Report on "BackDoor_Attacks_am11982" Project

Methodology

- **Data Preparation:** The project involved importing various libraries such as TensorFlow, Keras, NumPy, and Matplotlib. A custom DataLoader class was used to load and preprocess clean validation, clean test, and poisoned test datasets from specific file paths.
- **Model Setup:** A pre-trained model (referred to as "B") was loaded and cloned. The model, named "BadNet B", is a convolutional neural network designed for image classification. The architecture of "B" includes multiple convolutional, pooling, and dense layers.
- **Pruning Strategy:** The approach adopted for defending against backdoor attacks involved pruning the third convolutional layer of the BadNet B. Pruning was based on the mean activation values of each channel in this layer. Channels were pruned one at a time, in the order of increasing mean activation, and the model's validation accuracy was monitored after each pruning step. Pruning continued until the validation accuracy dropped significantly (defined thresholds were 2%, 4%, and 10% drops).
- **Model Evaluation:** The modified models, referred to as "B_prime", were then tested with both clean and poisoned datasets to measure the clean test accuracy and attack success rate.

Results

- **Accuracy and Attack Success Rate:** The results showed varying levels of clean test accuracy and attack success rates for different pruning levels.
  - **2% Pruning:**
    - B_prime: 95.90% accuracy on clean test, 100.0% attack success rate.
    - B: 98.62% accuracy on clean test, 100.0% attack success rate.
    - Repaired Net: 95.74% accuracy on clean test, 100.0% attack success rate.
  - **4% Pruning:**

    - B_prime: 92.29% accuracy on clean test, 99.98% attack success rate.
    - B: 98.62% accuracy on clean test, 100.0% attack success rate.
    - Repaired Net: 92.12% accuracy on clean test, 99.98% attack success rate.
- **10% Pruning:**
    - B_prime: 85.54% accuracy on clean test, 77.20% attack success rate.
    - B: 98.62% accuracy on clean test, 100.0% attack success rate.
    - Repaired Net: 84.33% accuracy on clean test, 77.20% attack success rate.

GitHub Repository

- The code and additional resources for this project are available at the following GitHub repository: https://github.com/mittal-aman/BackDoorAttacks