Adversarial Learning Jufeng Gao Week 12 25/05/2022

Motivation

Learning to create robust neural networks that can survive attacks from hackers.

Extremely reliable models are necessary for:

* Autonomous driving
* Surgical robot
* Air traffic control
* Medical diagnosis, etc
* Adversarial Learning research techniques can help with this.

**Example of evasion**

Models may be trained for specific words and each word will have a score. A threshold will then decide if the total score of the words are enough to classify as malicious or not.

Adversarial examples - artificially created.

Synthetic examples:

Purposefully injecting some noise into your images can make then harder to be detected.

**Adversarial ML**

Adversarial machine learning is a technique that attempts to fool models

Examples of model attack

Applying different levels of noise to an image can trick a model into thinking the images are different

**How to recognise differences**

L2-norm – looks at all pixels for the average

L-INFINITY – looks at one pixel, the maximum difference

**Attack approach**

We update inputs instead of parameters in the Adversarial attack

Fast Gradient Sign Method (FGSM)

**White Box vs Black Box**

If we know he network parameters then the attack is a white box attack, and if we don’t know them then it is a black box attack.

**Black Box attack**

If you have the training data of the target network, then we can train a proxy model and use the trained proxy network to generate attack objects.

Ensemble attack – combining multiple networks together for an attack on a particular model. This increases the attack success rate.

One pixel attacks can be successful at fooling networks.

**Attacks in the physical world**

Attackers need to find perturbations that generalize beyond a single image since images move in the real world.

It is desirable to craft perturbations that are comprised mostly of colours reproducible by the printer. The printer needs to be able to accurately print the noise on the glasses.

**Back-door attacks**

Attacks in the training phase are called back-door attacks

**Passive Defense**

Adding a filter to your model can make an attack signal less harmful and not effective on your model.

**Proactive Defense**

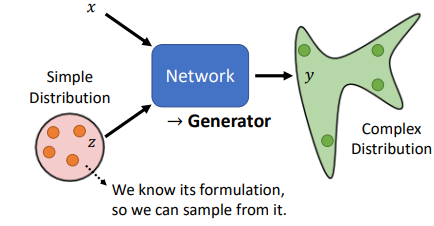
Training a model that is robust to adversarial attack

Requires applying the attack sample (adversarial labels) to the image for each of the training data.

Each image needs to be predicted one by one before the adversarial labels are applied to them. (Heavy computation)

**Adversarial Learning in Generative modelling – Generative Adversarial Networks**

Network as Generator



Generators may be able to help generate faces for example in a dataset, based on existing data

**Progressive GAN**