Generalization Improvements in Deepfake Detection for Real World Applications

Dan-Cristian Stanciu
AI Multimedia Lab
National University of Science and Technology POLITEHNICA Bucharest
Bucharest, Romania
https://orcid.org/0009-0008-4561-4328

Bogdan Ionescu
AI Multimedia Lab
National University of Science and Technology POLITEHNICA Bucharest
Bucharest, Romania
https://orcid.org/0000-0003-4112-5769

***Abstract*—Deepfakes are the result of the incredible developments of generative models from the last few years. They are images, videos, audio files that portray a person and are generated by artificial intelligence. Although data generation can be used for a lot of good purposes, there are also some malicious ways in which they can be used, like extortion or identity theft. Therefore, there is an increasing focus on detecting deepfakes in real-world conditions, with come really promising results. At the same time, the biggest problem with deepfake detectors is the lack of generalization: they perform well on test datasets but perform poorly in real life. This work presents our progress in increasing the generalization of deepfake detectors, lowering the model parameters and improving the performance of current deepfake detectors in the context of real world applications. Finally, this paper presents potential future implementations that can solve some of the shortcomings of current state-of-the-art approaches.**

***Index Terms*—deepfake detection, image forensics, multimedia**

I. Introduction

Nowadays, there are several means to produce images, videos, audio or text, many of which are open-source. It is easier than ever to generate a Deepfake, while having little to none programming or machine learning experience. This is why it is essential that we find ways to fight disinformation through the means of Deepfakes. This paper explores deepfake detection using a variety of architectures and improvements while emphasizing the challenge of creating models that generalize effectively across varied deepfake types. Our contributions are listed as follows:

- Summarizing the current state of the art in deepfake detection, while highlighting its challenges and shortcomings
- Presenting our progress in deepfake detection, focused on improving performance and generalization with resource-efficient models, augmentation and features.
- Outlining our future approaches and plans

II. Related work

A. Deepfake detection approaches

Some of the best-performing deepfake detectors are approaches like [1]–[3]. They all use complex deep learning solutions and architecture for deepfake detection, on some of the most realistic deepfake generation datasets, like CelebDF [4] and DFDC [5].

One of the best-performing approaches is EFNB4 + SBIs [1], which focuses on blending artifacts. When a face image is generated, it can be pasted on a real body of a person, changing the person's identity. This is called identity change deepfake. The majority of the datasets in the state of the art are focused on identity change, as it is one of the most common ways to generate a deepfake. Another approach is PCL + I2G [2], which focuses on generating new deepfakes for the identity-change task and learning to detect in-image inconsistency. Lastly, a 3D R50-FTCN [3] can be used to detect inconsistency in deepfake video. That is based on a large combination of 3D CNN and Video transformer architectures, and it hinges on the fact that many deepfakes are generated image-by-image, not as a whole video. This way, inconsistencies in video are always possible.

B. Discussion regarding current approaches

Current deepfake generators perform very well on research datasets like CelebDF or DFDC, but they perform much worse in real-life scenarios. One of the reasons for that is the fact that there is a large variety of possible scenarios, while the majority of the state-of-the-art approaches are focused on only one or two. For example, [1], [2] focus on the identity change deepfakes, which derive their weaknesses from the blending algorithm used to paste the deepfake onto the face. Therefore, a completely generated image, or one that generates the face and some background will not be detected easily. Basically, a lot of approaches do not focus on the deepfake itself, but rather on artifacts generated by simpler operations.

On the other hand, there are a lot of approaches that focus on video deepfakes, based on intra-frame consistency, like [3]. Those approaches do not work on images and are also very easy to fool with video compression algorithms, which can be found on every website nowadays.

Although current approaches are good at one specific thing, none of them are truly general, and very few focus on finding inconsistencies in the generated image itself. For this reason, the approaches presented in the next chapter are focused on

1st Workshop on Artificial Intelligence for Multimedia, AI Multimedia Lab, CAMPUS Research Institute, POLITEHNICA Bucharest, Bucharest, Romania, 8 November, 2024

two things: (1) improving performance without increasing model size and (2) improving performance of deepfake detectors in a multitude of conditions, focusing on the one common element in all of them: the generated image.

III. Contributions

A. Detecting Deepfakes Using Capsule Networks

To enhance deepfake detection accuracy while minimizing computational costs, we employed Capsule Networks (CapsNet) [6], an architecture that retains spatial hierarchies better than traditional convolutional neural networks (CNNs). Our modified CapsNet architecture specifically removes pooling layers to maintain image detail, increases the number of primary capsules for enriched data representation, and optimizes the routing algorithm through additional iterations. This architecture achieved a best-in-class AUC of 99.88% on the CelebDF dataset with 22 million parameters and maintained comparable performance (99.56% AUC) with a reduced model configuration of only 6.4 million parameters. These results underscore the scalability and efficiency of CapsNet models for deepfake detection, especially in scenarios where processing resources may be limited.

