Robust Physical-World Attacks on Deep Learning Models

Ravi Raju

12-27-2018

1 Motivation

This paper seeks to propose an attack algorithm which introduces perturbations which are analogous those in the physical world. This again deviates away from the min distance approach that we see in FGSM, CW attack, JSMA, etc.

2 Solution

The actual method is called Robust Physical Perturbations which generate physical perturbations for physical-world objects. An apt example of this can be seen in Figure 2. Essentially, they calculate the perturbation with the optimization method detailed below and then they apply a mask over the original image to obtain the correct position. They also apply the same type of transformations for the image to the perturbation in the formulation.

$$NPS = \sum_{\hat{p} \in R(\delta)} \prod_{p' \in P} |\hat{p} - p'|$$

where this is the fabrication error which is based on the Non-Printability Score (NPS). More specifically, printable colors (RGB triples) P and a set $R(\delta)$ of (unique) RGB triples used in the perturbation that need to be printed out in physical world.

$$\underset{\delta}{\text{minimize}} \quad \lambda \|M_x \cdot \delta\|_p + NPS + \mathbb{E}_{x_i \sim X^V} J(f_\theta(x_i + T_i(M_x \cdot \delta)), y^*).$$

T is the alignment function that maps transformations on the object to transformations on the perturbation.

3 Results

They conducted a test on an actual test drive scenario. They tested this on 2 types of CNNs and found that 100 % success on the smaller classifier and 87.5 % on the larger classifier.

4 Discussion/Takeaway

Basicallu the important thing to look at what is the minimum classification distance in these examples? Are CW and optimization based methods the way to solve these kind of problems? Do you just solve this with adverserial training? How do perturbation correspond to methods like the above mentioned?