Adversarial Attack

Network光是正确率高是不够的，需要抵抗攻击。

Example of Attack

在一张猫的照片上加入一个非常小的扰动（向量每一维加一个非常小的杂信）--Attacked Image

我们希望被攻击后的图片输入后答案改变。

无目标的攻击：输出不是猫就行

有目标的攻击：输出不是猫且为确定的错误答案

How to Attack?

Non-targeted: $x^* = arg \min_{d(x^0,x)\leqslant\epsilon} L(x), L(x) = -e(y,\hat{y})$.

Target: $L(x) = -e(y,\hat{y}) + e\left(y, y^{target}\right)$

Attack Approach:

$$x^* = arg \min L(x)$$

Gradient Descent:

Start from original image x^0

For $t = 1$ to T:

$x^t \leftarrow x^{t-1} - \eta g$

if $d\left(x^0, x^t\right) > \epsilon$:

$x^t \leftarrow fix\left(x^t\right)$

Fast Gradient Sign Method(FGSM):

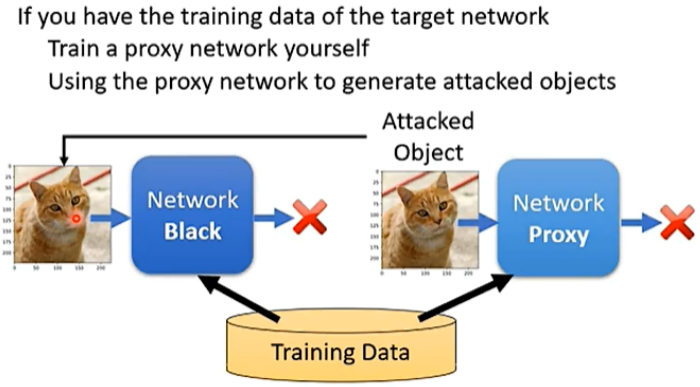
Start from original image x^0

For $t = 1$

$x^t \leftarrow x^{t-1} - \eta g$(g要么为1要么为-1,$\eta$为$\epsilon$)

White Box(知道模型参数) v.s. Black Box(不知道模型参数)

Black Box Attack:



What if we do not know the training data? 向模型中丢入输入，得到输出，将得到的输入输出拿去训练一个模型。

one pixel attack; universal adversarial attack

Attack in the Physical Word: eg. 人脸识别系统
被动防御：模型不动，给图片加一个filter（eg. 模糊化）；压缩再解压缩；Generator；
主动防御：

Proactive Defense

Adversarial Training

Training a model that is robust to adversarial attack.

Given training set $\mathcal{X} = \{(\boldsymbol{x}^1, \hat{y}^1), (\boldsymbol{x}^2, \hat{y}^2), \cdots, (\boldsymbol{x}^N, \hat{y}^y)\}$

Using $\mathcal{X}$ to train your model

For $n = 1$ to N

Find adversarial input $\tilde{\boldsymbol{x}}^n$ given $\boldsymbol{x}^n$ by an attack algorithm

We have new training data

Find the problem

$$\mathcal{X}' = \{(\tilde{\boldsymbol{x}}^1, \hat{y}^1), (\tilde{\boldsymbol{x}}^2, \hat{y}^2), \cdots, (\tilde{\boldsymbol{x}}^N, \hat{y}^y)\}$$

Using both $\mathcal{X}$ and $\mathcal{X}'$ to update your model