Adversarial attacks via Sequential Quadratic Programming

Théo Beuzeville[1,3] Pierre Boudier[2] Alfredo Buttari[3,5] Serge Gratton[4] Theo Mary[5,6] Stéphane Pralet[1] Erkan Ulker[1]

[1]ATOS

[2]NVIDIA

[3]IRIT

[4]Toulouse INP-ENSEEIHT

[5]CNRS

[6]LIP6

Sparse Days, June 2022

Adversarial attacks

The approach consists in computing the smallest norm perturbation on input data such that, for a given input x, the pertubed input is misclassified by the neural network, that is, it erroneously affects the input to a given class j instead of the expected one.

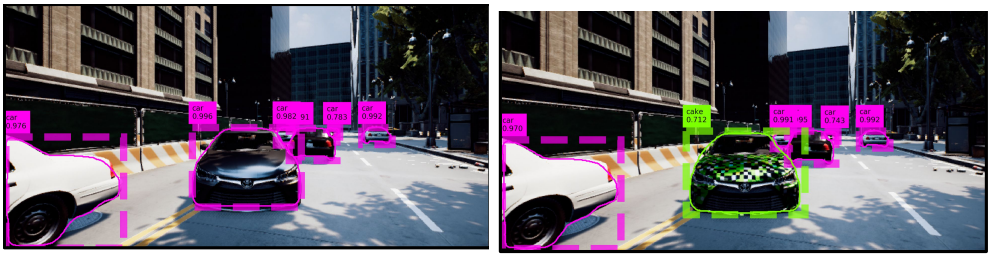

Figure: The camouflage fools the Mask R-CNN object detector (on the right), whereas plain colors (on the left) is being correctly detected.

Adversarial attacks

Mathematically, the adversarial perturbation is defined as the solution of the following minimization problem:

$$\begin{aligned} &\textbf{Solve} \\ &\min \|\Delta x\|^2 \\ &\textbf{subject to} \\ &C(x + \Delta x) = j. \end{aligned} \tag{1}$$

Where j is the target class and $C(x + \Delta x)$ the class of the perturbed image.

Adversarial attacks

This problem being difficult to solve, most of the approaches that generate adversarial examples commonly resort to solving the following problem:

Solve

$$\min c\|\Delta x\|^2 + \mathcal{L}(x + \Delta x, j), \tag{2}$$

where $\mathcal{L}$ is a given loss of the image with respect to a given target class.

Adversarial attacks

Solve

$$\min \|\Delta x\|^2 \tag{3}$$

subject to

$$C(x + \Delta x) = j.$$

We introduce a novel method to try and solve this problem more accurately using second order information.

Thank you for your attention and see you at the poster corner.