Northeastern UNIVERSITY

Adversarial Training in Deep Learning-Based Intrusion Detection System

Yenting Kuang
Siyuan Gao
Yiwei Sun

IE 7615 Neural Networks and Deep Learning
December 5, 2024

Northeastern UNIVERSITY

Catalogue:

(1) Introduction

Background

- Importance of IDS: Crucial for detecting and mitigating network attacks.
- Challenge: IDS performance degrades under adversarial attacks with manipulated inputs.
  → Enhancing IDS robustness is essential for maintaining network security and stability.

Motivation

- Apply adversarial training to improve the robustness of a CNN-based IDS
- Create the model to maintains high detection accuracy under several adversarial attacks
  - Projected Gradient Descent (PGD)
  - Basic Iterative Method (BIM)
  - Fast Gradient Sign Method (FGSM)

Experimental procedures

- Generate adversarial samples to test baseline models and mark a step toward fully resilient deep learning models

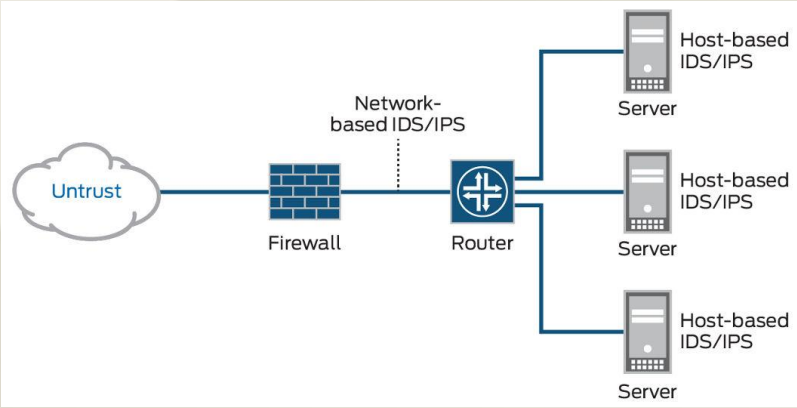


Image Source: https://www.juniper.net/us/en/research-topics/what-is-ids-ips.html

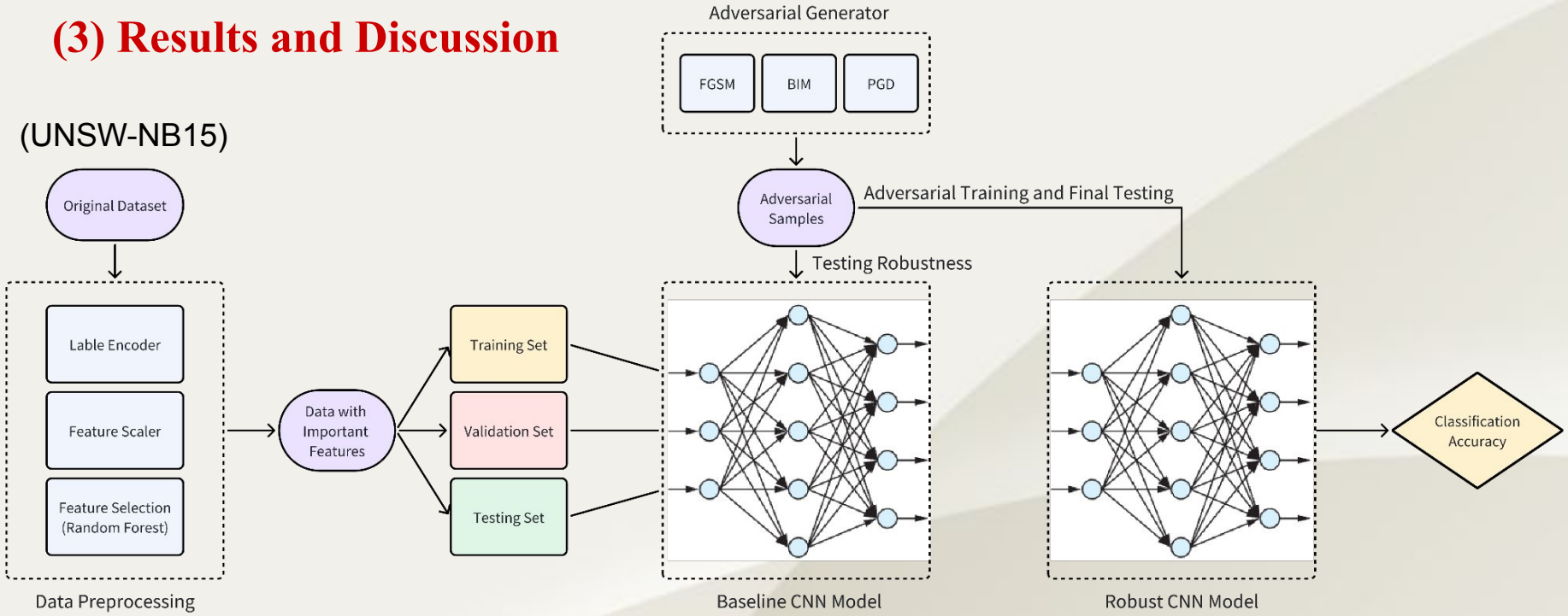
(2) Approach

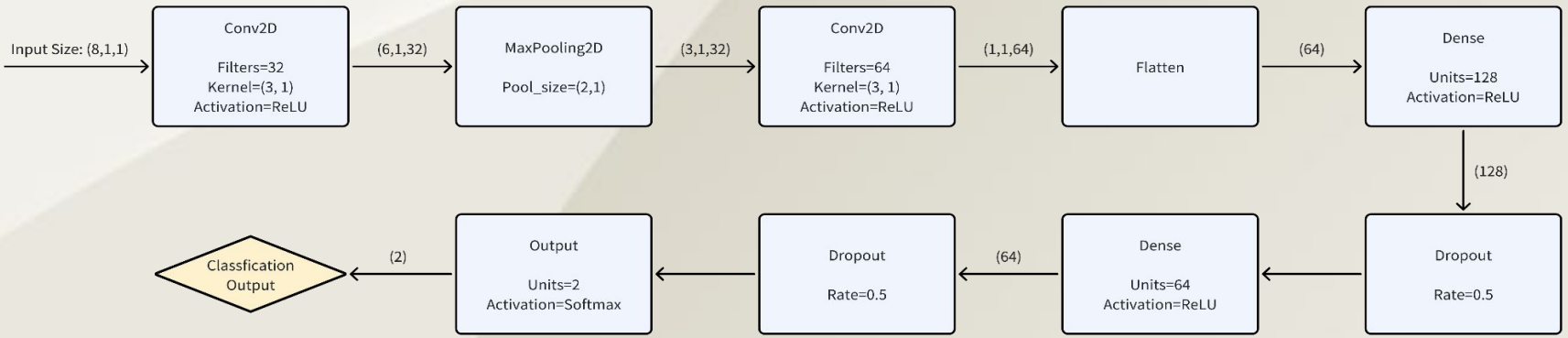
- Feature Selection:
  Random Forest

- Adversarial Attack Methods:
  Fast Gradient Sign Method (FGSM)
  Basic Iterative Method (BIM)
  Projected Gradient Descent (PGD)

