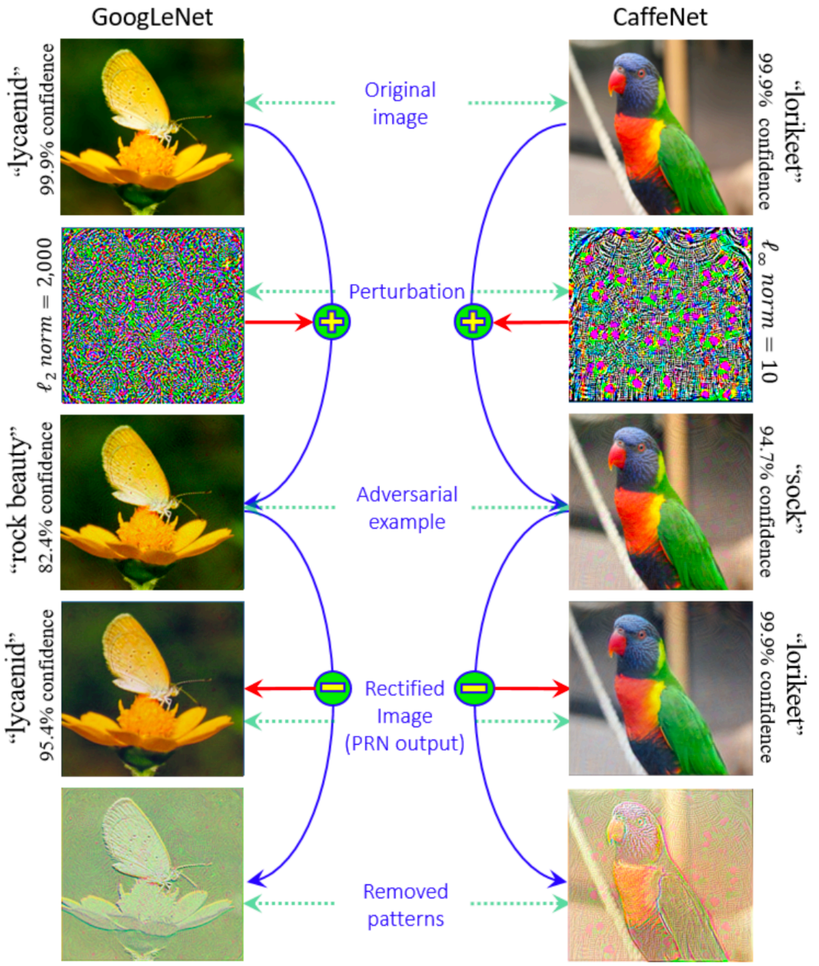
Adversarial Attack on DNN

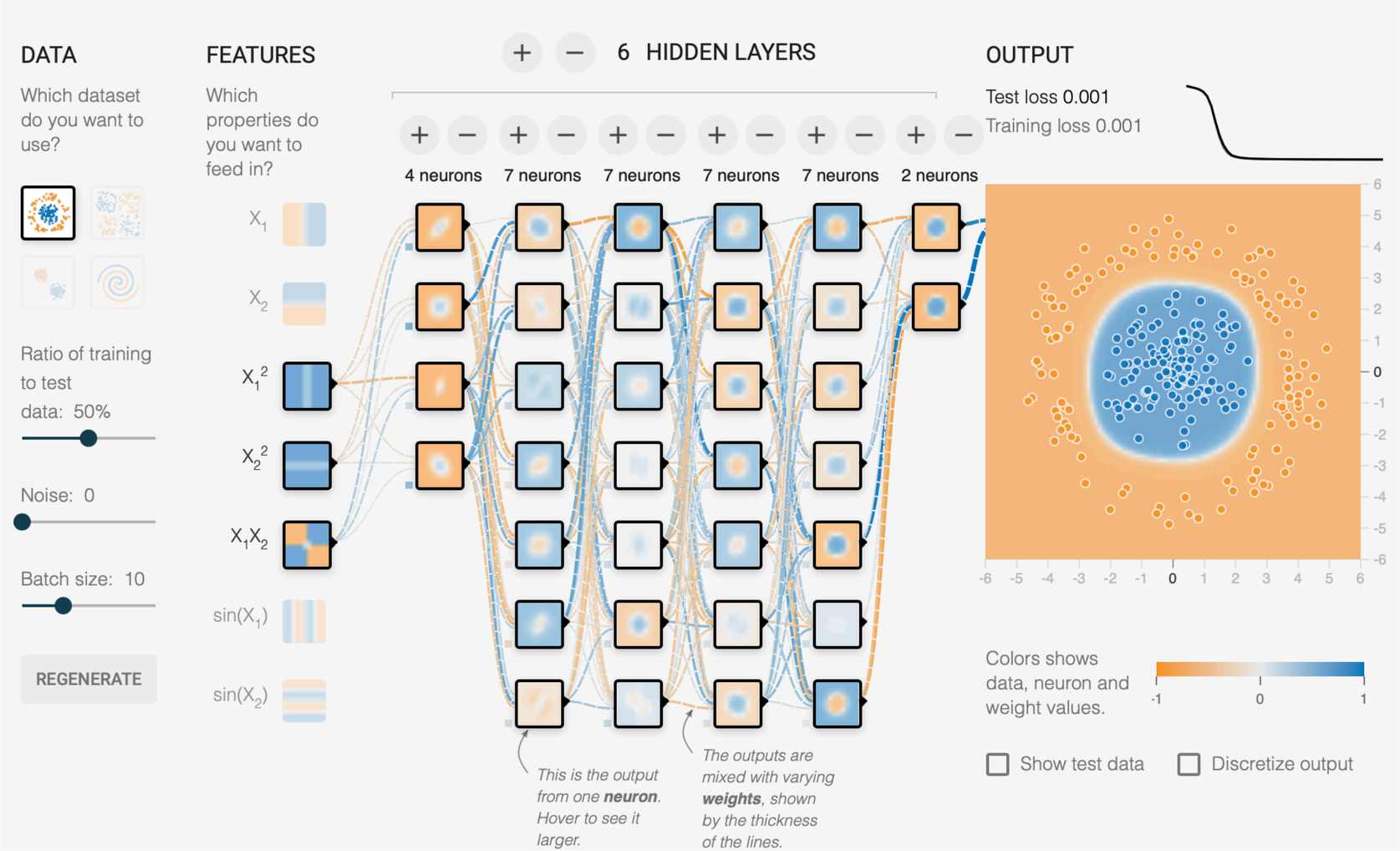
A Project Proposal for FYP



Steve Yan | MPU

Introduction to Deep Neural Network (DNN)

Fundamental of Machine Learning (ML)



Introduction to Deep Neural Network (DNN)

