Distinguishing Multimodal DeepFakes from Natural Images: The ELSA Challenge on DeepFake Detection

Lorenzo Baraldi[1,*], Marcella Cornia[1] and Rita Cucchiara[1,2]

[1]University of Modena and Reggio Emilia, Modena, Italy
[2]IIT-CNR, Pisa, Italy

Abstract

Images generated through AI algorithms are becoming increasingly realistic and popular, raising concerns on their improper usage and on their impact on society. The ELSA European Project aims at building a Lighthouse on Secure and Safe AI; among its initiatives, the Multimedia Use Case is building a unique challenge on deepfake detection, which aims at empowering the community with tools and benchmarks for defending against generated content and its improper usage. This paper describes the context of the challenge and the benchmark associated with it. The challenge will offer a unique opportunity to increase the interest of the community towards the development of novel deepfake detection approaches. Through the organization of workshops and dissemination events in major conferences, the Use Case will also create a line of networking opportunities for researchers interested in deepfake detection and build a community around the Use Case itself.

Keywords

Multimodal Deepfakes, Deepfake Detection, ELSA Project

1. Introduction

Machine-generated images are becoming more and more popular in the digital world, thanks to the spread of Deep Learning models that can generate visual data like Generative Adversarial Networks [1, 2] and Diffusion Models [3, 4]. While image generation tools can be employed for lawful goals (*e.g.*, to assist content creators, generate simulated datasets, or enable multimodal interactive applications), there is a growing concern that they might also be used for illegal and malicious purposes, such as the forgery of natural images, the generation of images in support of fake news, misogyny or revenge porn. While the results obtained in the past few years contained artifacts which made generated images easily recognizable, today's results are way less recognizable from a pure perceptual point of view. In this context, assessing the authenticity of fake images becomes a fundamental goal for security and for guaranteeing a degree of trustworthiness of AI algorithms. There is a growing need, therefore, to develop automated methods which can assess the authenticity of images (and, in general, multimodal content), and which can follow the constant evolution of generative models, which become more realistic over time.

The ELSA European Project[1] is a virtual center of excellence that will spearhead efforts in foundational safe and secure artificial intelligence (AI) methodology research. It consists of a large and growing network of top European experts in AI and machine learning and is to promote the development and deployment of cutting-edge AI solutions in the future and make Europe the world's lighthouse of AI. ELSA builds on and extends the existing internationally recognized and excellently positioned ELLIS (European Laboratory for Learning and Intelligent Systems) network of excellence, and comprises six use cases.

In particular, the *ELSA Use Case on Multimedia* focuses on the development of benchmarks and tools for Fake data Understanding and Detection, with the final goal of protecting from visual disinformation and misuse of generated images, and to monitor the progress of existing and proposed solutions for detection. It will investigate novel ways of understanding and detecting fake data, through new machine learning approaches capable of mixing syntactic and perceptive analysis. Also, the Use Case promotes the creation of a competition on deepfake detection which is connected to the ELSA grand challenge of "Human in the loop decision making". This will monitor and evaluate the development of algorithms for deepfake detection, in terms of efficacy, explainability and human oversight, by enabling domain experts to validate and improve results in a human-in-the-loop fashion. The Use Case will be connected to existing initiatives and will include the creation of new datasets for the aforementioned topics.

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy
*Corresponding author.
lorenzo.baraldi@unimore.it (L. Baraldi); marcella.cornia@unimore.it (M. Cornia); rita.cucchiara@unimore.it (R. Cucchiara)
0000-0001-5125-4957 (L. Baraldi); 0000-0001-9640-9385 (M. Cornia); 0000-0002-2239-283X (R. Cucchiara)

[1]Funded as part of the HORIZON-CL4-2021-HUMAN-01 call, Grant Agreement 101070617. See https://www.elsa-ai.eu/.

2. Context and Motivation

Machine-generated images are populating our digital world. Some of them are positive examples of creativity, design, and business, while some are instead fraudulent. Examples from the first class are real images which have been modified either by a human or by a machine learning tool for lawful or harmless goals (*e.g.* for fun, entertainment, communication, simulation, research), while examples of the second class are images tampered for illegal and criminal goals or machine-generated images for misinformation and other illegal goals.

