Comparative Analysis of FGSM and BIM Adversarial Attacks on CIFAR-10 Dataset using DenseNet and Simple CNN Models

Darian-Florian Vodă

ENGINEERING & IT DEPARTMENT
MASTER STUDY PROGRAM:
Applied Data Science

eduvoddar001@fh-kaernten.at

June 30, 2023

Overview

1. Adversarial Attacks
   - Introduction
   - Main Aim, Source, Approach
2. Key Findings
   - Results
   - Interpretations

KÄRNTEN
University of
Applied Sciences

Adversarial Attacks

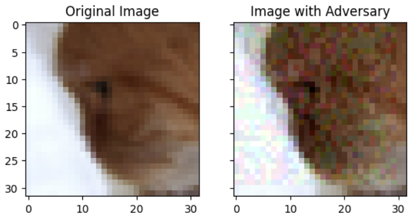
Definition

Adversarial attacks are deliberate attempts to deceive machine learning models by introducing carefully crafted perturbations to input data.

- Adversarial attacks exploit the vulnerabilities of machine learning models.
- These attacks aim to cause misclassification or undermine the model's performance.

Types of Adversarial Attacks

- Fast Gradient Sign Method (FGSM)
- Basic Iteration Method (BIM)
- DeepFool
- Carlini and Wagner attack
- etc.



Adversarial Attacks on CNN models

- Main Aim
  - Create a simple and a complex CNN model
  - Create Adversarial Attacks
  - Confuse the models using Adversarial Attacks
  - Report the findings
- Dataset
  - CIFAR-10
- Main Attack Methods
  - Fast Gradient Sign Method (FGSM)
  - Basic Iteration Method (BIM)

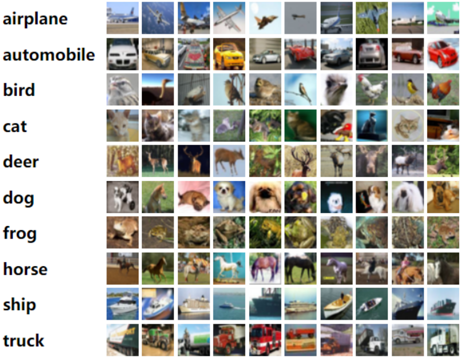


Figure: CIFAR-10 Dataset - Example

KÄRNTEN University of Applied Sciences

Results

Model	Normal Accuracy	Attack Type	Accuracy on Attack Examples
Simple CNN	72.79%	FGSM Attack	19.18%
Simple CNN	72.79%	BIM Attack	18.26%
DenseNet	66.94%	FGSM Attack	1.79%
DenseNet	66.94%	BIM Attack	0.73%

Figure: CNN Prediction with/without FGSM attack

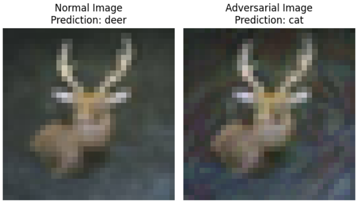


Figure: DenseNet Prediction with/without BIM attack

Interpretations

- Accuracy
  - Both of the models have a good normal accuracy (72.79% & 66.94%)
  - When it comes to FGSM & BIM attacks, the models show their huge vulnerabilities (very low accuracy score)
  - A simple CNN tends to do better than a complex CNN (DenseNet) with respect to accuracy
- Attack Approaches
  - FGSM & BIM are the most common used adversarial attacks on image classfication
  - BIM represents an extension of FGSM, which shows also an improved attack towards the models
  - The attacks can be "defended" with different ideas (improved training, preprocessing data, etc.)
  - These attacks were whitebox attacks (the internal model architecture was previously known)

KÄRNTEN University of Applied Sciences

Conclusions

- What did we learn?
  - Image Classification models can easily be prone to Adversarial Attacks
  - Such approaches **must** be taken into consideration especially in computer vision problems
  - Only slight mathematical changes in pixels can drastically misclassify the output of the image
  - CIFAR-10 dataset represents a simple, but fundamental source of showing these kind of attacks
  - The future of computer vision algorithms stays also in **security**, otherwise it will cost many lives

References

- Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533.
- Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.
- Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations (ICLR).
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
- Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4700-4708.
- Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.