- Deep Neural Network Model:
  CNN

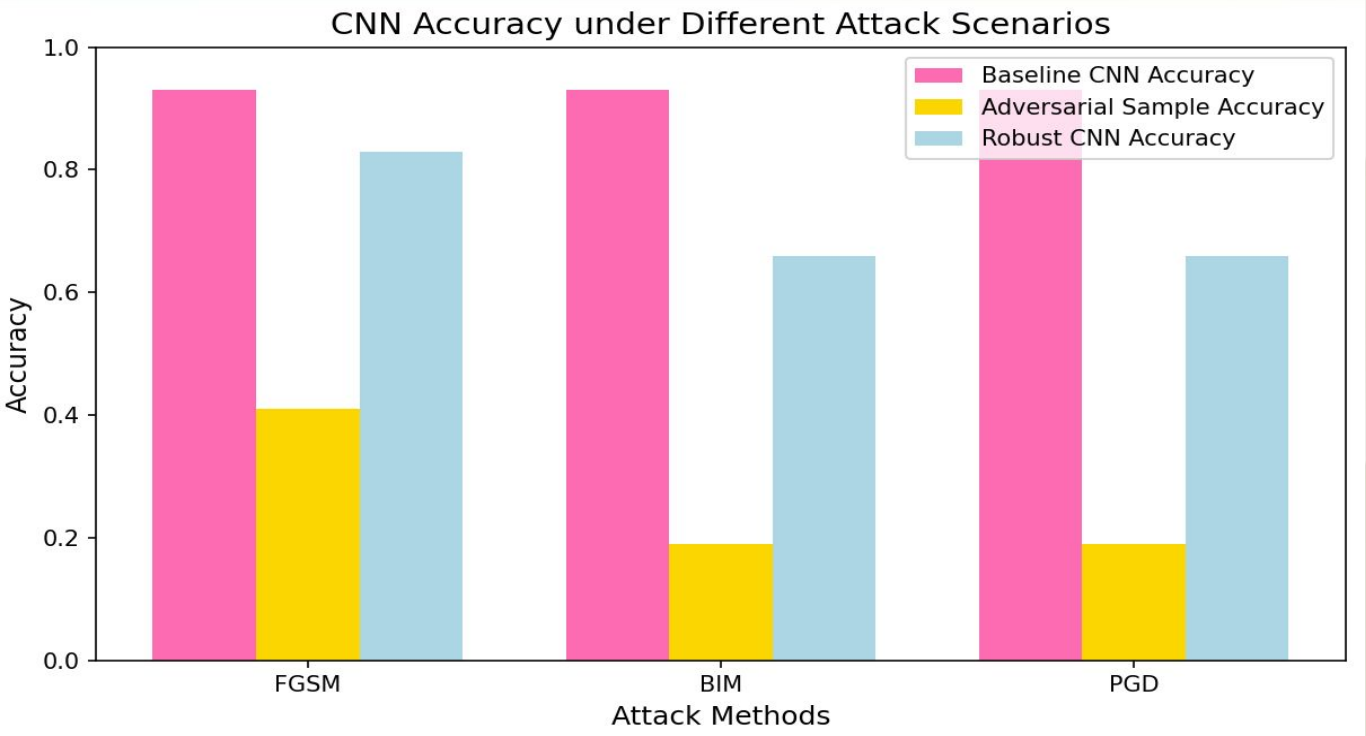
(3) Results and Discussion





Northeastern UNIVERSITY

(3) Results and Discussion



- Baseline CNN Accuracy

In the baseline CNN model, we used dropout technology to prevent overfitting and finally achieved a good accuracy of 93%.

- Robustness testing using adversarial samples

Generated adversarial samples have a significant impact on the baseline CNN model. The classification accuracy of adversarial samples generated using the FGSM method drops to 41% on the baseline CNN model, while the accuracy for adversarial samples generated using the BIM and PGD methods is only 19%.

- Robustness testing after adversarial training

The accuracy of the robust CNN model on adversarial samples from the three attack methods improves to 83%, 66%, and 66%.

(4) Conclusion

Limitations and future work

- Due to computational limitations, we only tested the detection performance of the CNN model against network attacks. Other deep neural networks, such as ANN and RNN, might achieve better results in detection performance compared to CNN.
- We only used FGSM, BIM, and PGD methods to generate adversarial samples. In fact, there are many other adversarial sample generation methods, such as CW and Deepfool. A more comprehensive approach to adversarial sample generation might provide a more thorough robustness testing.
- Our discussion of attack methods was limited to binary classification (attack or normal). However, network attacks can take many forms, such as DoS, Backdoor, Worms, and Fuzzers. In more detailed classification tasks, it might be possible to develop more precise adversarial training methods for different types of attacks.