With the constant growth of new Deep Learning models for image generation, the detection of such deepfakes poses several scientific challenges and threats that must be addressed in order to develop accurate algorithms, and to ensure that the proposed benchmark is effective in promoting scientific advancement. The ELSA Challenge on DeepFake detection considers, therefore, the following challenges:

- *the increasing realism of the generation*: deepfake images can be very similar to real images, making it difficult to distinguish between the two. To address this challenge, the Use Case needs to develop methods for generating and collecting high-quality training data, so as to allow researchers participating in the benchmark to design algorithms that can detect deepfakes even when they are highly realistic.
- *The emergence of new image generation algorithms over time.* With the increasing availability of high-quality image generation tools and novel deep learning techniques, the number of deepfakes being produced is growing rapidly, making it difficult for detection algorithms to keep up. Further, during the duration of the project, novel image generation algorithms will likely appear and be based on novel techniques, thus possibly requiring different detection algorithms with respect to the previous state of the art. Further, they will also likely increase their generation quality, thus increasing the level of difficulty in detection.
- *Dataset biases and data coverage*: a deepfake detection solution needs to be robust to different data distribution and to be able to generalise well according to semantic categories, lighting conditions, backgrounds and textual prompts. The benchmark will therefore need to address these challenges by promoting evaluation datasets which share such requirements.
- *Proper evaluation*: proper evaluation is crucial to sustain and promote the development of effective detection solutions. It is important to use rigorous quantitative evaluation metrics to ensure that the proposed methods are able to detect deepfakes with high accuracy and low false positive rate. Additionally, it's important to evaluate the methods against a diverse set of deepfake images that have been created using different techniques, to ensure that they can generalise to real-world scenarios. In addition, it is crucial to evaluate performance in a human-in-the-loop fashion, by comparing with the performance of human evaluators.
- *Explainability*: to build trust in deepfake detection models, it is important for the solution to be explainable, allowing for transparency in how it makes its decisions.

3. The ELSA Benchmark on DeepFake Detection

The benchmark will support two different tasks for deepfake detection, namely: (i) *the recognition of deepfakes generated by ordinary text-to-image models*, in which the entire image is generated at once and, therefore, every pixel of the image is generated; (ii) *the recognition of partially altered images*, *i.e.*, from inpainting models or text-to-image models which only generate a portion of the image.

3.1. Data

The collection and generation of data is a crucial step for the development of the benchmark. For this reason, we are generating a large dataset of generated images using state-of-the-art generators, and supporting both the aforementioned tasks. This will serve at both ensuring that the competition runs on a sufficiently large amount of data, capable of covering a significant portion of the distribution of generated images, and at ensuring that we cover a diverse number of state-of-the-art generators. To this aim, we will generate data with multiple generators, and release different updates of the dataset by adding new generators which will appear over time.

Pre-existing datasets. The community has developed different datasets over the time for deepfake detection. Most existing datasets focus on faces, in both static images [5, 6] and videos [7, 8], and do not consider other semantic classes. The Use Case will, instead, concentrate the benchmark on natural and static images. To this end, we will leverage a subset of the recently released DiffusionDB dataset [9], which currently is the only dataset including state-of-the-art generators and natural images. This will be inserted into the competition to allow participants to benchmark their existing solutions on a benchmark which is already recognized in literature and which has not been designed as part of the Use Case. Finally, we will closely monitor the release of other datasets which satisfy the same constraint, and add them to the benchmark whenever available.

