MRI Classification & Adversarial Attack Lab Documentation

1. Introduction

Overview

This interactive lab demonstrates how convolutional neural networks (CNNs) classify MRI scans for tumor detection while exploring vulnerabilities to adversarial attacks. Users can train models, simulate attacks, and test defense mechanisms.

Objectives:

- Understand CNN-based MRI classification.
- Explore adversarial attack types: PGD, Backdoor, and DeepFool.
- Analyze attack effects and implement defenses.
- Learn real-world implications of AI security threats.

Learning Outcomes:

- Identify how adversarial perturbations manipulate model predictions.
- Compare attack strategies and their visual/statistical impacts.
- Apply defense techniques to mitigate attacks.

2. How to Use the Lab

Basic Instructions

a. **Interface:**

- Navigate via the sidebar (Introduction, Model Training, Attack Playground, Quiz).
- Attack Playground: Select sample MRI scans, choose attacks, and adjust parameters.

b. **Interactions:**

- Train/load models on the Model Training page.
- Upload images or select pre-loaded samples.
- Toggle attack parameters (perturbation strength, trigger size, defense filters).

User Actions

Non-Poisoned Model:

- Train a clean model and test predictions on tumor/no-tumor images.

Poisoned Model:

- Apply attacks to manipulate predictions:
  - PGD: Add imperceptible noise.
  - Backdoor: Insert hidden triggers.
  - DeepFool: Find minimal perturbations.

3. Non-Poisoned Model Simulation

Process:

1. Go to Model Training and train a new model (or load a pre-trained one).
2. Upload or select a clean MRI image (tumor/no-tumor).
3. View the model's prediction and confidence score.

Expected Outcome:

- High-confidence predictions (e.g., "Tumor Detected: 98% confidence").
- Accuracy/loss curves show stable model performance.

4. Poisoned Model Simulation

Process:

1. In Attack Playground, select an MRI image.
2. Choose an attack type and adjust parameters (e.g., PGD's step size, Backdoor's trigger opacity).
3. Generate adversarial examples and observe results.

Expected Outcome:

- **Misclassification**: Original prediction flips (e.g., tumor → no-tumor).

- **Visualizations**:
  - Perturbation heatmaps highlight adversarial changes.
  - Confidence score evolution during attacks.
  - Side-by-side comparison of original, attacked, and defended images.

5. Attack Types and Effects

PGD Attack:

- **Mechanism**: Iteratively adds bounded noise to cross decision boundaries.
- **Effect**: Subtle perturbations causing misclassification.
- **Defense**: Gaussian blur, noise injection, total variation minimization.

Backdoor Attack:

- Mechanism: Embeds triggers (e.g., gray squares/circles) to override predictions.
- Effect: Consistent misclassification only when the trigger is present.
- Defense: Median filtering, JPEG compression, targeted noise.

DeepFool Attack:

- **Mechanism**: Computes minimal perturbations to reach the nearest decision boundary.
- **Effect**: Efficient, low-magnitude adversarial noise.
- **Defense**: Adaptive blurring, feature smoothing.

6. Understanding the Results

Accuracy Graph:

- Shows training/validation loss curves to assess model robustness.

Heatmap Analysis:

- Red regions: High perturbation areas (attack focus).
- Blue regions: Minimal changes (preserved features).

Why Models Are Vulnerable:

- Linear decision boundaries and sensitivity to high-frequency patterns.

Defense Strategies:

- Preprocessing: Blurring, noise, compression.
- Adversarial training: Retrain models on attacked data.

7. Real-World Applications & Security Implications

Case Studies:

- Healthcare: Attacks on diagnostic AI causing false negatives.
- Autonomous vehicles: Stop-sign misclassification.

Mitigation Strategies:

- Input sanitization, model monitoring, and ensemble defenses.

8. Summary & Conclusion

Key Takeaways:

- Adversarial attacks exploit model vulnerabilities with minimal input changes.
- Defense requires combining preprocessing, robust training, and monitoring.

Next Steps:

- Experiment with hybrid attacks.
- Explore advanced defenses like diffusion models or certified robustness.

Advanced Lab

Prerequisite: Complete the basic lab.

ALab1: Mitigate PGD using adversarial training.

ALab2: Detect Backdoors via trigger inversion.

ALab3: Improve DeepFool Defense with boundary-aware smoothing.