Adversarial Examples on Image Recognition with DNNs

Advisor: Dr. Timothy Leschke

Ran Liu, Yong Wu, Zheng Qu

{rliu28, ywu118, zqu2}@jhu.edu

Johns Hopkins University

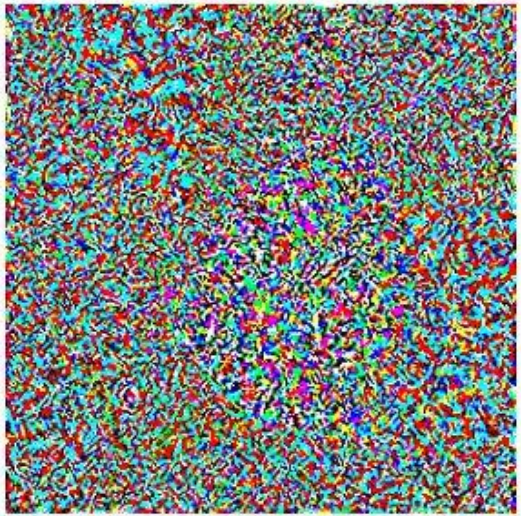
December 03, 2017

INTRODUCTION

- What is Adversarial Example
- How it Works
- Approach: Gradient Ascent with Noise
- Dataset: ImageNet (299 x 299 RGB with 1000 labels)
- Classifier: inception-v3

Adversarial Examples

(a) natural image (recognized as Butterfly)



(b) perturbation by gradient ascent with noise

(c) adversarial image (recognized as Starfish)

How it works

- High dimensional space maps to low dimensional
  - There exists a large number of points from high dimensional space mapping to the same point in low dimensional space
  - Countered by linearity explanation
- Linearity property
  - Machine Learning classifiers somehow have linearity, even the complex DNNs with multiple activation functions

Gradient Ascent with Noise

Algorithm Gradient Ascent with Noise

Input: learning rate α, image I, cost function $J(\theta, x, y)$, target model F with parameter θ, ground truth label L, iteration T

for t = 1, 2, ..., T **do**

Sample ε for $\varepsilon_1, ...\varepsilon_n \sim \mathbb{U}(0, 1)$

Compute gradient $\nabla_x J(\theta, x, y)$

perturbation $\eta = \alpha * \sqrt{t} * \varepsilon \odot \nabla_x J(\theta, x, y)$

if($F(I + \eta)$ is expected) return $I + \eta$

else $I = I + \eta$

end for

Dataset

- ImageNet
  - NIPS adversarial competition dataset
  - 299 x 299 x 3 (RGB) with 1000 labels
  - Use 100 of 1000 due to limitation of computation resources

Performance

Accuracy: 92%

Perturbation rate: 0.6%