State of the Art on: Deepfake Detection

Samuele Pino, samuele.pino@mail.polimi.it

1. Introduction to the research topic

Media manipulation exists since the times of the first analog photographs but the switch to digital media, besides bringing great technical benefits, made the manipulation process easier. Although today we can enjoy entire movies synthesized by computer graphics, at the same time threats linked to the misuse of such technologies are frighteningly increasing.

It is of crucial importance to always be able to tell apart real images and videos from synthesized ones, but sometimes it can be a hard task. Fake media are computer generated or manipulated images or videos with the precise goal to fool human eye: for this reasons sophisticated computational techniques like Deep Learning exist, and are still being studied, to perform this task with a higher accuracy than humans.

Table 1 shows the most prestigious journals and conferences on the research area, according to their H-index[1].

1.1. Preliminaries

To understand the main related works to the topic it is useful to briefly describe some typical deep learning approaches to problems. In particular some recurrent topics are Convolutional Neural Networks, Recurrent Neural Networks and Generative Adversarial Networks.

An Artificial Neural Network (ANN) is a computational model loosely inspired to the animals' brain. It can be seen as a layered graph consisting of neurons (nodes) and weighted connections (archs). Each neuron is a computational unit: it takes as input values from the incoming connections, performs an operation (like weighted sum), undergoes some non linear "activation function" and outputs a value. Usually the information flow goes from the input neurons layer, passing through possible "hidden" layers, to the output layer. Neural Networks are useful in many fields and can perform different tasks, from generation of data to classification.

A Neural Network is said to be "deep" when it contains more than one hidden layer, this usually adds generalization ability to the network, expanding the function space that it can approximate.

A Convolutional Neural Network (CNN) is a deep network where some layers (usually the very first) perform a 2D convolution operation over the neurons of the previous layers. They are the most common approach when

[1] H-index is defined as the maximum value of h such that the journal has published h papers each cited at least h times (source data from Scopus). H5-index is H-index evaluated only on the publications of the last 5 years (source data from Google Scholar).

Journal name	H-index
IEEE Transactions on Pattern Analysis and Machine Intelligence	288
IEEE Transactions on Neural Networks and Learning Systems	161
Pattern Recognition	160
Journal of Machine Learning Research	147
Conference name	**H5-index**
CVPR : IEEE/CVF Conference on Computer Vision and Pattern Recognition	240
NIPS : Neural Information Processing Systems (NIPS)	169
ECCV : European Conference on Computer Vision	137
ICML : International Conference on Machine Learning (ICML)	135

Table 1: Most prestigious journals and conferences on Machine Learning (source: `www.guide2research.com`).

dealing with images as input data. The convolution reshapes the information that flow through the neural network gradually moving from spatial so semantic features as the network gets deeper.

A Recurrent Neural Network (RNN) is a, usually shallow, network with a connection from the output layer to the input one. The input data is usually a sequence and the network is fed with one element of the sequence at the time, together with some information from the output of the previous element. This type of architecture is commonly used when dealing with temporal information. The most common and powerful type of RNN is Long Short Term Memory (LSTM).

A Generative Adversarial Network (GAN) is a system consisting of two networks, a *generative* one, that tries to recreate candidates statistically similar to the data present in the training set, and a *discriminative* network, that tries to distinguish true data from the one generated by the previous network. The discriminator is usually a convolutional networks while the generator is a deconvolutional network.

A large number of deep learning frameworks is available to build your model or use existing ones, some examples are TensorFlow, Keras, Caffe and PyTorch.[2]

1.2. Research topic

In the recent years new image and video manipulation techniques popped up on the internet, seriously challenging the classic manual manipulation detection methods. Such techniques take advantage of deep neural networks to skip the manual editing phase and automatically generate results so realistic to be almost indistinguishable from real ones to the naked eye.

Today, the most sophisticated techniques use Generative Adversarial Networks (GANs) together with Convolutional Neural Networks. Implementations are in somes case very popular, like the mobile application FaceApp[3], the open source FaceSwap[4] or the Face2Face approach [1].

In general, when we speak about facial manipulation, we can distinguish 4 types (from the lightest manipulation to heaviest): facial expression manipulation, facial attributes manipulation, face identity swap, whole face synthesis. We are going to focus on face identity swap, i.e. when the original face of a video is replaced with the face of someone else, but keeping the expression and movements of the original one.

The most popular face swapping technique in videos is called DeepFake, it started appearing in 2017 and it emulates the face of a source individual in different light and pose conditions, putting it on top of the face of a different target individual. The videos resulting from this approach are sometimes incredibly realistic. Deepfakes are generated through deep neural networks (hence the name) specially trained on datasets of video representing the source individual.

