AIMH Lab for Trustworthy AI

Nicola Messina, Fabio Carrara, Davide Alessandro Coccomini, Fabrizio Falchi, Claudio Gennaro, 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 Trustworthy AI. Artificial Intelligence is becoming more and more pervasive in our society, controlling recommendation systems in social platforms as well as safety-critical systems like autonomous vehicles. In order to be safe and trustworthy, these systems require to be easily interpretable and transparent. On the other hand, it is important to spot fake examples forged by malicious AI generative models to fool humans (through *fake news* or *deepfakes*) or other AI systems (through *adversarial examples*). This is required to enforce an ethical use of these powerful new technologies. Driven by these concerns, this paper presents three crucial research directions contributing to the study and the development of techniques for reliable, resilient, and explainable deep learning methods. Namely, we report the laboratory activities on the detection of adversarial examples, the use of attentive models as a way towards explainable deep learning, and the detection of deepfakes in social platforms.

1 Introduction

In a society increasingly permeated by algorithms and data-driven methods, there is a growing interest in making AI systems robust and reliable. The astonishing growth of AI, and Deep Learning in particular, enabled the deployment of many AI-powered applications in the real world, defining new business models for making these technologies easy accessible and, in the end, impacting people's everyday lives. Despite the unquestioned performance reached by deep learning methods in many fields — from image classification and natural language processing to medical imaging and self-driving — some important issues must be addressed to reach trustworthiness. It is now well known that neural networks are poorly explainable, vulnerable to imperceptible attacks and absorb the many biases present in the data with which they are trained. These technologies are on the right track for controlling delicate and safety-critical application scenarios such as health services and self-driving infrastructures. For this reason, it is important to study and develop AI methods that are robust, explainable, and reliable when deployed in the real world.

In this paper, we present some of the research carried out by the Artificial Intelligence for Media and Humanities (AIMH) laboratory of the ISTI-CNR, focusing on the study and development of trustworthy deep learning methods. Specifically, we discuss the work carried out in three important and very active areas of research: (a) the detection of adversarial examples in Convolutional Neural Networks (CNNs), (b) the use of attentive models — Transformers in particular — to obtain more powerful and explainable systems, and (c) the detection of deepfakes, that comprise fake images, videos, or pieces of texts used to polarize or poison public debates.

2 Research Themes

2.1 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, such as 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 (Figure 1). 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 lays 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. 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

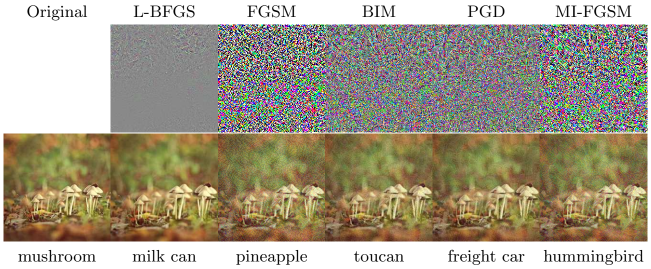


Figure 1: Examples of adversarial perturbation (on top) and inputs (on bottom) generated by different crafting algorithms. Image from [Carrara *et al.*, 2018].

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.

2.2 Transformers and Attention

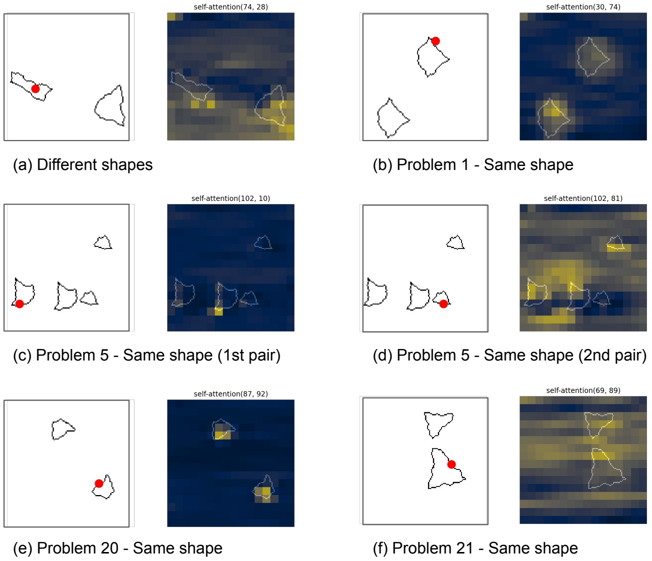
Modeling attention is particularly interesting from the perspective of interpretability, because it allows us to directly inspect the internal working of the deep learning architectures. In the context of visual-language tasks, the hypothesis is that the magnitude of attention weights correlates with how relevant a specific region of input is for the prediction of output at each position in a sequence [Chaudhari *et al.*, 2021].

