On the Use of Implicit Representations for Deepfake Detection

Miguel Leão
https://visteam.isr.uc.pt/team/120/
Nuno Gonçalves
https://visteam.isr.uc.pt/team/nuno-goncalves-2/

Institute of Systems and Robotics
University of Coimbra
Portugal

INTRODUCTION

The developments in home computers, united with the thousands upon thousands of images/videos of individuals present on the Internet, allowed for the proliferation of deepfaked media affecting the lives of private individuals and the dangerous spread of misinformation. Current state-of-the art detection methods show impressive results. However the development of improved generation methods overcomes them, as there are generalization difficulties.

Following the logic of forensic approaches in both the color and frequency space, this work investigates the use of the implicit space in the problem of deepfake detection. Implicit representations have recently offered new research avenues for image analysis, translating a scene usually in a coordinates-based representation, that allows for detailed reconstructions of the original. Using Sinusoidal Representation Networks (SIRENs) [9] the video frames of the Deepfake Detection Challenge Dataset (DFDC) [2] were translated to the implicit space and analyzed.

This work uses Fréchet Video Distance (FVD) [10] between the original DFDC videos and their respective SIREN reconstruction, to show a significant difference in the average FVDs of the bonafide and deepfake pairs. We expect that this work might open new avenues of research for the deepfake detection problem.

METHOD

Implicit representation

An image is represented as a function $I : \Omega \subset R^2 \rightarrow C$, where Ω is the image's domain and C is the color space. The image is then parameterized with a coordinate-based neural network $I_\theta : R^2 \rightarrow C$ with parameters θ. To train the neural image I_θ so that it approximates I, the model optimizes the following objective:

$$\int_\Omega (I - I_\theta)^2 \, dx.$$

The coordinate-based network is a sinusoidal multilayer perceptron (MLP) $f_\theta(p) : R^n \rightarrow R^m$, defined as a composition of d sinusoidal layers:

$$f_\theta(x) = W_d \circ f_{d-1} \circ \cdots \circ f_0(x) + b_d,$$

where each layer $f_i(x_i) = \sin(W_i x_i + b_i) = x_{i+1}$, with $W_i \in R^{n_{i+1} \times n_i}$ being the weight matrices, and $b_i \in R^{n_{i+1}}$ being the biases. The collection of these parameters defines θ. The integer d denotes the depth of the network, and n_i refers to the width of the layers.

With the neural image defined by θ, the RGB values for any pixel of a reconstructed image are given by the value of f_θ at x coordinates.

Through the method used in [8], the neural images of each frame of the subject's face is obtained. The individual frames are then joined into a reconstructed video.

Distance between original and reconstructed videos

This article proposes to show that there is a difference between how reliable the neural reconstruction of a video is for bonafide and deepfake video cases, so that it can be used to detect the latter. This is measured through Fréchet Video Distance (FVD).

FVD is proposed as an improvement on common video analysis approaches such as Peak Signal-to-Noise-Ratio (PSNR) or Structural Similarity (SSIM) [11] claiming that these lack for the temporal coherence of the video, aside from the video quality itself. It is based on the principal of Fréchet Inception Distance (FID) [4], commonly used for image analysis, where the distance between the real world data distribution P_R and the distribution defined by the generative model P_G is defined by:

$$d(P_R, P_G) = min_{X,Y} E|X - Y|^2$$

where the minimization is over all random variables X and Y with distributions P_R and P_G respectively. With the data distribution being represented as a multivariate Gaussian using a suitable feature space, the previous equation can be solved as:

$$d(P_R, P_G) = |\mu_R - \mu_G|^2 + Tr(\Sigma_R + \Sigma_G - 2(\Sigma_R \Sigma_G)^{\frac{1}{2}})$$

where μ_R and μ_G are the means and Σ_R and Σ_G are the co-variance matrices of P_R and P_G. This representation is obtained from an Inflated 3D ConvNet (I3D) [1], and the distance between videos is obtained. In our work, we obtained the FVD through the implementation used in [3].

EXPERIMENTS AND RESULTS

Dataset

The Deepfake Detection Challenge (DFDC) [2] is a self-designated third generation dataset featuring 23,654 videos from 960 actors hired for this purpose, from which 104,500 fake videos are created using various deepfake creation methods.

These include a Deepfake Auto Encoder (DFAE) model with a shared encoder but two isolated decoders, one for each identity, and a Neural Talking Heads (NTH) [12] model comprised of a metalearning stage and a fine-tuning stage.

It also includes deepfakes generated from FSGAN [6] which applies an adversarial loss to generators for reenactment and inpainting, and trains additional generators for face segmentation and Poisson blending and StyleGAN [5] which is modified to produce a face swap between a given fixed identity descriptor onto a video by projecting this descriptor on the latent face space. Finally, certain videos from the previous categories are processed with a sharpening filter to improve the quality of the final video and certain videos receive vocal deepfakes as presented in [7].

SIREN reconstructions

The SIREN models were trained for 1000 epochs for each frame, resulting in a reconstruction that shows no differences to the naked eye, for both deepfake and bonafide videos, even for the ones scoring the highest FVD scores, as shown in figures 1.