B. Temporal Detection of Deepfakes on Selected Facial Features

Deepfake videos often display inconsistencies over time, particularly in key facial features, that static image-based models may overlook. To address this, we developed a CNN-LSTM [7] approach that leverages temporal information, focusing on specific facial landmarks such as the mouth, eyes, and nose. Our method begins by extracting frame-by-frame features through a CNN backbone (e.g., XceptionNet or ResNet), followed by temporal analysis using a 2-layer LSTM network over sequences of 60 frames. By restricting analysis to key facial regions, we enhance the model's ability to detect subtle temporal manipulations. This approach achieved a significant performance improvement, raising AUC from 83.6% to 97% on CelebDF. Additionally, it enables detection in cases where only a single facial region has been altered, proving the robustness and specificity of this targeted temporal detection technique.

C. Improving Generalization with Autoencoder-Based Augmentation

One major limitation in deepfake detection lies in the model's ability to generalize across diverse datasets, often due to overfitting to specific "generator fingerprints" left by deepfake generation methods. To tackle this, we introduced an autoencoder-based augmentation technique [8], [9] that improves generalization by minimizing model dependence on generator-specific artifacts. In our approach, we utilize a range of over 80 unique autoencoder configurations to regenerate training images, adding noise or alternative "fingerprints" that suppress model-specific features. This training paradigm effectively reduces overfitting, leading to improved cross-dataset performance: AUC increased by nearly 10% on CelebDF and by 2% on DFDC, when the model was trained on FaceForensics++. Further, this augmentation enhances robustness against common perturbations, such as lossy compression and adversarial attacks, showing over 3% and 2% improvement in AUC, respectively. This approach highlights a promising pathway toward more resilient, real-world-ready deepfake detection models.

D. Improving Generalization in Deepfake Detection via Augmentation with Recurrent Adversarial Attacks

In this paper [10], we introduce a novel framework to enhance generalization in deepfake detection through a recurrent adversarial augmentation approach. The primary contribution is a data augmentation framework that generates diverse deepfake samples solely from real images, exploiting adversarial attacks to create realistic but challenging instances that evade current detectors. Our approach centers on a recurrent training paradigm, wherein adversarial deepfakes are continuously produced and used to update the detector, therefore improving its robustness and adaptability. We validate our method on unseen datasets, showing substantial improvements in generalization without additional training data, achieving significant AUC increases on CelebDF and DFDC Preview. Additionally, we investigate enhancements to this framework, including varying generator architectures and incorporating more real data, demonstrating that these adjustments further amplify detector performance at a minimal computational cost. Our results demonstrate a notable improvement in generalization, with AUC gains of up to 10% on CelebDF and 9% on DFDC Preview, underscoring the effectiveness of our recurrent adversarial augmentation framework in enhancing deepfake detection across diverse datasets. This recurrent adversarial framework provides a flexible, model-agnostic solution, capable of operating on individual images, videos, and a wide range of deepfake types, positioning it as an adaptable tool for real-world deepfake detection applications.

IV. Conclusion and future work

In this paper, we highlighted out progress in the field of deepfake detection, focusing on real-world conditions. We presented the current state-of-the-art approaches for deepfake detection and highlighted some of their shortcomings, which we aim to solve, like generalization, focusing on a single type of deepfakes or the model sizes. We presented our contributions, aimed to solve those problems: scaling models while retaining performance using Capsule Networks, Tackling the temporal dimension using a CNN-LSTM architecture with facial features, and improving generalization by using a multitude of autoencoder architectures to simulate different deepfake generator and to allow the detector to learn the "fingerprints" of the convolution operation, or improving generalization by generating an unlimited number of pseudo-deepfakes based on real images, aimed to train any deepfake generator and improve its performance and resilience. For future developments, we are looking at extending the generalization methods to the temporal dimension and evaluating

their effect on the state-of-the-art approaches and in real-life, diverse conditions.

REFERENCES

[1] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 720–18 729.

[2] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for deepfake detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 023–15 033.

[3] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 044–15 054.

[4] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3207–3216.

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

[6] D.-C. Stanciu and B. Ionescu, "Uncovering the strength of capsule networks in deepfake detection," in *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation*, ser. MAD '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 69–77. [Online]. Available: https://doi.org/10.1145/3512732.3533581

[7] ——, "Deepfake video detection with facial features and long-short term memory deep networks," in *2021 International Symposium on Signals, Circuits and Systems (ISSCS)*, 2021, pp. 1–4.

[8] ——, "Autoencoder-based data augmentation for deepfake detection," in *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*, ser. MAD '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 19–27. [Online]. Available: https://doi.org/10.1145/3592572.3592840

[9] L.-D. Ştefan, D.-C. Stanciu, M. Dogariu, M. G. Constantin, A. C. Jitaru, and B. Ionescu, "Deepfake sentry: Harnessing ensemble intelligence for resilient detection and generalisation," 2024. [Online]. Available: https://arxiv.org/abs/2404.00114

[10] D.-C. Stanciu and B. Ionescu, "Improving generalization in deepfake detection via augmentation with recurrent adversarial attacks," in *Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation*, ser. MAD '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 46–54. [Online]. Available: https://doi.org/10.1145/3643491.3660291