Trustworthy AI @ PICUS Lab

Valerio La Gatta, Marco Postiglione, Stefano Marrone, Vincenzo Moscato Carlo Sansone and Giancarlo Sperlí

Department of Electrical Engineering and Information Technology (DIETI)
University of Naples Federico II, Via Claudio 21, 80125 Naples, Italy
first.last@unina.it

Abstract

The impact of AI on the industry has been so disrupting that it gave rise to a new wave of research and applications that goes under the name of *Industry 4.0*. As a consequence, today's developing trend is increasingly focusing on AI based data-driven approaches, mainly because leveraging user's data (such as location, action patterns, social information, etc.) can make applications able to adapt to them, enhancing the user experience. To this aim, these systems process enormous amounts of data too often rich in sensitive information. Although this opens unprecedented scenarios, it is important to note that its misuse (malicious or not) can lead to unintended consequences, such as unethical or unfair use (e.g. discriminating on the basis of ethnicity or gender), or used to harm people's privacy. We strongly believe that, since being just a (very powerful) tool, AI is not to blame for its misuse. Nonetheless, we claim that in order to develop a more ethical, fair and secure use of artificial intelligence, all the involved actors must have a very clear idea about some critical questions. In this work we briefly summarise the research activities at the PICUS (Pattern analysis and Intelligent Computation for mUltimedia Systems) Lab toward the realisation of a trustworthy AI.

1 Introduction

In recent years the term "Artificial Intelligence" (or AI) has become more and more an integral part of the daily life of all of us. We are increasingly dealing with *smart* mobile phones, *intelligent* voice assistants, *robotic* chats, etc. Our interaction with these "intelligent systems" has become so predominant and widespread that even the world of industry has begun to use such AI in factory life and logistics. The term artificial intelligence refers to the ability of a computer (an *artificial* entity) to perform functions resembling the typical reasoning of the human mind (i.e. *intelligence*). Nowadays, the term AI is widely abused, and one of the unpleasant effects of this spread with the mass audience is the confusion made with all its related terms, such as "Pattern Recognition", "Machine Learning", "Deep Learning", etc., too often used interchangeably. Indeed, what usually media refers to with AI is actually Machine Learning (ML), a term used to describe the ability of this kind of AI systems to *learn from examples*, just as we humans learn from experience.

Among all machine learning models, Artificial Neural Networks (ANN, often referred simply as Neural Network - NN) are definitely the branch that has been receiving the most media coverage since their parallel layered structure of computing elements (i.e. artificial neurons) is inspired to the human brain complex interconnected structure of biological neurons. More recently, the introduction of General Purpose GPU (GP-GPU) computing, the development of free and easy to use frameworks, the availability of huge labelled dataset and progresses in gradient-based optimisation, determined the uprising of *Deep Neural Networks*. The term "Deep Learning" (DL) refers to a particular subset of ANNs characterised, inter alia, by a very "depth structure" (i.e. made up of several layers). Another key aspect of deep models is their ability to autonomously learn the best or set of features for the task under analysis, to the point of even exceeding human capabilities in some tasks. This characteristic, known as *feature learning*, has played a key role in the recent spread of AI since allowed DL use also in domains lacking effective expert-designed features.

The impact of AI, and in particular of deep learning, on the industry has been so disrupting that it gave rise to a new wave of research and applications that goes under the name of *Industry 4.0*. This expression refers to the application of AI and cognitive computing to leverage an effective data exchange and processing in manufacturing technologies, services and transports, laying the foundation of what is commonly known as *the fourth industrial revolution*. As a consequence, today's developing trend is increasingly focusing on AI based data-driven approaches, mainly because leveraging user's data (such as location, action patterns, social information, etc.) can make applications able to adapt to them, enhancing the user experience. To this aim, tools like automatic image tagging (e.g. those based on face recognition), voice control, personalised advertising, etc. process enormous amounts of data (often remotely due to the huge computational effort required) too often rich in sensitive information. Although this opens unprecedented scenarios, it is important to note that its misuse (malicious or not) can lead to unintended consequen-

