AIMH Lab Approaches for Deepfake Detection

Davide Alessandro Coccomini[1,*], Roberto Caldelli[2,3], Andrea Esuli[1], Fabrizio Falchi[1], Claudio Gennaro[1], Nicola Messina[1] and Giuseppe Amato[1]

[1]*ISTI-CNR, via G. Moruzzi, 1, 56100, Pisa, Italy*

[2]*National Inter-University Consortium for Telecommunications (CNIT), viale Morgagni, 65, 50134, Florence, Italy*

[3]*Mercatorum University, 00186, Rome, Italy*

Abstract

The creation of highly realistic media known as deepfakes has been facilitated by the rapid development of artificial intelligence technologies, including deep learning algorithms, in recent years. Concerns about the increasing ease of creation and credibility of deepfakes have then been growing more and more, prompting researchers around the world to concentrate their efforts on the field of deepfake detection. In this same context, researchers at ISTI-CNR's AIMH Lab have conducted numerous researches, investigations and proposals to make their own contribution to combating this worrying phenomenon. In this paper, we present the main work carried out in the field of deepfake detection and synthetic content detection, conducted by our researchers and in collaboration with external organizations.

Keywords

Deepfake Detection, Syntethic Content Detection, Computer Vision, Deep Learning

1. Introduction

In recent years, there has been a rapid increase in the development of artificial intelligence technologies, including deep learning algorithms, that have led to the creation of highly realistic media manipulations known as deepfakes. Deepfakes refer to synthetic media generated using machine learning techniques designed to mimic the appearance and behaviour of real individuals in videos or images manipulating what they do and what they say.

While deepfakes have some potential positive applications, such as in the entertainment industry, they pose severe risks to society, including political, social, and economic threats. For instance, deepfakes can be used to spread disinformation, manipulate public opinion and damage personal reputations.

Given the potential harm caused by deepfakes, it is crucial to develop effective methods for detecting and mitigating them. In recent years, there has been a surge in research on deepfake detection techniques, and several deepfake detection tools have been developed. However, the field of deepfake detection is still in its initial phase and there is a need for further research and development to create more accurate and reliable detection methods.

For these reasons, AIMH Lab conducted some studies trying both to investigate the various deepfake detection solutions and also propose novel approaches capable of effectively managing the heterogeneity of real-world contexts.

2. Research Works in Deepfake Detection

In this section we present our works in Deepfake Detection and related fields, highlighting the contributions and discoveries made.

2.1. Convolutional Cross Vision Transformer

When we started to take our first steps in this field, we noticed a shortage of Vision Transformer-based deepfake detectors and even more so an almost total absence of hybrid architectures used for this purpose. The reason for this is certainly their very recent advent. We therefore wished to explore this untrodden field in the paper [1] in which we realised a hybrid architecture composed of a convolutional network, in particular, EfficientNet-B0, and a Cross Vision Transformer. The latter's internal attention mechanism, in our proposal, instead of working on the patches extracted from the images, acts on the features obtained from the EfficientNet. The advantage lies in the fact that these features are obtained from a learnable process that is refined in the training phase

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy

*Corresponding author.

davidealessandro.coccomini@isti.cnr.it (D. A. Coccomini); roberto.caldelli@unifi.it (R. Caldelli); andrea.esuli@isti.cnr.it (A. Esuli); fabrizio.falchi@isti.cnr.it (F. Falchi); claudio.gennaro@isti.cnr.it (C. Gennaro); nicola.messina@isti.cnr.it (N. Messina); giuseppe.amato@isti.cnr.it (G. Amato)

0000-0002-7703-3367 (D. A. Coccomini); 0000-0003-3471-1196 (R. Caldelli); 0000-0002-5725-4322 (A. Esuli); 0000-0001-6258-5313 (F. Falchi); 0000-0002-3715-149X (C. Gennaro); 0000-0003-3011-2487 (N. Messina); 0000-0003-0171-4315 (G. Amato)

CEUR Workshop Proceedings (CEUR-WS.org)

