Attack On Image Classification Network

Project Supervisor: Dr. Puneet Gupta
Saumya Mishra, Namrata Tiwari, Anil

M.S(Research)

13 September 2019

1 Introduction

Deep Learning is providing major breakthroughs in solving the problems that have withstood many attempts of machine learning and artificial intelligence community in the past. As a result, it is currently being used to decipher hard scientific problems at an unprecedented scale, e.g. in reconstruction of brain circuits ; analysis of mutations in DNA ; prediction of structure-activity of potential drug molecules, and analyzing the particle accelerator data.Deep neural networks have also become the preferred choice to solve many challenging tasks in speech recognition and natural language understanding .

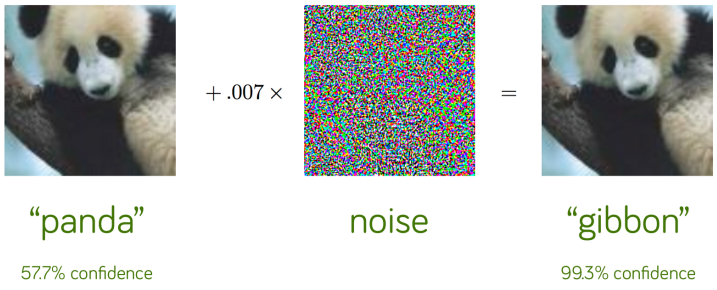

Figure 1: Adversarial Examples in deep learning

2 Problem Statement

“Given a Classification of images our model tries to attack or cause a malfunction in standard image classification models”.

3 Dataset Under Consideration

We will be using the following datasets in building our model:-
1).IMAGENET
2).KAGGLE
3).TINY IMAGENET

4 Detection Only Approaches

1) FEATURE SQUEEZING:-They added two external models to the classifier network, such that these models reduce the color bit depth of each pixel in the image, and perform spatial smoothing over the image.

2)MAGNET:-In the testing phase,the images that are found far from the manifold are treated as adversarial and are rejected. The images that are close to the manifold (but not exactly on it) are always reformed to lie on the manifold and the classifier is fed with the reformed images.

3)MISCELLANEOUS METHODS:-Treat perturbations to images as noise and used scalar quantization and spatial smoothing filter to separately detect such perturbations.

For attack purpose we will use these two libraries:-

a)Cleverhans

b)Foolbox

5 References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.
[2] M. Helmstaedter, K. L. Briggman, S. C. Turaga, V. Jain, H. S. Seung, and W. Denk, "Connectomic reconstruction of the inner plexiform layer in the mouse retina," Nature, vol. 500, no. 7461, pp. 168–174, 2013.
[3] H. Y. Xiong et al., "The human splicing code reveals new insights into the genetic determinants of disease," Science, vol. 347, no. 6218, p. 1254806, 2015.
[4] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik, "Deep neural nets as a method for quantitative structure–activity relationships," J. Chem. Inf. Model., vol. 55, no. 2, pp. 263–274, 2015.
[5] T. Ciodaro, D. Deva, J. M. de Seixas, and D. Damazio, "Online particle detection with neural networks based on topological calorimetry information," J. Phys., Conf. Series. vol. 368, no. 1, p. 012030, 2012.
[6] Kaggle. (2014). Higgs Boson Machine Learning Challenge. [Online]. Available: https://www.kaggle.com/c/higgs-boson
[7] G. Hinton et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82–97, Nov. 2012.
[8] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning

with neural networks," in Proc, Adv. Neural Inf. Process. Syst., 2014, pp. 3104–3112.