ces, such as unethical or unfair use (e.g. discriminating on the basis of ethnicity or gender), or used to harm people's privacy. Indeed, if on one hand, the industry is pushing toward a massive use of artificial intelligence enhanced solution, on the other it is not adequately supporting researches in end-to-end understating of capabilities and vulnerabilities of such systems. The results may be very (negatively) mediatic, especially when regarding borderline domains such those related to subjects privacy or to ethical and fairness, like users profiling, fake news generation, reliability of autonomous driving systems, etc.

We strongly believe that, since being just a (very powerful) tool, AI is not to blame for its misuse. Nonetheless, we claim that in order to develop a more ethical, fair and secure use of artificial intelligence, all the involved actors (in primis users, developers and legislators) must have a very clear idea about some critical questions, such as *"what is AI?"*, *"what are the ethical implications of its improper usage?"*, *"what are its capabilities and limits?"*, *"is it safe to use AI in critical domains?"*, and so on. Moreover, since AI is very likely to be an important part of our everyday life in the very next future, it is crucial to build trustworthy AI systems. In this work we thus briefly summarise our research activities intended to support, often by shedding lights on problems and issues, the realisation of a more trustworthy AI world.

2 The Need for Reproducible Research

The increasing spread of free and intuitive frameworks and affordable hardware supported researchers to explore the use of AI, and in particular of deep learning, in several application fields. If, on one hand, the availability of such frameworks allows developers to use the one they feel more comfortable with, on the other, it raises questions related to the reproducibility of the designed model across different hardware and software configurations, both at training and at inference time. This reproducibility assessment is important in order to determine whether the resulting model produces good or bad outcomes due to its capacity or just because of luckier or blunter environmental training conditions. As AI is being used in critical domains, more concerns this problem start arising, since reproducibility related issues could make authors claims harder to verify [Hutson, 2018]. Different initiatives have been deploying to promote results reproducibility, including the collection and release of big public datasets or the acknowledgement of a reproducibility label[1] or badges [2]. Unfortunately, when it comes to deep learning there is some non-determinisms intrinsically associated with the training procedure that unavoidably affect DL-based approaches reproducibility.

Although there are situations in which relaxing the reproducibility constraint is still acceptable (e.g. because what matters is the macro behaviour of a system), there are some *critical domains* in which this requirement is mandatory [Ras *et al.*, 2018], since unexpected behaviours could lead to potentially critical situations. A striking example is represented by e-Health [Amato *et al.*, 2019], a term that refers to all of healthcare practices supported by electronic means and elaborations. Indeed, although biomedical image processing [Piantadosi *et al.*, 2018a] is one of the research fields that has benefited the most by the rise of big data and deep learning [Gravina *et al.*, 2021] [Galli *et al.*, 2021], it is important to take into account some relevant differences between natural and medical images when using deep learning approaches, including the wide inter/intra patients variability, the need for elaborations respecting patients' privacy, the criticisms associated with false negative/positive in "sensitive" applications (e.g. tumours detection). Usually, to face this problem, researches take into account probabilistic considerations about the distribution of data or focus their attention on very huge datasets. However, this approach does not really fit the medical imaging analysis standards, with Computer-Aided Detection and Diagnosis systems (CAD) requiring demonstrable proofs of results effectiveness and repeatability. Our opinion is that *in these cases it is very important to clarify if and to what extent a DL based application is stable and repeatable over than effective*. Therefore, we recommended [Piantadosi *et al.*, 2018b] [Marrone *et al.*, 2019] to quantitatively highlight the reproducibility problem of CNN based approaches, proposing to overcome it by using statistical considerations. As a case of study, we considered our ICPR2018 [Piantadosi *et al.*, 2018c] proposal for the breast tissues segmentation in DCE-MRI by means of a 2D U-Net CNN [Ronneberger *et al.*, 2015], considering different configurations. The obtained results suggest that the reproducibility issue can be shifted from a strictly combinatorial problem to a statistical one, in order to validate the model robustness and stability more than its perfect outcomes predictability. Generally speaking, the randomness introduced by deep learning libraries, could impact outcomes of biomedical image processing application relying on deep learning approaches. Therefore, in order to avoid providing not totally reproducible claims, it is very important to shifts the attention from a pure performance point-of-view to a statistical validity of the obtained outcomes.