Model	AUC	F1-score	# params
ViT with distillation [5]	0.978	91.9%	373M
Selim EfficientNet B7 [6]†	0.972	90.6%	462M
Convolutional ViT [4]	0.843	77.0%	89M
Efficient ViT (our)	0.919	83.8%	109M
Conv. Cross ViT Wodajo CNN (our)	0.925	84.5%	142M
Conv. Cross ViT Eff.Net B0 - Avg (our)	0.947	85.6%	101M
Conv. Cross ViT Eff.Net B0 - Voting (our)	0.951	88.0%	101M

Table 1
Video-Level results of our models and other previous works DFDC test dataset. The symbol †indicates that the model uses an ensemble of 6 networks.

and that has therefore allowed us to obtain a state-of-the-art model on different datasets. In particular, the one we named Convolutional Cross Vision Transformer proved to be state-of-the-art in terms of accuracy and AUC on both DFDC[2] dataset (shown in Table 1) and FaceForensics++[3], two of the main datasets used in this field, comparing with previous works like [4, 5]. All this, with significantly fewer parameters than other solutions and therefore lighter, thanks to the exploitation of a hybrid architecture. The article also investigated the impact of certain implementation choices such as the number of frames considered per video and the management of multiple identities in the same scene, which were then the basis for some subsequent work.

2.2. Partecipation to the ICIAP 2021 competition

The previous paper presented at ICIAP 2021 was also used as the basis for participation in the Face Deepfake Detection challenge organised at that conference. During the competition, the ability of the participants' solutions to identify deepfakes 'in the wild' was assessed and their generalisation capability was investigated. The latter is one of the main problems in deepfake detection as normally deepfake detectors are very good at identifying deepfakes generated with methods used for the creation of the training set but practically useless when compared with content manipulated with novel techniques. Our method, based on what was presented in the previous section, placed fourth in the ranking and we subsequently produced a paper together with the competition organisers and other participants published in the Journal of Imaging and entitled ”The Face Deepfake Detection Challenge” [7].

2.3. On the generalization of Deepfake Detectors

Building on the experience of previous work, we focused on investigating the generalization ability of deepfake detectors by primarily aiming to validate various deep learning architectures in a cross-forgery context. The first work done in this regard was published at ICMR 2021 under the title of “Cross-Forgery Analysis of Vision Transformers and CNNs for Image Deepfake Detection” [8] and consists of a comparison between a convolutional network, the EfficientNet-V2-M, and a classical Vision Transformer, namely ViT-Base. These two models are based on very different concepts and structures, and it is in these differences that their peculiarities reside, which are reflected in different behaviour in deepfake detection. Our experiments were conducted on the ForgeryNet dataset [9] consisting of images manipulated with 15 different techniques. The models were trained on images manipulated with a specific method or a group of methods and then tested on images manipulated with all available methods. According to our results, the Vision Transformer turns out to have a significantly higher generalization ability than EfficientNet, which instead tends to store more of the specific artifacts introduced by deepfake generation algorithms and thus detects few images generated by other methods. This result is more pronounced in the presence of large masses of data, which allows the Vision Transformer to generalize even better to the deepfake concept.

This work conducted on images was then recently extended to the video portion of ForgeryNet submitted to the Journal of Imaging with the title of ”On the generalization of Deep Learning techniques for Video Deepfake Detection” [10]. In fact, not necessarily what has been discovered in images is reportable on videos since the anomalies that can be found in them can be both spatial and temporal in nature unlike images. In this case, the manipulation techniques used on videos are 9 divided into two macro-categories, ID-Replaced and ID-Remained. The architectures explored are the same as in the previous work but with the addition of a Swin Transformer. The latter was interesting to validate since it is a Transformer based on a hierarchical attention mechanism inspired by convolutional networks and therefore a sort of middle ground between the two architectures considered. According to our experiments, conducted similarly to the previous work, the EfficientNet performs better on a lower data regime while again the Vision Transformer, even on video, is more capable of generalization and less tied to the methods used to create the training set. The Swin Transformer, on the other hand, proves to be a good middle ground between the architectures hardly excelling on the others but achieving satisfactory performance on average.