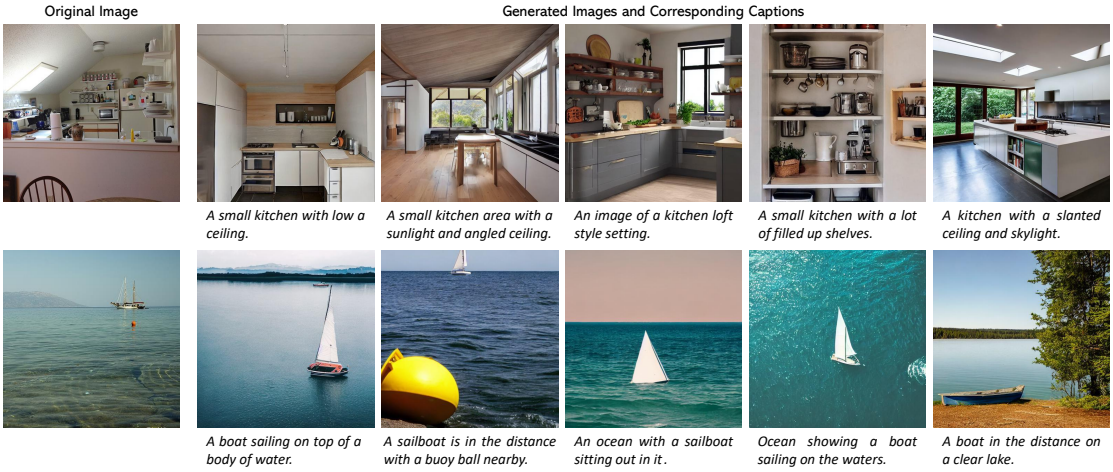

Figure 1: Sample images from the COCOFake dataset. The left-most column shows the original (real) image, while the remaining columns show fake images generated from each of the five COCO captions.

Generation of new data. In addition to identifying and selecting existing and relevant datasets, we will also generate new ones to tackle the specific objective of the challenge. A first result in this direction is the COCOFake dataset[2], which has been generated by UNIMORE leveraging the CINECA supercomputing facilities and which has been publicly released. The dataset consists of 650k images generated using Stable Diffusion v1, using textual prompts coming from the COCO dataset for image captioning (see Figure 1). As such, it contains clusters of five generated images sharing the same semantics and generated from five different textual prompts. In comparison with existing datasets for deepfake detection, it features more diversity, uniform coverage of semantic classes, and can easily be expanded to a larger scale.

Building on the basis of the COCO Fake dataset, we will extend it to support partially-generated images, multiple generators, and also increase the quantity of data by going beyond the prompts contained in the COCO dataset. The generation of the dataset will be carried out with a close collaboration between UNIMORE and Leonardo SpA and leveraging HPC facilities available at both partners. In particular, UNIMORE will exploit its connection with the NVIDIA AI Technical Centre of Modena, which provides computational support as well as access to HPC facilities at CINECA, the Italian Supercomputing Center.

In addition to generating full images, we will also tackle the aforementioned inpainting setting, in which only a portion of the image is altered, starting from a real picture. Also in this case we will generate tampered images using different generators and on a variety of semantic domains. In this case, in addition to releasing generated images, we will also release detection masks for a validation split of the dataset. Masks for the test split of the dataset will instead be kept private in order to avoid overfitting. Finally, the Use Case also plans to collect actual fake data from the web, to see how their quality evolves and how detectors perform over them.

3.2. Metrics

The benchmark aims at evaluating the progress of the community in the detection and handling of deepfakes. In terms of evaluation, in addition to automatic metrics, the challenge will consider off-line human evaluation, user studies, and will have a focus on evaluating the explainability of the proposed solutions. Overall, we plan to evaluate across two different settings:

- the recognition of deepfakes in which the entire image is generated at once. In terms of automatic evaluation, in this case the task can be considered as a binary classification one, and will be evaluated with standard accuracy (as main metric), logarithmic loss, AUROC, equal error rate (EER), Precision, Recall and F1-score. On the one hand, the accuracy value collects a general overview regarding how good the system is at discriminating between real and fake data. The Logarithmic loss, on the other hand, expresses how close the prediction probabilities are to the real data labels. By describing the model performance on a class-by-class basis, Precision and Recall can be used to assess the model competence in a real-world scenario, in which fake samples are unbalanced with respect to the real ones. F1-score is used to summarise Precision and Recall in a single

[2]https://github.com/aimagelab/ELSA_COCO-Fake

value, while AUROC is inserted to evaluate the behaviour of the model in terms of both FPR and TPR at different prediction thresholds. Focusing on error rates, EER allows identifying the point where the FPR and the FNR are equal;

- the recognition of partially altered or generated images. Again in terms of automatic evaluation, in this case the problem can be considered as a dense prediction task with the goal of predicting both fake and real areas. Performances will therefore be evaluated using the mIoU, that averages the intra class IoU (averaged among all instances of a class) in a single score. The metric will also be split across different semantic domains and different difficulty levels.

