Adversarial Attacks Explained

Simply put, the adversarial attack is a deceiving technique that is "fooling" machine learning models using a defective input. Adversarial machine learning is aimed to cause a malfunction of an ML model (think of a self-driving car that takes a stop sign as a speed limit or a Tesla autopilot car moving in the opposite direction from intended).

Adversarial attacks become possible because of inaccurate or misrepresenting data used during the training or using maliciously designed data for an already trained model. Let us delve into the nuts and bolts step-by-step.

Why are adversarial attacks dangerous?

While ML is a relatively new domain (less than ten years old), it is developing tremendously, gaining wide popularity across lots of industries. We will witness how non-internet sectors like agriculture, education, logistics, manufacturing, and energy sectors will bring up to $13 trillion of GDP growth by 2030, as per the McKinsey research. But adversarial attacks might cause severe problems across all these sectors.

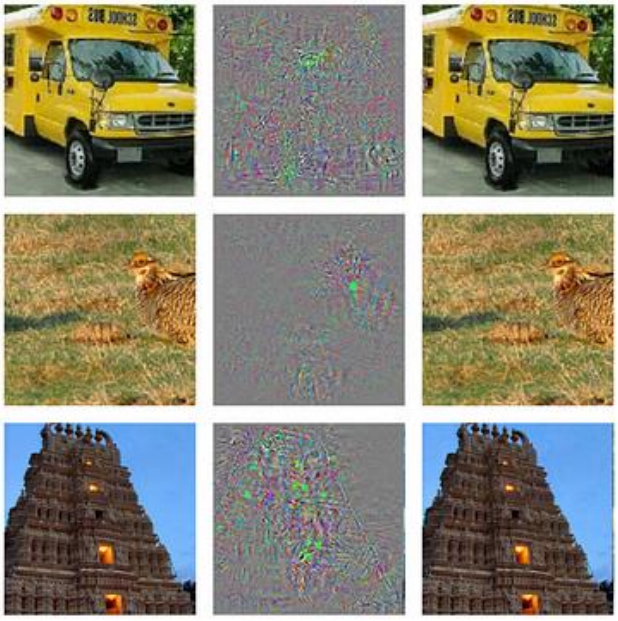
For example, research shows how adversarial attacks on medical machine learning can make ML algorithms classify benign moles as malignant. Consider the impact of such malicious actions at scale in any other business vertical.

How long did adversarial attacks enter the picture?

In 2014 there were no papers regarding the adversarial attacks on preprint server Arxiv.org. But as per the moment of writing (August 2021), there are around 1000 research papers on adversarial attacks and their examples. It seems like it is going to be the next arms race while AI adoption is rising globally. One of the first researches by Google and New York University, "*Intriguing properties of neural networks*," is dated by 2013 and has shed some light on the essence of adversarial attack.

Thus, the adversarial attack is an optical illusion for the ML model that misperceives the objects while not visible to the naked eye. Check out the following example:

Random images in the third row are perceived as images of an ostrich by the ML model. The images in the middle are examples of adversarial perturbation., i.e., noise added to the clean image in the first left row to create adversarial examples.



scíforce

What are the types of adversarial attacks?

Depending on the influence of the classifier (ML algorithm), security violation, and specificity, adversarial attacks could be subcategorized to “white-box” or “black-box” attacks. A white-box attack means that the attacker has access to the model’s parameters, and there is no access to parameters in case of a black-box attack.

What is under the hood of an adversarial attack? In general, adversarial attacks share the same idea. They use (sometimes approximated) knowledge about the model’s internal state to modify

input pixels to cause the greatest chance of error. In other words, a small perturbation changes the class label.

Mathematically, it looks like the following:

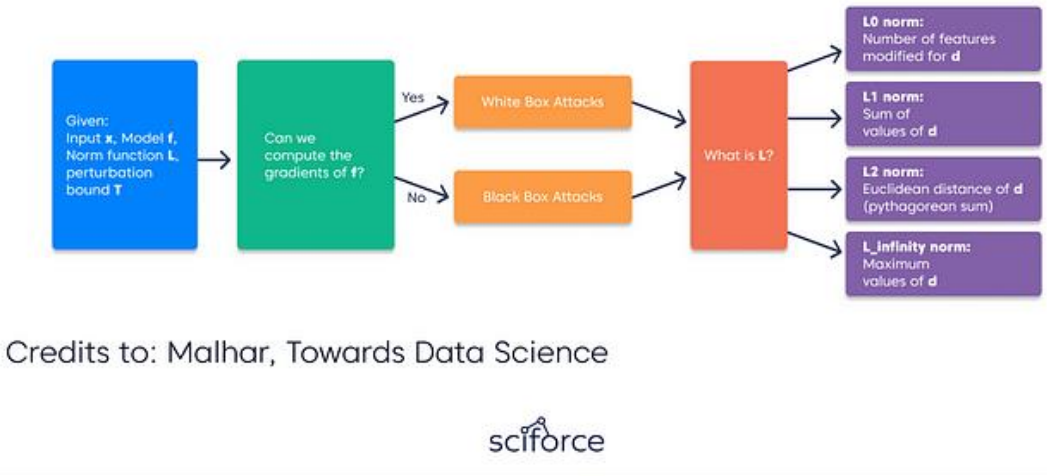
f(x+d)!=y

Model **f** using the input **x** can produce prediction **y**. But here, we have adversarial example **d** that leads to the prediction **y** that is not equal to the prediction of the model **f** with the input **x**.

