own, while it will have a low probability of interaction with opinions far from their own.

Simulations of the algorithmic bias model show several results that suggest that online platforms can have important effect on opinion formation and consensus in society. First, the number of opinion clusters grows when algorithmic bias grows (see illustration). This means that online platforms can favour fragmentation of opinions. Second, this leads also to polarisation, where the distance between the opinions of the people is larger compared to the situation without algorithmic bias. Third, the changes in opinion are much slower when the bias is in operation. Even when consensus is obtained, the time to reach it becomes very long. In practice, this means that it could take years for people to agree on an issue, being in a highly fragmented state while this occurs.

These results bring important evidence that algorithmic bias may affect outcomes of public debates and consensus in society. Thus, we believe measures are required to at least stop its effects, if not reverse them. Researchers are investigating means of promoting consensus to counteract for the algorithmic bias effects. In the meantime, users could be informed of the way platforms feed information and the fact that this could affect their opinions, and maybe the mechanisms implemented by the platforms could be slowly withdrawn.

Reference:
[1] Alina Sîrbu, et al.: "Algorithmic bias amplifies opinion polarization: A bounded confidence model", arXiv preprint arXiv:1803.02111, 2018. https://arxiv.org/abs/1803.02111

Please contact:
Alina Sîrbu,
University of Pisa, Italy
alina.sirbu@unipi.it

Detecting Adversarial Inputs by Looking in the Black Box

by Fabio Carrara, Fabrizio Falchi, Giuseppe Amato (ISTI-CNR), Rudy Becarelli and Roberto Caldelli (CNIT Research Unit at MICC – University of Florence)

The astonishing and cryptic effectiveness of Deep Neural Networks comes with the critical vulnerability to adversarial inputs — samples maliciously crafted to confuse and hinder machine learning models. Insights into the internal representations learned by deep models can help to explain their decisions and estimate their confidence, which can enable us to trace, characterise, and filter out adversarial attacks.

Machine learning and deep learning are pervading the application space in many directions. The ability of Deep Neural Network (DNN) to learn an optimised hierarchy of representations of the input has been proven in many sophisticated tasks, such as computer vision, natural language processing and automatic speech recognition. As a consequence, deep learning methodologies are increasingly tested in security- (e.g. malware detection, content moderation, biometric access control) and safety-aware (e.g. autonomous driving vehicles, medical diagnostics) applications in which their performance plays a critical role.

However, one of the main roadblocks to their adoption in these stringent contexts is the diffuse difficulty to ground the decision the model is taking. The phenomenon of adversarial inputs is a striking example of this problem. Adversarial inputs are carefully crafted samples (generated by an adversary — thus the name) that look authentic to human inspection, but cause the targeted model to misbehave (see Figure 1). Although they resemble legitimate inputs, the high non-linearity of DNNs permits maliciously added perturbations to steer at will the decisions the model takes without being noticed. Moreover, the generation of these malicious samples does not require a complete knowledge of the attacked system and is often efficient. This exposes systems with machine learning technologies to potential security threats.

Many techniques for increasing the model's robustness or removing the adversarial perturbations have been developed, but unfortunately, only a few provide effective countermeasures for specific attacks, while no or marginal mitigations exist for stronger attack models. Improving the explainability of models and getting deeper insights into their internals are fundamental steps toward effective defensive mechanisms for adversarial inputs and machine learning security in general.

To this end, in a joint effort between the AIMIR Research Group of ISTI-CNR and the CNIT Research Unit at MICC (University of Florence), we analysed the internal representations learned by deep neural networks and their evolution throughout the network when adversarial attacks are performed. Opening the "black box" permitted us to characterise the trace left in the activations throughout the layers of the network and discern adversarial inputs among authentic ones.

We recently proposed solutions for the detection of adversarial inputs in the context of large-scale image recognition with deep neural networks. The rationale of our approaches is to attach to each prediction of the model an authenticity score estimating how much the internal representations differ from expected ones (represented by the

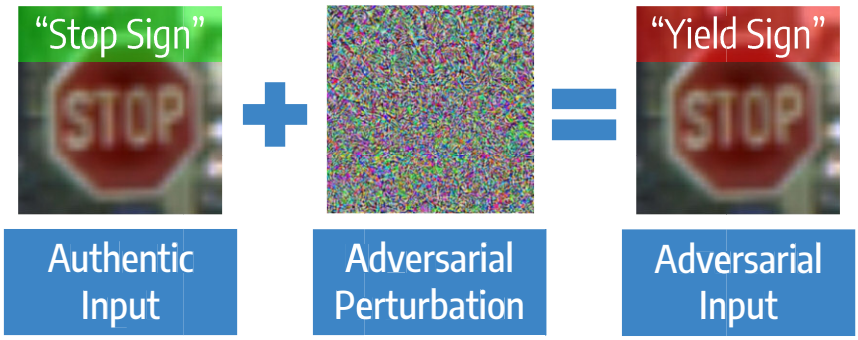

Figure 1:Example of a common adversarial attack on image classifiers. The adversarial perturbation added (magnified for visualization purposes) fools the network to predict a wrong class with high confidence.

