Problem: Adversarially Robust Classifier

Background:– An adversarial perturbation is a change in an image that is imperceptible to the human eye, but can influence a classifier to output a radically different probability distribution / class scores [3] as shown in Figure 1.

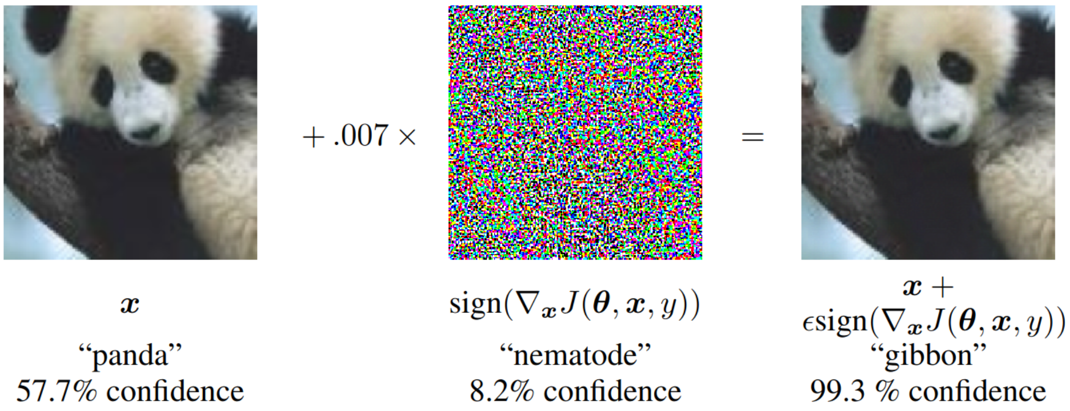
$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Figure 1: Adding a small perturbation to an image causes an adversarially non-robust classifier to dramatically change its class confidence scores. Image taken from [3].

Exercise:– (1) Your primary task is to design a classifier for the CIFAR-10 dataset that is robust to (i) the Fast Gradient Signed Method (FGSM) attack [3], (ii) the Projected Gradient Descent Attack [5], and (iii) L_∞ and L_2 momentum iterative attack [2]. (2) In order to ensure that your defence algorithm is not limited to the non-differentiable pre-processing based family, use the Backward Pass Differentiable Approximation [1] to perform an adaptive attack on your classifier. Ensure that your algorithm is also robust to the same.

Evaluation:– You have to perform all tests using images from the test split of the CIFAR-10 [4] dataset. For each of the attacks applied to your baseline classifier, you have to report the robustness improvement achieved by your defence algorithm(s) (you can use different defences against each of the attacks, but a more general defence would be considered more elegant). You also have to describe how and why your defence algorithm was able to achieve the performance it did against each of the individual attacks.

Deliverable:– You need to submit **a short report of maximum four A4 sides** on your work and **any tool or code** that you implemented to obtain results. In the report, you need to describe the novel or additional components of your work with an appropriate reference to the existing works and the reason of any particular decision you made in your work. Then you will show and discuss the results that you obtain and draw conclusions from there. You will also have to provide all the code and models that you have developed for this problem along with a documentation (in a README file) specifying the steps to setup the environment, run your code, and reproduce your results.

This report can be treated as an informal medium for communication, but should be sufficiently comprehensible. There is no need to follow any particular structure for writing it and you are not expected to write literature review. There is no word limit for the report and there will not be any penalty for crossing the 4 pages limit, but we are not obligated to look beyond the first 4 pages. Please, feel free to use a standard readable (font size 10pt) LaTeX template of your choice for the report.

Deadline:– You are expected to submit your deliverable on or before **Wednesday 18 May 2022**.

Question:– Any question, doubt, difficulty, discussion, submission of deliverable, please forward to **Dr Anjan Dutta** (anjan.dutta@surrey.ac.uk).

References

[1] Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018.

[2] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018.

[3] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *ICLR*, 2015.

[4] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). *Toronto*, 2009.

[5] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.