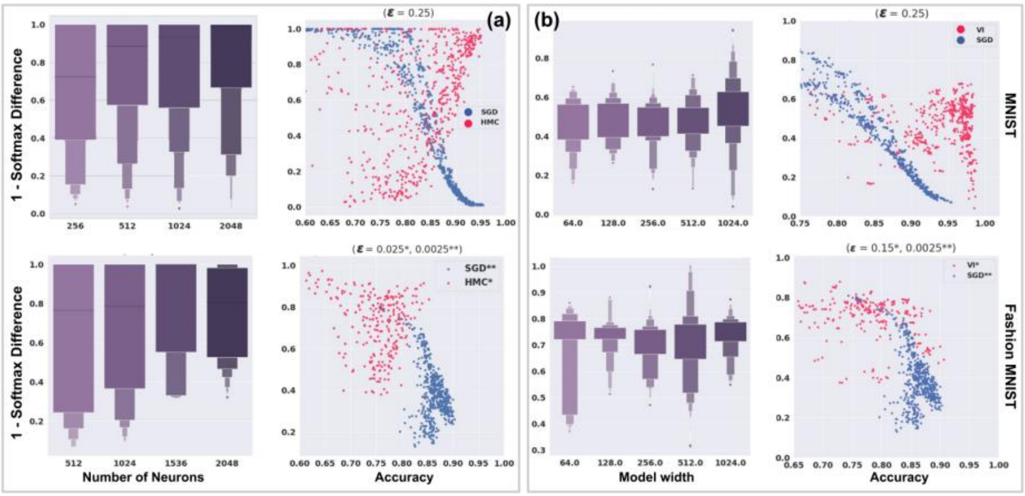
Press Release

Stronger neural networks: a method of not deceiving artificial intelligence

A joint study by the University of Trieste of Sissa and the University of Oxford has managed to understand the origin of fragility in the classification of objects by algorithms and to remedy it



Trieste, 15 October 2020

A modification imperceptible to the human eye can deceive sophisticated artificial intelligences by forcing them to make classification mistakes that a human being would never make, such as confusing a bus with an ostrich. A significant safety issue when it comes to applying deep learning, deep neural networks, to tools like self-driving cars.

A team of researchers and professors from the University of Trieste, Scuola Internazionale di Studi Superiori e Avanzati – SISSA and Oxford University, has nevertheless succeeded in showing a new way to make these neural networks more robust and difficult to deceive.

The study, "Robustness of Bayesian Neural Networks to Gradient-Based Attacks", whose authors are Ginevra Carbone, Matthew Wicker, Luca Laurenti, Andrea Patane, Luca Bortolussi, Guido Sanguinetti, was accepted by the Conference on Neural Information Processing Systems, The main artificial intelligence conference.

This confirms that, within the relatively new "data science" area, excellent partnerships are being consolidated in the regional context, with results already recognizable and resonances in the international context.

Another reason of note is that the first author, Ginevra Carbone, is a Phd student at the University of Trieste, one of the first students to have graduated in the Master’s degree in *“Data Science and Scientific Computing”*. As well, established, there are not many women who undertake this type of study in this field, and this, therefore, is a result of particular value.

Deep neural networks (using deep learning) are fundamental algorithms in modern artificial intelligence, thanks to their ability to recognize complex structures in huge masses of data. A striking example is the ability to recognize objects in natural images, or to translate (and generate) language. But these networks can be fooled: a hacker in possession of some information on the network (the so-called “loss function”) can easily insert small perturbations to the inputs, causing catastrophic failures in predictions.

For example, a change of a few pixels in an image can make the network classify a bus as an ostrich, a trap in which a human being would never fall. The vulnerability to these "adversarial attacks" has been for years a limit to the application of this type of networks to applications where security is critical. In the new research, adopting a geometric point of view and using results that come from statistical physics, the authors have been able to understand and demonstrate the mathematical origin of the problem and to show that in some cases it can be compensated using Bayesian techniques, using a set of statistically consistent networks that have the ability to compensate each other in the mistake they make, becoming more robust.

The paper is available at this link: https://arxiv.org/pdf/2002.04359.pdf

It should also be noted that some industry blogs, available at the links below, have already taken up, independently, the work mentioned in very flattering terms:
https://statsandai.wordpress.com/2020/10/08/research-highlights-robustness-of-bayesian-neural-networks-to-gradient-based-attacks/
https://www.programmersought.com/article/98695654326/

CONTATTI
UNIVERSITA' DI TRIESTE
Giampiero Viezzoli
✉ giampiero.viezzoli@amm.units.it
T +39 040 5583042
M +39 3204365043

CONTATTI
SISSA
Alessandro Tavecchio
✉ atavecch@sissa.it

M +39 333 6877130

Marina D'Alessandro
✉ mdalessa@sissa.it
T +39 040 3787231
M +39 349 2885935