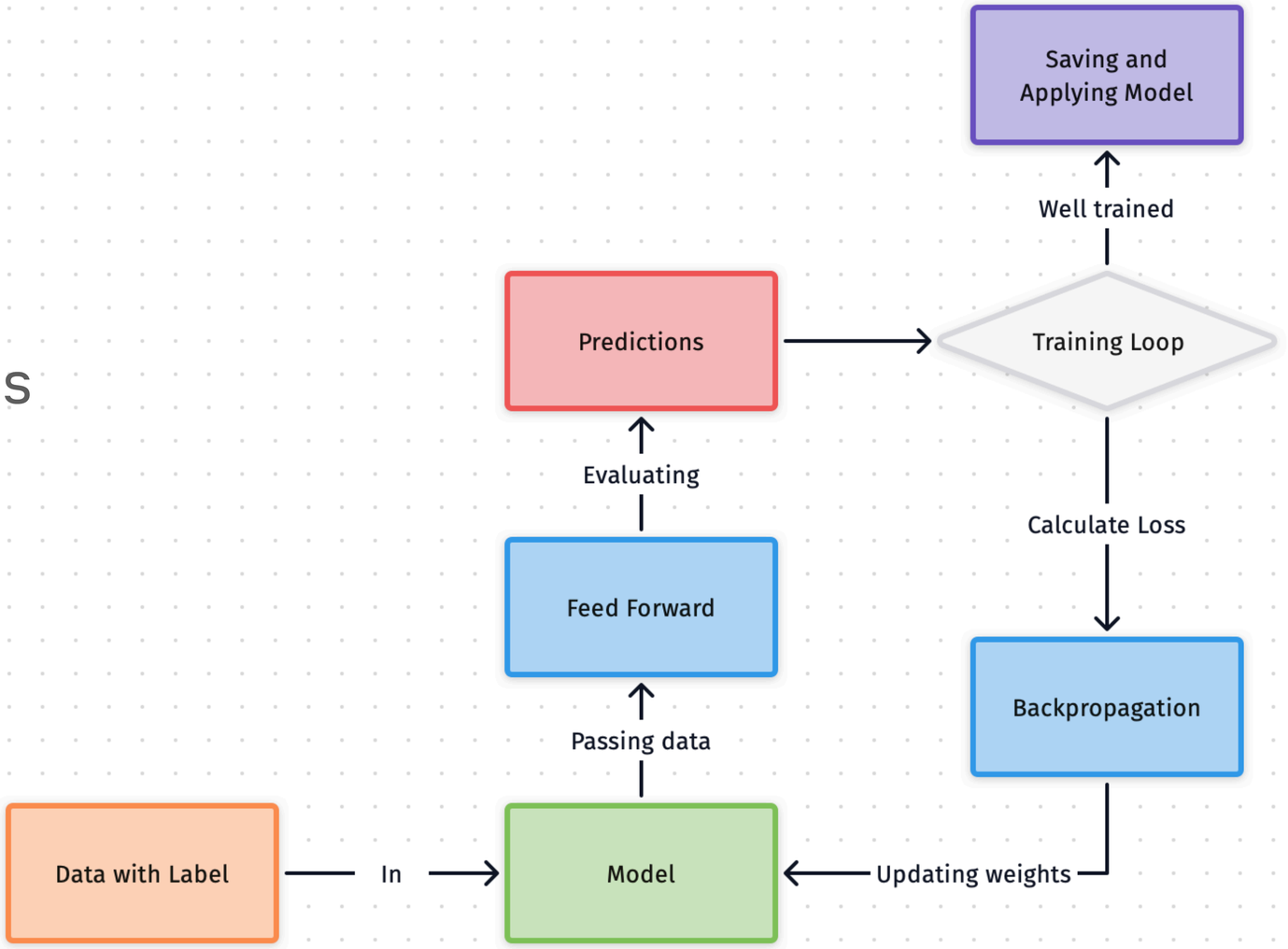
Training process

Feed Forward

- Weight and bias
- Activation functions

Back-prop

- Gradient Descent



Introduction to Adversarial Attack

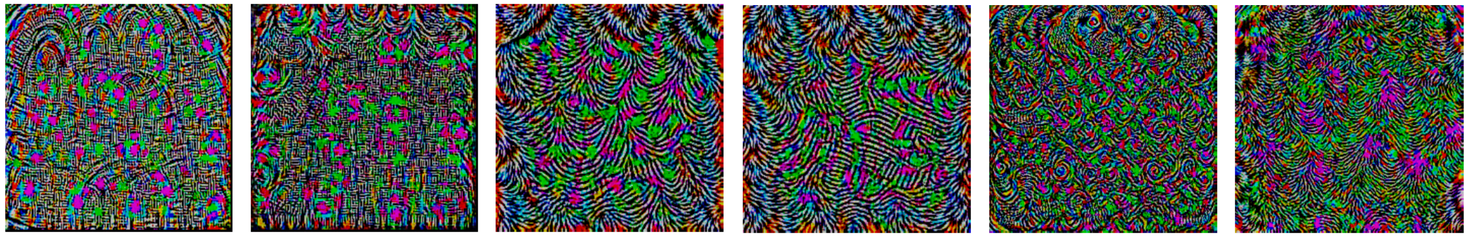
Usually in Computer Vision

