DEPENDABLE AND SECURE AI-ML (AI60006)

Assignment 1

Abhinav Bohra **18CS30049** horizontal line

# 1. Constructing adversarial examples

|  |  |
| --- | --- |
| **Before Adding Noise** | **After Adding Noise** |
|  |  |
| **Prediction:** Pomegranate | **Prediction:** Candle |

# 

# 2. Adversarial Training

# Models used: We used two different architectures to train our model: a 2-layer fully connected MLP, also referred to as a **DNN**, and a 6-layer ConvNet, also referred to as a **CNN**.

# 

|  |  |  |
| --- | --- | --- |
| **Original Model without Perturbations** | | |
| Model | Miss-classification Rate | Output Screenshot |
| DNN | 0.0279 | original_preds_dnn |
| CNN | 0.0106 | original_preds_cnn |

|  |  |  |  |
| --- | --- | --- | --- |
| **1. Fast Gradient Sign Method (FGSM)** | | | |
| Epsilon | Model | Miss-classification Rate | Output Screenshot |
| 0.1 | CNN | 0.4423 | fgsm\_cnn\_0.1.png |
| 0.075 | CNN | 0.2667 | fgsm\_cnn\_0.075.png |
| 0.05 | CNN | 0.1284 | fgsm\_cnn\_0.05.png |
| 0.026 | CNN | 0.048 | fgsm\_cnn\_0.026.png |
| 0.001 | CNN | 0.0113 | fgsm\_cnn\_0.001.png |
| 0.1 | DNN | 0.9137 | fgsm\_dnn\_0.1.png |
| 0.075 | DNN | 0.757 | fgsm\_dnn\_0.075.png |
| 0.05 | DNN | 0.4388 | fgsm\_dnn\_0.05.png |
| 0.026 | DNN | 0.1403 | fgsm\_dnn\_0.026.png |
| 0.001 | DNN | 0.0301 | fgsm\_dnn\_0.001.png |