One of the most promising attentive architectures proposed in recent years is the Transformer model [Vaswani *et al.*, 2017], which computes attention between two sequences as scaled dot-products between *keys* from the source sequence and *queries* from the target one. Although this model had great success in Natural Language Processing (NLP), recently it defined the state-of-the-art in multi-modal processing. Our team has been active in this promising field, defining some important milestones in multi-modal processing and exploiting Transformers in the pure visual domain.

Multi-modal Processing

We designed efficient and effective Transformer-based models for cross-modal visual-textual retrieval. This task consists in finding images most related to a given textual query or vice-versa. The works in [Messina *et al.*, 2021d; Messina *et al.*, 2021b] process images and texts as sets of tokens. These tokens are contextualized using the Transformer self-attention mechanism and then projected into the same concept space, where the affinity between a visual and a textual concept can be easily computed using cosine similarity. In [Messina *et al.*, 2021c] we explored some approaches to sparsify and index the resulting global and local contextualized features for performing large-scale cross-modal retrieval using off-the-shelf search engines.

We found another application of multi-modal Transformers in the detection of persuasion techniques in memes from social networks. Social networks play a critical role in our society, as 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-trivial to understand if the shared contents are free of subtle meanings altering people's judgments. In [Messina *et al.*, 2021e], we tackle the problem of recognizing which kind of disinformation technique is used to forge memes for a disinformation campaign. In particular, we propose an architecture based on the Transformer architecture model for processing both the textual and visual inputs from the meme. This architecture, which we call DVTT (Double Visual Textual Transformer), comprises two full Transformer networks working respectively on images and texts. Our proposed model could reach top positions on the *SemEval 2021 Task 6*[1] publicly available leaderboard.



(a) Different shapes

(b) Problem 1 - Same shape

(c) Problem 5 - Same shape (1st pair)

(d) Problem 5 - Same shape (2nd pair)

(e) Problem 20 - Same shape

(f) Problem 21 - Same shape

Figure 2: Attention visualization for the same-different tasks. Attention is used to inspect the salient image patches with respect to a particular query image patch (identified by the red dot). Image from [Messina *et al.*, 2021a] .

Vision Transformers

Transformers are recently reshaping the computer vision world. Some works demonstrated astonishing results in image classification using fully-Transformer architectures [Dosovitskiy *et al.*, 2020]. Nevertheless, very recently, mixed Transformer-Convolutional models are defining new standards in image processing. In fact, the attentive processing of convolutional features enables us to easily visualize the global correlation among every pair of visual patches. We recently exploited the synergy between Transformers and Convolutional Neural Networks for tackling abstract visual reasoning [Messina *et al.*, 2021a] and video deepfake detection [Coccomini *et al.*, 2021] tasks.

In particular, in [Messina *et al.*, 2021a] we tackled apparently a trivial set of tasks called *same-different* tasks. The same-different tasks consist of understanding if the shapes

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

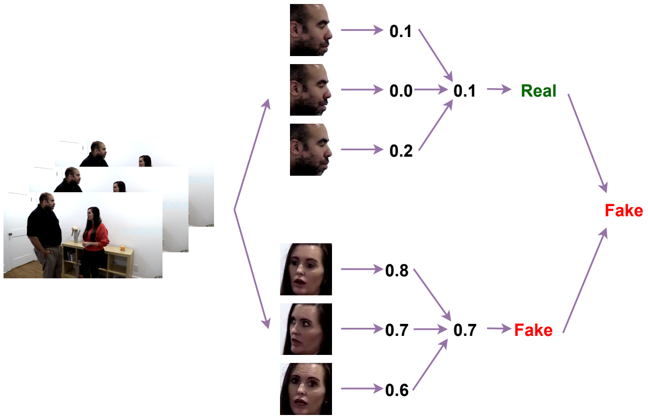


Figure 3: Video deepfake inference strategy. Faces are detected across multiple frames. Then, a voting algorithm decides if at least one of them was manipulated. Image from [Coccomini *et al.*, 2021].

in an image satisfy a certain unknown rule that must be automatically inferred from data. Despite the straightforward 2D synthetic figures used for this research, the same-different tasks aim at testing the reasoning abilities of deep neural networks and their proficiency to relate distant zones in an image. Given the difficulties of CNN-based methods on this task, in this work, we proposed to use a Recurrent Transformer network to perform high-level reasoning, using the features extracted by a simple upstream CNN. We obtained remarkable accuracies on the SVRT dataset using a relatively small training set, demonstrating the generalization abilities of the proposed model. Furthermore, the learned attention maps clearly indicate which image patches the model is attending to answer correctly (Figure 2).