(a) Person (low) (b) Sports ball (low) (c) Untargeted (low)

(d) Person (high) (e) Sports ball (high) (f) Untargeted (high)

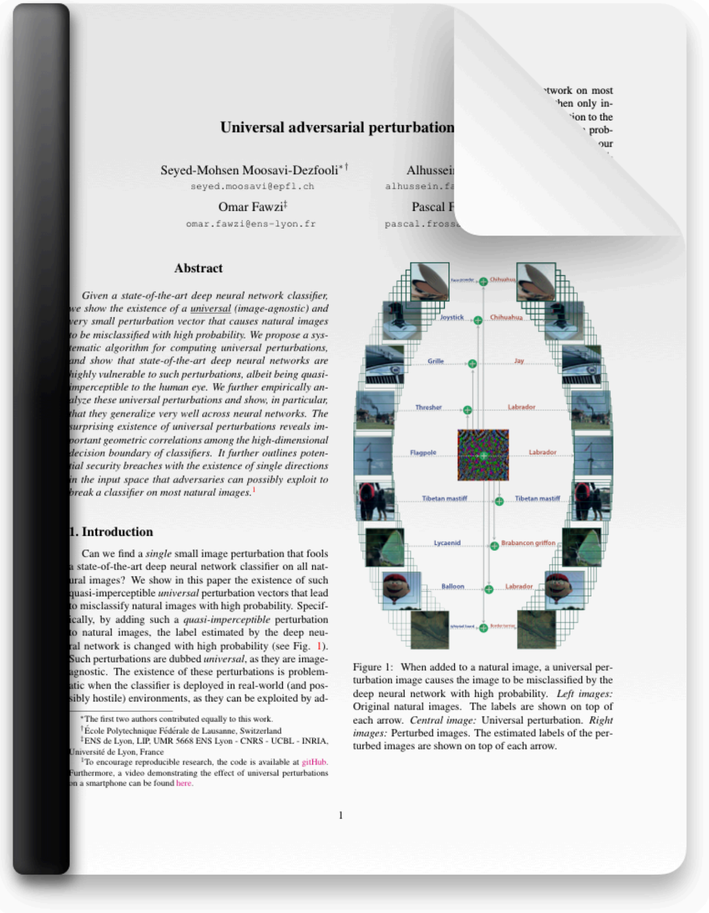
Current Research

Universal Perturbation [Moosavi-Dezfooli et al.]



