ADVERSARIAL EXAMPLES IN THE PHYSICAL WORLD

# 论文核心内容：

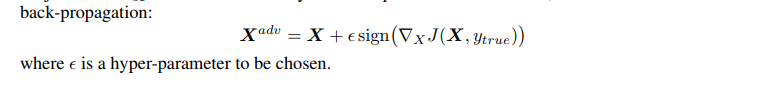
* 提出了迭代FGSM的方法，以及least likely 攻击的方法。
* 做了一系统的实验，验证在真实的物理环境中，那些直接使用数字图像生成的对抗样本是否依然具有使分类器分类错误的能力。为什么要有这么疑问呢？这是因为，之前的机器学习对抗工作都是直接在数字图像本身做的操作，这些对抗后的图像会直接调用api输入到神经网络里面去。但是在现实中，图像都是基于感光设备采集的，这些感光设备在生成图像的时候，其实是引入噪声的，例如环境光的影响、摄像头离图片的位置的影响，这些影响是很有可能把攻击者千方百计加入的微小扰动的作用给破坏掉的。作者把这种感光设备采集图像中引入噪声的过程叫做 photo transformation，然后这些transformation又可以看作是改变光照、模糊、加入噪声的这些子变化的合成，并做实验量化这些transformation的影响。

# Iterative FGSM

## Basic Iterative FGSM

作者对原始的FGSM方法做了一点推广，提出了迭代的FGSM.

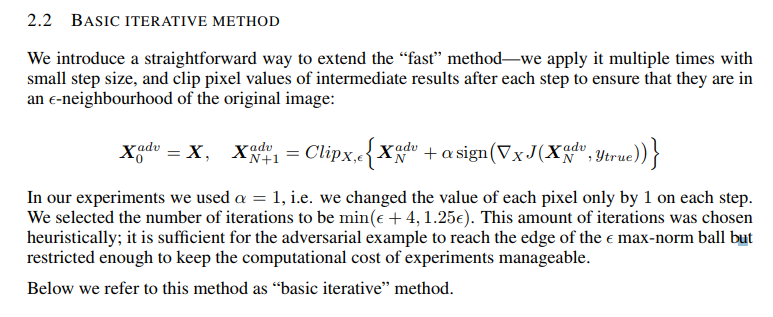
原始的FGSM:



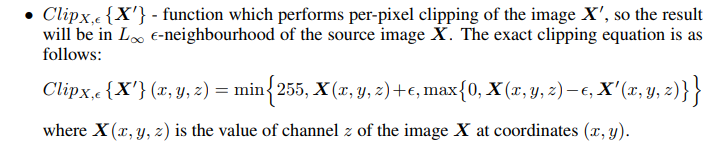
FGSM里面，每次寻找对抗样本的时候，直接是 ，这样是把每个像素点都变化了 这么大。这个改动其实是挺大，因为每个像素点都动了。我们考虑FGSM提出的理论假设：假设目标损失函数与x之间是近似线性的，即，然后我们想往x加上一个小的扰动，使得变的最大。而,这玩意要最大，那当时是直接。

但是 这个线性假设不一定是正确的呀，如果J和x不是线性的，那么是否存在之间的某个扰动，使得J增加的也很大？此时x的修改量就可以小于 了。

于是，作者就提出迭代的方式来找各个像素点的扰动，而不是一次性所有像素都改那么多的想法。



其中有个Clip截断函数：



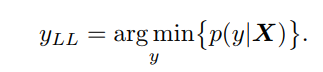
这个截断函数写的花里胡哨的，但是表达的意思说穿了就一文不值，这个函数等价于：

它其实是想说X’的每个像素都得在X对应像素的邻域，同时又得是有意义的图片。

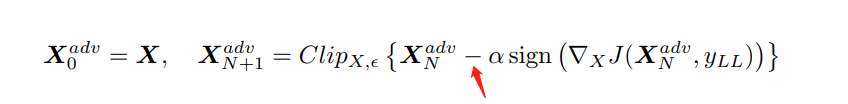
于是这个迭代的含义就是：每次在上一步的对抗样本的基础上，各个像素增长（当然也可以减少），然后再执行裁剪，保证新样本的各个像素都在X的各个像素的邻域内。这种迭代的方法是有可能在各个像素变化小于 的情况下找到对抗样本的，如果找不到大不了最差的效果就跟原始的FGSM一样。

