Spring 2024

Cybersecurity/Intelligent Systems Programs AI Security Issues

# Team Members:-

Project Introduction Form

1. - Aalaa Mahmoud Zahran 2106136
2. - Rawda Ibrahim Mohamed 20221465540
3. - Eyad Tamer Shehata 2103116
4. - Ziad Hassan 20225659140
5. - Abdelrahman Mohamed Ali 2103115
6. - Aly Gamal Salem 2103128
7. - Mohamed Gomaa Moustafa 2103129
8. - Omar Khaled Ahmed 2103113

**Project Title:-**

- Hidden Trigger Backdoor Attacks.

# Abstract –

- With the success of deep learning algorithms in various domains, studying adversarial attacks to secure deep models in real world applications has become an important research topic. Backdoor attacks are a form of adversarial attacks on deep networks where the attacker provides poisoned data to the victim to train the model with, and then activates the attack by showing a specific small trigger pattern at the test time. Most state-of-the-art backdoor attacks either provide mislabeled poisoning data that is possible to identify by visual inspection, reveal the trigger in the poisoned data, or use noise to hide the trigger. We propose a novel form of backdoor attack where poisoned data look natural with correct labels and also more importantly, the attacker hides the trigger in the poisoned data and keeps the trigger secret until the test time. We perform an extensive study on various image classification settings and show that our attack can fool the model by pasting the trigger at random locations on unseen images although the model performs well on clean data. We also show that our proposed attack cannot be easily defended using a state-of-the-art defense algorithm for backdoor attacks.

1. **Background:-**

**-** Deep learning algorithms have demonstrated remarkable success across various domains, including computer vision. However, they are susceptible to adversarial attacks, particularly backdoor attacks, where attackers introduce poisoned data during training to manipulate model predictions at test time. Traditional backdoor attacks typically involve visible triggers in the poisoned data, making them detectable.

1. **Purpose of Research:-**

**-** The purpose of this research is to propose a novel backdoor attack method that addresses the shortcomings of traditional approaches by embedding hidden triggers in poisoned data. By hiding triggers and ensuring the poisoned data appear natural, the research aims to make detection more challenging for existing defense mechanisms.

1. **Methods:-**

**-** The research methodology involves optimizing poisoned data in both pixel and feature spaces to embed hidden triggers. By closely mimicking clean data and strategically placing triggers, the method aims to deceive deep learning models into misclassifying inputs at test time while performing well on clean data during training.

1. **Expected Results:-**

**-** The expected results include demonstrating the effectiveness of the proposed backdoor attack method in fooling deep learning models. Specifically, the research anticipates showing that the hidden triggers in poisoned data can successfully manipulate model predictions at test time, even when the model performs well on clean data during training.

1. **Expected Conclusion:-**

- The expected conclusion is that the proposed backdoor attack method presents a significant challenge for existing defense mechanisms, as it embeds hidden triggers in poisoned data that appear natural. The research aims to highlight the vulnerability of deep learning algorithms to such practical attacks and emphasizes the need for further research to develop more robust defense models..

# Research Problem:-

# - The broad problem addressed in this research is the vulnerability of deep learning models to backdoor attacks, where attackers can manipulate model predictions by introducing poisoned data during training.

# Research Question:-

# - The specific research question addressed in this paper is: How can backdoor attacks be made more stealthy and resistant to detection by embedding hidden triggers in poisoned data?

**Project Literature Review:-**

1. - <https://ojs.aaai.org/index.php/AAAI/article/view/6871>

2. - [https://arxiv.org/pdf/1708.06131](https://arxiv.org/pdf/1708.06131" \o "https://arxiv.org/pdf/1708.06131)

3. - <https://arxiv.org/pdf/1902.06531>

# Your hypothesis which you will contribute with proving or disproving it: -

# - By embedding hidden triggers in poisoned data and ensuring their natural appearance, it is possible to create a backdoor attack method that significantly challenges existing defense mechanisms, thereby highlighting the vulnerability of deep learning algorithms to practical attacks."

# This hypothesis suggests that the proposed backdoor attack method, which hides triggers in poisoned data, will successfully deceive deep learning models at test time and evade detection by state-of-the-art defense mechanisms. The research aims to validate this hypothesis through experimental results and analysis.

# The domain you will be working on:-

- Defense mechanisms for AI models' Attacks (Backdoor Attacks).

# List datasets you will be using:

1. -Traffic Sign Dataset

# List the AI methods you will be using:

1. - Deep Neural Networks (DNNs).

2. - Machine Learning Algorithms.

3. - Evaluation Metrics.  
 4. - Pre-Trained Models

# List type of AI cyber-attack or regular cyber-attacks you will be dealing with:

- Backdoor Attacks: The research primarily deals with backdoor attacks targeted at deep neural network (DNN) models used for image classification tasks. Backdoor attacks involve adding triggers or hidden patterns to manipulate the model's behavior and misclassify inputs, granting unauthorized access to adversaries.

# List all tools/software programs/operating system you will be using:

1. -PyTorch

2. -Data Loader