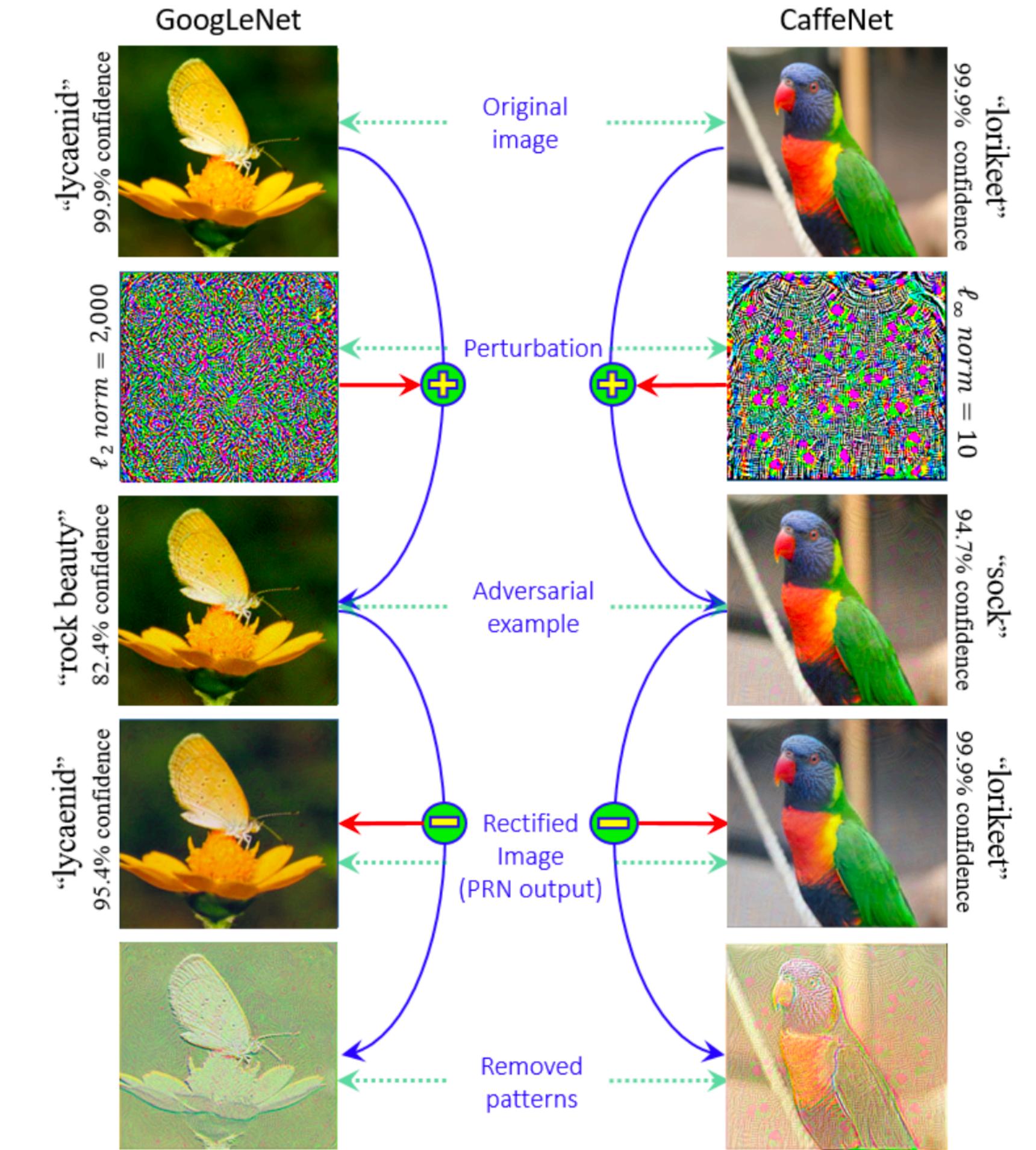


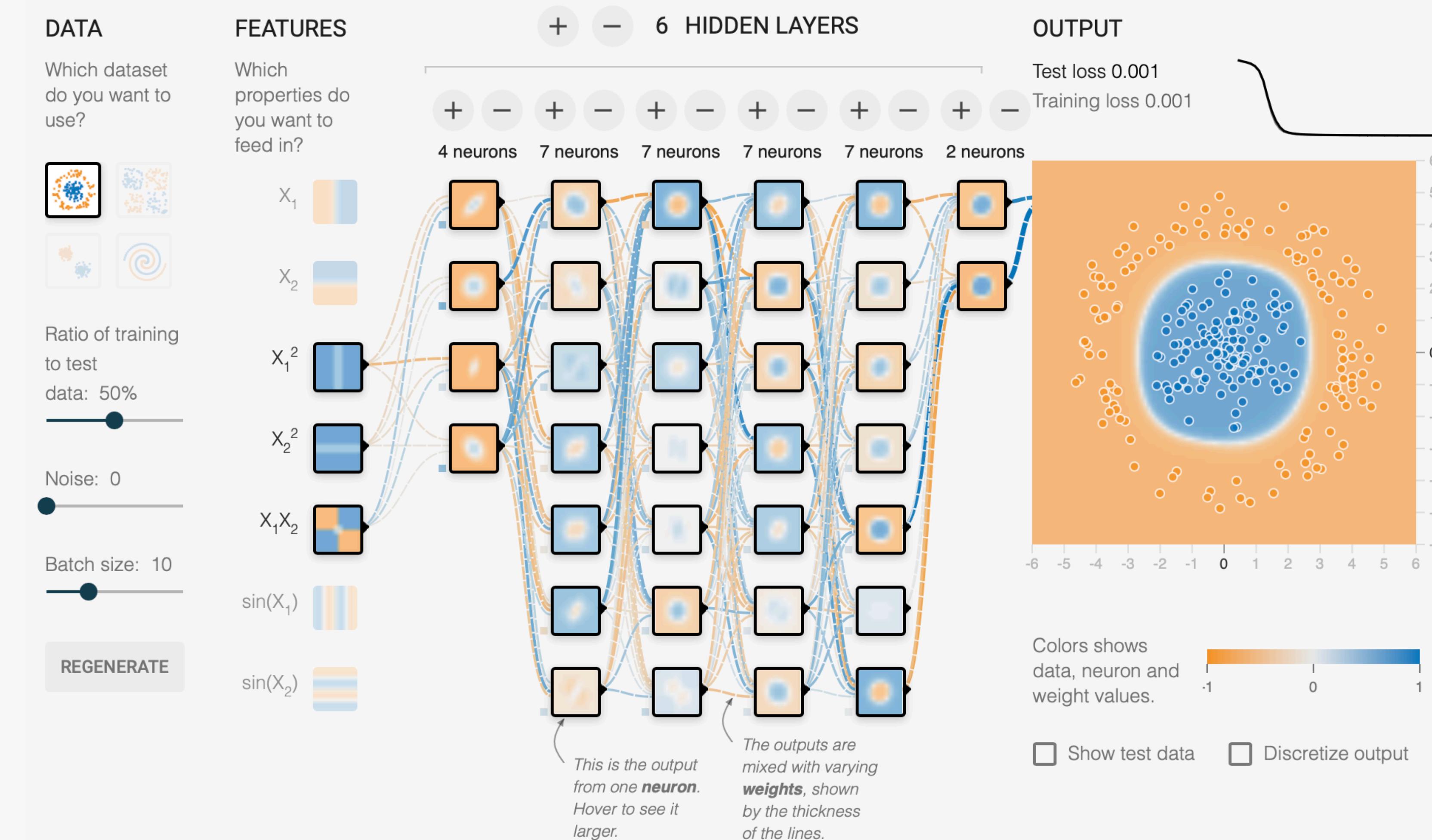
# Adversarial Attack on DNN

A Project Proposal for FYP



# Introduction to Deep Neural Network (DNN)

## Fundamental of Machine Learning (ML)



# Introduction to Deep Neural Network (DNN)

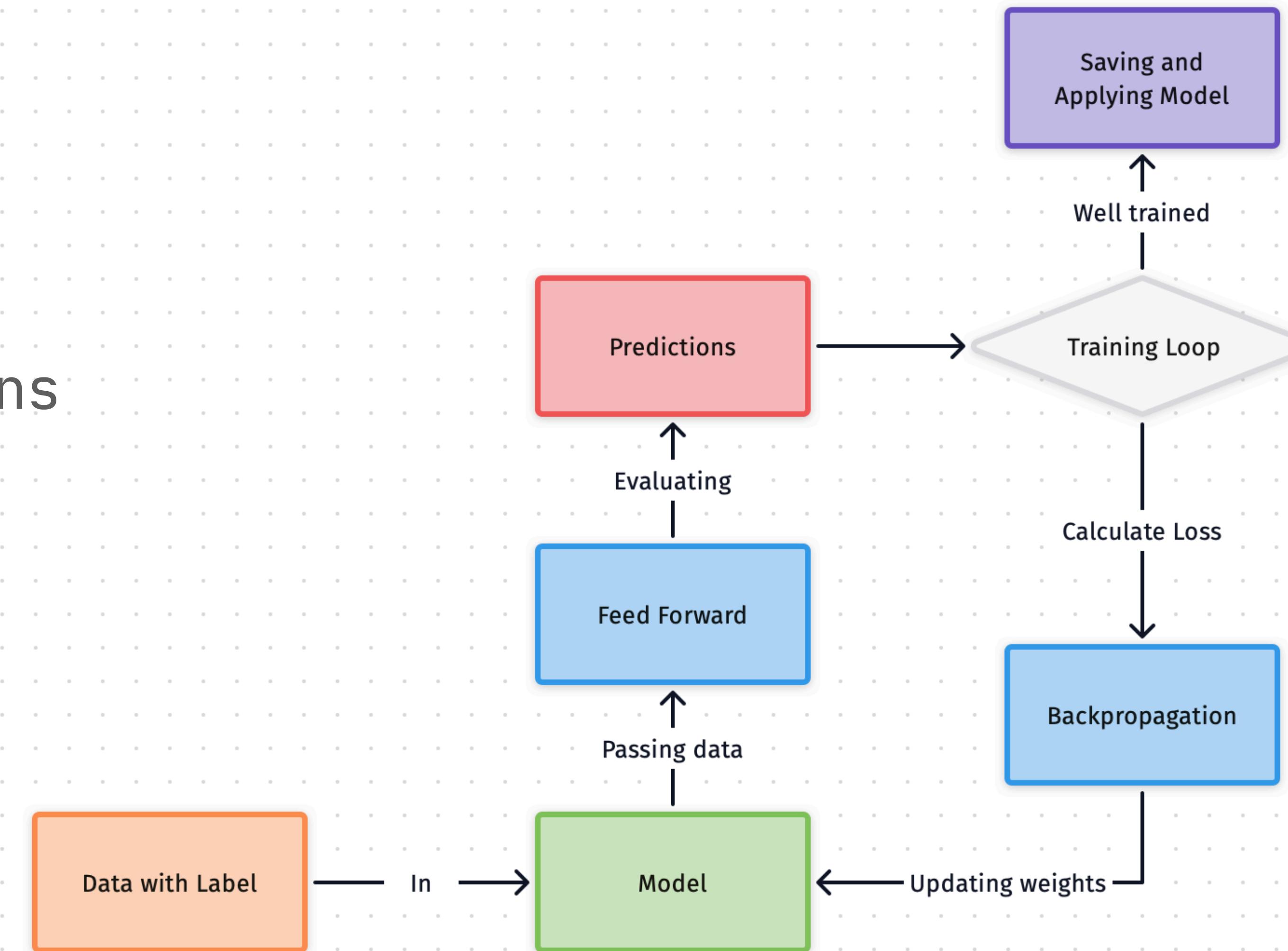
## Training process

### 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.]

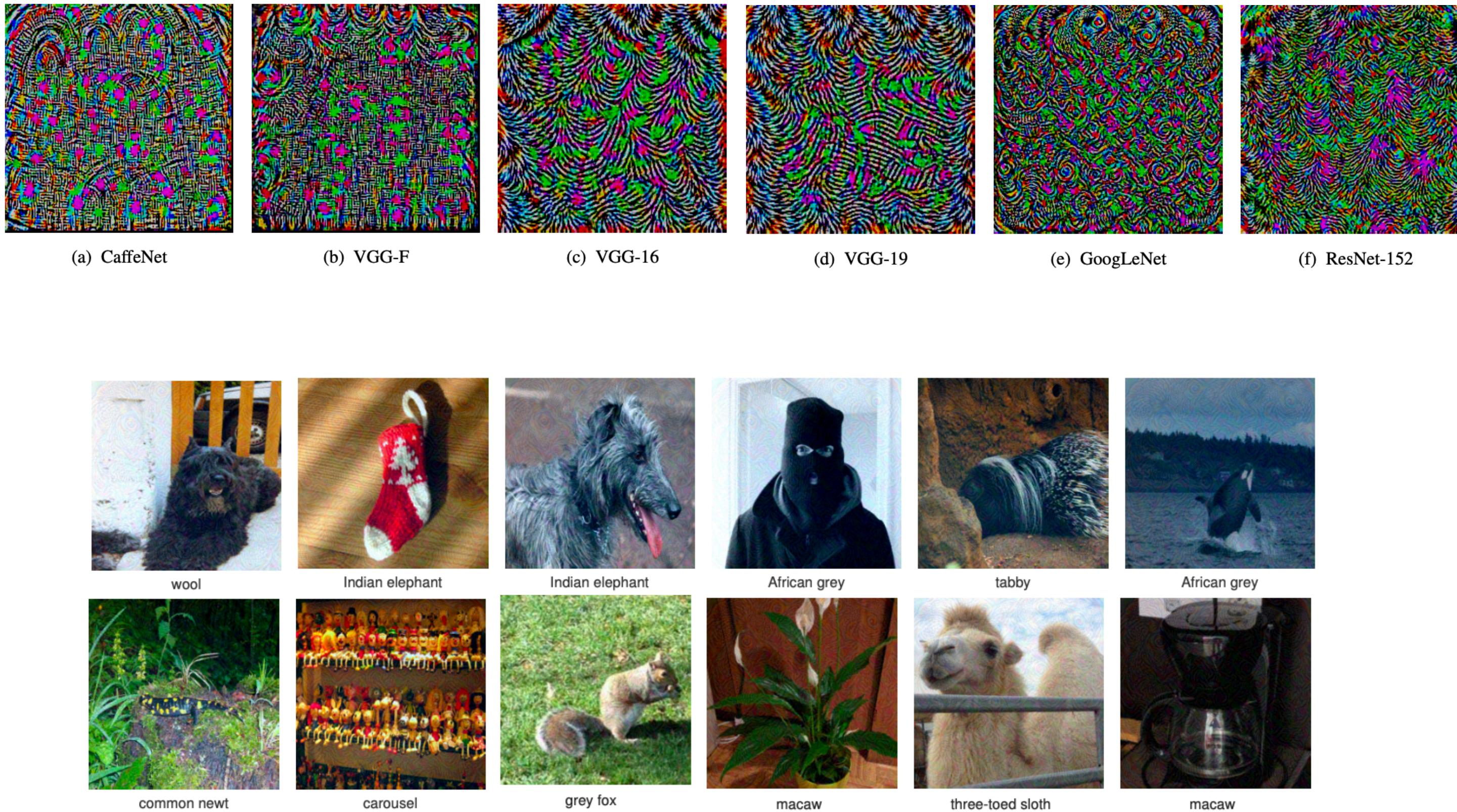


Figure 3: Examples of perturbed images and their corresponding labels. The first 8 images belong