Instead, in [Coccomini *et al.*, 2021] we combined various types of Vision Transformers with a convolutional EfficientNet B0 used as a feature extractor, for tackling the task of video deepfake detection. We better detail this contribution in the next section.

2.3 Deepfake Detection

Deepfake detection is a critical task in 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 to contaminate public debate. Deepfake images and videos can be used to harm important and strategic public figures. For this reason, it is very important to detect them to stop their diffusion promptly. Although many methods focused on image deepfake detection, in [Coccomini *et al.*, 2021] we tackled deepfake detection in videos. The challenge is identifying if there are people having their faces replaced or manipulated. In particular, we used a mixed Transformer-Convolutional model to attend the face patches. Unlike 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 analyzed 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 (Figure 3). 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. A video deepfake detector could also be used on a large scale. 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 the detection of deepfake 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. To show the challenges that TweepFake poses and provide a solid baseline of detection techniques, we also evaluated 13 different deepfake text detection methods. Some 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 good word representations for both text representation-based and fine-tuning-based detection techniques.

3 Projects

AI4Media

A Centre of Excellence delivering next-generation AI Research and Training at the service of Media, Society, and Democracy.

4 Challenges

There are major challenges to address to reach a robust, reliable, and explainable AI.

A considerable effort is being made in joining deep learning methods with symbolic AI. As pointed out in an article recently published by Turing Award winners Yoshua Bengio, Yann LeCun, and Geoffrey Hinton [Bengio *et al.*, 2021], deep learning demonstrated astonishing abilities in tasks involving fast and instinctive thinking, such as detecting salient objects in an image. Nevertheless, it still has problems in slow and analogical reasoning. This second way of deriving knowledge has been addressed in the past using symbols and logic inference rules. Trying to merge perception using deep learning systems and reasoning via symbolic methods would result in an interesting research direction, oriented at creating more fair and explainable models. This would result in stronger systems, also more resilient to external attacks.

Despite being popularly used to shed light on the inner working of black-box neural networks, using attention weights for model explainability remains an area of active research. Future extensions to this paradigm would be to include external knowledge, accessing it through analogous attention mechanisms. This would enable to instantly retrieve relevant common-sense or historical facts from a pre-built knowledge base, increasing the awareness of current visual and linguistic models. This would help both in cross-modal information retrieval systems and deepfake detection models, where, for example, knowing who a certain person is or when a certain event happened makes the difference between fake and not-fake news.

References

[Bengio *et al.*, 2021] Y. Bengio, Y. Lecun, e G. Hinton. Deep learning for ai. *Commun. ACM*, 64(7):58–65, jun 2021.

[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 *15th Int. Work. 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.

[Chaudhari *et al.*, 2021] S. Chaudhari, V. Mithal, G. Polatkan, e R. Ramanath. An attentive survey of attention models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(5):1–32, 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.

[Dosovitskiy *et al.*, 2020] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[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.

[Messina *et al.*, 2021a] N. Messina, G. Amato, F. Carrara, C. Gennaro, e F. Falchi. Recurrent vision transformer for solving visual reasoning problems. *arXiv preprint arXiv:2111.14576*, 2021.

[Messina *et al.*, 2021b] N. Messina, G. Amato, A. Esuli, F. Falchi, C. Gennaro, e S. Marchand-Maillet. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(4):1–23, 2021.

[Messina *et al.*, 2021c] N. Messina, G. Amato, F. Falchi, C. Gennaro, e S. Marchand-Maillet. Towards efficient cross-modal visual textual retrieval using transformer-encoder deep features. *arXiv preprint arXiv:2106.00358*, 2021.

[Messina *et al.*, 2021d] N. Messina, F. Falchi, A. Esuli, e G. Amato. Transformer reasoning network for image-text matching and retrieval. In *25th Int. Conf. on Pattern Recognition (ICPR)*, pages 5222–5229. IEEE, 2021.

[Messina *et al.*, 2021e] 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 Int. Work. 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.