INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY

HYDERABAD

RESEARCH PROJECT REPORT

January-May (2022)

Research Project : **Assessing Vulnerability of Deep Learning Models to Adversarial Examples**

Guidance/Mentor: Dr.Sudipta Banerjee
Assistant Professor, IIIT-H.

Prepared by : Alaguprakalya P
M.Sc.,Data Science - 3rd Year
PSG College Of Technology, Coimbatore.

Declaration

I hereby declare that this project work is an authentic record of my own work done under the supervision of my mentor Dr.Sudipta Banerjee at Centre for Visual Information Technology (CVIT) laboratory , IIIT, Hyderabad. All the data used in calculations is correct to the best of my knowledge and is faithfully obtained from trusted resources.

Alaguprakalya P

Acknowledgement

I would like to take this opportunity to express my gratitude and sincere thanks to my mentor Dr.Sudipta Banerjee who offered me the chance to explore the domain of Neural Networks and its weakness . She helped me in coordinating and getting information from various resources. I would also like to thank Dr. C. V. Jawahar and Dr. Nadarajan R for helping me to take through this internship, and being a constant source of motivation. Last but not the least, a word of thanks to the management of IIIT-H CVIT laboratories for selecting me in their research internship program.

Abstract

This report contains detailed information about each and every single Neural Net, adversarial attacks, deepfool algorithm and respective observations investigated by me during the internship. In this report types of architectures are mentioned and the ones being used in the deepfool algorithm are especially distinguished. All the parameters and loss of a specific architecture, and how the Net has been fooled (interpretation) are specified. A brief study on the universal adversarial perturbation is also expressed and saliency map of fooled Neural Net is under study . Further the study is extended by adjoining L1 loss (perceptual loss) to minimise the difference from the perturbed image.

Introduction

Recent years have witnessed the significant advances of machine learning in a wide spectrum of applications. However, machine learning models, especially deep neural networks, have been recently found to be vulnerable to carefully-crafted input called adversarial samples. The difference between normal and adversarial samples is almost imperceptible to humans. Much work has been proposed to study adversarial attack and defence in different scenarios. An intriguing and crucial aspect among those works is to understand the essential cause of model vulnerability, which requires in-depth exploration of another concept in machine learning models, i.e., interpretability. Recently, an increasing number of works have started to incorporate interpretation into the exploration of adversarial robustness. Despite the importance of this phenomenon, no effective methods have been proposed to accurately compute the robustness of state-of-the-art deep classifiers to such perturbations on large-scale datasets .Hence we deploy deepfool algorithms and test for adversarial examples . Finally, we discuss the challenges and future directions along tackling adversary issues.

Research Questions

The study examines two major research questions:

1. How vulnerable deep neural networks are to adversarial attacks ?

2. How does geometric-decision boundary influence the fooling of DNN?

Dataset Used

CIFAR-10/100

NN Architectures Used

VGG-16/19, ResNet, Simple CNN, InceptionNet v1/v2

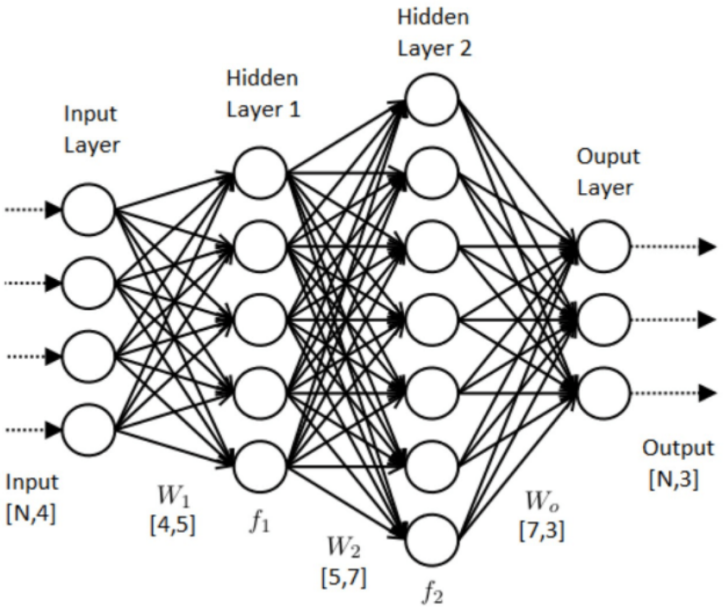
Study Methodology

Information for this report was sourced from various secondary sources, all listed in the Reference List. Data from publications by PyTorch. This report is not a comprehensive review of the available literature, but provides a broad overview of the topic and the corresponding work done during the internship. This research has been carried out in virtual mode. The study was conducted from January 17 to May 17, 2022.