to the ILSVRC 2012 validation set, and the last 4 are images taken by a mobile phone camera. See supp. material for the original images.

**Universal adversarial perturbation**

Seyed-Mohsen Moosavi-Dezfooli<sup>\*†</sup>  
seyed.moosavi@epfl.ch

Omar Fawzi<sup>‡</sup>  
omar.fawzi@ens-lyon.fr

Alhussein Farhadi  
alhussein.farhadi@epfl.ch

Pascal Frossard  
pascal.frossard@epfl.ch

**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.<sup>1</sup>

**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.  
<sup>†</sup>Ecole Polytechnique Federale de Lausanne, Switzerland  
<sup>‡</sup>ENS de Lyon, LIP, UMR 5668 ENS Lyon - CNRS - UCBL - INRIA, Universite de Lyon, France

<sup>1</sup>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](#).

# Current Research

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



■ 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<sup>2</sup>. An unperturbed model is classified correctly as a turtle nearly 100% of the time.

Anish Athalye <sup>\*12</sup> Logan Engstrom <sup>\*12</sup> Andrew Ilyas <sup>\*12</sup> F

**Synthesizing Robust Adversarial Examples**

Information for the <sup>\*</sup>rsar-  
n

**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.

<sup>1</sup>Equal contribution <sup>2</sup>Massachusetts Institute of Technology -LabSix. Correspondence to: Anish Athalye <aathalye@mit.edu>.

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

Figure 1. Randomly sampled poses of a 3D-printed turtle adversarially perturbed to classify as a rifle at every viewpoint<sup>2</sup>. 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.