L(d)<T,

L is a generic function that measures the norm of d, and T stands for the upper bound of this norm.

Bearing that in mind, you can encounter a set of algorithms that could generate such perturbations:



Credits to: Malhar, Towards Data Science

Let us define what stands behind the different types of perturbations. L stands for perturbation bound that measures the size of perturbation d, usually Lp norm is used:

L0 norm: it implies modifying the exact number of features of the input. In reality, only a tiny piece of the information is modified, but it can deceive the overall system. Take a real-world example — DNN classifier misperceives the small part on the STOP sign and generates command of going further instead of stop moving:

The left image shows real graffiti on a Stop sign something that most humans would not think in suspicions. The right image shows our a physical perturbation applied to a Stop sign. We design our perturbation to mimic graffity, and thus "hide in the human psyche."

Source: Robust Physical-World Attacks on Deep Learning Visual Classification (Kevin Eykholt et al.)

sciforce

L1 norm: implies the total sum of all perturbation values involved. In reality, you can not encounter this type of attack often. To quote Pin-Yu Chen et al.: ”However, despite the fact that L1 distortion accounts for the total variation and encourages sparsity in the perturbation, little has been developed for crafting L1-based adversarial examples.”

L2 norm: implies upper bounding the Euclidean distance (Pythagorean distance) of the perturbation d. In other words, it is the squared difference between the images X and Z (calculate the distance between the images X and Z for each pixel, and then sum all pixels.

Examples of this type include the Carlini and Wagner attack is the most effective white-box attack in researches.

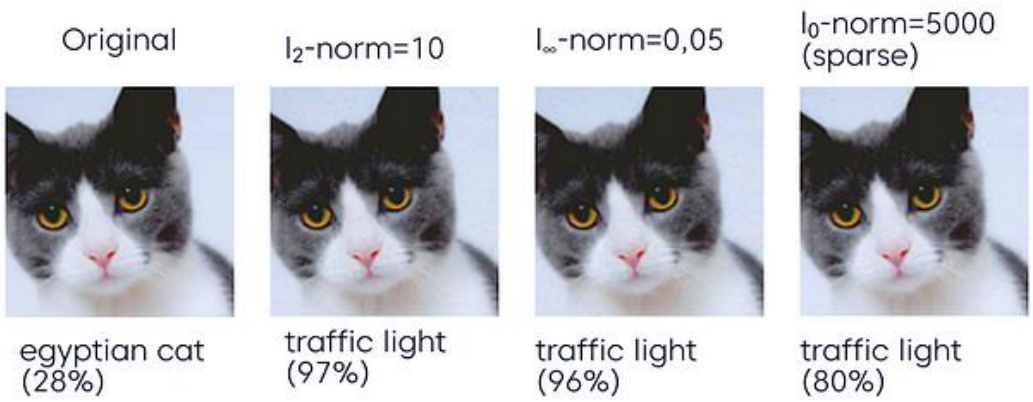
Example of an adversarial image on the ImageNet dataset. The sample **x** is recognized as a Curly-coated retriever. Adding a perturbation **δ** we obtain an adversarial image that is classified as a microwave (with $||\delta||_2$=0.7).

Source: Decoupling Direction and Norm for Efficient Gradient-Based L2 Adversarial Attacks and Defenses (Jérôme Rony et al.)

sciforce

L norm: implies the maximum value of perturbation d. They are represented in the researches the most given their robust optimization and mathematical convenience.

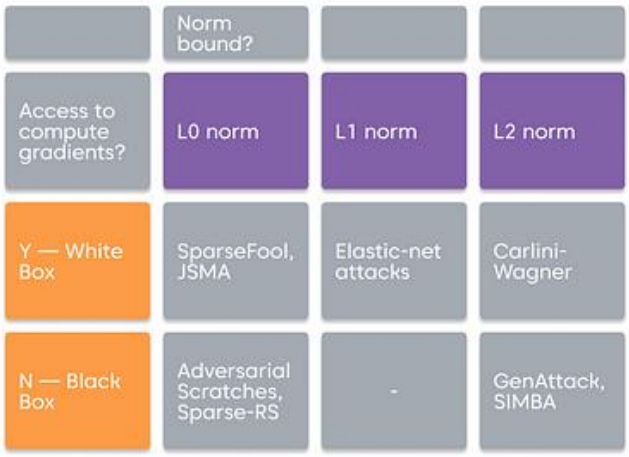
Adversarial examples with different norm constraints formed via the projected gradient method (Madry et al., 2017) on a Resnet50, along with the distance between the base image and the adversarial example, and the top class label.



Source: Are adversarial examples inevitable? (Ali Shafahi et al.)

sciforce

Based on the type of attack (white-box or black-box and perturbation bound, adversarial attacks could be categorized further. Check out this table concerning the types f the attack developed by Malhar:



The types of adversarial attacks, considering the access to compute guidelines and perturbation bound. Credits to Malhar

sciforce