PhD in Information Technology and Electrical Engineering
Università degli Studi di Napoli Federico II

PhD Student: Alessandro Pianese

Cycle: XXXVIII

Training and Research Activities Report

Year: First

Tutor: prof. Giovanni Poggi

Co-Tutor:

Date: December 11, 2023

Finanziato dall'Unione europea
NextGenerationEU

MUR
Ministero dell'Università e della Ricerca

Training and Research Activities Report

PhD in Information Technology and Electrical Engineering

Cycle: XXXVIII **Author: Alessandro Pianese**

1. Information:

- **PhD student: Alessandro Pianese**
- **DR number: DR996974**
- **Date of birth: 04/05/1996**
- **Master Science degree: Computer Science – Intelligent Systems and Visual Computing**
- **University: University of Groningen**
- **Doctoral Cycle: XXXVIII**
- **Scholarship type: *PNRR CN1 - HPC***
- **Tutor: Giovanni Poggi**
- **Co-tutor:**

2. Study and training activities:

Activity	Type[1]	Hours	Credits	Dates	Organizer	Certificate[2]
Using Deep Learning Properly	Course	10	4	10.01.23-24.01.23	DIETI ITEE PhD	Y
How to Boost your PhD	Course	16	4	11.01.23-01.03.23	DIETI ITEE - ICTH - CQB PhD programs	Y
Visione per Sistemi Robotici	Course	72	9	07.03.23-09.06.23	University of Naples, Federico II	Y
Summer School on Metaverse Technologies	Doctoral School	50	5	18.09.23-22.09.23	IEEE SPS, University of Cagliari	Y
Advances on Multimodal Machine Learning Solutions for Speech Processing Tasks and Emotion Recognition	Seminar	1	0.2	19.01.23	ITEE SPS	Y
The Super Neuron Model – A new generation of ANN-based Machine Learning and Applications.	Seminar	1	0.2	09.02.23	EURASIP JIVP	Y
Human Centric Visual Analysis - Hand, Gesture, Pose, Action, and Beyond	Seminar	1	0.2	13.02.23	IEEE SPS	Y
What's Up with Image and Video Forensics?	Seminar	1	0.2	02.03.23	EURASIP	Y

Unleashing the Power of LLMs: a Historical perspective on Generative AI	Seminar	1	0.2	02.03.23	Dr. Stefano Marrone	Y
Statistical Multimedia Security and Forensics	Seminar	20	4	08.05.23-12.05.23	University of Trento	Y
Computational Disinformation Symposium	Seminar	7	1.4	06.06.23	NYU Tandon School of Engineering	Y
Prompting in Vision	Seminar	3	0.2	19.06.23	CVPR 2023	Y
Scientific Integrity Verification through Image Forensics	Seminar	1	0.2	06.07.23	IEEE SPS	Y
Progressive JPEGs in the Wild: Implications for Information Hiding and Forensics	Seminar	1	0.2	09/11/2023	Magdeburg University	Y
Study of state of the art audio/video deepfake detection models Running experiments of said models Attendance to weekly technical meetings	Research		5.2	01.01.23-28.02.23		**N**
Studying audio generation through diffusion models Studying state of the art of audio deepfake detection Generation of synthetic audios using diffusion models Analysis of general deepfake audio characteristics	Research		5.6	01.03.23-30.04.23		**N**
Participation to Computer Vision and Pattern Recognition Conference (CVPR) 2023. Date: 18/06/23 - 22/06/2023 Attendance to weekly	Research		3	01.05.23-30.06.23		**N**

technical meetings. Experiments regarding the improvements of the method described in workshop paper “Audio-Visual Person-of-Interest Deepfake Detection”						
Study of state of the art audio/video deepfake detection models Research of new datasets that can be employed Trained several audio deepfake detection models on different lengths of audio Tested Teacher/Student training on said models Attendance to weekly technical meetings	Research		6.5	01.07.23-31.08.23		**N**
Investigated the use of attention transformers for deepfake detection Execution of experiments related to the point above Benchmarked the POI method for a demo Investigated SOTA 3DMM features extractors	Research		5	01.09.23-31.10.23		**N**
Studying text-prompting for audio classification Studying large pretrained models Experimenting for audio deepfake detection with state of the art models	Research		8	01.11.23-31.12.23		**N**

1) Courses, Seminar, Doctoral School, Research, Tutorship
2) Choose: Y or N

2.1. Study and training activities - credits earned

	Courses	Seminars	Research	Tutorship	Total
Bimonth 1	4	0.8	5.2	-	10
Bimonth 2	4	0.4	5.6	-	10
Bimonth 3	9	4	3	-	16
Bimonth 4	-	0.2	6.5	-	6.8
Bimonth 5	-	-	5	-	5
Bimonth 6	-	0.2	8	-	8.2
Total	17	5.6	33.3	-	56
Expected	30 - 70	10 - 30	80 - 140	0 – 4.8	

3. Research activity:

Synthetic media generation has become a key technology in many industrial applications, from film production to the video game industry. Facial manipulations, however, also pose a serious and growing threat to our society, of which financial fraud and disinformation campaigns are just a few examples. With the advancement of such technology, there is a steady increase in the level of photorealism, as more and more methods of video manipulation emerge. In particular, the term deepfake, which is often associated with face-swapping, has now become associated with negative implications.

Currently, the deepfake term has taken on an even broader meaning, including a variety of possible video manipulations: speech can be synthesized in anyone’s voice, face expression can be modified, the identity of a person can be swapped with another, even altering what they are saying. Dealing with such a large spectrum of manipulations is the main challenge for current video deepfake detectors. In fact, it is particularly difficult to develop a method that can detect multiple known manipulation methods at the same time; this is only exacerbated when targeting unknown methods that were not part of any training samples.

