P. Breckon: GANomaly: Semi-supervised Anomaly Detection via Adversarial Training, Asian Conference on Computer Vision (ACCV), 2019

[16] K. Zhang, Z. Zhang, Z. Li, Y. Qiao: Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks, IEEE Signal Processing Letters, 2016