AIMH Lab for Cybersecurity

Claudio Vairo, Davide Alessandro Coccomini, Fabrizio Falchi, Claudio Gennaro, Fabio Valerio Massoli, Nicola Messina, Giuseppe Amato

Artificial Intelligence for Media and Humanities laboratory
Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", CNR
<name.surname>@isti.cnr.it

Abstract

In this short paper, we report the activities of the Artificial Intelligence for Media and Humanities (AIMH) laboratory of the ISTI-CNR related to Cybersecurity. We discuss about our active research fields, their applications and challenges. We focus on face recognition and detection of adversarial examples and deep fakes. We also present our activities on the detection of persuasion techniques combining image and text analysis.

1 Introduction

The massive growth of recent AI technologies (like Deep Learning, Convolutional Neural Networks), has led to significant improvements in the development of solutions and applications based on multimedia analysis, like face recognition. However, this results in new and dangerous threats to Cybersecurity, for the intrinsic characteristics of these new technologies. For example, machine learning models, including deep learning methods, are highly vulnerable to adversarial examples which can lead to misclassification with high confidence of the attacked model by simply applying a small intentional perturbation in the input. In most cases, the difference between the original and perturbed image is imperceptible to a human observer. Another severe and increasingly growing problem raised by the advent of these new AI technologies is the creation of deep fakes, for example, images, videos, or news. These deep fakes are almost impossible to be detected or distinguished by humans and can lead to misinformation or attacks on people. Also, the persuasion of ideas, thoughts, or political beliefs through social networks can be manipulated in a way to alter other people's judgment ability. This can lead to severe consequences for society.

Lots of research is being conducted in order to address these issues and to better exploit all the benefits that these new technologies can bring. In this paper, we present some of the research work done in the last few years by the Artificial Intelligence for Media and Humanities (AIMH) laboratory of the ISTI-CNR in Pisa related to Cybersecurity. In particular, in Section 2.1 we describe the activities carried out in the field of Face Recognition applied to security and surveillance; in Sections 2.2, 2.3, and 2.4 we present the work done for the detection of, respectively, adversarial attacks, deep fakes, and persuasion techniques in social networks; Section 3 briefly presents some of the research projects where we applied the solution described; finally, in Section 4, we discuss some of the challenges that still are open in these fields and worth additional research activities.

2 Research Themes

2.1 Face Recognition

Face recognition is an important task in security and surveillance. With the advent of deep learning-based methods, face recognition algorithms have become more effective and efficient. We have studied techniques to perform face recognition in different application scenarios and contexts. In particular, we investigate the issues of implementing a smart surveillance system for buildings by using embedded devices as smart cameras to perform face recognition [Amato *et al.*, 2018a; Kavalionak *et al.*, 2019; Barsocchi *et al.*, 2018]. Most of the current commercial video surveillance systems rely on a classical client/server architecture to perform face and object recognition. In order to support the more complex and advanced video surveillance systems proposed in the last years, companies are required to invest resources to maintain the servers dedicated to the recognition tasks. We propose a novel distributed protocol for a face recognition system that exploits the computational capabilities of the surveillance devices (i.e. smart cameras) to perform the recognition of the person.

We also implemented an intrusion detection system for embedded devices that is based on facial recognition [Amato *et al.*, 2018b]. Our system is composed of smart cameras (i.e. video cameras capable of processing and analyzing the acquired data) to monitor the access to restricted areas of a building, like office rooms, by using a facial recognition algorithm implemented with a Deep Learning approach. Face recognition is performed using a knn classifier on features extracted from a 50-layers Residual Network (ResNet-50) trained on the VGGFace2 dataset. Our solution is aimed at embedded platforms that do not exploit the computational power of the GPUs. In particular, we deployed our system on a Raspberry Pi. An example of an intrusion detection output is reported in Figure 1.

