MACHINE LEARNING AGAINST ADVERSARIAL SAMPLES

PHILLIP F. AGUILERA

WHAT IS MACHINE LEARNING?

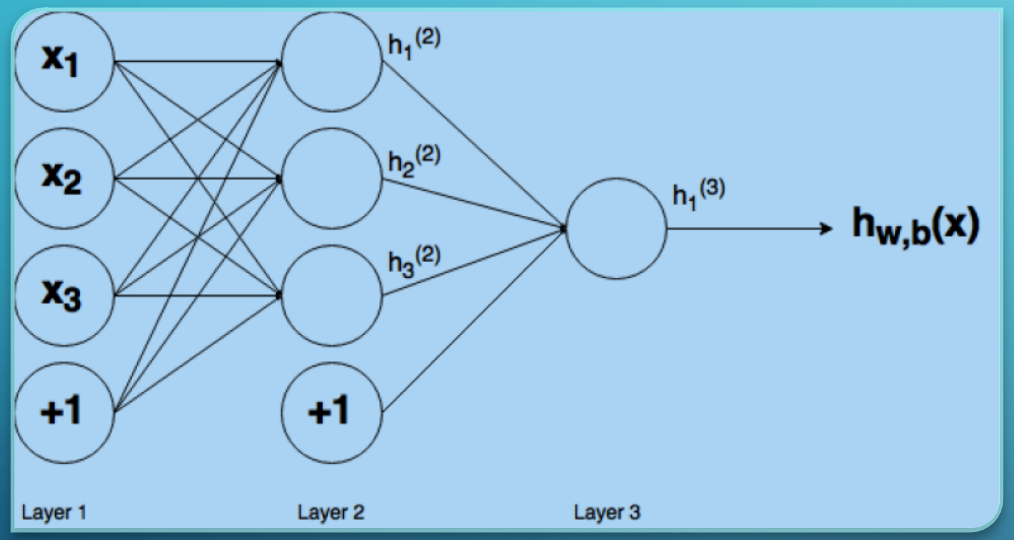
- Machine Learning (from henceforth “ML”) is a subset of Artificial Intelligence that utilizes algorithms to perform some task learned from data that is not explicitly programmed.
- There are several types of ML, such as: Supervised, Unsupervised, and Deep Learning, just to name a few.
- Supervised ML utilizes Artificial Neural Networks at its core to train and “learn” from data.

- https://www.ibm.com/topics/machine-learning

WHY MACHINE LEARNING?

- In a typical computer program, the programmer writes rules that act on a set of input data and produce expected output.
- In ML, the program inputs data and expected output which, in turn, yields the rules that act on the input data.
- In this way, characteristics and features can be calculated by the ML program on the data itself, which then can further generalized to process larger and larger amounts of data.
- Of course, with increasing amounts of data, the ML model can trained to learn even more and produce more accurate results with more general data.

WHAT IS AN ARTIFICIAL NEURAL NETWORK?



- An Artificial Neural Network (from henceforth "ANN") is essentially a function that is modeled after the biological Neural Network that is present in human brains.
- ANNs consist of nodes (neurons) that are connected to one another. These nodes have an input from one or more nodes and an output to one or more nodes.
- Taking the analogy further, when particular signals are sent to a node, it "activates". When such connections are established and the desired result is achieved, the connection between such nodes are strengthened, which facilitates "learning".

- https://adventuresinmachinelearning.com/

HOW DOES AN ARTIFICIAL NEURAL NETWORK "LEARN"?

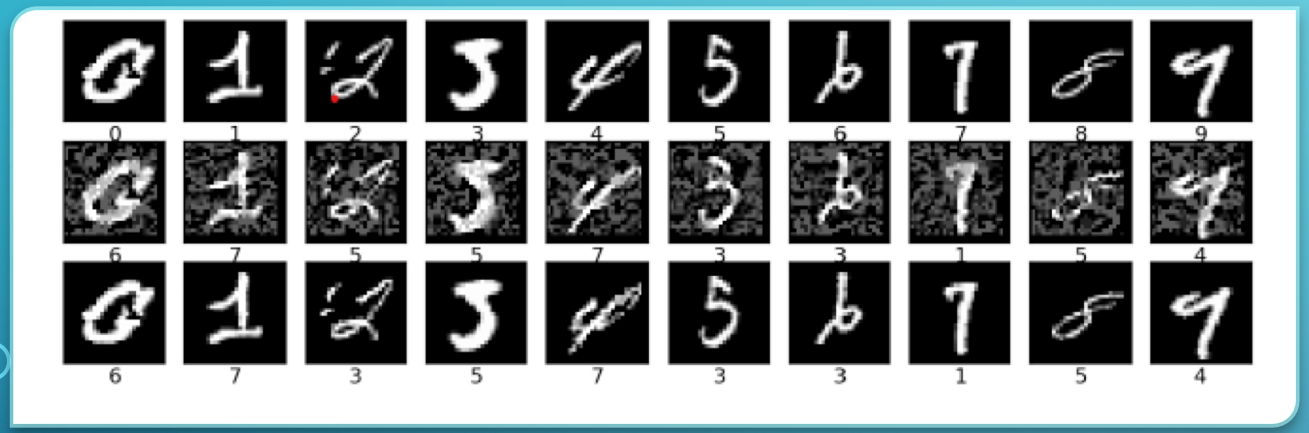
- ANNs must undergo training to be able to learn and yield output that is expected.
- As stated, an ANN is a function, and can thus be represented by *F: X -> Y*, where X is the input and Y is the output. Input is "fed" to the ANN and it produces an output that is compared to the *expected* output.
- The difference between the given output and the expected output is utilized to "tweak" the ANN to improve its "intelligence". This is done by two techniques known as Backwards Propagation (or "Backpropagation") and Gradient Descent. Specifically, the purpose of Backpropagation with Gradient Descent is to find the best tweaks to optimize the ANN to increase its accuracy of what is expected for them to produce.