To summarize, in light of the many experiments conducted in this work, the Vision Transformer and its variants are more suitable to be used as the basis for deepfake detectors to be applied in the real world. The continuing emergence of new deepfake generation techniques

Model	Identities	Accuracy	AUC
SlowFast R-50† [12]	1	82.59	90.86
SlowFast R-50‡ [12]	1	**88.78**	93.88
X3D-M‡ [13]	1	87.93	93.75
MINTIME-EF	1	81.92	90.13
MINTIME-EF	2	82.28	90.45
MINTIME-EF	3	82.05	90.28
MINTIME-XC	1	85.96	93.20
MINTIME-XC	2	87.64	**94.25**
MINTIME-XC	3	86.98	94.10

Table 2
Video-Level Evaluation on ForgeryNet Validation Set. The identities column represents the number of considered identities during the inference. † Indicate that the model has been trained in our setup. ‡ Indicate that the result is taken from [9].

forces detectors to untether themselves from any attempt to memorize the specific anomalies of a technique and abstract the concept of anomaly so that they can recognize deepfakes regardless of the manipulation technique.

2.4. MINTIME: Multi-Identity Size Invariant TimeSformer

The challenges encountered during our previous research resulted in a paper conducted in collaboration with CERTH Thessaloniki and it will be submitted to an international journal, entitled ”MINTIME: Multi-Identity Size Invariant Deepfake Detector”[11]. In this work, we identified some frequent problems in the real world and tried to develop a detector capable of handling them effectively. The main novelties introduced by this work are:

- Development of a new deepfake detection model capable of capturing both spatial and temporal anomalies
- Capability of managing multiple identities within the same scene through the introduction of new techniques of attention, positional embedding and input sequence generation
- Robustness in changing the ratio between the area of the face and that of the entire frame through the introduction of size embedding capable of inducing the real original size of the face with respect to the entire scene

This is also the first deepfake detection work based on TimeSformer and the first time this architecture has been combined in the realisation of a transformer-convolutional hybrid model.

Our approach achieved state-of-the-art results on several datasets outperforming other methods like [14, 15, 16, 17] as shown in Table 2 and in particular when tested on multi-identity videos only as can be seen in Table 3. Our approach also exposed outstanding results in terms of generalization which can be seen in cross-forgery evaluation presented in Table 4. From the interpretation of the attention maps it is also possible to trace which of the multiple identities, if any, has been manipulated, making the approach more usable in the real world.

Model	Accuracy	AUC
SlowFast R-50 [12]	72.63	80.92
MINTIME-EF	81.21	89.56
MINTIME-XC	**86.68**	**94.12**

Table 3
Evaluation on multi-identity only videos of ForgeryNet Validation Set. The models are all trained in our setup.

		ID-replaced		ID-remained	
		Accuracy	AUC	Accuracy	AUC
X3D-M[13]	ID-replaced	87.92	92.91	55.25	65.59
	ID-remained	55.93	62.87	88.85	95.40
SlowFast R-50[12]	ID-replaced	**88.26**	92.88	52.64	64.83
	ID-remained	52.70	61.50	87.96	95.47
MINTIME-EF	ID-replaced	80.18	83.86	79.03	86.98
	ID-remained	63.13	66.26	89.22	95.02
MINTIME-XC	ID-replaced	86.58	**93.66**	**84.02**	**88.43**
	ID-remained	**64.01**	**68.53**	**92.08**	**97.26**

Table 4
Cross-Forgery Evaluation on ForgeryNet Validation Set. X3D-M and SlowFast R-50 results are taken from [9].

