(a) Image classified as a squirrel.

(b) Perturbation by an adversarial attack.

(c) Image classified as a parking meter.

Believing in AI

TakeAIM 2019 Runner-Up:

Giuseppe Ughi
University of Oxford

Would you rely on an self-driving car which could steer right into the opposite traffic lane because of some imperceptible dots on the dividing line [1]? Would you allow the police to automatically fine people via a recognition algorithms implemented on CCTV cameras, if you knew that a slight noise in the picture could convince the algorithms that you are breaking the law while normally walking down a sidewalk?

This kind of susceptibility of AI based algorithms is one of the main reasons why society is reluctant to automate sensitive tasks; the advantages of executing a task more efficiently are overshadowed by the risk of attacks which leverage on this weakness to imperceptible perturbations. Thus, understanding how feasible these attacks are in practice, is central to the full deployment of AI.

In my research I am studying how different attacks compare when they are trying to convince an unknown AI algorithm that an image depicts a targeted class through some small perturbations, see attached Figure. These attacks are allowed to query multiple times the black box algorithm on the confidence with which it classifies an image as different objects.

The achieved results show that the less an image is allowed to be perturbed, the more the attacks based on traditional mathematical optimisation routines are successful, as the one I implemented. Thus the risk of attacks is real and we must focus on generating AI algorithms that are mathematically guaranteed to be robust.

Sponsors

dyson

EPSRC
Engineering and Physical Sciences
Research Council

nag

NATS

Mondelēz
International

TUI

The Smith Institute, enabled by the generous sponsorship of our leading corporate partners, ran the TakeAIM competition in 2019 to make visible the crucial role that mathematics will increasingly play in all aspects of our lives. The competition was open to undergraduate and postgraduate students working in the mathematical sciences. First prize was £1,000 of Apple or Amazon vouchers, with second prize winners receiving £200 and 8 runners-up receiving £25 in their choice vouchers.