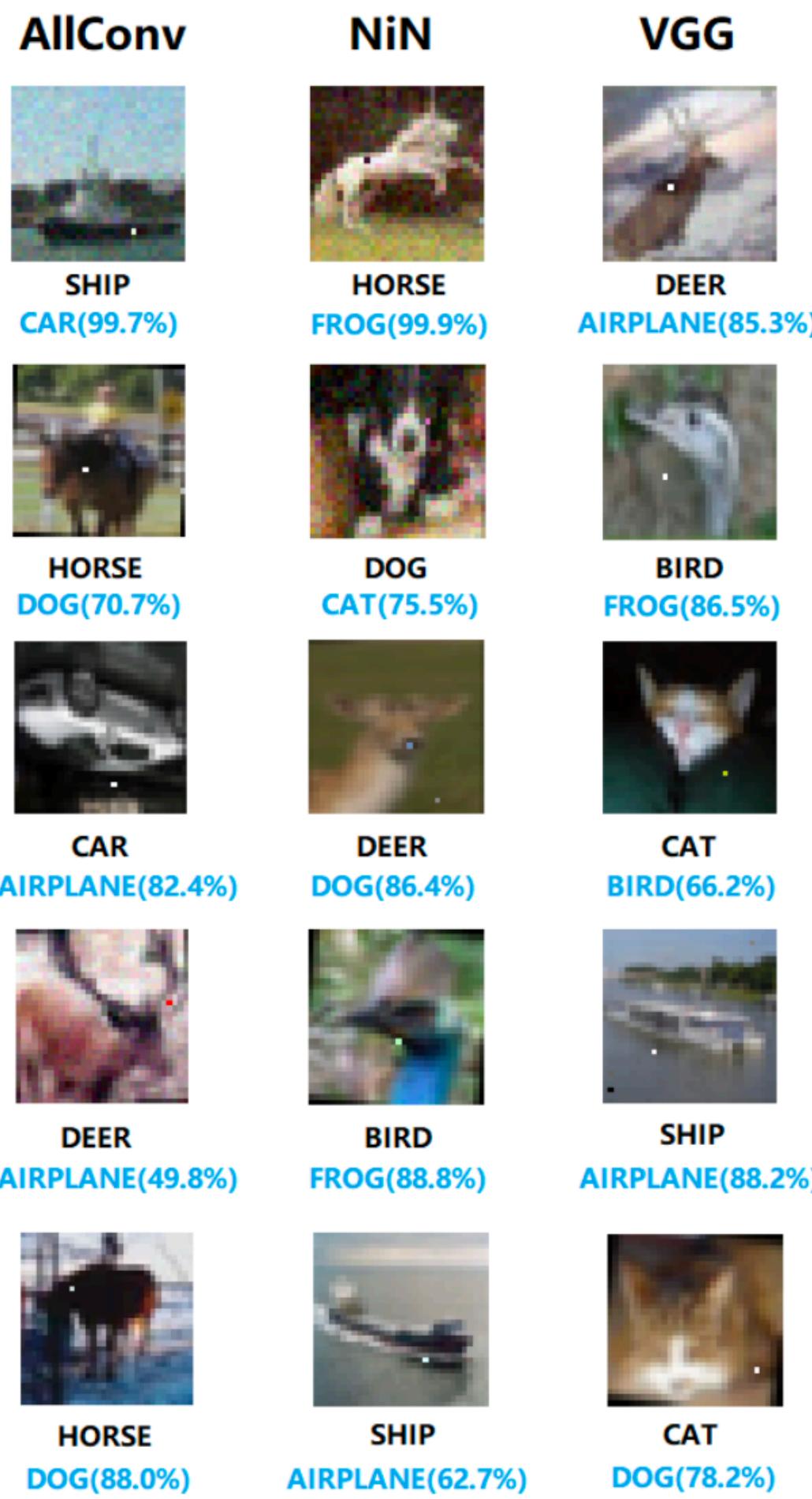
**1.1. Challenges**

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

<sup>2</sup>See <https://youtu.be/YXy6oX1iNoA> 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.]



Cup(16.48%)  
Soup Bowl(16.74%)  
Bassinet(16.59%)  
Paper Towel(16.21%)  
Teapot(24.99%)  
Joystick(37.39%)  
Hamster(35.79%)  
Nipple(42.36%)



**Abstract**—Recent research has revealed that the output of Deep Neural Networks (DNN) can be easily altered by adding relatively small perturbations to the input vector. In this paper, we analyze an attack in an extremely limited scenario where only one pixel can be modified. For that we propose a novel method for generating one-pixel adversarial perturbations based on differential evolution (DE). It requires less adversarial information (a black-box attack) and can fool more types of networks due to the inherent features of DE. The results show that 67.97% of the natural images in Kaggle CIFAR-10 test dataset and 16.04% of the ImageNet (ILSVRC 2012) test images can be perturbed to at least one target class by modifying just one pixel with 74.03% and 22.91% confidence on average. We also show the same vulnerability on the original CIFAR-10 dataset. Thus, the proposed attack explores a different take on adversarial machine learning in an extreme limited scenario, showing that current DNNs are also vulnerable to such low dimension attacks. Besides, we also illustrate an important application of DE (or broadly speaking, evolutionary computation) in the domain of adversarial machine learning: creating tools that can effectively generate low-cost adversarial attacks against neural networks for evaluating robustness.

