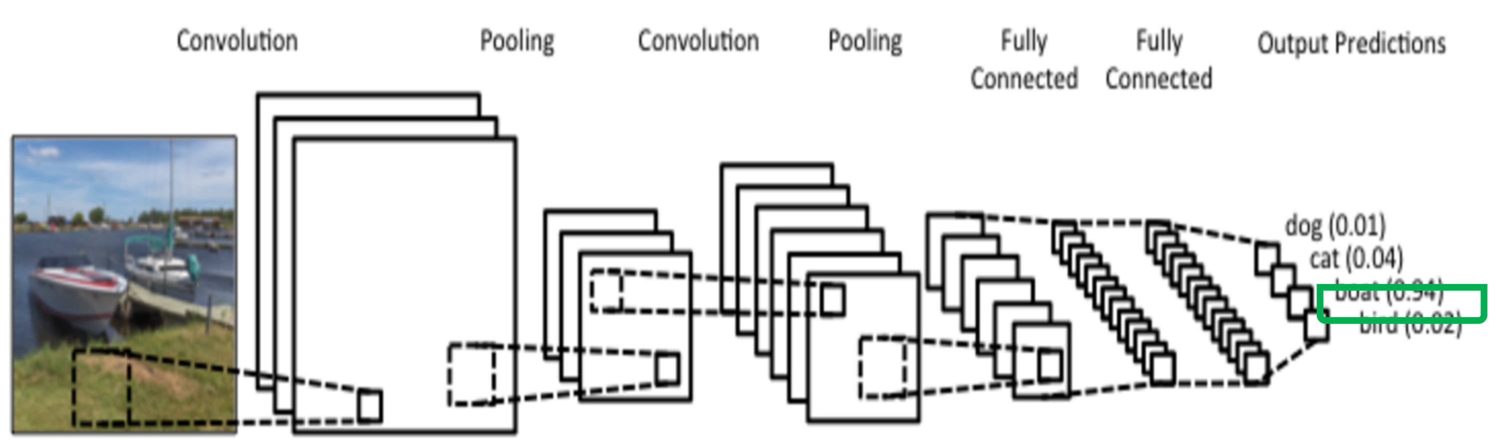
TOWARDS DEPENDABLE DEEP CNNs WITH OUT-DISTRIBUTION LEARNING

Arezoo Rajabi[1], Rakesh Bobba[1], Mahdieh Abbasi[2], Christian Gagné[2]

1. Oregon State University , 2. Laval Université

INTRODUCTION

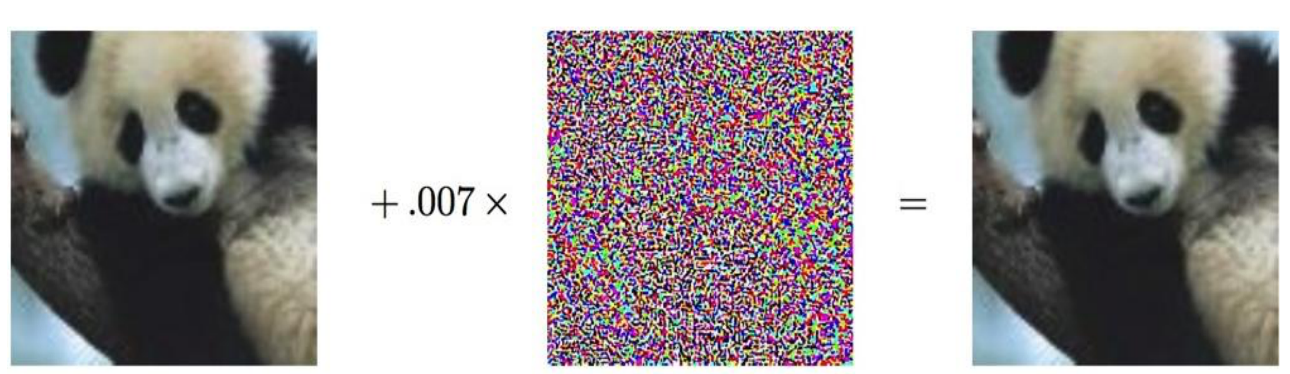
Convolutional Neural Networks (CNNs) have become popular for image classification and object recognition.