2.5. Syntethic Media Detection

Deepfakes are among the applications of deep learning that have caused most concern to date, however, recently not only people’s faces have been subject to manipulation. In fact, many techniques are emerging that make it possible to generate images of any subject, for instance from a text describing it. This poses major challenges as it will be increasingly difficult to distinguish between synthetic and real content. The ability to differentiate between synthetic and real images is essential for preserving the integrity of information and safeguarding individuals against the malicious use of synthetic media. As text-to-image methods become increasingly prevalent and accessible to the general public, society is moving toward a point where a significant amount of online content is synthetic, blurring the line between reality and fiction. In our recent work titled ”Detecting Images generated by Diffusers”[18], we present an initial attempt to distinguish between generated and real images based on the image itself and the associated text used to describe and generate it. Additionally, we analyze the image and text’s peculiarities that may result in a more or less credible image that is difficult to identify. Our analysis focused on detecting content generated by text-to-image systems, specifically Stable Diffusion and GLIDE. We tested various classifiers, including MLPs and Convolutional Neural

Networks, and found that traditional deep learning models can easily distinguish images generated with these systems once they have seen examples in the training set. However, when tested for generalization ability, they were rarely able to identify images generated by methods other than those used in the training set, highlighting a significant issue for these systems' real-world adoption. We also conducted an analysis of the correlation between the credibility of generated images and their category, as well as the composition of their associated captions. Our experiments found that images generated by both generators are more credible when they depict inanimate objects, resulting in greater classifier error. In contrast, images depicting people, animals, or animate subjects, in general, are easier to identify. Moreover, there appears to be no strong correlation between the sentence's linguistic composition and the models' classification ability.

3. Related Projects

3.1. AI4Media

The AI4Media project is a European Union-funded initiative that aims to advance the state-of-the-art in artificial intelligence (AI) and machine learning (ML) technologies and their application in the media industry. The project focuses on developing innovative tools and techniques for improving media production, distribution, and consumption, with the ultimate goal of enhancing the quality, diversity, and accessibility of media content while maintaining high ethical and social standards. The project brings together leading research institutions, media organizations, and technology companies across Europe, including universities, broadcasters, publishers, and startups. Through collaborative research and development efforts, the project aims to address various challenges and opportunities in the media industry, such as the creation of personalized content, the detection and prevention of fake news and disinformation, and the improvement of accessibility and user experience. Partners in this project can join the AI4Media Fellowship programme in which young researchers can visit the sites of other members and collaborate on joint research projects. In particular, Davide Alessandro Coccomini, PhD student at the University of Pisa and research associate at ISTI-CNR, spent a two-month exchange period at CERTH in Thessaloniki. During this period and the following months, research was conducted on the realisation of a deepfake detector capable of handling various real-world challenging situations.

3.2. SERICS

The SERICS Foundation, focused on Security and Rights in Cyberspace, has been established in compliance with the principles and legal framework of the participation foundation, which is part of the broader category of foundations governed by the Civil Code and related laws. Its primary objective is scientific and technological research, and as such, it has been designated as the implementing party for the "SERICS - Security and Rights in CyberSpace" Partnership, which is funded under the Public Notice "for the presentation of Proposals for the creation of "Partnerships extended to universities research centers, companies for the funding of basic research projects" - as part of the National Recovery and Resilience Plan, Mission 4 "Education and Research" - Component 2 "From Research to Enterprise" - Investment 1. 3, funded by the European Union - NextGenerationEU - Notice no. 341 of 15.3.2022.

4. Conclusions and Future Work

In this paper, we illustrated the work conducted by ISTI-CNR's AIMH laboratory in the field of deepfake detection. The first steps taken in this field are relatively recent, but within a short period of time the team of researchers has achieved excellent results in various contexts by exploring and discovering peculiarities, solving open problems and posing new ones, all in collaboration with the fervent group of researchers actively working in this field. In the future, the aim is to extend the work that has been started and to continue the collaboration with other institutions in order to achieve an increasingly precise, robust and real-world-friendly deepfake detector, but also to investigate the field of synthetic content detection in greater depth, so as to counter misinformation and contribute to the maintenance of network security for all users.

Acknowledgments

This work was partially supported by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU and by FAIR (PE00000013) funded by the European Commission under the NextGenerationEU. It was also partially supported by AI4Media project, funded by the EC (H2020 - Contract n. 951911).

References

[1] D. A. Coccomini, N. Messina, C. Gennaro, F. Falchi, Combining efficientnet and vision transformers for video deepfake detection, in: Image Analysis and Processing (ICIAP 2022) - Part III, Springer, 2022, p. 219–229. URL: https://doi.org/10.1007/978-3-031-06433-3_19. doi:10.1007/978-3-031-06433-3_19.

