APPELS À SUJETS DOCTORANTS ET POST-DOCTORANTS 3IA 2020

- Title of the proposed topic (en anglais): *Robust Detection of DeepFakes*
- PhD or post-doc: PhD and a Post-doc
- Research axis of the 3IA: l'IA et les territoires intelligents et sécurisés
- Supervisor (name, affiliation, email): Antitza Dantcheva, INRIA Sophia Antipolis, antitza.dantcheva@inria.fr
- Potential co-supervisor (name, affiliation): Francois Bremond, INRIA Sophia Antipolis, francois.bremond@inria.fr
- The laboratory and/or research group: STARS Team of Inria
- The description of the topic:

Manipulated images and videos, i.e., deepfakes have become increasingly realistic due to the tremendous progress of deep convolutional neural networks (CNNs). While forgery was associated with a slow, painstaking process usually reserved for experts, deep learning and related *manipulation-technologies* are streamlined to reduce costs, time and skill needed to doctor images and videos.
The *manipulation scenario* of interest in this proposal has to do with a face image of a *target person* being superimposed to a video of a *source person*, widely accepted and referred to as *deepfake*.

While technically intriguing, such progress raises a number of social concerns. In particular, such manipulations can fabricate animations of subjects involved in actions that have not taken place and such manipulated data can be circumvented nowadays rapidly via social media. Hence, we cannot trust anymore, what we see or hear on video, as deepfakes betray sight and sound, the two predominantly trusted human innate senses, posing a threat of distorting what is perceived as reality. To further fuel concern, deepfake techniques have become open to the public via phone applications such as FaceApp[1] and ZAO[2].
We differentiate two cases of concern: the first one has to do with deepfakes being perceived as real, and the second relates to real videos being misdetected for fake, the latter referred to as *liar's dividend*. Such concerns

[1] https://apps.apple.com/gb/app/faceapp-ai-face-editor/id1180884341
[2] https://apps.apple.come/cn/app/id146519927

necessitate the introduction of robust and reliable methods for fake image and video detection.

Motivated by the above, we propose research which studies detection of manipulated videos.

We note that the detection of deepfakes is challenging for several reasons: (i) it evolves a “cat-and-mouse-game” between the adversary and the system designer, (ii) deep models are highly domain-specific and likely yield big performance degradation in cross-domain deployments, especially with large train-test domain gap. The challenge (i) has to do with the fact that by improving the deepfake-detection mechanism, generative-networks can be improved accordingly, which in turn can be beneficial in improving again the detector. This results in this game, which can never be won. The challenge (ii) indicates that detectors trained on known manipulation techniques generalize poorly to tampering methods outside of the training set, which we show in a very recent work [1].
Considering this, we intend to investigate three strategies of detection by designing algorithms that can successfully generalize onto unknown manipulations.

(a) **3D CNNs.** Our intuition is that current state of the art forgery detection techniques omit a pertinent clue, namely temporal information by investigating only spatial information [7][8][9][10][11][12]. In our extensive work on video generation [2][3][4][5][6], we have found out that generative models have exhibited difficulties in *preserving appearance* throughout generated videos, as well as *motion consistency* and we intend to exploit these factors in deepfake detection.

(b) **Few-shot learning.** We will focus on learning *talking-patterns* pertaining to a set of enrolled subjects. We intend to classify the integrity of videos by analyzing the likelihood that a portrayed talking-behavior to belong to an enrolled person. By few-shot learning we intend to generalize this method onto unseen subjects.

(c) **Generated noise.** Images and videos acquired by sensors incorporate a *noise-pattern* specific to the sensor (caused by the physical properties of the sensor). Hence, such noise-pattern is absent in generated data. In this context, we intend to study (a) whether such noise-pattern is instrumental in classifying deepfakes, (b) whether such noise pattern can be added onto generated data and hence renders videos not detectable, see (a), as well as (c) whether generative

adversarial networks incorporate a generative pixel-level noise into generated data.

References

[1] Wang, Yaohui; Dantcheva, Antitza
A video is worth more than 1000 lies. Comparing 3DCNN approaches for detecting deepfakes
FG'20, 15th IEEE International Conference on Automatic Face and Gesture Recognition, May 18-22, 2020, Buenos Aires, Argentina.

[2] Wang, Yaohui; Bilinski, Piotr; Bremond, Francois; Dantcheva, Antitza
ImaGINator: conditional spatio-temporal GAN for video generation
WACV'20, Winter Conference on Applications of Computer Vision, March 2-5, 2020, Aspen, USA.

[3] Wang, Yaohui; Bilinski, Piotr; Bremond, Francois; Dantcheva, Antitza
G^3AN: This video does not exist. Disentangling motion and appearance for video generation
arXiv preprint arXiv:1912.05523, 2019.

[4] Wang, Yaohui; Dantcheva, Antitza; Bremond, Francois
From attribute-labels to faces: face generation using a conditional generative adversarial network
ECCVW'18, 5th Women in Computer Vision (WiCV) Workshop in conjunction with the European Conference on Computer Vision, September 9, 2018, Munich, Germany.

[5] Wang, Yaohui; Dantcheva, Antitza; Broutart, Jean Claude; Robert, Philippe; Bremond, Francois; Bilinski, Piotr
Comparing methods for assessment of facial dynamics in patients with major neurocognitive disorders
ECCVW'18, 6th International Workshop on Assistive Computer Vision and Robotics (ACVR) in conjunction with the European Conference on Computer Vision, September 9, 2018, Munich, Germany.

[6] Wang, Yaohui; Dantcheva, Antitza; Bremond, Francois
From attributes to faces: a conditional generative adversarial network for face generation
BIOSIG'18, 17th International Conference of the Biometrics Special Interest Group, September 26-28, 2018, Darmstadt, Germany.

[7] J. Fridrich and J. Kodovsky,
Rich models for steganalysis of digitalimages
IEEE Transactions on Information Forensics and Security,vol. 7, no. 3, pp. 868–882, 2012.

[8] A. Roessler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, andM. Nießner
Faceforensics++: Learning to detect manipulated facial images
arXiv preprint arXiv:1901.08971, 2019.

[9] D. Cozzolino, G. Poggi, and L. Verdoliva
Recasting residual-basedlocal descriptors as convolutional neural networks: an application toimage forgery detection
in Proceedings of the 5th ACM Workshopon Information Hiding and Multimedia Security.ACM, 2017, pp.159–164.

[10] B. Bayar and M. C. Stamm
A deep learning approach to universalimage manipulation detection using a new convolutional layer
In Proceedings of the 4th ACM Workshop on Information Hiding andMultimedia Security. ACM, 2016, pp. 5–10.

[11] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen
Distinguishing computer graphics from natural images using convolution neural networks
In 2017 IEEE Workshop on Information Forensics and Security (WIFS). IEEE, 2017, pp. 1–6.

[12] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen

Mesonet: a compact facial video forgery detection network
In 2018 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2018, pp. 1–7.