(a) CaffeNet (b) VGG-F (c) VGG-16 (d) VGG-19 (e) GoogLeNet (f) ResNet-152

Figure 3: Examples of perturbed images and their corresponding labels. The first 8 images belong to the ILSVRC 2012 validation set, and the last 4 are images taken by a mobile phone camera. See supp. material for the original images.



Universal adversarial perturbation

Seyed-Mohsen Moosavi-Dezfooli*†
seyed.moosavi@epfl.ch

Omar Fawzi‡
omar.fawzi@ens-lyon.fr

Abstract

Given a state-of-the-art deep neural network classifier, we show the existence of a universal (image-agnostic) and very small perturbation vector that causes natural images to be misclassified with high probability. We propose a systematic algorithm for computing universal perturbations, and show that state-of-the-art deep neural networks are highly vulnerable to such perturbations, albeit being quasi-imperceptible to the human eye. We further empirically analyze these universal perturbations and show, in particular, that they generalize very well across neural networks. The surprising existence of universal perturbations reveals important geometric correlations among the high-dimensional decision boundary of classifiers. It further outlines potential security breaches with the existence of single directions in the input space that adversaries can possibly exploit to break a classifier on most natural images.[1]

1. Introduction

Can we find a *single* small image perturbation that fools a state-of-the-art deep neural network classifier on all natural images? We show in this paper the existence of such quasi-imperceptible *universal* perturbation vectors that lead to misclassify natural images with high probability. Specifically, by adding such a *quasi-imperceptible* perturbation to natural images, the label estimated by the deep neural network is changed with high probability (see Fig. 1). Such perturbations are dubbed *universal*, as they are image-agnostic. The existence of these perturbations is problematic when the classifier is deployed in real-world (and possibly hostile) environments, as they can be exploited by ad-

*The first two authors contributed equally to this work.
†École Polytechnique Fédérale de Lausanne, Switzerland
‡ENS de Lyon, LIP, UMR 5668 ENS Lyon - CNRS - UCBL - INRIA, Université de Lyon, France
[1]To encourage reproducible research, the code is available at gitHub. Furthermore, a video demonstrating the effect of universal perturbations on a smartphone can be found here.

1

Figure 1: When added to a natural image, a universal perturbation image causes the image to be misclassified by the deep neural network with high probability. *Left images:* Original natural images. The labels are shown on top of each arrow. *Central image:* Universal perturbation. *Right images:* Perturbed images. The estimated labels of the perturbed images are shown on top of each arrow.

Current Research

3D Printing fools DNN [Anish Athalye Et al.]