3 Explainable Artificial Intelligence

In recent years, sophisticated AI models are being applied in real contexts, assisting humans in the most disparate domains. However, the increase of their performance has been accompanied by an increase of complexity at a level that no one, neither their designers, can interpret the inner workings leading to decisions. Consequently, the current focus on the *eXplainable AI (XAI)* is driven by the need to enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners. Within this context, we propose a novel *local* and *model-agnostic* XAI methodology, named *Pivot-Aided Space Transformation for Local Explanations (PASTLE)* [La Gatta *et al.*, 2021], which aims to explain any model by not only providing features' contribution to its predictions but also adding information about how those predictions would change if feature changes were applied.

[1] https://rrpr2018.sciencesconf.org/

[2] https://www.acm.org/publications/policies/artifact-review-badging

4 Deep Approximate Computing

One of the common misconceptions with numerical computing is related to the concepts of "correct" and "approximate" computation: the former term is often erroneously (at list in the numerical computing field) used as synonym for "closed-form solution", while the latter tends to be perceived as "not precise" or "roughly estimated". The error lays its foundations in the erroneous belief (from non-expert people) of computers being able to directly interface with the real world (continuous in its nature), totally forgetting about their intrinsically discrete nature. Indeed, when a natural signal (e.g. an image, a sound, an earthquake trace, etc) needs to be processed by means of a computer, the first step is to make it discrete (e.g. by using quantization). Some applications have the property of being *resilient*, meaning that they are robust to noise (e.g. due to error, to discretization, etc.) in the data. This characteristic is very useful in situations where an *approximate computation* (e.g. by representing rational numbers by using single-precision floating-point variables instead of double precision ones) allows to perform the computation in less time or to deploy it on embedded hardware [Amato *et al.*, 2017]. Deep learning is clearly one of the fields that can benefit from approximate computing since, by definition, once trained they show an impressive generalisation ability (i.e. resiliency to an error in the data or to intrinsic randomness). One of the most adopted solutions in this regard is to quantize the learnt weights [Han *et al.*, 2015], with the aim of both reducing the required memory and to speed-up the inference stage. The limitation of this solution is in the fact that the obtained network is not necessarily the most compact possible one since quantization operates on all the weights similarly. Therefore, we proposed [Marrone *et al.*, 2021] to face the problem from a different perspective: instead of reducing the weights or look for the perfect approximated network, we investigate whether is it possible to remove whole neurons without substantially affecting the network performance. The whole idea is based on the fact that (almost) all CNNs can be seen as a stack of neurons specialised for the feature engineering (the ones in the convolutional layers) and neurons intended for the classification task (the ones in the fully connected layers). According to this point of view, the fully connected layers can be seen as a standard multilayer perceptron (MLP) classifier that relies on features extracted from the convolutional layers (although it is worth noting that all layers are trained together); in fact using AlexNet as example, it has $2,334,080$ parameters in the convolutional layers and $58,631,144$ parameters in the fully connected ones because AlexNet was originally intended to face the 1000 classes of ImageNet, therefore, after the feature engineering, a great representational capacity is required.

