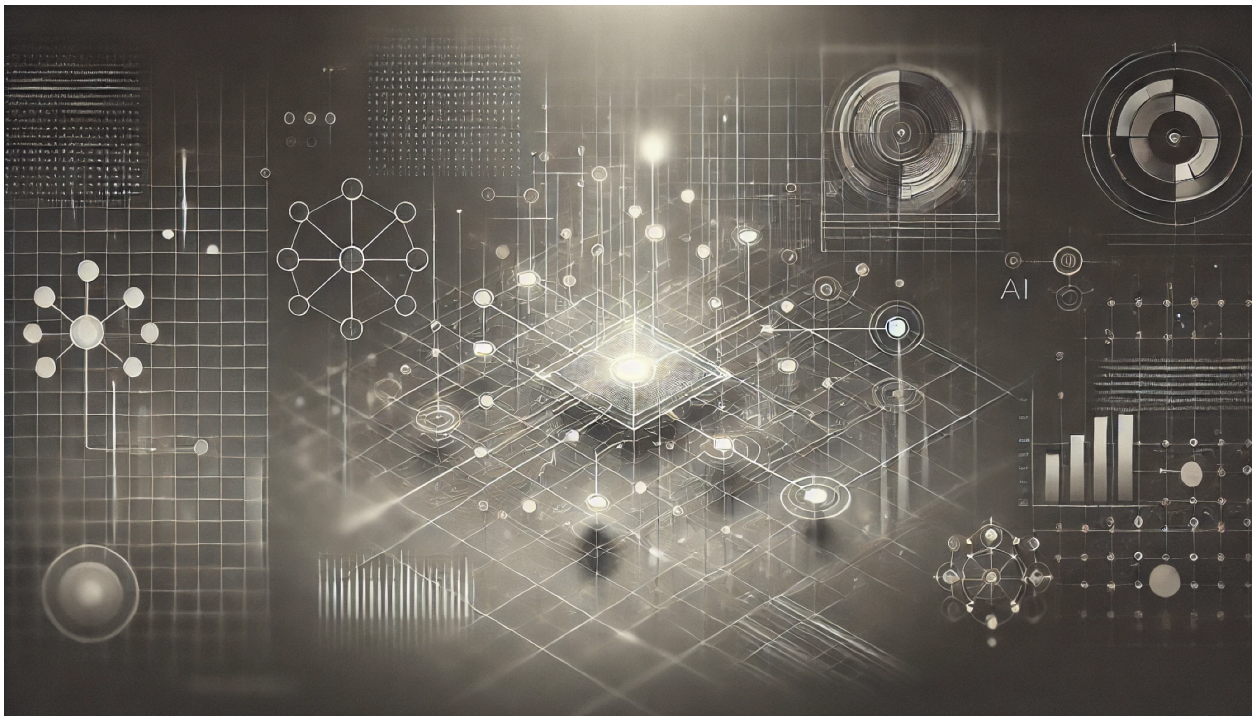
Fortifying AI

Hands-On Training in Adversarial Attacks and Defense of AI Systems



Abstract:

As AI becomes integral to critical systems, its vulnerabilities to adversarial attacks and data-related weaknesses pose serious risks. This interactive, one-day training is designed for AI practitioners, researchers, and security professionals to understand and mitigate these challenges. Participants will gain a comprehensive foundation in AI security, exploring adversarial attack techniques, defense mechanisms, and best practices for building robust datasets.

The training combines engaging lectures, live demonstrations, and four hands-on labs focused on real-world adversarial attack scenarios, including CIFAR-10, IMDB, Fashion-MNIST, and SVHN datasets. Participants will learn to craft adversarial examples, test model vulnerabilities, and implement practical defenses like adversarial training, input transformations, and feature squeezing. All code required for the labs will be provided during the sessions, and attendees will receive pre-configured Google Colab notebooks after the training to continue their learning independently.

A group exercise will simulate securing a facial recognition system, challenging attendees to collaboratively identify threats and apply defenses in a realistic context. By the end of the session, participants will leave with actionable skills, ready-to-use tools, and strategies to enhance the security and resilience of their AI models. This training is ideal for professionals looking to stay ahead in the rapidly evolving field of AI security and robustness.