ADVERSARIAL SAMPLES

- One major issue in ML is the output of “false positives”.
- Recent studies have shown that ML models “are vulnerable to adversarial samples that are close to their original samples to human eyes but misclassified by Deep Neural Network (DNN).”
- This is a major problem as ML is used in a variety of fields, such as the medical industry, self – driving cars, malware detection, navigation, and many other fields.

- “Robust Machine Learning against Adversarial Samples at Test Time”

AN EXAMPLE OF ADVERSARIAL SAMPLES



- The researchers use an image classifier to conduct their research on adversarial samples to ML models.
- The image to the right shows three rows of the digits 0 through 9.
- The top row shows three images with their respective correct labels.
- However, the bottom two rows offer images that are adversarial. They have been slightly modified to “throw off” the classifier and produce incorrect results.

TYPES OF ADVERSARIAL ATTACKS

- As stated, an adversarial attack / sample is one that can easily be recognized by human eyes, but the ML model fails to classify it correctly.
- Therefore, there exist two attacks that can mislead a ML model: targeted and non - targeted.
- In an untargeted attack, there is no specified label that the attacker wishes to target for the sample that passes through a specified ML model.
- A targeted attack consists of the attacker aiming to mislead the ML model with an adversarial sample and an incorrect target label.

SOLUTIONS TO TARGETED ATTACKS

- There have been many research studies that attempt to produce an adversary sample given a sample and an incorrect label to better train the ML model. In this way, ML models are able to learn to recognize samples with correct target labels.
- Such a method is the Limited – memory Broyden Fletcher Goldfarb Shanno (L-BFGS) that aims to find such an adversarial sample with the box – constrained optimization problem.
- Other methods include: Fast Gradient Sign Method, DeepFool, and Carlini and Wagner (C & W), which are all attacks that aim to increase the defense of the ML model.

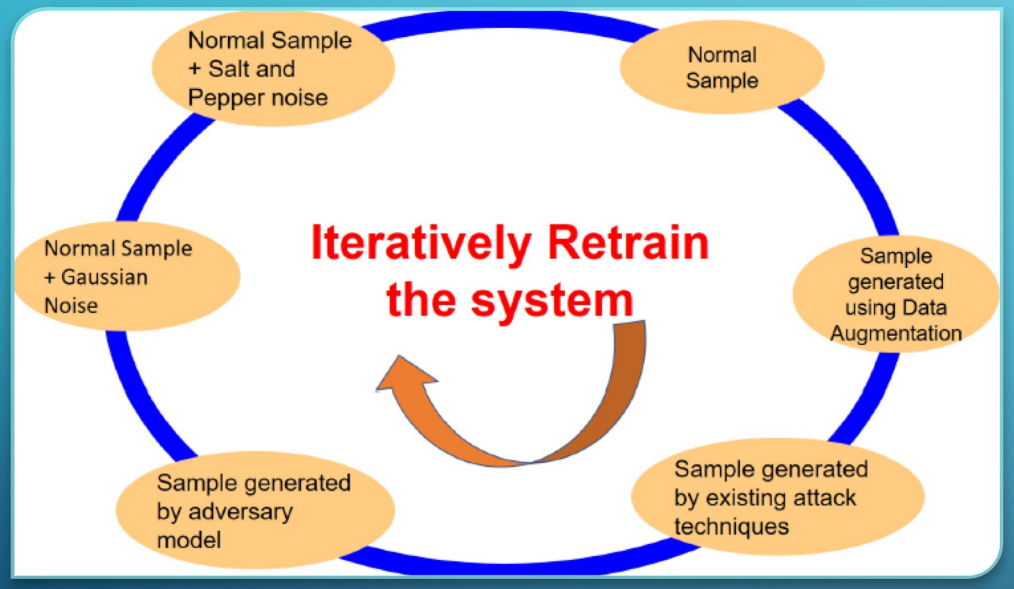
MACHINE LEARNING CRITERIA

- In order for a ML model to learn from data, careful attention must be paid to the following criteria:
- - Data: The data needs to be in large quantities and representative of what is desired for the ML model to yield.
- - Neural Network Architecture: How the ANN is built and constructed is also crucial as it is where the rules and characteristics are formulated.
- - Computational Power: Machine Learning requires much computational powers due to the heavy amount of matrix and gradient calculations.
- These three criteria need to be considered when constructing an efficient ML model with strong performance.

A PROPOSED SOLUTION

- One such group of researchers have proposed a solution that could assist in preventing these false positives.
- Their solution essentially consists of retraining the ANN with samples that slightly differ from the original input. If the samples are above a chosen threshold, then the ANN is tweaked to accept that sample with the correct label.
- This results in the ANN to separate important features of input from non – important features; this enables the ANN to reject input with incorrect labels while simultaneously still accepting input with accurate labels.

AN ITERATIVE APPROACH TO TRAINING



- The proposal offered by these researchers performed an iterative training approach.
- At each step of the training, a new adversarial sample is generated by different methods to improve the ML model.
- As seen at the beginning, one of these training methods involve adding “noise” to better train the model.

EVALUATION

Attack	FGSM	C&W	DeepFool
Original	29%	7%	29%
Robust Classifier	91%	70%	91%

- The researchers chose to use an image classifier to perform their research.
- The performance of the ML model classifier before and after the proposed training can be seen in the table.
- One can see that the accuracy of the classifier has improved dramatically.