As a case of study, we focused on transfer learning, by *proposing to further adapt deep CNNs by performing a sizing of hidden layers (i.e. those between the input and the output layers) when using the fine-tuning strategy*. With the term sizing, we refer to the use of a suitable strategy to reduce the number of used neurons, without significantly affecting the network performance. This can be achieved in two different ways: by reducing the synapses (i.e. connections between neurons), or by removing whole neurons. We analysed the latter approach, focusing in particular only on neurons in the fully connected layers since, as previously discussed, are the most numerous ones. To measure the suitability of the sizing strategy for approximate computing purposes, it is important to measure the resiliency of the *"sized"* network, with respect to the *"original"* one, on a given task. Both the network depth and the considered task might affect the approach effectiveness: the former, due to the different ratio between the numbers of neurons in the convolutional and in the fully connected layers; the latter due to the different number of classes, directly related to the number of neurons in the classification layer. Therefore, we considered two different networks on three classification problems differing in terms of number of classes, number of samples and image resolution.

Although this work must be considered just as a proofs-of-concept, results show that it could be possible to reduce the number of neurons in the hidden layer without statistically affect the network performance, but with a significant reduction of the number of parameters (up to $\sim 27\%$ for AlexNet and $\sim 11\%$ for Vgg19) and required memory occupation (up to $\sim 41\%$ for AlexNet and $\sim 10\%$ for Vgg19). This would imply that, though fine-tuning can already be effectively used in many different contexts, other investigations are needed to develop improvements that allow unleashing its full potential. Moreover, the possibility of reducing the memory occupation without a direct impact on the network classification ability open new scenarios toward the application of powerful CNNs on embedded devices, as already done with Random Forest classifier [Barbareschi *et al.*, 2017].

Given those results, we also investigated whether the shape of the network itself contributes to the effectiveness adversarial perturbation attacks. Interestingly, results show that, apart for a single combination, sized CNNs needs a "stronger" adversarial noise to be mislead. This further motivate other investigations in the direction of CNNs layer sizing, since the reported analysis seems to suggest that the adversarial perturbation problem can be faced (or at least limited) by reducing the number of neurons/connections that actively take part in the network decision problem. This results, although preliminary, represent a novel contribution in the field of adversarial defense strategies since CNN sizing does not relay on the analysis of the data but could make CNNs intrinsically more robust by changing the neurons connection pattern. This will help the user in "sizing" the network accordingly to the desired levels of performance/robustness they needs to obtain on the basis of the risk associated with the task.

5 Fairness and Privacy in Face Analysis

Privacy threats are usually perceived as far from us or, even worse, able to affect only our digital alter-ego. Unfortunately, this is a false belief. Indeed, especially when it comes to derive subjects' soft biometrics from data spontaneously published by target users (e.g. a profile picture on a social media, a blog post, a product review, etc.), there are many very effective and sneaky (i.e. not perceived by users) attacks able to extract our soft biometrics with a single glance in a real environment, maybe also without our explicit consensus

[Kosinski *et al.*, 2013]. We thus investigated [Marrone e Sansone, 2019] if and to what extent adversarial perturbation can be used to protect subjects against unwanted soft biometric detection by automatic means. The idea is to create an adversarial patch [Brown *et al.*, 2017] that, once printed, is able to fool CNNs by simply "wearing" it, in the shape, for example, of a sticker, a clip or a pendant. By definition, an adversarial perturbation should be as invisible as possible and this constraint is usually met since the injected noise is distributed over the whole image. On the contrary, in our case, it is strongly preferable to trade a very visible perturbation in exchange for having the opportunity to apply it on a very limited portion of the image. As a case of study, we will focus on the generation of a *general adversarial patch* specifically designed to fool a CNN for ethnicity recognition in a real-world application. In particular, the aim is to create a patch that works across different subjects, including those never seen by the CNN. Results shows that the approach is effective also when the used patch size is reduced. It is interesting to note that the performance drop associated with the patch size is more significant for black subjects: this could be due to many factor, including problems related to a smaller diversity of the dataset (the number of black subjects is almost the half of white ones) or, on the contrary, to a greater variation between black subject images. It is worth noting that our approach can work in a real world scenario, by printing the patch and using against any live camera acquisition.

