Indian Institute of Technology, Jodhpur

Dependable AI | Assignment 2

Topic : Adversarial Attacks - Attack, Detection and Mitigation

Total Marks : 130+ 50 Bonus + 20 Marks(Report) + 30 Marks(Viva)

Submission Policy and Requirements :

- Any kind of plagiarism is not accepted. We will strictly follow institute policies for plagiarism.
- Recommended programming languages: Python + Keras/TensorFlow/PyTorch.
- You may use any external libraries or GitHub codes. However, the evaluation will test your knowledge of the algorithm and the choice of hyperparameters. Do cite the libraries/codes.
- Submission should include: Working code for each of the parts separately and a report to show the analysis of results in each of the parts.

Assessment criterion:

The assessment will be done on the basis of the following components:

- Working codes
- Analysis and clarity of results (drawing comparisons across different parts) & clarity of the report
- Understanding the theoretical concepts and the choice of hyperparameters.

Guidelines for Submission:

- A single report(pdf) for all questions.
- Mention all the relevant results, comparisons as asked or wherever required for better understanding for the results.
- A single zip file containing the report, codes and readme if required

---

AIM: Understand and Perform different types of Adversarial Attacks, Detection and Mitigation.

Dataset: D-10 Dataset(**CLICK** for Dataset) - 10 Class Animal Dataset

Reference Paper: **Adversarial Attacks and Defenses in Deep Learning**

a. Take any Deep Model (like VGG16, VGG19, ResNet50, GoogleNet etc.) and train the D-10 Dataset from scratch(no pre-initialised weights) and **report the accuracy, training-testing loss graph, classification report**(use sklearn library for assistance). **[20 Marks]**

b. Perform both **targeted** and **untargeted** attacks(on testing set) for any **3 attacks** among the following. **[30 Marks]**

   i. L-BFGS algorithm
   Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. 2013. arXiv:1312.6199.

   ii. Fast gradient sign method
   Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. 2014. arXiv:1412.6572.

   iii. BIM and PGD
   Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world. 2016. arXiv:1607.02533.

   iv. Momentum iterative attack
   Dong Y, Liao F, Pang T, Su H, Zhu J, Hu X, et al. Boosting adversarial attacks with momentum. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA; 2018. p. 9185–193.

v. Distributionally adversarial attack
Zheng T, Chen C, Ren K. Distributionally adversarial attack. 2018. arXiv:1808.05537.

vi. Carlini and Wagner attack
Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proceedings of the 2017 IEEE Symposium on Security and Privacy; 2017 May 22–26; San Jose, CA, USA; 2017. p. 39–57.

vii. Jacobian-based saliency map approach
Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The limitations of deep learning in adversarial settings. In: Proceedings of the 2016 IEEE European Symposium on Security and Privacy; 2016 Mar 21–24; Saarbrucken, Germany; 2016. p. 372–87.

c. Report SSIM(Structural Similarity) for the predictions. Infer **[10 Marks]**

d. **Detection**:Identify at least **2 measures** that can be used as a metric **to detect adversarial perturbation** and **compare the results. Explain the measures in detail.** Reference **[20 Marks]**

e. Perturb all images in the training and testing set (using any 1 attack of your choice), train the model with perturbed training set image and then perform 10-class classification. Provide the **classification report** (use sklearn library for assistance), **compare the training and testing accuracy** with the previous case. **Report SSIM** for this case too and **compare the results** with the previous case. Infer **[30 Marks]**

f. **Mitigation:** Use perturbed test images(from previous question) and perform JPEG compression at 2 different compression rates, compare the **classification report**(for the model trained with original image set) of this case with all 2 previous cases i.e. original test samples and perturbed samples in tabular format. Infer **[20 Marks]**

Bonus [50 Marks]

Come up with a visible attack for a face that can fool the face recognition model.

Example:

Applying beard costumes over the face can cause a person to fake the identity. It has also been used to fool machine learning models to fake their identity. Your task is to come up with a machine learning model to generate a beard for the faces so that after mapping them to the human faces, the identity of the person can be fooled by naked eyes as well as by face detection model. (You may refer to any other article apart from the one mentioned below, but remember to cite the work properly. NOTE: Do not copy/submit already implemented paper. Reference can be taken from paper)

Refer these articles:

Paper: Intuitive, Interactive Beard and Hair Synthesis With Generative Models

Blog:https://blog.insightdatascience.com/generating-custom-photo-realistic-faces-using-ai-d170b1b59255