As a result, current SOTA detectors, trained over large datasets of deepfake and pristine videos, often show an unsatisfactory cross-dataset performance, clearly highlighting the limitations of the supervised deep learning based approach. In addition, performance often drops dramatically under a more challenging scenario, such as low-quality videos. Such conditions, however, are commonplace in the real world where videos are mostly disseminated through social networks where they are further compressed with the loss of relevant audio-visual information. It is also worth pointing out that deep learning-based methods are vulnerable to adversarial attacks with detection performance that degrades sharply even in a black-box scenario [1, 2, 3].

A possible solution to gain generalization and robustness is to shift to a completely different paradigm, training models only on real videos, with the goal to detect manipulated videos based on their anomalous behavior [4]. This approach turns out to be particularly effective if the characterization of pristine faces is based on semantic features, such as soft biometrics, leading to Person-of-Interest (POI) based detection [5, 6, 7, 8, 9, 10]. First papers on this topic exploited specific face and head movement patterns [5, 6, 8], inconsistencies between mouth shape dynamics and spoken phonemes [11] or cues related to the specific words uttered by the identity [7] or inconsistencies between inner and outer face regions [10]. These methods present some limitations: in particular, they rely on video-only features, sometimes complemented by categorical information, neglecting precious audio information. Moreover, they often need several hours of videos of the identity under test.

During this year, we developed and proposed a new person-of-interest (POI) deepfake detector called POI-Forensics. The key feature of our approach is the use of a multi-modal analysis. More specifically we rely on a contrastive learning approach and train an audio and video network so that the learned representation characterizes temporal segments of the same identity close to one-another but far from each-other for different identities. At test time, we compute similarity indices between the features extracted from the video under analysis and those extracted by a set of POI-based reference videos. We also include joint audio-video similarity indices that have been shown to improve the discrimination ability of the detector.

Overall the proposed method ensures a number of important benefits:

- **Generalization**. Since training does not rely on fake videos, our detector works equally well on known and unknown manipulations. This ensures good generalization to attacks not seen during training (as none are seen during training). The detector can deal with any manipulation method (face swapping, facial reenactment or anything else) with performance that depends only on the fidelity of the video, not on specific inconsistencies or artifacts of the manipulation method.
- **Flexibility**. Due to our multi-modal approach, we can detect also video-only and speech-only manipulations, and even the swapping of a real audio track on the real video of the original identity. When the manipulation involves video and audio jointly, the joint-modality analysis improves performance.
- **Robustness**. The detector is robust to many challenging conditions frequently encountered in real scenarios, where videos are compressed or even maliciously attacked.
- **No need of re-training**. Training does not require videos of the POI under test, hence re-training is not needed when testing new identities, and only a few short POI videos (around 10 minutes) are necessary at test time.

In addition to this, we are looking into increasing our efforts for the audio-only segment of our experiments. Recently, the importance of audio deepfake detection techniques has increased due to the proliferation of audio spoofing methods and the widespread diffusion of voice spoofed tracks in the current year. At the same time, the deep learning community is leaning towards the idea that huge pre-trained models are good at generalizing to diverse tasks and we are seeing the increase in complexity for audio feature extraction models: the well-established Wav2Vec 2.0 [12] offers weights with up to 2 billion parameters. This trends has the potential of working alongside the POI focus we mentioned earlier due to the common characteristics of training only on real data, robustness and the avoidance of re-training.

References:

[1] D. Cozzolino, J. Thies, A. Rossler, M. Nießner, and L. Verdoliva. SpoC: Spoofing camera fingerprints. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2021.

[2] S. Hussain, P. Neekhara, M. Jere, F. Koushanfar, and J. McAuley. Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. In IEEE Winter conference on Applications of Computer Vision (WACV), 2021.

[3] P. Neekhara, B. Dolhansky, J. Bitton, and Cristian Canton Ferrer. Adversarial threats to deepfake detection: A practical perspective. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[4] D. Cozzolino, G. Poggi, and L. Verdoliva. Extracting camera-based fingerprints for video forensics. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019

[5] S. Agarwal, H. Farid, T. El-Gaaly, and S. Lim. Detecting deep-fake videos from appearance and behavior. In IEEE international workshop on information forensics and security (WIFS), 2020.

[6] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li. Protecting world leaders against deep fakes. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019.

[7] S. Agarwal, L. Hu, E. Ng, T. Darrell, H. Li, and A. Rohrbach. Watch those words: Video falsification detection using wordconditioned facial motion. In IEEE Winter conference on Applications of Computer Vision (WACV), 2023.

[8] M. Bohácek and H. Farid. Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. In Proceedings of the National Academy of Sciences, 2022.

[9] D. Cozzolino, A. Rossler, J. Thies, M. Nießner, and L. Verdoliva. ID-Reveal: Identity-aware DeepFake Video Detection. In IEEE International Conference on Computer Vision (ICCV), 2021.

[10] X. Dong, J. Bao, D. Chen, T. Zhang, W. Zhang, N. Yu, D. Chen, F. Wen, and B. Guo. Protecting celebrities from deepfake with identity consistency transformer. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[11] S. Agarwal, H. Farid, O. Fried, and M. Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2020.

[12] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In Advances in neural information processing systems, 2020.

4. Research products:

Conference Paper:

[P1] Cozzolino, D., **Pianese, A**., Nießner, M., & Verdoliva, L. (2023). Audio-visual person-of-interest deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

5. Conferences and seminars attended

IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)

- Dates: 18.06.2024 - 22.06.2023
- Place: Vancouver, British Columbia, Canada
- Co-Author of Paper "Audio-Visual Person-of-Interest Deepfake Detection" published in the Workshop on Media Forensics (**WMF**)

6. Activity abroad:

None

7. Tutorship

None