**Index Terms**—Differential Evolution, Convolutional Neural Network, Information Security, Image Recognition.

### I. INTRODUCTION

In the domain of image recognition, DNN-based approach has outperformed traditional image processing techniques, achieving even human-competitive results [25]. However, several studies have revealed that artificial perturbations on natural images can easily make DNN misclassify and accordingly proposed effective algorithms for generating such samples called “adversarial images” [7][11][18][24]. A common idea for creating adversarial images is adding a tiny amount of well-tuned additive perturbation, which is expected to be imperceptible to human eyes, to a correctly classified natural image. Such modification can cause the classifier to label the modified image as a completely different class. Unfortunately, most of the previous attacks did not consider extremely limited scenarios for adversarial attacks, namely the modifications might be excessive (i.e., the amount of modified pixels is fairly large) such that it may be perceptible to human eyes (see Figure 3 for an example). Additionally, investigating adversarial images created under extremely limited scenarios might give

new insights about the geometrical characteristics and overall behavior of DNN’s model in high dimensional space [9]. For example, the characteristics of adversarial images close to the decision boundaries can help describing the boundaries’ shape.

Authors are with the Graduate School/Faculty of Information Science and Electrical Engineering, Kyushu University, Japan. The third author is also affiliated to Advanced Telecommunications Research Institute International (ATR).

The official version of this article has been published in IEEE Transactions on Evolutionary Computation [65], which can be accessed through the following link: <https://ieeexplore.ieee.org/abstract/document/8601309>

\*Both authors have equal contribution.

In this paper, by perturbing only one pixel with differential evolution, we propose a black-box DNN attack in a scenario where the only information available is the probability labels (Figure 1 and 2). Our proposal has mainly the following advantages compared to previous works:

1) The proposed attack explores a different take on adversarial machine learning in an extreme limited scenario, showing that current DNNs are also vulnerable to such low dimension attacks.

2) The proposed attack requires less adversarial information (a black-box attack) and can fool more types of networks due to the inherent features of DE.

3) The proposed attack is based on differential evolution, which is a well-known evolutionary computation algorithm.

4) The proposed attack is efficient and can be applied to various DNN models.

# 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.

$$\begin{aligned} & \text{---} \\ & \quad \mathbf{x} \\ & \quad \text{“panda”} \\ & \quad 57.7\% \text{ confidence} \\ & + .007 \times \\ & \quad \text{sign}(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, y)) \\ & \quad \text{“nematode”} \\ & \quad 8.2\% \text{ confidence} \\ & = \\ & \quad \mathbf{x} + \\ & \quad \epsilon \text{sign}(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, y)) \\ & \quad \text{“gibbon”} \\ & \quad 99.3 \% \text{ confidence} \end{aligned}$$