model's training set). In [1], such a score is obtained by analysing the neighbourhood of the input with a nearest-neighbour search in the activation space of a particular layer. Our experiments on adversarial detection permitted us to identify the internal activations which are influenced the most by common adversarial attacks and to filter out most of the spurious predictions in the basic zero-knowledge attack model (see [L1]).

Building on this idea, in [2] we improved our detection scheme considering the entire evolution of activations throughout the network. An evolution map is built by tracing the positions an input occupies in the feature spaces of each layer with respect to most common reference points (identified by looking to training set inputs). Experiments showed that adversarial inputs usually tend to deviate from reference points leading to different activation traces in the network with respect to authentic inputs (see Figure 2). Thus, conditioning our detector on such information permitted us to obtain remarkable detection performance under commonly used attacks.

We plan to extend our analysis in order to fully characterise the effect of adversarial attacks on internal activations even in stricter attack models, i.e. when the attacker is aware of defensive systems and tries to circumvent it.

Despite our experimentation on adversarial input detection, both the presented approaches actually aim to cope with a broader problem, which is assigning a confidence to a model's decision by explaining it in terms of the observed training data. We believe this is a promising direction for reliable and dependable AI.

Links:
[L1]
http://deepfeatures.org/adversarials/

References:
[1] Carrara et al.: "Adversarial image detection in deep neural networks", Multimedia Tools and Applications, 1-21, 2018
[2] Carrara et al.: "Adversarial examples detection in features distance spaces", ECCV 2018 Workshops, 2018.

Please contact:
Fabio Carrara, ISTI-CNR, Italy
fabio.carrara@isti.cnr.it

Roberto Caldelli, CNIT Research Unit at MICC – University of Florence, Italy
roberto.caldelli@unifi.it

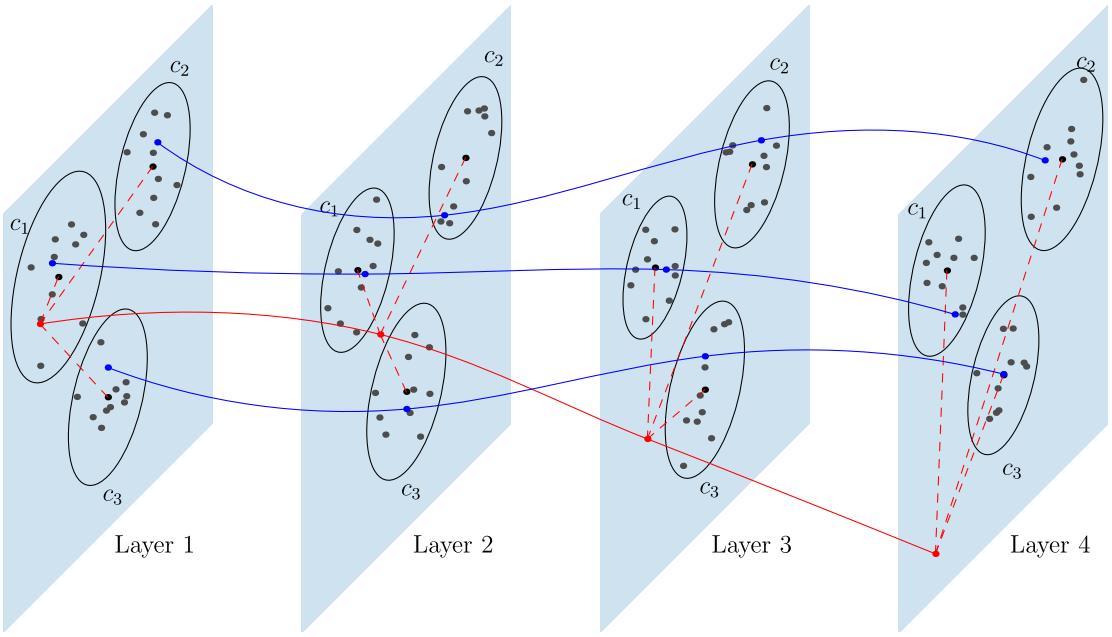

Figure 2: Conceptualisation of the evolution of features while traversing the network. Each plane represents a feature space defined by the activations of a particular layer of the deep neural network. Circles on the features space represent clusters of features belonging to a specific class. Blue trajectories 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.