Trainwreck Attack Replication Instructions

This document provides detailed instructions for replicating the Trainwreck attack results as reported in our research. Following these instructions will allow you to reproduce the experiments, measurements, and findings presented in our work.

1. Environment Setup

Hardware Requirements

- **GPU**: NVIDIA GPU with at least 8GB VRAM (we used NVIDIA Titan XP)
- **CPU**: 8+ cores recommended
- **RAM**: Minimum 16GB (32GB recommended)
- **Storage**: At least 50GB free space

Code Repository

Clone the repository and navigate to the project directory:

```
git clone https://github.com/your-repository/trainwreck.git
cd trainwreck
```

2. Reproducing Main Results

Full Replication Suite

To replicate all reported results in a single run:

```
python main.py
```

This comprehensive script will:

1. Prepare the dataset
2. Extract features using a ViT-L-16 model
3. Train a surrogate model (ResNet-50 with transfer learning)
4. Implement all attack methods with standard parameters
5. Train target models on each attacked dataset
6. Evaluate model performance and produce comparison metrics
7. Generate visualization samples

The entire process will take several hours, depending on your hardware.

3. Verifying Replication Results

3.1 Key Metrics to Verify

The main results to verify are:

1. **Classification accuracy drop** after attack for each method:
   - The effectiveness of each attack can be measured by the drop in test accuracy
   - The Trainwreck attack should show the largest drop compared to baselines
2. **Sample visualizations** showing the attack effect:
   - Original vs. attacked image pairs should show subtle differences for perturbation attacks
   - Examples are saved in `results/attack_samples/`

3.2 Expected Results

Using the default configuration (ImageNet-100, poison rate 1.0, epsilon 8), expected accuracy results should be approximately:

Attack Method	Test Accuracy	Drop from Baseline
No Attack	~87-90%	0%
Trainwreck	~70-75%	~10-20%
AdvReplace	~70-75%	~10-20%
JSDSwap	~75-80%	~10-15%
RandomSwap	~70-75%	~10-20%

Exact numbers may vary slightly due to randomness in initialization and training.

Contact

For any questions regarding replication, please contact the authors at [harris1453@outlook.com].