# 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

Content Filtering

Extending  
attack to NLP

Working out an  
evaluation standard for  
attack and defense

Data Augmentation

Working with GAN

Finding new  
Defense Methods

Enhancing models  
robustness

# Risk Identification

## Embedding Trojan pattern

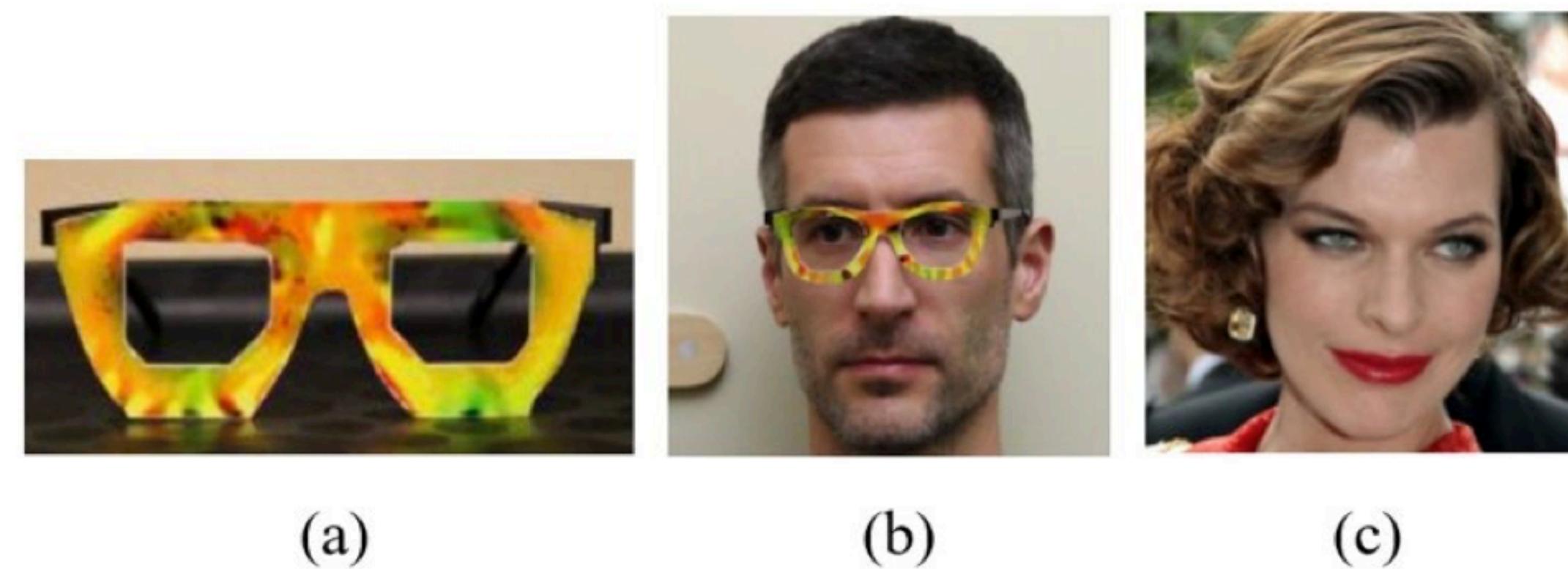
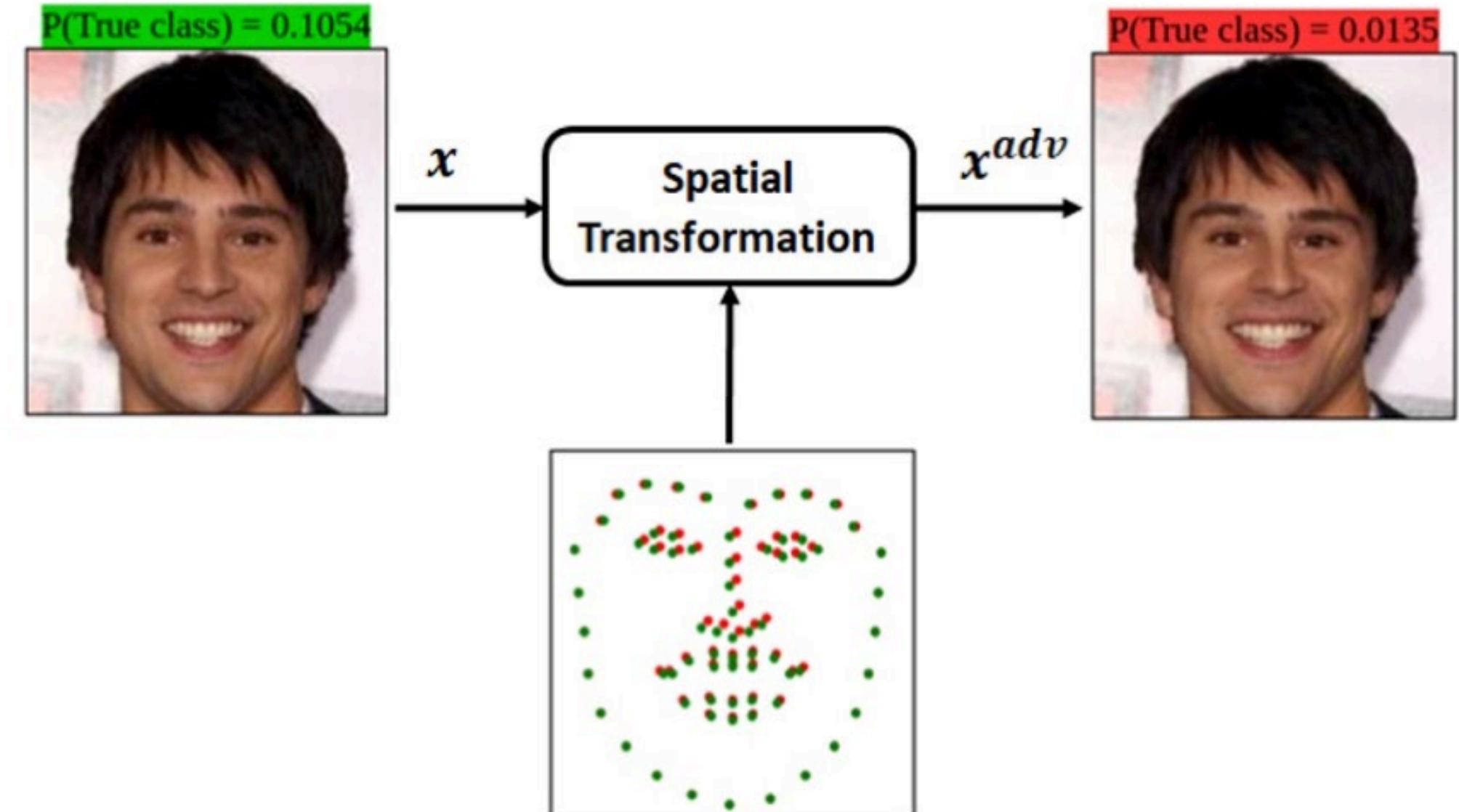
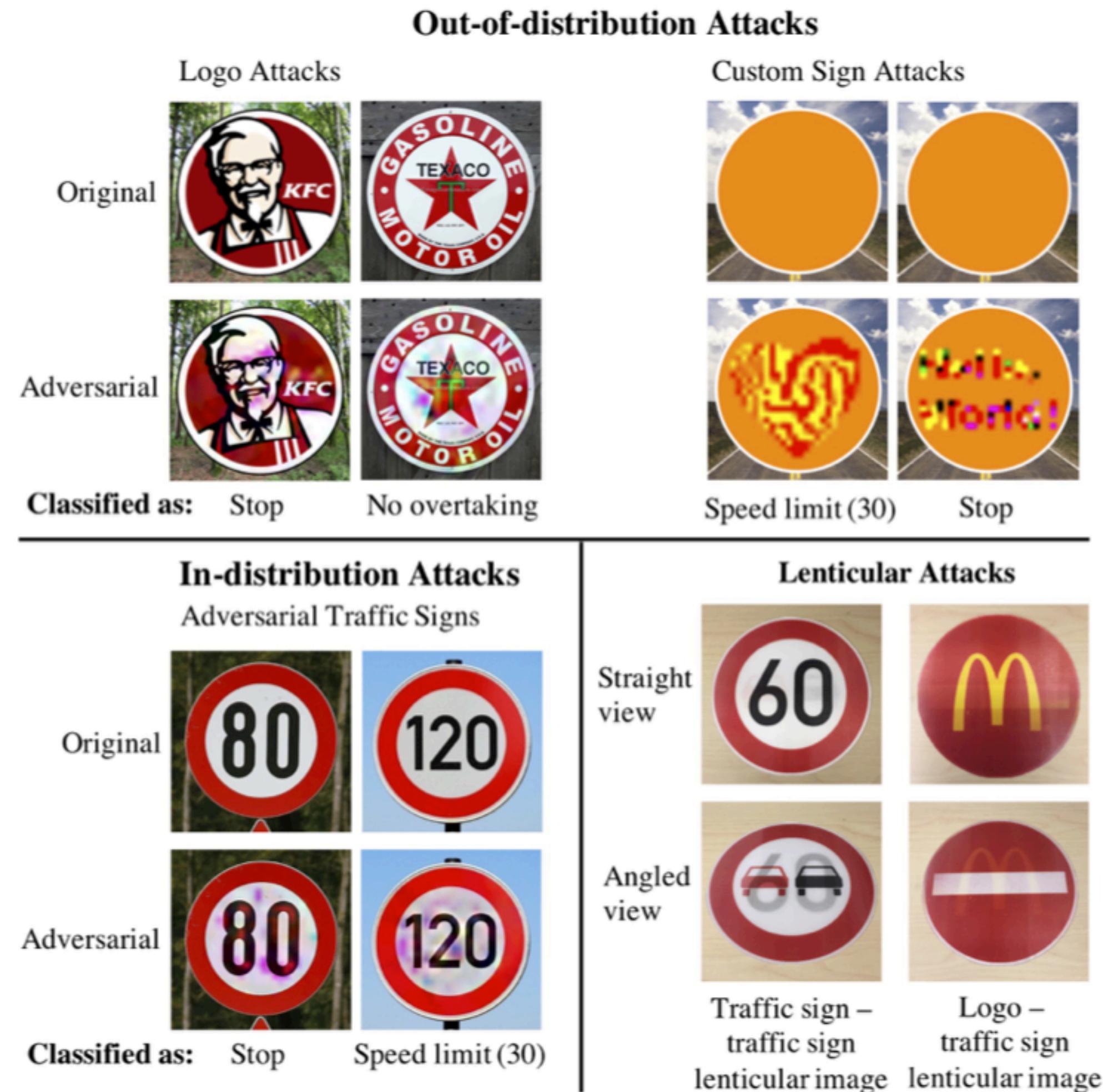


Fig. 5. The eyeglass frames (a) were used by Lujo Bauer (b) to impersonate Milla Jovovich (c) (Sharif et al., 2016).