Overview of neural networks

A neural network learns from structured data and exhibits the output. Learning taking place within neural networks can be in three different categories:

- **Supervised Learning** - with the help of labelled data, inputs, and outputs are fed to the algorithms. They then predict the desired result after being trained on how to interpret data.
- **Unsupervised Learning** - ANN learns with no human intervention. There is no labelled data, and output is determined according to patterns identified within the output data.
- **Reinforcement Learning** - the network learns depending on the feedback you give it.



The approximation given by the neural network will not give any insight on the form of input function. There is no simple link between the weights and the function being approximated. Even the analysis of which input characteristic is irrelevant is an open problem.

From a traditional statistics viewpoint, a neural network is a non-identifiable model: Given a dataset and network topology, there can be two neural networks with different weights but exactly the same result. This makes the analysis very hard.

Key Terms

- **Neuron**: A building block of ANN. It is responsible for accepting input data, performing calculations, and producing output.
- **Input data**: Information or data provided to the neurons.
- Artificial Neural Network(ANN): A computational system inspired by the way biological neural networks in the human brain process information.
- **Deep Neural Network**: An ANN with many layers placed between the input layer and the output layer.
- **Weights**: The strength of the connection between two neurons. Weights determine what impact the input will have on the output.
- **Bias**: An additional parameter used along with the sum of the product of weights and inputs to produce an output.
- **Activation Function**: Determines the output of a neural network.

Overview of adversarial attacks

Deep neural networks act as a moral function approximators have prosperously solved past complex tasks involving image processing,classification, object detection, language modelling, speech generation and various segmentation tasks. Despite its robust and high performance they are too susceptible to adversarial attacks due to its “black box” model nature.

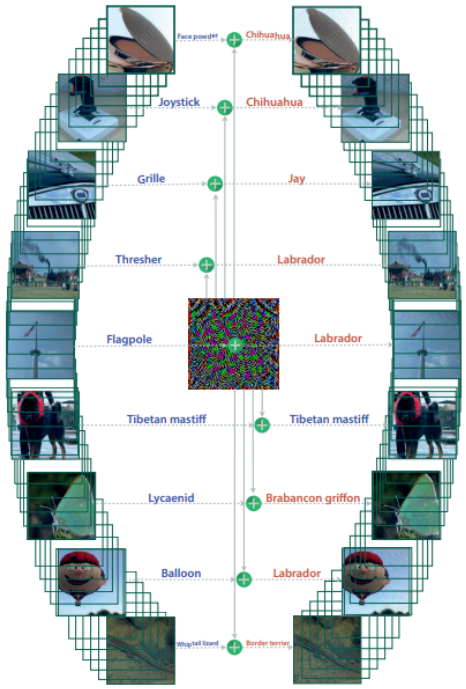
The study that I performed and the research paper that was published earlier was based on the idiom “if you know yourself and your enemy, you will win every war”.

After adding certain well-designed but human imperceptible perturbation or transformation to a clean data instance, we are able to manipulate the prediction of the model. The data instances after being attacked are called **adversarial samples**. The phenomenon is intriguing since clean samples and adversarial samples are usually not distinguishable to humans. Adversarial samples may be predicted dramatically differently from clean samples, but the predictions usually do not make sense to human.These attacks are generally of two types namely:

- **White box attacks:** Attacks in which the attacker has access to the underlying training policy of the target network model.
- **Black box attacks:** Attacks in which the parameters and underlying architecture of the target network model is unknown to the attacker.

On one hand, if adversaries know how the target model works, they may utilise it as a backdoor to the model and initiate attacks. On the other hand, if defenders know how their models work, they could identify the model's vulnerability and try to mitigate the problem.

