Introduction

Neural network first caught many people’s attention in imageNet contest in 2012. Neural network increased accuracy to 85% from 75%. The following year, it is increased to 89%. From no one used neural network to everyone uses the neural network. Today we have 97% accuracy in using deep neural network (DNN). So the problem of image recognition by are artificial intelligence is solved. However, there is one catch(Carlini.N,2017).

The catch is that the DNN is really easy to be fooled. Here is an example. An image of a dog is classified as a hummingbird. Research studies by Google Brian, which is a deep learning artificial intelligence research team at Google, showed that any machine learning classifier can be tricked to give wrong predictions. The action of designing an input in a specific way to get the wrong result from the model is called an adversarial attack (Roman Trusov,2017). The input image is the adversarial image. This image is created by adding a tiny amount of perturbation, which is not so imperceptible to human eyes. After zooming into figure 1, a small amount of perturbation led to misclassify a dog as a hummingbird.



Figure 1



figure2

How does it possible? DNN models consist of transformation. Most of those transformations are sometimes very sensitive to a small change. Think of the DNN as a set of high-dimensional decision boundary. When an input is not perfect, if the decision boundary is too simple and linear, mostly it leads to misclassify.

Harnessing this sensitivity is a way to better understand and product robust algorithm in AI security. This paper aims to demonstrate the vulnerability of DNN by presenting some extreme scenarios - one pixel attack. As shown in figure3, only one pixel was perturbed, the classification was wrong in each image. Although, there is no profound defense to the attack as of current state, the investigation of one-pixel attack may shield lights on the behavior of DNN. Ultimately, it leads to the discussion of the security implications to future solution.



This paper proposed one pixel attack in a scenario where the only information available is the probability labels. Comparing to previous work, this proposal showed its effectiveness of successful attack rate up to 73%, its simplicity of semi-black-box which only required probability label no need inner information, and its flexibility in attacking more models, especially the networks that are not differentiable and the gradient calculation is difficult.

With the intension of creating an adversarial attack for better understanding the security of DNN, one pixel attack should be considered. Two main reasons: 1) a new way of exploring the high dimensional DNN by using fewer and lower dimensional slices. It is different from previous work, where perturbation was done by adding small value to each pixel. 2) a measure of perceptiveness to demonstrate the severity of one-pixel attack as comparing to a few pixel examples.

Related works

The sensitivity to well-turned artificial perturbation were investigated in various related work.

1. First perturbation was crafted by several gradient-based algorithms using back- propagation for obtaining gradient information. (C. Szegedy et al. )
2. Fast gradient sign algorithm for calculating effective perturbation

It was with the hypotheses of the linearity and high-dimensions of inputs were the main reason of vulnerability. (I.J.Goodfellow et al.)

1. Greedy perturbation searching method by assuming the linearity of DNN decision boundary (S.M Moosavi-Dezfooli et al.) Jacobian matrix to build “Adversarial Saliency Map” which indicates the effectiveness of conducting a fixed length perturbation through the direction of each axis (N. Papernot et al. )
2. The images can hardly be recognized by human eyes but nevertheless classified by the network with high confi- dence. (A. Nguyen et al. )
3. Several black-box attacks that require no internal knowledge about the target systems such as gradients, have also been proposed. only utilized it as a starting point to derive a further semi black-box attack which needs to modify more pixels (N. Narodytska et al)
4. Both natural and random images are found to be vulnerable to adversarial perturbation. Assuming these images are evenly distributed, it suggests that most data points in the input space are gathered near to the boundaries. ( A. Fawzi, S. M. Moosavi Dezfooli, and P. Frossard.)
5. A curvature analysis region along most directions around natural images are flat with only few directions where the space is curved and the images are sensitive to perturbation. (A. Fawzi et al.)
6. Universal perturbations (i.e. a perturbation that when added to any natural image can generate adversarial samples with high effectiveness) were shown possible and to achieve a high effectiveness when compared to random perturbation. This indicates that the diversity of boundaries might be low while the boundaries’ shapes near different data points are similar (S. M. Moosavi Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard.)

Reference:

<https://blog.xix.ai/how-adversarial-attacks-work-87495b81da2d>

<https://arxiv.org/pdf/1710.08864.pdf>

Carlini.N,2017 <https://www.youtube.com/watch?v=yIXNL88JBWQ>