The technique has become popular among non-technical people thanks to mobile applications (e.g. FakeApp), letting everyone forge their fake contents without any specific background knowledge, although the results are still poor due to the limited computational power of a smartphone. However, given the proper computational speed, a large enough dataset of videos depicting the person and a good software, it is possible to obtain results that are indistinguishable from a real video for the human eye.

Researchers are currently working intensively on this field, developing every day better techniques to detect such forgeries in videos. The research field is still fresh and it is growing fast, also helped by investments and competitions.

The reasons why the research exploded are obvious and are mainly about privacy, reputation, politics and public security. The first time deepfakes appeared on the internet, they were faces of famous people applied to adult videos, with the potential of creating a high reputation damage to the forgeries' victims. Moreover, imagine a scenario of high tension between two countries: what would happen if these techniques were used to show one of the two leaders declaring a war action against the opponent?

It is therefore of the highest importance to counteract the development of these forgeries techniques with tools that can reliably discern real and fake videos.

[2] `https://www.tensorflow.org/` – `https://keras.io/` – `https://caffe.berkeleyvision.org/` – `https://pytorch.org/`

[3] `https://faceapp.com`

[4] `https://github.com/deepfakes/faceswap`

Authors (year)	Detected features	Classifier Architecture	Mask	Time aware
Afchar et al. (2018) [2]	Mesoscopic (learned)	CNN (MesoNet)	-	-
Korshunov and Marcel (2018) [3]	Lip sync	CNN + LSTM	-	Yes
Güera and Delp (2018) [4]	(learned)	CNN + LSTM	-	Yes
Li et al. (2018) [5]	Eye blinking	CNN + LSTM	-	Yes
Li and Lyu (2018) [6]	Affine transformations	CNN	-	-
Yang et al. (2019) [7]	Head-face poses	Face landmarks + SVM	-	-
Rössler et al. (2019) [8]	(learned)	CNN (XceptionNet)	-	-
Matern et al. (2019) [9]	Visual artifacts	Preprocessing + FFNN	-	-
Nguyen et al. (2019) [10]	(learned)	CNN	Yes	-
Stehouwer et al. (2019) [11]	(learned)	CNN + Attention	Yes	-
Agarwal et al. (2019) [12]	Face/head actions	OpenFace2 + SVM	-	Yes
Sabir et al. (2019) [13]	(learned)	CNN + BRNN	-	Yes
Nguyen et al. (2019) [14]	(learned)	CNN + CapsNet	-	-
Li et al. (2019) [15]	Blending boundaries	CNN	Yes	-

Table 2: Summary of detection methods. The columns address: what are the features detected (specified if not learned), the architecture of the classifier, wheter a gray-scale mask is produced as output with the location of the forged parts, whether the algorithm uses temporal information.

2. Main related works

2.1. Classification of the main related works

One approach to the problem consists in manually selecting a set of specific features that can help to discriminate forged videos from real ones (chosen features). Evaluating the presence, absence or intensity of such feature it is possible to classify the video under analysis with a certain degree of accuracy.

Other approaches prefer to let the model detect and learn the features by itself, in a supervised environment (learned features). Often these types of approaches make use of a CNN, in some cases together with an RNN model.

Another important dimension of classification is whether the algorithm takes into consideration time information (time aware) or not. This does not simply mean to average the results obtained from each frame independently, but to take into consideration also the frames order and/or their relative distances.

A summary of the works described in the next section is presented in Table 2.

2.2. Brief description of the main related works

Afchar et al. proposed Mesonet-4 and MesoInception-4 [2], two networks composed by a low number of layers to catch information at the mesoscopic level of the image. The former consists of 4 convolutional and 1 fully connected, while in the latter the first two layers are replaced with inception modules. The idea is to avoid getting semantic information with very deep networks and at the same time to drop the low level information such as noise in the image, useless due to video compression. The experiments showed an accuracy of 0.984 (averaging frames predictions in video) on their own dataset and an important robustness also when tested on unseen datasets, like FaceForensics++ with accuracy of 0.98 [16].

Korshunov and Marcel focused on lip sync inconsistencies [3] considering Mel-Frequency Cepstral Coefficients (MFCCs) as audio features and mouth landmarks as image features. Features dimensionality was reduced through Principal Component Analysis (PCA) and an LSTM was used to classify real and fake videos.

Also Güera and Delp in [4] use a temporal-aware approach by using a CNN (InceptionV3 pretrained on ImageNet) to extract frame-level features, giving them to an LSTM and then to two fully connected layers for classification. Experiments were performed on a proprietary database with an accuracy of 0.971.