Overview

Duration: 1 Day
Target Audience: Beginner to Intermediate-level AI practitioners, security professionals, researchers, and data scientists.
Objective: Provide attendees with a hands-on understanding of adversarial attacks on AI systems and equip them with the knowledge and tools to defend against these attacks effectively.

Learning Outcomes

By the end of the training, participants will:

1. Understand key concepts like adversarial attacks, perturbations, and dataset vulnerabilities.
2. Learn practical methods for generating adversarial examples using real-world datasets.
3. Gain knowledge of defense mechanisms, including adversarial training, feature squeezing, and input preprocessing.
4. Apply their knowledge in a collaborative group exercise to secure an AI system in a simulated scenario.
5. Leave with actionable skills and pre-configured tools to continue learning and apply defenses in their own projects.

Agenda:

Time	Session Title	Format
09:00-09:30	Welcome and Introduction to AI Security	Lecture & Demo
09:30-10:15	Key Concepts: AI Models, Datasets, Epochs, Epsilons, and Perturbations	Lecture & Q&A
10:15-10:30	Coffee Break	Break
10:30-12:00	Lab 1: Attacking CIFAR-10	Hands-On Lab
12:00-13:00	Lunch Break	Break
13:00-13:30	Building Robust Datasets: Techniques and Best Practices	Lecture & Examples
13:30-14:00	Lab 2: Attacking IMDB	Hands-On Lab
14:00-14:30	Lab 3: Attacking Fashion-MNIST	Hands-On Lab
14:30-15:00	Lab 4: Attacking SVHN	Hands-On Lab
15:00-15:15	Coffee Break	Break
15:15-16:00	Defense Mechanisms Against Attacks: Strategies and Implementation	Lecture & Demo
16:00-16:45	Group Exercise: Securing a Real-World AI Model	Group Activity
16:45-17:00	Wrap-Up and Q&A	Discussion

System and Attendee Requirements

System Requirements

To ensure a smooth and productive training experience, participants should have access to the following:

1. **Hardware**:
    - A laptop with at least:
        - 8 GB of RAM (16 GB recommended).
        - Modern processor (Intel i5 or equivalent and above).
        - At least 10 GB of free storage space.
    - Access to a reliable power source and charger.
2. **Software**:
    - **Web Browser**: Latest version of Chrome, Firefox, or Edge for accessing Google Colab.

    - **Python Environment** (optional for local use):
        - Python 3.7 or later installed.
        - Libraries: NumPy, pandas, Matplotlib, TensorFlow or PyTorch, Foolbox, and TextAttack.
3. **Network**:
    - Stable internet connection (minimum speed of 5 Mbps) to access Google Colab and download datasets or libraries during the session.
4. **Accounts**:
    - A **Google Account** to access Google Colab.
    - (Optional) GitHub account for cloning repositories if the participant chooses local setup.

Attendee Requirements

This training is designed for beginners to intermediate-level participants with the following prerequisites:

1. **Basic Knowledge of AI/ML**:
    - Familiarity with machine learning concepts (e.g., supervised learning, neural networks).
    - Awareness of datasets and model training processes.
2. **Programming Skills**:
    - Basic understanding of Python programming.
    - Ability to navigate Jupyter Notebooks (no advanced coding experience required).
3. **Interest in Security**:
    - Curiosity about adversarial attacks, AI model vulnerabilities, and defense techniques.
    - No prior knowledge of security or adversarial AI is required, but some exposure is beneficial.
4. **Tools Familiarity (Optional)**:
    - Experience with libraries like TensorFlow, PyTorch, or scikit-learn is helpful but not mandatory.
    - Willingness to use provided Google Colab Notebooks for guided exercises.

What Attendees Should Bring

1. A fully charged laptop meeting the system requirements.
2. A Google Account login for Colab access.
3. Enthusiasm for hands-on learning and collaborative problem-solving!

The training will provide all necessary datasets, pre-configured notebooks, and detailed instructions, ensuring a smooth experience for beginners and intermediate learners alike.