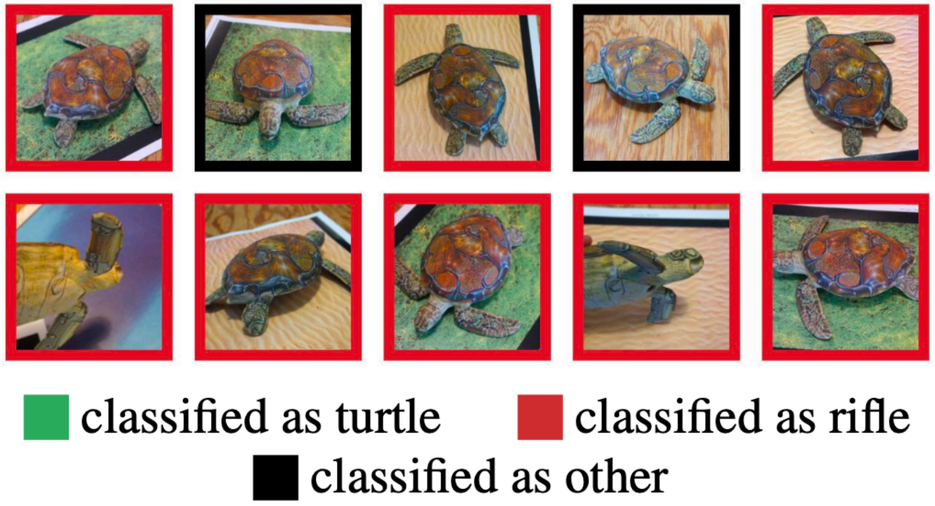
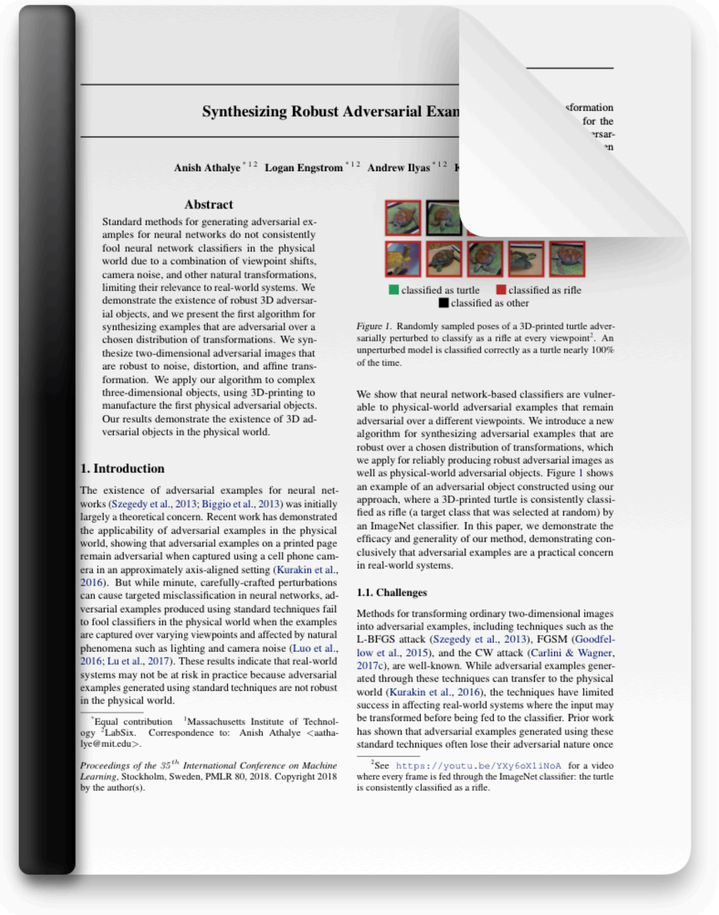


Figure 1. Randomly sampled poses of a 3D-printed turtle adversarially perturbed to classify as a rifle at every viewpoint[2]. An unperturbed model is classified correctly as a turtle nearly 100% of the time.



Synthesizing Robust Adversarial Exan

Anish Athalye [*12] Logan Engstrom [*12] Andrew Ilyas [*12]

Abstract

Standard methods for generating adversarial examples for neural networks do not consistently fool neural network classifiers in the physical world due to a combination of viewpoint shifts, camera noise, and other natural transformations, limiting their relevance to real-world systems. We demonstrate the existence of robust 3D adversarial objects, and we present the first algorithm for synthesizing examples that are adversarial over a chosen distribution of transformations. We synthesize two-dimensional adversarial images that are robust to noise, distortion, and affine transformation. We apply our algorithm to complex three-dimensional objects, using 3D-printing to manufacture the first physical adversarial objects. Our results demonstrate the existence of 3D adversarial objects in the physical world.

1. Introduction

The existence of adversarial examples for neural networks (Szegedy et al., 2013; Biggio et al., 2013) was initially largely a theoretical concern. Recent work has demonstrated the applicability of adversarial examples in the physical world, showing that adversarial examples on a printed page remain adversarial when captured using a cell phone camera in an approximately axis-aligned setting (Kurakin et al., 2016). But while minute, carefully-crafted perturbations can cause targeted misclassification in neural networks, adversarial examples produced using standard techniques fail to fool classifiers in the physical world when the examples are captured over varying viewpoints and affected by natural phenomena such as lighting and camera noise (Luo et al., 2016; Lu et al., 2017). These results indicate that real-world systems may not be at risk in practice because adversarial examples generated using standard techniques are not robust in the physical world.

[*]Equal contribution [1]Massachusetts Institute of Technology [2]LabSix. Correspondence to: Anish Athalye <aathalye@mit.edu>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

classified as turtle classified as rifle
classified as other

Figure 1. Randomly sampled poses of a 3D-printed turtle adversarially perturbed to classify as a rifle at every viewpoint[2]. An unperturbed model is classified correctly as a turtle nearly 100% of the time.

We show that neural network-based classifiers are vulnerable to physical-world adversarial examples that remain adversarial over a different viewpoints. We introduce a new algorithm for synthesizing adversarial examples that are robust over a chosen distribution of transformations, which we apply for reliably producing robust adversarial images as well as physical-world adversarial objects. Figure 1 shows an example of an adversarial object constructed using our approach, where a 3D-printed turtle is consistently classified as rifle (a target class that was selected at random) by an ImageNet classifier. In this paper, we demonstrate the efficacy and generality of our method, demonstrating conclusively that adversarial examples are a practical concern in real-world systems.