One of the early examples of techniques to spot deepfakes was presented by by Li et al. at the beginning of 2018 [5]. It was based on the analysis of eye blinking of the subject: due to the scarcity of training samples in which the subject had closed eyes, the generated video had an unusual blinking pattern or no blinking. As expected, as soon as the problem was exposed, new manipulation algorithm came out with the blinking feature integrated, making the approach obsolete [12].

The same authors proposed later a new method detecting inconsistent head poses [7]. This is possible since the generated face applied to the original video is characterized by errors that can be revealed estimating its 3D pose from landmarks. In a real video, whether we consider all the landmarks or only the ones near the center of the face, we expect to get similar results. The two estimation are instead very different in deepfakes, where the center of the face comes from the synthesizer, as shown in their experiments. They manage to obtain an AUROC measure of 0.89 at the frame level and 0.974 at the video level (averaging the predictions on frames) on UADFV [7] dataset. However this model, pretrained on that specific dataset, cannot generalize well when tested on other datasets as shown in [16].

Always Li and Lyu in [6] proposed later a new method that detects traces of affine transformations (i.e., scaling, rotation and shearing) on the face: in fact the current DeepFake algorithm, due to computational constraints, tend to generate the new faces at a fixed resolution and then to warp it to match the original face. This process leaves distinctive artifacts in the output video (like resolution inconsistency between face area and its surroundings) detectable through a CNN. The performed experiments show great results, with an AUC of 0.974 on UADFV dataset and 0.999 on DeepFakeTIMIT (LQ) [5]. Training data is created using simple image processing on original images, saving computational time and avoiding over-fit to a specific DeepFake generator since such artifacts are common to several sources of DeepFakes.

In [8] Rössler et al. analyze five detection systems: (i) hand-crafted Steganalysis features by [17] coupled with a CNN, (ii) constrained convolutional network is designed to suppress high level contents of the image [18], (iii) a CNN with a pooling layer computing simple statistics (mean, variance, maximum and minimum) [19], (iv) MesoInception-4 [2], (v) XceptionNet [20] pre-trained on ImageNet database [21]. The pipeline consists in a preliminary extraction of the face rectangle, followed by the classification with one of the said methods. They train evaluate the five systems on their proposed dataset FaceForensics++ achieving the best accuracy of 0.993 with the XceptionNet architecture. Differently from other benchmarks, the one proposed here uses different levels of compression and video quality, in order to simulate a more realistic and closer scenario to the social networks compressing process.

Matern et al. in [9] show how to expose Deepfake manipulations with the detection of simple artifacts like imprecise estimation of incident light on the face (since it must be from the original image to the manipulated one) in the areas of nose and eyes, imprecise estimation of the face geometry around its border and in the eyebrows, while occluded parts of the face like strands of hair and teeth are badly modeled. For the detection, full face, eyes and teeth are detected, then 16 features are extracted through convolution kernels and used for classification. Only using teeth and eyes features they obtain an AUC of 0.851 on their own dataset and 0.702 on UADFV. [16]

Nguyen et al. in [10] designed a CNN that uses multitask learning to detect manipulated images and locate the manipulated region. The architecture is characterized by a Y shape that enables sharing of information in the early layers of the network (encoder) and then splits into two branches (decoders): one is used to segment the image while the other reconstruct the input (to improve segmentation performances). The network also outputs the probability of the image being spoofed. The model needs only few samples for fine tuning on other datasets, but the performances are not the greatest. The results are an accuracy of 0.93 when trained on FaceForensics and tested on the same dataset, while 0.84 when finetuned and tested on FaceForensics++, and 0.658 when tested on UADFV (no finetuning) [16].

Stehouwer et al. demonstrate in [11] how the usage of a CNN with an attention mechanism can improve the classification. The attention map they propose has a flexible architecture since it can be attached to existing backbone networks. The model, besides classifying the input image as fake or real, also outputs a mask that locates the predicted fake areas of the image. Their experiments are performed with DFFD (combination of different

[5] https://www.idiap.ch/dataset/deepfaketimit

datasets, among which also FaceForensics++), achieving and AUC of 0.994 on the identity swap (deepfake) samples.