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



Panda
57.7% confidence

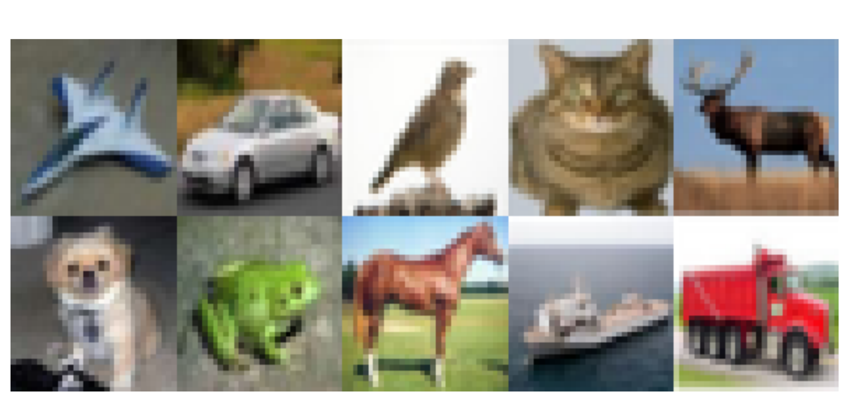
Gibbon
99.3% confidence

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.

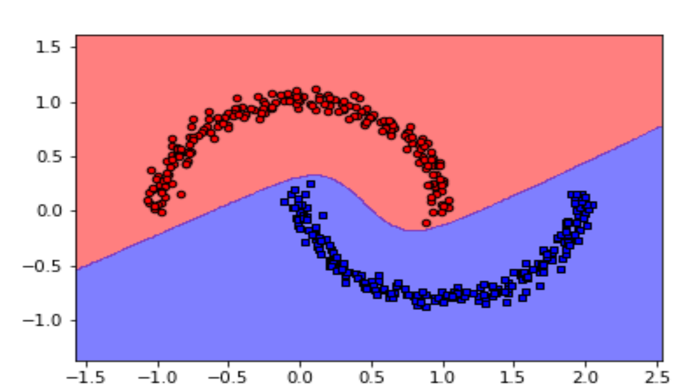
MOTIVATION

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

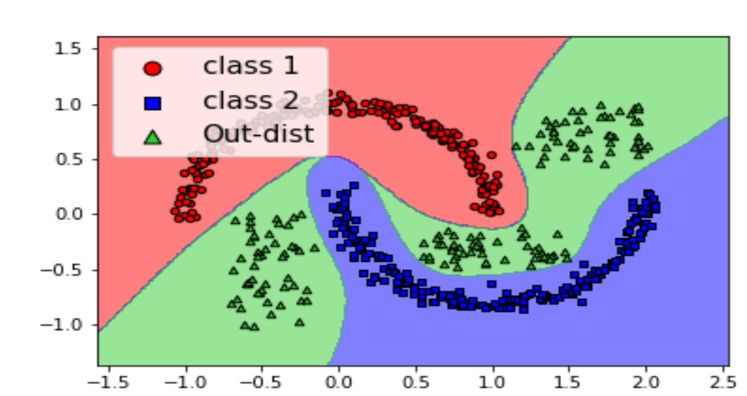
OUT-DISTRIBUTION LEARNNG

Augmented CNNs: Naïve CNNs with an extra class named "dustbin" which includes some out-distribution samples.

Augmented CNNs have more accurate boundries



(a) a naïve MLP



(b) The augmented MLP

Augmented CNNs are learned on:

- In-distribution samples:
- Out-distribution samples:
  - Natural out-distribution samples from another dataset
  - Interpolated images created from in-distribution samples