Figure 1: Comparison between an original frame (left) from a video, it's SIREN reconstruction (center) and their difference (right), for bonafide cases in green and deepfake cases in red.

Although the reconstructions do not show visible differences when analyzed, it is possible to find the areas in the image where the reconstructions is not perfect. Analyzing these areas together with additional information from the scene can give insights into the problem.

This would greatly benefit from a labeling effort on the dataset to properly analyze if and how different conditions affect the SIREN representation for bonafide and deepfake cases.

Testing the hypothesis

The average FVD scores obtained show that SIREN reconstructions for the bonafide videos have lower fidelity to their original video than in the deepfake videos, achieving higher FVD scores, as shown in figure 1. Higher FVD scores mean higher distance between video pairs i.e. worse reconstructions.

To confirm that the data shows that SIREN reconstructions could be used for deepfake detection, specifically that the neural reconstructions of bonafide material have lower fidelity than deepfake reconstructions, a one tail significance test is conducted. First the null hypothesis is defined as there is no difference between the distribution of FVD scores for the original and SIREN reconstruction of bonafide videos and deepfake videos, or that the FVD scores for bonafide videos are lower then the deepfake scores, i.e. SIRENS achieve better reconstructions on bonafide videos than deepfake ones:

$$H_0 : \mu FVD_{bonafide} <= \mu FVD_{deepfake}$$

and present the alternative hypothesis that the bonafide video pairs score higher FVDs, therefore have worse fidelity, than the deepfake video pairs:

$$H_a : \mu FVD_{bonafide} > \mu FVD_{deepfake}$$

conducting the test with a significance level of $\alpha = 0.01$, the p-value result is equal to $1.145e-5$ giving $p < \alpha$, thus rejecting the null hypothesis.

Discussion

The data shows that SIREN reconstructions bonafide videos have lower fidelity than the reconstructions of deepfake videos. This could suggest that bonafide videos contain richer information, which is lost during manipulation.

The fact that the standard deviation for deepfake scores is lower, may also indicates a process of homogenization of the information. The DFDC is a large dataset, so as previously mentioned, its full translation into neural representations would, at this pace, take a not practical amount of time. However by expanding this research over more videos from the dataset, it would give a clearer idea of how different attributes from these videos might affect the neural representation, or how certain deepfake generative methods impact the image.

This result could contribute to the development of a system for detecting deepfakes by learning how to distinguish between bonafide videos and deepfake videos by analyzing the original video and its neural reconstruction.

However, there is still work to be done in this area, particularly in understanding the influence of certain factors like resolution, the context of the video (e.g. how much of the frame does the face occupy), among other elements.

CONCLUSION AND FUTURE WORK

Conclusion

This article presented the hypothesis of using implicit representations of facial videos to distinguish between bonafide and deepfake videos. Carrying out this first analysis with videos reconstructed from the SIREN representation, the FVD value between the original videos and their reconstructions was measured. These values were use to test the hypothesis that the bonafide reconstructions have lower fidelity to their original material when compared to the reconstruction of deepfake videos. Carrying out a significance test at a significance level of 99%, we were able to show that the null hypothesis was rejected. Although these are initial results, the hypothesis that we can use implicit representations to detect deepfakes seems promising.

Future work

Having reached these conclusions, it is necessary to consider how to proceed. The end result of this research is expected to achieve state-of-the-art deepfake detection. There are still a number of obstacles to overcome, with problems such as data volume. It is still required to test if all frames from a video are required to achieve satisfactory results. This, among a battery of ablation tests, will be conducted as to conceive the "ideal" conditions to proceed with research.

While this paper revolves around videos reconstructed from their SIREN representations, it is to show a discernible distinction between deepfakes and bonafide material. Future work will be conducted as much as possible with the implicit representation itself.

REFERENCES

[1] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. doi: 10.1109/CVPR.2017.502.

[2] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton-Ferrer. The deepfake detection challenge dataset. *ArXiv*, abs/2006.07397, 2020.

[3] Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the Content Bias in Fréchet Video Distance . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7277–7288, Los Alamitos, CA, USA, June 2024. IEEE Computer Society. doi: 10.1109/CVPR52733.2024.00695.

[4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[5] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019. doi: 10.1109/CVPR.2019.00453.

[6] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7183–7192, 2019. doi: 10.1109/ICCV.2019.00728.

[7] Adam Polyak, Lior Wolf, and Yaniv Taigman. Tts skins: Speaker conversion via asr. In *Interspeech*, 2019.

[8] Guilherme Schardong, Tiago Novello, Hallison Paz, Iurii Medvedev, Vinicius Da Silva, Luiz Velho, and Nuno Goncalves. Neural Implicit Morphing of Face Images . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7321–7330, Los Alamitos, CA, USA, June 2024. IEEE Computer Society. doi: 10.1109/CVPR52733.2024.00699.

[9] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7462–7473. Curran Associates, Inc., 2020.

[10] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. In *DGS@ICLR*, 2019.

[11] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

[12] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9458–9467, 2019. doi: 10.1109/ICCV.2019.00955.