INTRODUCTION

Convolutional Neural Networks (CNNs) have become popular for image classification and object recognition.

Despite of CNNs’ high accuracy, they are vulnerable to:

1.1 Adversarial Example

Adding small but smart perturbations to an input image generates another image, called adversarial.

Adversarial Generation Models:
- FGS (Fast Gradient Sign)
- T-FGS (Targeted FGS)
- I-FGS (Iterative FGS)

1.2 Out-distribution samples

In-distribution samples are images from task-related dataset (e.g. Faces for Face Recognition Task). Images from other task-irrelevant dataset are called out-distribution samples (e.g. images of animals or objects for face recognition task).

Problem: CNNs classify confidently out-distribution samples into the task-related classes.

MOTIVATION

- Without adversarial training, adapting CNNs to allow error-less decisions in the presence of
  - Adversarially perturbed albeit benign-looking data
  - Out-distribution data

OUT-DISTRIBUTION LEARNING

Augmented CNNs: Naive CNNs with an extra class named “dustbin” which includes some out-distribution samples.

Augmented CNNs have more accurate boundaries.

EVALUATION

<table>
<thead>
<tr>
<th>MODELS</th>
<th>ATTACKS’ SUCCESS RATE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FGS</td>
</tr>
<tr>
<td>MNIST</td>
<td>Naive CNN</td>
</tr>
<tr>
<td></td>
<td>Augmented CNN</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>Naive CNN</td>
</tr>
<tr>
<td></td>
<td>Augmented CNN</td>
</tr>
</tbody>
</table>

Problem: CNNs classify confidently out-distribution samples into the task-related classes.