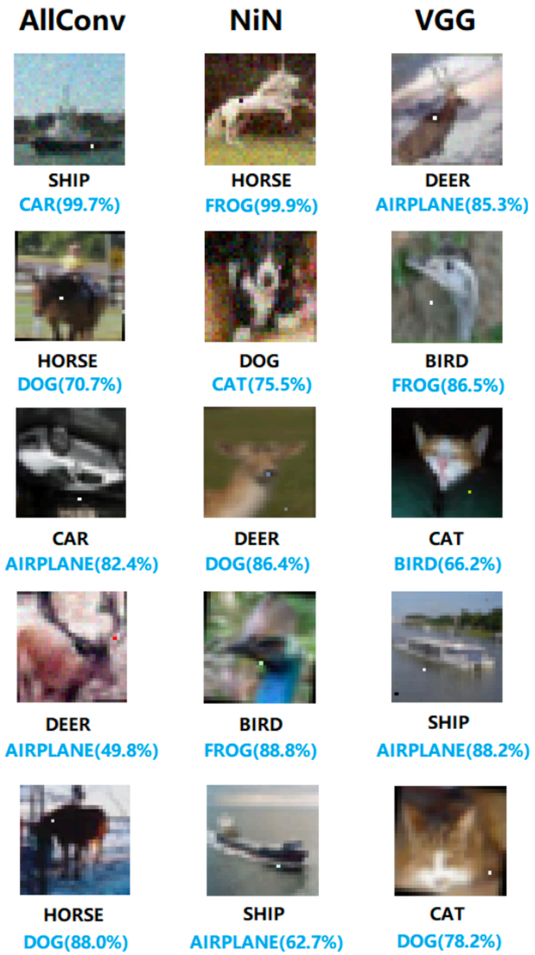
1.1. Challenges

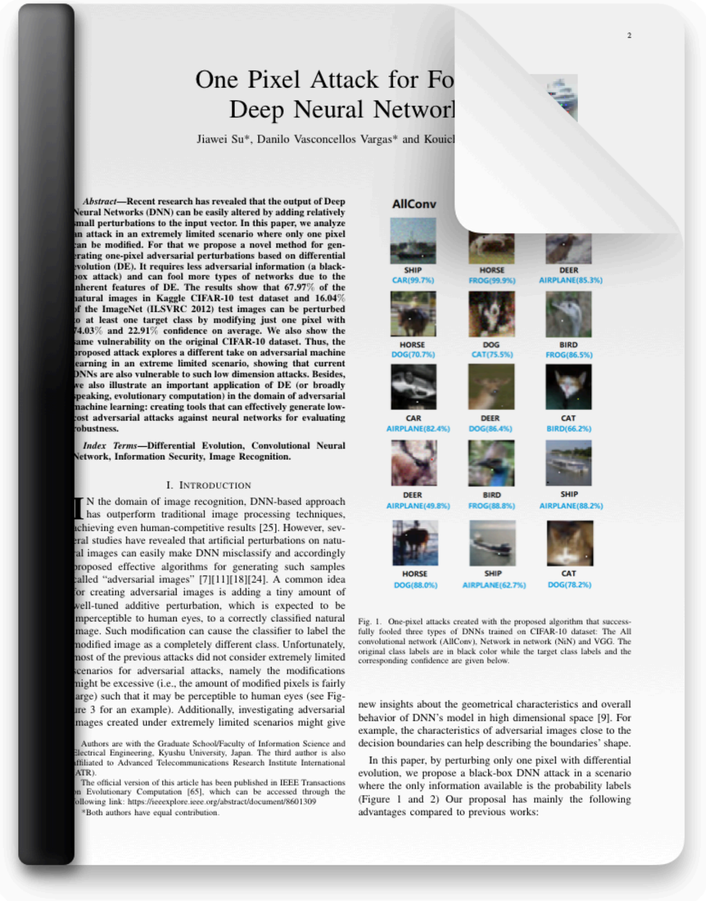
Methods for transforming ordinary two-dimensional images into adversarial examples, including techniques such as the L-BFGS attack (Szegedy et al., 2013), FGSM (Goodfellow et al., 2015), and the CW attack (Carlini & Wagner, 2017c), are well-known. While adversarial examples generated through these techniques can transfer to the physical world (Kurakin et al., 2016), the techniques have limited success in affecting real-world systems where the input may be transformed before being fed to the classifier. Prior work has shown that adversarial examples generated using these standard techniques often lose their adversarial nature once

[2]See https://youtu.be/YXy6oXlINoA for a video where every frame is fed through the ImageNet classifier: the turtle is consistently classified as a rifle.