We also conducted some studies in the forensics field. In particular, we studied how deep learning techniques can be

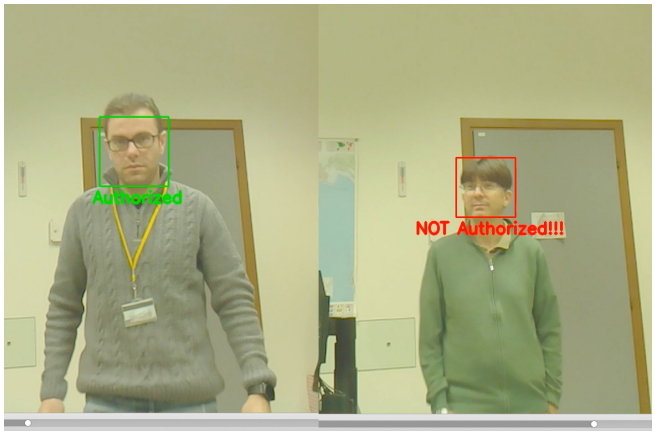

Figure 1: Example of intrusion detection implemented on a Raspberry Pi.

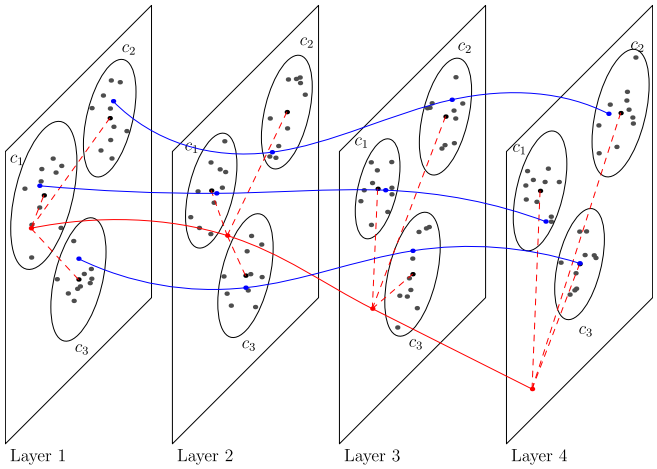

Figure 2: Example of the evolution of features while traversing the network that illustrates our hypothesis. Each plane represents a feature space defined by the activations of a particular layer of the deep neural network. Circles on the features space represent clusters of features belonging to a specific class. Blue trajectories represent authentic inputs belonging to three different classes, and the red trajectory represent an adversarial input. We rely on the distances in the feature space (red dashed lines) between the input and some reference points representatives of the classes to encode the evolution of the activations. Image courtesy of [Carrara *et al.*, 2018].

applied to the face verification problem to be used in forensics [Amato *et al.*, 2019] and we compared the accuracy performance of face verification implemented with deep learning techniques and by using the distance of facial landmarks [Amato *et al.*, 2018c]. In fact, the last one is the method currently used in trials as proof, but its recognition accuracy is quite low. Deep learning-based face verification approach, on the other hand, achieves very high accuracy performance, but its usage as proof in trials has still some concerns. This is still an open problem and it is worth further investigation.

More recently, we started investigating the problem of multi-scale [Massoli *et al.*, 2019] and multi-resolution [Amato *et al.*, 2020] in face recognition. In fact, high-resolution images are usually used for training CNN models, and for this reason, their discrimination ability is usually degraded when they are tested against low-resolution images. Thus, Low-Resolution Face Recognition remains an open challenge for deep learning models. Such a scenario is of particular interest for surveillance systems in which it usually happens that a low-resolution probe has to be matched with higher resolution images. We studied this problem in the context of images acquired by surveillance drones in [Amato *et al.*, 2020; Ferro *et al.*, 2020], where weather conditions, link wind, pose a severe limit on image stability and the distance the drones fly is typically higher than ground cameras. This translates into a degraded resolution of the face images. Multi-resolution is crucial also in the context of adversarial attacks [Massoli *et al.*, 2020], as we will see in Section 2.2.

2.2 Adversarial Examples

Adversarial attacks pose great challenges to deep learning models. In fact, it is well known that deep learning methods can be easily fooled by adversarial examples. This kind of attack is particularly harmful in safety-critical scenarios — for example, self driving — where the vision system must be robust to ad-hoc crafted external perturbations. An adversarial example is a malicious input typically created applying a small but intentional perturbation, such that the attacked model misclassifies it with high confidence. We have been active in detecting adversarial examples in the context of image classification. In particular, in [Carrara *et al.*, 2017; Carrara *et al.*, 2019b; Caldelli *et al.*, 2019] we analyzed the hidden layers activation of CNNs to spot adversarial examples. This method relies on the assumption that layer activations lay on a different feature subspace when the CNNs are fed with adversarial examples. Going a step further, in [Carrara *et al.*, 2018] we argued that the representations of adversarial inputs follow a different evolution, or *trajectory*, with respect to genuine inputs, and we defined a distance-based embedding of features to efficiently encode this information (see Figure 2). We trained an LSTM network that analyzes the sequence of deep features embedded in a distance space to detect adversarial examples.

