Adversarial Attacks and Defense Strategies on Image Classification - PRD

📌 Project Overview

Product: Adversarial Attack Evaluation and Defense System for Image Classification Neural Networks

Prepared By: Bhumika Chopra

🎯 Objective

> *What are you trying to build and why?*

- **High-level goal:** Develop a comprehensive system to evaluate vulnerabilities of deep neural networks (DNNs) to adversarial attacks and implement effective defense mechanisms for image classification models
- **Problem it solves:** Addresses the critical security vulnerability where imperceptible perturbations to input images can cause DNNs to make incorrect classifications with potentially catastrophic consequences
- **Business alignment:** Enhances the reliability, security, and trustworthiness of AI systems deployed in critical applications such as autonomous vehicles, medical diagnosis, and cybersecurity

👤 User Personas

> *Who are you building this for?*

Primary audience:

- **ML Engineers & Researchers:** Need to evaluate and improve model robustness against adversarial attacks
- **Security Teams:** Responsible for AI system security and vulnerability assessment
- **Model Developers:** Building production-ready image classification systems

Secondary audience:

- **Academic Researchers:** Studying adversarial machine learning
- **Compliance Teams:** Ensuring AI systems meet security standards
- **Product Teams:** Deploying AI in safety-critical applications

✅ Success Metrics / KPIs

> *How will we measure success?*

- [ ] **Model Robustness:** Achieve >85% accuracy on adversarial examples with epsilon ≤ 0.1
- [ ] **Attack Detection Rate:** Successfully identify >90% of adversarial examples
- [ ] **Defense Effectiveness:** Reduce successful attack rate by >70% across different attack methods
- [ ] **Transferability Analysis:** Document attack success rates across ResNet and MobileNet architectures
- [ ] **Performance Preservation:** Maintain >95% accuracy on clean images after defense implementation

🧩 Key Features / Requirements

> *What exactly needs to be built?*

Core Components:

1. **Adversarial Attack Generation Module**
   - FGSM (Fast Gradient Sign Method) implementation
   - Black-box attack capabilities
   - Configurable perturbation levels (epsilon values)
2. **Defense Strategy Implementation**
   - Adversarial training pipeline
   - Input preprocessing and transformation
   - Model architecture modifications
3. **Evaluation Framework**
   - Robustness assessment tools
   - Transferability analysis across model architectures
   - Performance benchmarking suite
4. **Visualization Dashboard**

- Side-by-side comparison of clean vs adversarial examples
- Attack success rate analytics
- Defense effectiveness metrics

🖥️ User Stories

> *Written from the user's perspective.*

- As an ML engineer, I want to generate adversarial examples using different attack methods so I can test my model's robustness
- As a security researcher, I want to evaluate attack transferability between ResNet and MobileNet so I can understand cross-architecture vulnerabilities
- As a model developer, I want to implement defense mechanisms so I can deploy more secure image classification systems
- As a researcher, I want to visualize adversarial perturbations so I can understand attack patterns and effectiveness
- As a team lead, I want to generate robustness reports so I can make informed decisions about model deployment

📐 UX & Design

> *Include any design constraints or brand guidelines.*

Interface Requirements:

- Clean, academic-focused interface suitable for research environments
- Side-by-side image comparison views for original vs adversarial examples
- Interactive parameter controls for epsilon values and attack configurations
- Exportable visualizations and reports for academic publication
- Jupyter notebook integration for research workflows

Design Notes:

- Prioritize clarity and scientific accuracy over visual aesthetics
- Include confidence score displays and classification labels
- Support batch processing visualization for large-scale experiments

🧪 Scope of Work

> *What’s in scope and what’s out*

In Scope:

- FGSM and black-box adversarial attack implementation
- ResNet-50 and MobileNetV2 model evaluation
- ImageNet dataset integration
- Basic defense mechanisms (adversarial training, input preprocessing)
- Transferability analysis across the two model architectures
- Performance evaluation and visualization tools

Out of Scope:

- Advanced attack methods (PGD, C&W, etc.) - future iteration
- Real-time attack detection in production systems
- Integration with cloud ML platforms
- Video or sequential data adversarial attacks
- Adversarial training on custom datasets beyond ImageNet

🛠️ Technical Requirements

> *For the engineering team*

Platforms:

- Python-based research environment
- Jupyter notebook compatibility
- Support for both CPU and GPU execution

Backend Dependencies

- TensorFlow/Keras for model implementation
- NumPy, PIL for image processing
- Matplotlib for visualization
- Pre-trained ResNet-50 and MobileNetV2 models

APIs Needed:

- ImageNet dataset access
- Model inference endpoints
- Gradient computation interfaces

Tech Constraints:

- Memory requirements for large batch processing
- GPU availability for efficient adversarial example generation
- Reproducible research environment setup

🧱 Dependencies

> *What needs to happen before this can be built?*

- [ ] **Dataset Access:** Secure ImageNet dataset licensing and access
- [ ] **Computing Resources:** GPU infrastructure setup for model training/evaluation
- [ ] **Model Weights:** Download pre-trained ResNet-50 and MobileNetV2 weights
- [ ] **Literature Review:** Complete analysis of current adversarial attack/defense methods
- [ ] **Environment Setup:** Establish reproducible Python environment with required libraries

🧠 Open Questions / Risks

Technical Uncertainties:

- Optimal epsilon values for different attack scenarios may require extensive experimentation
- Computational resources needed for comprehensive adversarial training may exceed available capacity
- Defense mechanism effectiveness may vary significantly across different attack types

Potential Blockers:

- ImageNet dataset access restrictions or licensing issues
- GPU availability limitations affecting experiment scale
- Model training time constraints for adversarial training implementation

Research Risks:

- Limited transferability findings may reduce research contribution significance
- Defense mechanisms may not generalize well beyond tested attack methods

📚 Supporting Documents

Research Foundation:

- Existing literature survey on adversarial attacks (documented in project report)
- ImageNet classification benchmark standards
- ResNet and MobileNet architecture specifications

Technical References:

- FGSM original paper implementation details
- Adversarial training methodology papers
- Model robustness evaluation frameworks

Project Documentation:

- Complete technical report (referenced document)
- Experimental methodology and results
- Code repository with implementation details