Agarwal et al. in [12] focus instead on DeepFake applied to famous people only, like world leaders: this enables them take advantage of the large amount of video data available on the subjects. They propose a technique that models facial expression and typical movements of the subject's speaking pattern. Apparently, deepfakes violate such correlation, therefore they can be used to spot forged videos. The first step is the tracking of 18 facial actions (like cheek raiser, nose wrinkler, etc.) and 4 head movement features, resulting in a total feature vector of dimension 190 for a 10-seconds video that, if the video is real, should be characteristic of the person of interest (POI). This feature vector is fed to a SVM classifier trained on the specific POI. The positive part of dataset for the experiments is built collecting videos of the POIs speaking in formal settings (e.g. news and public speech) and facing towards the camera, while the negative part was created using a GAN trained on the POI. This approach has the advantage of not exploiting low level features, therefore being resilient to laundering (noise or blur adding) and it is able to detect not only deep fakes, but also detect even only facial expression manipulation. The average accuracy in face swap detection over a testing on 4 different POIs is 0.95.

In [13] Sabir et al. propose a convolutional model together with a bidirectional recurrent model, similar to [4] but training it from scratch. The model exploits discrepancies between frames, achieving a state of the art accuracy of 0.969 on the deepfake samples of FaceForensics++ dataset (LQ).

In [14] Nguyen et al. uses a CapsNet for detection on the faces extracted from frames. The network has a first part consisting of the first 3 layers of a VGG network trained on ILSVRC dataset (not too deep to take information about the object detection task), then a layer of primary capsules and the final capsules. The capsule part is the one which is fine tuned on the fake detection task. The scores of all frames in the video are finally averaged. During the training phase the paper also uses some regularization techniques to reduce the overfitting phenomenon. The network uses the consensus of different capsules on the last but one layer to make a prediction on the last one, this gives a degree of robustness in the sense that if a capsule fails to detect the manipulation the network can rely on the others. Moreover the network is not limited to binary classification, so in the experiments it has been trained on FaceForensics++ dataset to output is an image is 'real', 'Deepfake', 'Face2Face' or 'FaceSwap'. The network achieves similar results to the XceptionNet described in [8] but with about one fifth of the parameters.

Li et al. in [15] generalize the manipulation detection by using the fact that most forgery techniques have in their pipeline a merging step when the fake face is applied to the real image. This merging carries inconsistencies between the low level features (noise and artifacts) of the face and the surroundings. By training a CNN they produce a gray-scale "X-Ray" version of the input image where the white pixels correspond to the blending border while all the rest is black. With the same network they predict also if the image is real or fake. They achieve great generalization results in several datasets while training in only one. Despite the generalization skill, the network does not perform as well as other networks when they are trained on the specific tested dataset.

2.3. Discussion

Many new algorithms have been proposed in the last couple of years, and more are going to be published in near future. Several approaches are shown to obtain outstanding accuracy results, especially when tested on a familiar dataset, but they seem to struggle when facing new manipulation techniques. It is not by chance that the most recent works try to address the generalization problem, achieving some interesting result, but it's still a long way to go. Another issue is understanding the reason why in certain models a video is predicted as fake, especially when the model is a complex deep neural networks. While some methods use carefully chosen features, others delegate the feature detection and extraction to a machine learning process, which gives us little clue about the internal mechanisms that lead to the decision.

References

[1] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2387–2395, June 2016. doi: 10.1109/CVPR.2016.262.

[2] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, 2018.

[3] Pavel Korshunov and Sebastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection, 2018.

[4] D. Güera and E. J. Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, Nov 2018. doi: 10.1109/AVSS.2018.8639163.

[5] Y. Li, M. Chang, and S. Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, 2018.

[6] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 46–52, 2018.

[7] X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265, 2019.

[8] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images, 01 2019.

[9] F. Matern, C. Riess, and M. Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92, Jan 2019. doi: 10.1109/WACVW.2019.00020.

[10] Huy H. Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos, 2019.

[11] Joel Stehouwer, Hao Dang, Feng Liu, Xiaoming Liu, and Anil Jain. On the detection of digital face manipulation, 2019.

[12] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[13] E. Sabir, K. Cheng, A. Jaiswal, W. Abdalmageed, I. Masi, and P. Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. In *Workshop on Applications of Computer Vision and Pattern Recognition to Media Forensics with CVPR*, 2019.

[14] H. H. Nguyen, J. Yamagishi, and I. Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2307–2311, 2019.

[15] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection, 2019.

[16] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection, 2020.

[17] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Recasting residual-based local descriptors as convolutional neural networks: An application to image forgery detection. In *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, page 159–164, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450350617. doi: 10.1145/3082031.3083247. URL https://doi.org/10.1145/3082031.3083247.

[18] Belhassen Bayar and Matthew C. Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, page 5–10, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342902. doi: 10.1145/2909827.2930786. URL https://doi.org/10.1145/2909827.2930786.

[19] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen. Distinguishing computer graphics from natural images using convolution neural networks. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pages 1–6, Dec 2017. doi: 10.1109/WIFS.2017.8267647.

[20] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2017.

[21] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848.