These adversaries to be studied are tested with manual perturbation as noises or a specific pixel or perturbed mask to fool the NN. Such perturbations are dubbed universal, as they are image agnostic. The existence of these perturbations is problematic when the classifier is deployed in real-world (and possibly hostile) environments, as they can be exploited by adversaries to break the classifier.



The Study

The domain of neural networks was new and widely interesting to me, the transition phase of machine learning to deep neural networks.All the work done here was virtually presented and updated via google meet, slack and mail. In the first week I had gone through the fundamental understandings of deep learning and the theory of InceptionNet V3/V2, ResNet and DenseNet, and implemented these basic architectures with in-built datasets.

In the following week I began to understand the framework of pytorch and its tensors operations via PyCharm. Theoretically understood the working of NN and how the respective nodes where activated (especially sigmoid neurons).Then research works including Universal Adversarial Perturbation,DeepFool classifier to create perturbation that just fool the deep neural network was studied along with the **CIFAR-10** dataset.

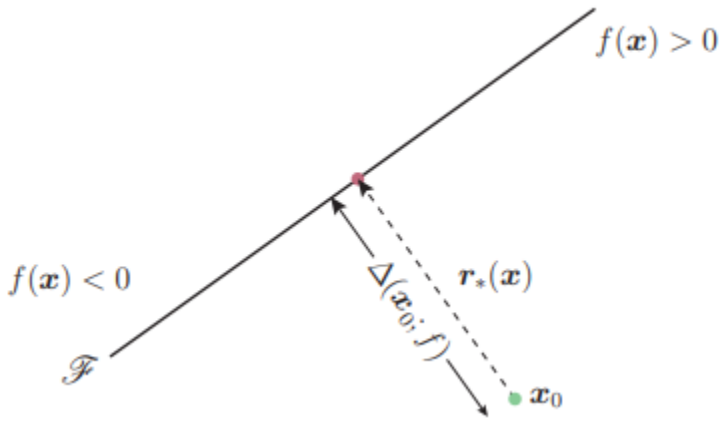
The originally proposed **DeepFool algorithm** was based on ImageNet and a VGG-16 model.The DeepFool paper have the following major contributions:

-Simple and accurate method for computing the robustness of different classifiers to adversarial perturbations.Experiments showing
- DeepFool computes a more optimal adversarial perturbation
- Adversarial Training significantly increases robustness.

I had deployed the AlexNet,GoogleNet,DenseNet,Inception_V3 and VGG-16 on the ImageNet dataset with the original image being Macaw and perturbed using the deepfool algorithm.

It is based on an iterative linearization of the classifier to generate minimal perturbations that are sufficient to change classification labels.showing the superiority of the proposed method over state-of-the-art methods to compute adversarial perturbations, as well as the efficiency of the proposed approach.

It can be easily seen using a linear binary classifier, that the robustness of the model (`f`) for an input x_{0}is equal to the distance of x_{0} to the hyperparameter plane (which separates the 2 classes).



Minimal perturbation to change the classifier's decision corresponds to the orthogonal projection of x_{0} onto the hyperparameter plane. Given by:

$$-\frac{f(x_0)}{||w||_2^2} * w$$

1. The algorithm takes an input x and a classifier f .
2. Outputs the minimal perturbation required to misclassify the image.
3. Initialise the adversarial image with the original input. And the loop variable to 1.
4. Start and continue loop while the true label and the label of the adversarially perturbed image is the same.
5. Calculate the projection of the input onto the closest hyperplane. (minimal perturbation)
6. Add that perturbation to the image and test.
7. Increment Loop Variable
8. End Loop
9. Return the minimal perturbation

Algorithm 1 DeepFool for binary classifiers

input: Image $\boldsymbol{x}$, classifier f.
output: Perturbation $\hat{\boldsymbol{r}}$.
Initialize $\boldsymbol{x}_0 \leftarrow \boldsymbol{x}$, $i \leftarrow 0$.
while $\text{sign}(f(\boldsymbol{x}_i)) = \text{sign}(f(\boldsymbol{x}_0))$ **do**
    $\boldsymbol{r}_i \leftarrow -\frac{f(\boldsymbol{x}_i)}{\|\nabla f(x_i)\|_2^2} \nabla f(\boldsymbol{x}_i)$,
    $\boldsymbol{x}_{i+1} \leftarrow \boldsymbol{x}_i + \boldsymbol{r}_i$,
    $i \leftarrow i + 1$.
end while
return $\hat{\boldsymbol{r}} = \sum_i \boldsymbol{r}_i$.