Performance of the approaches participating in the benchmark will be assessed in terms of recognition accuracy over both settings and by varying the generator. To measure the robustness of the approaches and their performance on out-of-distribution data, we also plan to evaluate on images belonging to separate semantic domains.

Beyond automatic evaluation metrics, off-line human evaluations and user studies will be carried out to compare the quality of deepfake detection tools with the human performance. Finally, we will add explainability metrics. This will help to measure the degree of explainability and trustworthiness of the developed fake detection approaches, and also ensure that they are correctly grounded with respect to the input data. In detail, we will employ the ADCC metric, a tool to evaluate XAI methods [10]. It is a single-valued score which measures the goodness of a saliency map, in terms of Coherency, Complexity, and Average Drop in confidence score. We plan to use it as a proxy to justify the explainability results, for both the two aforementioned different settings. In particular, we intend to use a normalised version of the ADCC for the partially-altered scenario, in which a score for each segmented area is required and evaluated.

3.3. The competition

The competition will run through the ELSA Benchmark Platform. Here, participants will get access to the datasets of the competition as well as the performances of baseline approaches. In addition to that, they will be able to participate in the benchmark, by submitting their own deepfake detection approach and have it evaluated on the Use Case datasets. The submission platform will be kept constantly open and will evaluate the results obtained by the participant on-the-fly, without scheduling specific submission or evaluation rounds.

The first step for users who want to submit their detector is to run their approach on the test splits of the datasets of the Use Case. Then, users can upload their predictions in a JSON-formatted text file, which will then be read by the platform to compute the final metrics and include the approach in the appropriate leaderboards. We will set up a separate competition track (and, therefore, leaderboard) for each generator and dataset release. If feasible, we will set appropriate submission limits (*e.g.*, number of submissions from the same participant per time) in order, again, to avoid parameter tuning on the test sets.

In addition to simply submitting prediction results, users will also be to submit their full pipeline including their source code and /or executables (e.g. through a Docker Container) and pre-trained models (*e.g.* using PyTorch, Tensorflow or ONNX formats) so to make their approach publicly available to the community. As previously mentioned, the datasets of this Use Case will constantly be evolving over time to keep them up-to-date with the evolution of deepfake generators. Methods submitted from participants who choose to release their source code will be automatically run over the different dataset releases, so to measure the evolution of their performance with respect to novel generators.

3.4. Baselines and Research Directions

All tracks of the use case competition will include state-of-the-art approaches for deepfake detection as baseline reference approaches. To this aim, we will also involve researchers from the community to disseminate the competition and propose the inclusion of relevant approaches from the literature, together with their source code and pipelines. To add further baselines which can be later employed for the development of novel approaches, we will also investigate the role of state-of-the-art multimodal features for deepfake classification.

To this aim, we are planning to test features coming from vision-only and vision-and-language backbones trained in supervised or contrastive settings, *e.g.* ResNets and ViT models pretrained on ImageNet, OpenAI WIT LAION-400M, and LAION-2B. These features, although being high-level features with a natural focus on semantics, are still capable of linearly separating generated and realistic data, at least with the current state of the art of the generators. This is shown qualitatively in Figure 2, where we train two linear projections on top of visual-semantic features trained à-la-CLIP. The first linear projection is trained to separate real and fake data (through a supervised contrastive objective), whilst the second is trained to preserve semantics (again, through a supervised contrastive objective where elements belonging to the same semantic cluster are attracted).