[2] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C. C. Ferrer, The deepfake detection challenge (dfdc) dataset, arXiv preprint arXiv:2006.07397 (2020).

[3] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: Learning to detect manipulated facial images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1–11.

[4] D. Wodajo, S. Atnafu, Deepfake video detection using convolutional vision transformer, arXiv preprint arXiv:2102.11126 (2021).

[5] Y.-J. Heo, Y.-J. Choi, Y.-W. Lee, B.-G. Kim, Deepfake detection scheme based on vision transformer and distillation, arXiv preprint arXiv:2104.01353 (2021).

[6] S. Seferbekov, Dfdc 1st place solution, 2020. URL: "https://github.com/selimsef/dfdc_deepfake_challenge".

[7] L. Guarnera, O. Giudice, F. Guarnera, A. Ortis, G. Puglisi, A. Paratore, L. M. Q. Bui, M. Fontani, D. A. Coccomini, R. Caldelli, F. Falchi, C. Gennaro, N. Messina, G. Amato, G. Perelli, S. Concas, C. Cuccu, G. Orrù, G. L. Marcialis, S. Battiato, The face deepfake detection challenge, Journal of Imaging 8 (2022). URL: https://www.mdpi.com/2313-433X/8/10/263. doi:10.3390/jimaging8100263.

[8] D. A. Coccomini, R. Caldelli, F. Falchi, C. Gennaro, G. Amato, Cross-forgery analysis of Vision Transformers and CNNs for deepfake image detection, in: Proceedings of the 1st International Workshop on Multimedia AI against Disinformation, MAD '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 52–58. URL: https://doi.org/10.1145/3512732.3533582. doi:10.1145/3512732.3533582.

[9] Y. He, B. Gan, S. Chen, Y. Zhou, G. Yin, L. Song, L. Sheng, J. Shao, Z. Liu, Forgerynet: A versatile benchmark for comprehensive forgery analysis, in: CPVR, 2021, pp. 4358–4367. doi:10.1109/CVPR46437.2021.00434.

[10] D. A. Coccomini, R. Caldelli, F. Falchi, C. Gennaro, On the generalization of deep learning models in video deepfake detection (2023). URL: https://www.preprints.org/manuscript/202303.0161/v1. doi:https://doi.org/10.20944/preprints202303.0161.v1.

[11] D. A. Coccomini, G. K. Zilos, G. Amato, R. Caldelli, F. Falchi, S. Papadopoulos, C. Gennaro, Mintime: Multi-identity size-invariant video deepfake detection, 2022. URL: https://arxiv.org/abs/2211.10996. doi:10.48550/ARXIV.2211.10996.

[12] C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, in: ICCV, 2019.

[13] C. Feichtenhofer, X3d: Expanding architectures for efficient video recognition, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 200–210. doi:10.1109/CVPR42600.2020.00028.

[14] Y. Zheng, J. Bao, D. Chen, M. Zeng, F. Wen, Exploring temporal coherence for more general video face forgery detection, in: ICCV, 2021, pp. 15024–15034. doi:10.1109/ICCV48922.2021.01477.

[15] A. Haliassos, K. Vougioukas, S. Petridis, M. Pantic, Lips don't lie: A generalisable and robust approach to face forgery detection, in: CVPR, 2021, pp. 5037–5047. doi:10.1109/CVPR46437.2021.00500.

[16] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, P. Natarajan, Recurrent convolutional strategies for face manipulation detection in videos, in: CVPR Workshops, 2019.

[17] H. H. Nguyen, F. Fang, J. Yamagishi, I. Echizen, Multi-task learning for detecting and segmenting manipulated facial images and videos, in: 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), 2019, pp. 1–8. doi:10.1109/BTAS46853.2019.9185974.

[18] D. A. Coccomini, A. Esuli, F. Falchi, C. Gennaro, G. Amato, Detecting images generated by diffusers, 2023. URL: https://arxiv.org/abs/2303.05275. doi:10.48550/ARXIV.2303.05275.