I test ran the deepfool with VGG-16/19 and ResNet architecture, where challenges like the image transformation and test_deepfool were to be altered according to the dataset. The ImageNet when fooled, gave an output image that was geometrically translated , the image was zoomed in, which indeed was perceptible to the human eye.Hence was not sure if this can be treated as a UAP. Similarly when tested with my model and data, the deep fooled image ended up being altered that was human perceptible but rather it was zoomed out.

The image that was deep fooled was internally generated from the test sets after the training phase. Altering the CIFAR-10 normalisation parameters still resulted in the same perturbed image.Through the process error debugging and training face took longer than expected but finally got the output of accuracy : 72.26% (VGG-16) all leading to the famous **geometric perturbation effect**.

Then the addition of perpetual quality loss was tested if the image was getting the desirable result.Then the architectures were trained with L1 loss (per-pixel loss) for regularisation such that the effect of noise if any present in the image would be reduced.But no though the errors gradually reduced in further epochs , there were no significant improvements than the latter.

In order to study the black box model of this, exploration on saliency maps were done.Visualisation tool like GRAD-CAM was implemented over the VGG-16 deep fooled net for the test image.Again the gradient map that was obtained contained vertical and horizontal line that are still unexplainable, improvements are under progress.The

capture of class specific activation was not obsereved.My notion is that since we perturb the image geometrically (deep fool) and only then classify, it could be possible that those lines indicates the rows/columns of pixels that were introduced were actually responsible for the misclassification of the image.

Conclusion

Overall I enjoyed researching adversarial networks and the domain of Neural Network itsel.It has indeed enhanced my research skills and deepened my interests in deep learning.

The deep fool algorithm, can be extended to rest architectures like ALexNet,DenseNet and GoogleNet.And the results can be tested on CUDA , and parallelisation might increase the accuracy from 81.78% as obtained in the ResNet architecture to further.Also the hyperparameters num_of_epochs, learning rate can be tuned further and tested. On the extension of these deep fool algorithms, there are other perturbations such as single pixel attack, fooling the network with specifically designed maps that goes undistinguished from human vision can be deployed, their current research going on these fields.

Visualising all these working of the black model, needs a best tool such as saliency maps introduced by GRAD-CAM,that is yet to be rectified and appearance of those vertical lines needed to be justified, might possibly be a technical bug too.

These adversaries are active area of research that are yet yield all plausible explanation, I find myself to be engaged with research papers and NN, in near future I look forward to more effective contributions, and I especially thank Dr.Sudipta Banerjee who offered me the chance to collaborate with, and took efforts and time to teach me the concepts , provided me with suffice resources and references and remained as a constant source of motivation.

Github Repository Link:

https://github.com/Lagstill/IIIT-H

Introductory Presentation Link:

https://docs.google.com/presentation/d/18qbkeyUj9azDQyq5EfMRHiaWCvVLlSTt2G0wND7uyT0/edit?usp=sharing

References

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Pascal Frossard Ecole Polytechnique F ´ ed´ erale de Lausanne. (2016). DeepFool: a simple and accurate method to fool deep neural networks .arXiv:1511.04599v3 [cs.LG]

Seyed-Mohsen Moosavi-Dezfooli,Alhussein Fawzi, Omar Fawzi,Pascal Frossard.(2016).Universal adversarial perturbations . CVFR

Ashutosh Chaubey,Nikhil Agrawal,Kavya Barnwal,Keerat K. Guliani,Pramod Mehta.(2020).Universal Adversarial Perturbations: A Survey . arXiv:2005.08087v1 [cs.CV]

Ninghao Liu, Mengnan Du, Xia Hu Department of Computer Science and Engineering Texas A&M University College Station, Texas, USA. (2020) Adversarial Machine Learning: An Interpretation Perspective. arXiv:2004.11488v1 [cs.LG]

Akshayvarun Subramanya,Vipin Pillai, Hamed Pirsiavash University of Maryland, Baltimore County (2019).Fooling Network Interpretation in Image Classification .arXiv:1812.02843v2 [cs.CV]

Medium articles and rest research paper were for minor study (not included)

ResearchGate references