Current Research

One Pixel Attack [Jiawei Su Et al.]



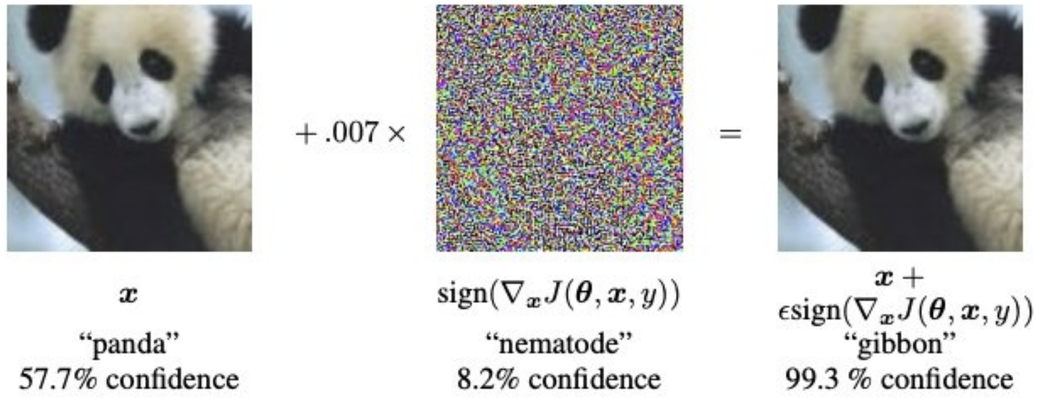


Current Research

Fast Gradient Signed Method (FGSM)

is used to generate adversarial examples **to minimize the maximum amount of perturbation added to any pixel of the image** to cause misclassification.

- **Advantages:** Comparably efficient computing times.
- **Disadvantages:** Perturbations are added to every feature.



Gap Identification

DNN can be attacked

Filling the Gap

“We believe, only the attack algorithms are more advanced will leads to a better defense.”