We also investigated the robustness of recent ODE networks [Carrara *et al.*, 2021; Carrara *et al.*, 2019a]. ODE networks define a continuous hidden state that can be formalized using parametric ordinary differential equations. In particular, we show that Neural ODE are natively more robust to adversarial attacks with respect to state-of-the-art residual networks, and some of their intrinsic properties, such as adaptive computation cost, open new directions to further increase the robustness of deep-learned models.

Given the experience we already discussed related to face recognition and cross-resolution in particular, we developed specific approach for adversarial faces [Massoli *et al.*, 2021] and cross-resolution face recognition adversarial tacks [Massoli *et al.*, 2020].

2.3 Deep Fake Detection

Deep fake detection is a critical task in the modern society, where increasingly powerful generative methods are used to craft fake images, videos, or fake news through *social bots*.

All this ad-hoc generated content is spread on the web usually via social networks, and it is used to propagate misinformation and fake news, with the aim of contaminating public debate. Deep fake images and videos can be used to harm important and strategic public figures. For this reason, it is very important to promptly detect them to stop their diffusion. Although many methods focused on image deep fake detection, in [Coccomini *et al.*, 2021] we tackled deep fake detection in videos. The challenge is identifying if there are people having their face replaced or manipulated. In particular, we used a mixed Transformer-Convolutional model to attend the face patches. Differently from current state-of-the-art approaches, we use neither distillation nor ensemble methods, and we obtained remarkable results on the DeepFake Detection Challenge (DFDC) and on FaceForensics++ datasets. In addition to proposing new hybrid architectures to deal with deepfake video detection, alternative approaches to efficient and effective inference were analysed in this study. Indeed, at inference time, the faces from different frames of the video are independently analyzed, grouped and a simple voting algorithm is used to decide if the video shot was altered or not. With the proposed approach it is possible to better manage situations such as the presence of several people in the same video where only one has been manipulated so as to counter false negatives or attacks aimed at deceiving the detector. A Video Deepfake Detector could also be used on a large scale and therefore a short study was also carried out to identify the optimal number of faces to be classified within a video to achieve the best ratio of reliability and scalability of classification.

We have been also involved in research related to detection of deep fake tweets [Fagni *et al.*, 2021]. Despite the critical importance, few works tackled the detection of machine-generated texts on social networks like Twitter or Facebook. With the aim of helping the research in this detection field, in this work we collected the first dataset of real deepfake tweets, TweepFake. It is real in the sense that each deepfake tweet was actually posted on Twitter by social bots. With the aim of showing the challenges that TweepFake poses and providing a solid baseline of detection techniques, we also evaluated 13 different deepfake text detection methods. Some of the detectors exploit text representations as inputs to machine-learning classifiers, others are based on deep learning networks, and others rely on the fine-tuning of transformer-based classifiers. A comprehensive analysis of these techniques showed how the newest and more sophisticated generative methods based on the transformer architecture (e.g., GPT-2) can produce high-quality short texts, difficult to unmask also for expert human annotators. Additionally, the transformer-based language models provide very good word representations for both text representation-based and fine-tuning based detection techniques.