Riferimenti bibliografici

[Amato *et al.*, 2017] Flora Amato, Mario Barbareschi, Giovanni Cozzolino, Antonino Mazzeo, Nicola Mazzocca, e Antonio Tammaro. Outperforming image segmentation by exploiting approximate k-means algorithms. In *International Conference on Optimization and Decision Science*, pages 31–38. Springer, 2017.

[Amato *et al.*, 2019] Flora Amato, Stefano Marrone, Vincenzo Moscato, Gabriele Piantadosi, Antonio Picariello, e Carlo Sansone. Holmes: ehealth in the big data and deep learning era. *Information*, 10(2):34, 2019.

[Barbareschi *et al.*, 2017] Mario Barbareschi, Cristina Papa, e Carlo Sansone. Approximate decision tree-based multiple classifier systems. In *International Conference on Optimization and Decision Science*, pages 39–47. Springer, 2017.

[Brown *et al.*, 2017] TB Brown, D Mané, A Roy, M Abadi, e J Gilmer. Adversarial patch. arxiv e-prints (dec. 2017). *arXiv preprint cs.CV/1712.09665*, 1(2):4, 2017.

[Galli *et al.*, 2021] Antonio Galli, Stefano Marrone, Gabriele Piantadosi, Mario Sansone, e Carlo Sansone. A pipelined tracer-aware approach for lesion segmentation in breast dce-mri. *Journal of Imaging*, 7(12):276, 2021.

[Gravina *et al.*, 2021] Michela Gravina, Stefano Marrone, Mario Sansone, e Carlo Sansone. Dae-cnn: exploiting and disentangling contrast agent effects for breast lesions classification in dce-mri. *Pattern Recognition Letters*, 145:67–73, 2021.

[Han *et al.*, 2015] Song Han, Huizi Mao, e William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[Hutson, 2018] Matthew Hutson. Artificial intelligence faces reproducibility crisis, 2018.

[Kosinski *et al.*, 2013] Michal Kosinski, David Stillwell, e Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[La Gatta *et al.*, 2021] Valerio La Gatta, Vincenzo Moscato, Marco Postiglione, e Giancarlo Sperlì. Pastle: Pivot-aided space transformation for local explanations. *Pattern Recognition Letters*, 149:67–74, 2021.

[Marrone *et al.*, 2019] Stefano Marrone, Stefano Olivieri, Gabriele Piantadosi, e Carlo Sansone. Reproducibility of deep cnn for biomedical image processing across frameworks and architectures. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2019.

[Marrone *et al.*, 2021] Stefano Marrone, Cristina Papa, e Carlo Sansone. Effects of hidden layer sizing on cnn fine-tuning. *Future Generation Computer Systems*, 118:48–55, 2021.

[Marrone e Sansone, 2019] Stefano Marrone e Carlo Sansone. An adversarial perturbation approach against cnn-based soft biometrics detection. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[Piantadosi *et al.*, 2018a] Gabriele Piantadosi, Stefano Marrone, Roberta Fusco, Mario Sansone, e Carlo Sansone. Comprehensive computer-aided diagnosis for breast t1-weighted dce-mri through quantitative dynamical features and spatio-temporal local binary patterns. *IET Computer Vision*, 12(7):1007–1017, 2018.

[Piantadosi *et al.*, 2018b] Gabriele Piantadosi, Stefano Marrone, e Carlo Sansone. On reproducibility of deep convolutional neural networks approaches. In *International Workshop on Reproducible Research in Pattern Recognition*, pages 104–109. Springer, 2018.

[Piantadosi *et al.*, 2018c] Gabriele Piantadosi, Mario Sansone, e Carlo Sansone. Breast segmentation in mri via u-net deep convolutional neural networks. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3917–3922. IEEE, 2018.

[Ras *et al.*, 2018] Gabriëlle Ras, Marcel van Gerven, e Pim Haselager. Explanation methods in deep learning: Users, values, concerns and challenges. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pages 19–36. Springer, 2018.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, e Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.