Perspective Applications

Data Augmentation

Content Filtering

Extending attack to NLP

Working with GAN

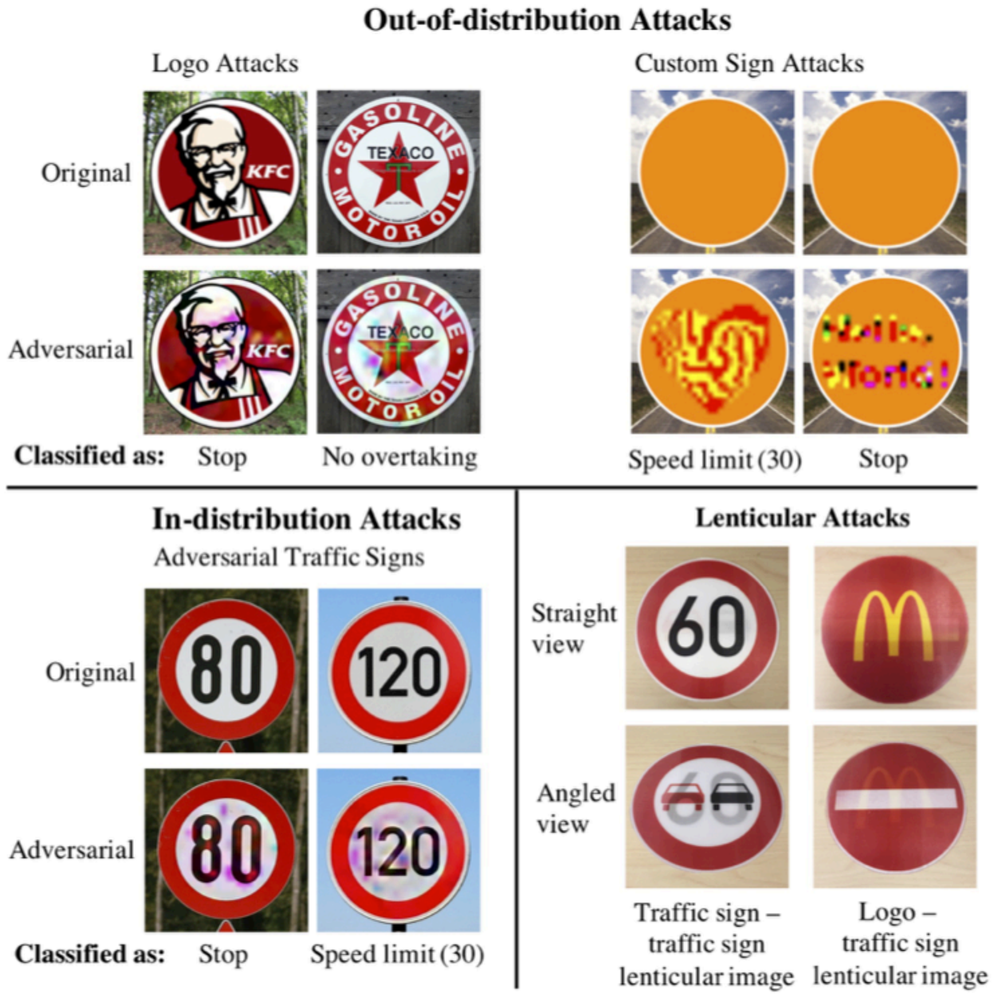
Finding new Defense Methods

Working out an evaluation standard for attack and defense

Enhancing models robustness

Risk Identification

Embedding Trojan patten



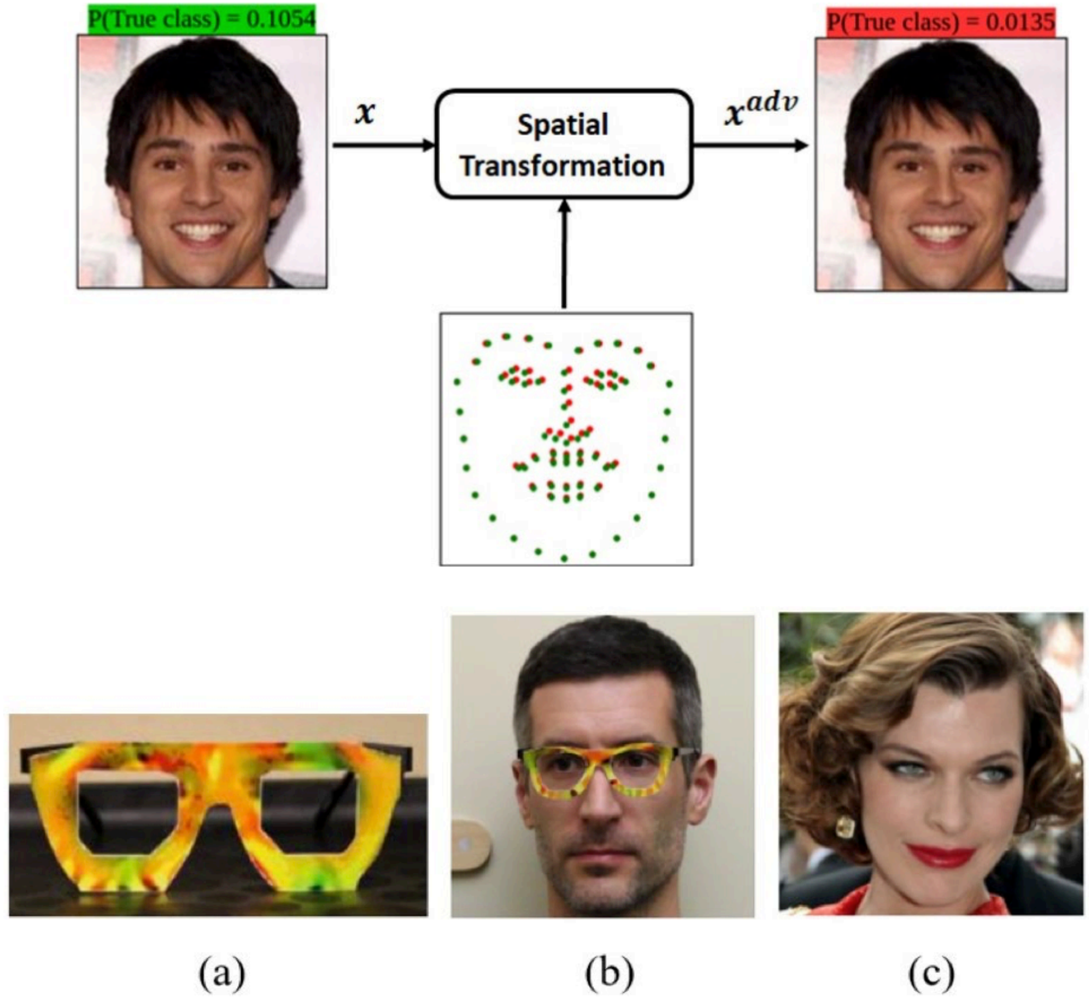


Fig. 5. The eyeglass frames (a) were used by Lujo Bauer (b) to impersonate Milla Jovovich (c) (Sharif et al., 2016).