2.4 Detection of Persuasion Techniques

Social networks play a critical role in our society. Nowadays, most of the ideas, thoughts, and political beliefs are shared through the internet using social platforms like Twitter, Facebook, or Instagram. Although these online services enable information to be spread efficiently and effectively, it is non-

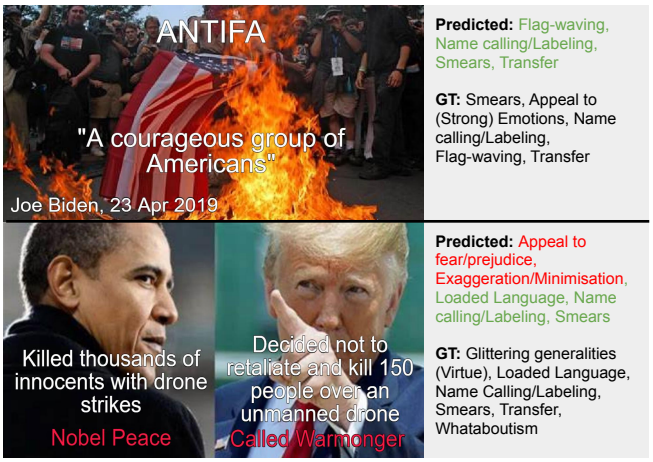

Figure 3: Predictions from the DVTT model for the detection of persuasion techniques in memes from social networks. In green, the true positives labels; in red, the false positives labels. Image courtesy of [Messina *et al.*, 2021].

trivial to understand if the shared contents are free of subtle meanings altering people's judgment abilities.

In [Messina *et al.*, 2021] we tackle the problem of recognizing which kind of disinformation technique is used to forge *memes* for a disinformation campaign. Memes are small yet effective units of information able to spread cultural ideas, symbols, or practices and usually exist under the form of pictures, possibly with overlaid text. Memes are created so that they can propagate rapidly and reach a large number of users; for this reason, they are one of the most popular types of content used in an online disinformation campaign In particular, we proposed an architecture based on the well-established Transformer architecture model [Vaswani *et al.*, 2017] for processing both the textual and visual inputs from the meme. This architecture, called DVTT (Double Visual Textual Transformer), comprises two full Transformer networks working respectively on images and texts; each of these Transformers is conditioned on the other modality. We consider this task as a multi-label classification problem, where text and/or images from the meme are processed, and probabilities of presence of each possible persuasion technique are returned as a result. Our proposed model reached remarkable results on the publicly available leaderboard of the *SemEval 2021 Task 6* challenge[1]. Two output examples from our network are reported in Figure 3.

3 Projects

AI4Media

A Centre of Excellence delivering next generation AI Research and Training at the service of Media, Society and Democracy. The project has a specific task on "Manipulation and synthetic content detection in multimedia".

AI4CHSites

Artificial Intelligence for monitoring Cultural Heritage Sites. Funded by the Tuscany Region, has CNR, INERA and Opera

[1]https://propaganda.math.unipd.it/semeval2021task6/index.html

della Primaziale Pisana has partners. Prototypes are tested on the Square of Miracles in Pisa including the Leaning Tower.

4 Challenges

Concerning the detection of persuasion techniques in social media contents, there are many interesting research directions opened. For example, there are cases where it is probably necessary to access more contextual information to detect the more subtle persuasions. In order to solve this issue, it would be necessary to access external data to effectively reason on the complex common sense and historical facts hidden behind the most complex and deep memes. For this reason, it would be interesting to leverage the attention mechanisms of the Transformer to integrate the data with a knowledge base of historical facts to create a more suitable context.

References

[Amato *et al.*, 2018a] G. Amato, P. Barsocchi, F. Falchi, E. Ferro, C. Gennaro, G. R. Leone, D. Moroni, O. Salvetti, e C. Vairo. Towards multimodal surveillance for smart building security. In *Multidisciplinary Digital Publishing Institute Proceedings*, volume 2, page 95, 2018.

[Amato *et al.*, 2018b] G. Amato, F. Carrara, F. Falchi, C. Gennaro, e C. Vairo. Facial-based intrusion detection system with deep learning in embedded devices. In *Proceedings of the 2018 International Conference on Sensors, Signal and Image Processing*, pages 64–68, 2018.

[Amato *et al.*, 2018c] G. Amato, F. Falchi, C. Gennaro, e C. Vairo. A comparison of face verification with facial landmarks and deep features. In *10th Intl. Conference on Advances in Multimedia (MMEDIA)*, pages 1–6, 2018.

[Amato *et al.*, 2019] G. Amato, F. Falchi, C. Gennaro, F. V. Massoli, N. Passalis, A. Tefas, A. Trivilini, e C. Vairo. Face verification and recognition for digital forensics and information security. In *7th Intl Symposium on Digital Forensics and Security (ISDFS)*, pages 1–6. IEEE, 2019.

