Review: Building Adversary Resistant Deep Neural Networks with Random Feature Nullification

Summary

The paper presents the random feature nullification to preserve the DNN's integrity against adversarial samples. Based on the discussion on the related work to increase the DNN's resistance to adversarial samples, the non-deterministic technique introduced is a nature progression from the current available techniques of adversarial training and model complexity enhancement. (weakness of adversarial training and model complexity enhancement here?) The paper points out that to be able to put up a strong resistance against adversarial attacks, the model architecture has to be one makes it impossible to generate adversarial samples even when it is disclosed.

Positivity

Negativity

- Could explore deeper on how attackers to bypass the random feature nullification technique other than the traditional method used by them to produce adversarial samples.
- could have tested it on CIFAR dataset
- there are papers that the author cited from that discussed about adversarial examples in other datasets.

References

[Greenemeier, 2016] Greenemeier, L. (2016). Deadly tesla crash exposes confusion over automated driving.