The content-style separation approach described above separates state-of-the-art features into two components, one related to low-level features that are useful for

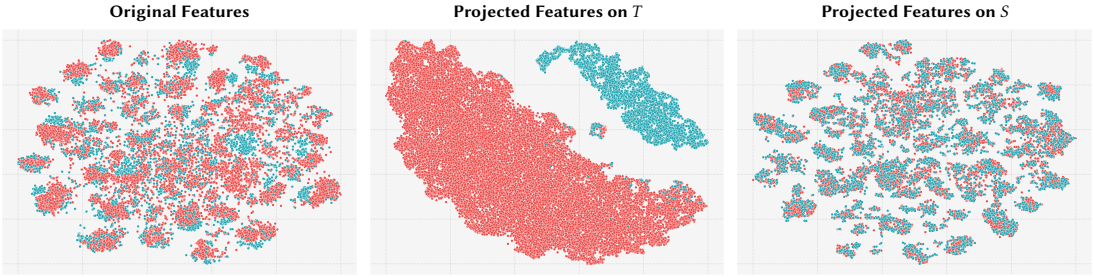

Figure 2: t-SNE visualizations over the validation set using the original visual features from the OpenCLIP ViT-B/32 LAION-2B backbone (left), the features projected on the T space (style) after disentanglement (middle), and the features projected on the S space (semantics) after disentanglement (right). Red dots indicate fake images, blue dots indicate real images.

verifying the authenticity of data and the other related to the semantics contained in the data itself. While the first component can be employed as a deepfake detector, the second one only focuses on semantics and therefore reflects the ideal case where a generator does not induce different low-level statistics with respect to real images. This will also be an interesting space to analyse, where deepfakes can be detected only on the basis of their semantics and ignoring perceptual cues.

This challenge will offer a unique opportunity to the community to increase the interest on deepfakes and the development of novel deepfake detection approaches. In addition to the research directions outlined above, we plan to develop deepfake detection approaches based on both low- and high-level features, and integrating multiple modalities, such as vision and text. This will also be helpful for the development of misinformation detection approaches, in which a textual verification of a fact is needed. Finally, an important research direction will be that of extending such approaches to the case of videos.

3.5. Networking Actions

The challenge on deepfake detection is directed to the Computer Vision, Machine Learning, Natural Language Processing and Multimedia communities. We plan to focus the networking actions primarily on the Computer Vision community and in international conferences like CVPR, ECCV and ICCV, which nevertheless have strong Multimedia components. The first networking action that has been performed is that of organizing workshops on deepfake detection at a major Computer Vision conference. Two workshops will be organized at ICCV 2023 and in ACM Multimedia 2023.

In addition, the same workshop will be proposed in other major conferences and in the following years, to create a line of networking opportunities for researchers interested in deepfake detection and build a community around the Use Case. Clearly, the workshops will also be an opportunity to showcase the results of the competition associated with the Use Case and promote further submissions from the community.

Acknowledgments

The research activities described in this work are partially supported by the Horizon Europe project "European Lighthouse on Safe and Secure AI (ELSA)" (HORIZON-CL4-2021-HUMAN-01-03), co-funded by the European Union.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Communications of the ACM 63 (2020).

[2] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784 (2014).

[3] P. Dhariwal, A. Nichol, Diffusion models beat gans on image synthesis, 2021.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: CVPR, 2022.

[5] R. Durall, M. Keuper, F.-J. Pfreundt, J. Keuper, Unmasking deepfakes with simple features, arXiv preprint arXiv:1911.00686 (2019).

[6] Y. Zhang, Z. Yin, Y. Li, G. Yin, J. Yan, J. Shao, Z. Liu, Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations, in: ECCV, 2020.

[7] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: Learning to detect manipulated facial images, in: ICCV, 2019.

[8] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-df: A large-scale challenging dataset for deepfake forensics, in: CVPR, 2020.

[9] Z. J. Wang, E. Montoya, D. Munechika, H. Yang, B. Hoover, D. H. Chau, Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models, arXiv preprint arXiv:2210.14896 (2022).

[10] S. Poppi, M. Cornia, L. Baraldi, R. Cucchiara, Revisiting the evaluation of class activation mapping for explainability: A novel metric and experimental analysis, in: CVPR Workshops, 2021.