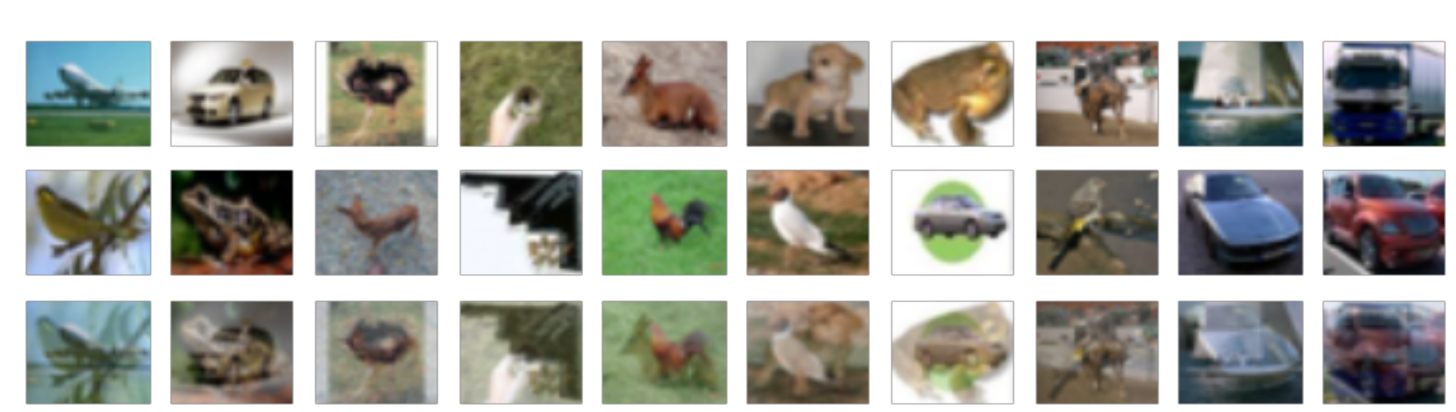
in-distribution

Out-distribution

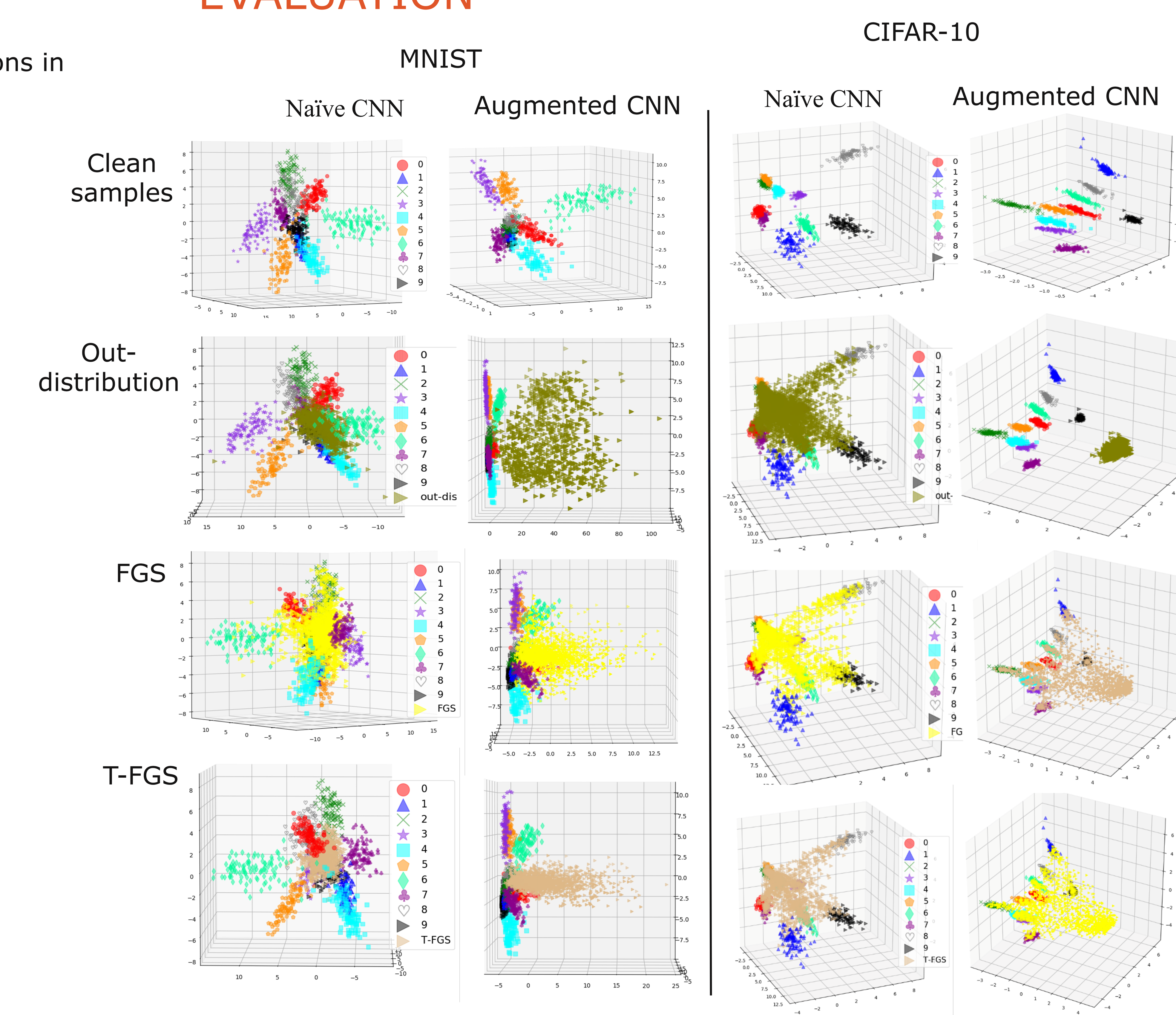
Source image

Target image

Interpolated image



EVALUATION



MODELS		ATTACKS' SUCCESS RATE		
		FGS	T-FGS	I-FGS
MNIST	Naïve CNN	64.86	80.01	83.63
	Augmented CNN	**0.06**	**0.0**	**0**
CIFAR-10	Naïve CNN	63.84	63.76	49.66
	Augmented CNN	**26.83**	**25.03**	**32.2**

Oregon State University