[Amato *et al.*, 2020] G. Amato, F. Falchi, C. Gennaro, F. V. Massoli, e C. Vairo. Multi-resolution face recognition with drones. In *2020 3rd International Conference on Sensors, Signal and Image Processing*, pages 13–18, 2020.

[Barsocchi *et al.*, 2018] P. Barsocchi, A. Calabrò, E. Ferro, C. Gennaro, E. Marchetti, e C. Vairo. Boosting a low-cost smart home environment with usage and access control rules. *Sensors*, 18(6):1886, 2018.

[Caldelli *et al.*, 2019] R. Caldelli, R. Becarelli, F. Carrara, F. Falchi, e G. Amato. Exploiting cnn layer activations to improve adversarial image classification. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2289–2293. IEEE, 2019.

[Carrara *et al.*, 2017] F. Carrara, F. Falchi, R. Caldelli, G. Amato, R. Fumarola, e R. Becarelli. Detecting adversarial example attacks to deep neural networks. In *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*, pages 1–7, 2017.

[Carrara *et al.*, 2018] F. Carrara, R. Becarelli, R. Caldelli, F. Falchi, e G. Amato. Adversarial examples detection in features distance spaces. In *European Conference on Computer Vision (ECCV) Workshops*, 2018.

[Carrara *et al.*, 2019a] F. Carrara, R. Caldelli, F. Falchi, e G. Amato. On the robustness to adversarial examples of neural ode image classifiers. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2019.

[Carrara *et al.*, 2019b] F. Carrara, F. Falchi, R. Caldelli, G. Amato, e R. Becarelli. Adversarial image detection in deep neural networks. *Multimedia Tools and Applications*, 78(3):2815–2835, 2019.

[Carrara *et al.*, 2021] F. Carrara, R. Caldelli, F. Falchi, e G. Amato. Defending neural ode image classifiers from adversarial attacks with tolerance randomization. In *International Conference on Pattern Recognition*, pages 425–438. Springer, 2021.

[Coccomini *et al.*, 2021] D. Coccomini, N. Messina, C. Gennaro, e F. Falchi. Combining efficientnet and vision transformers for video deepfake detection. *arXiv preprint arXiv:2107.02612*, 2021.

[Fagni *et al.*, 2021] T. Fagni, F. Falchi, M. Gambini, A. Martella, e M. Tesconi. Tweepfake: About detecting deepfake tweets. *Plos one*, 16(5):e0251415, 2021.

[Ferro *et al.*, 2020] E. Ferro, C. Gennaro, A. Nordio, F. Paonessa, C. Vairo, G. Virone, A. Argentieri, A. Berton, e A. Bragagnini. 5g-enabled security scenarios for unmanned aircraft: Experimentation in urban environment. *Drones*, 4(2):22, 2020.

[Kavalionak *et al.*, 2019] H. Kavalionak, C. Gennaro, G. Amato, C. Vairo, C. Perciante, C. Meghini, e F. Falchi. Distributed video surveillance using smart cameras. *Journal of Grid Computing*, 17(1):59–77, 2019.

[Massoli *et al.*, 2019] F. V. Massoli, G. Amato, F. Falchi, C. Gennaro, e C. Vairo. Improving multi-scale face recognition using vggface2. In *Intl Conference on Image Analysis and Processing*, pages 21–29. Springer, 2019.

[Massoli *et al.*, 2020] F. V. Massoli, F. Falchi, e G. Amato. Cross-resolution face recognition adversarial attacks. *Pattern Recognition Letters*, 140:222–229, 2020.

[Massoli *et al.*, 2021] F. V. Massoli, F. Carrara, G. Amato, e F. Falchi. Detection of face recognition adversarial attacks. *Computer Vision and Image Understanding*, 202:103103, 2021.

[Messina *et al.*, 2021] N. Messina, F. Falchi, C. Gennaro, e G. Amato. Aimh at semeval-2021 task 6: multimodal classification using an ensemble of transformer models. In *15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1020–1026, 2021.

[Vaswani *et al.*, 2017] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, e I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.