## Iterative Least-likely Class FGSM

最不相似的FGSM,其实是有目标攻击，它把目标设置为原样本最不可能的标签，然后执行有目标攻击。

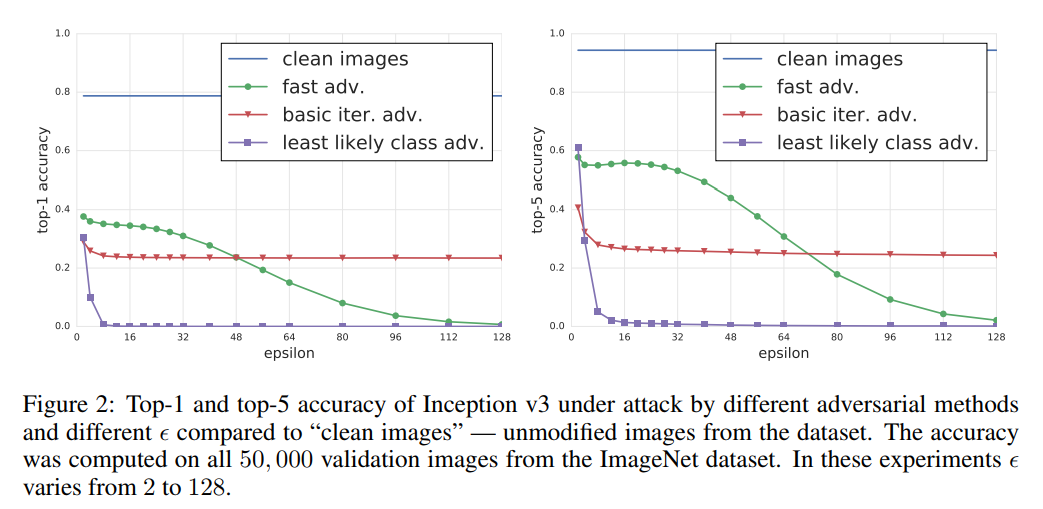


再迭代攻击：

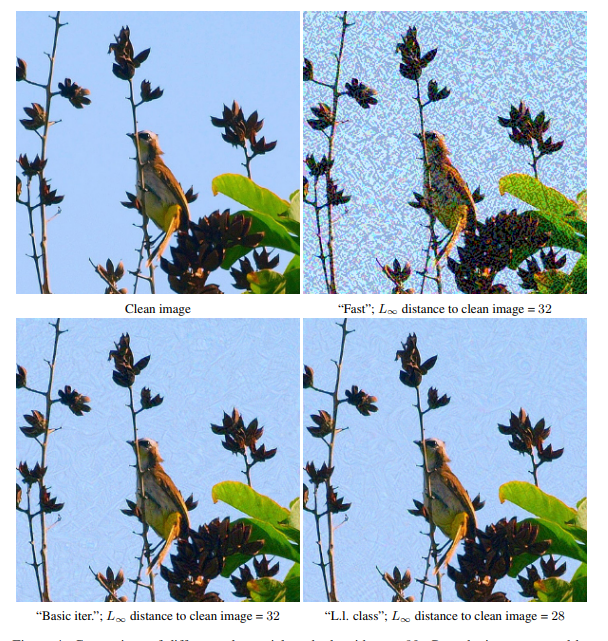


注意，这里是减去扰动，其原因和FGSM的定向攻击是一样的。

# 迭代FGSM的攻击效果



可以看到，当 比较小的，迭代攻击效果是更原始的FGSM好的。而且Least-likely 的攻击效果最好，从图片上看也能看出来least-likely的改动是最小的。

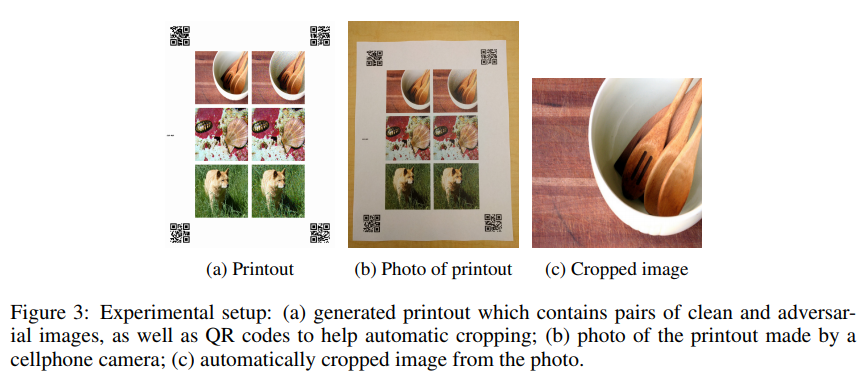


# 现实攻击

作者想实验一下这些生成出来的对抗图片，被照相机拍照后是否还具有对抗效果。注意！作者没有对照相机对图片的扰动进行建模。而是直接拿着这些对抗图片的拍照结果放进模型试试。

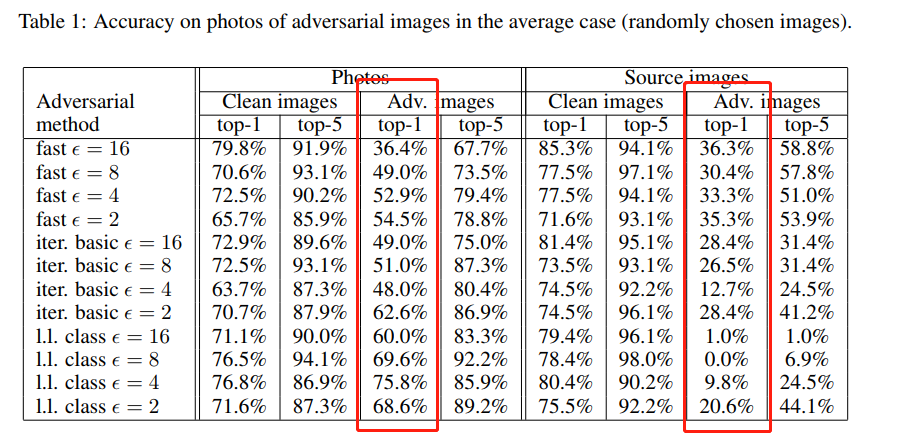
## 实验设置

不得不说，这篇论文的实验设置还是很系统化的。

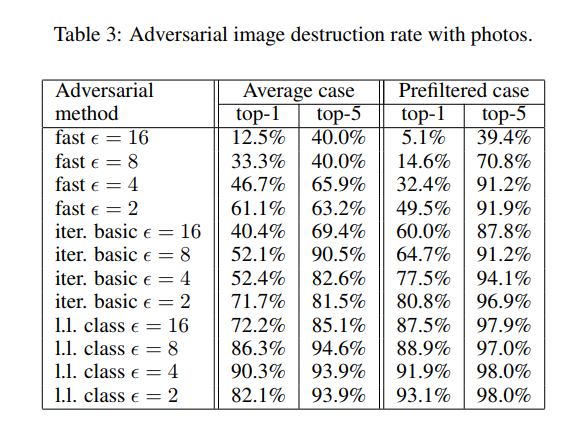


步骤：

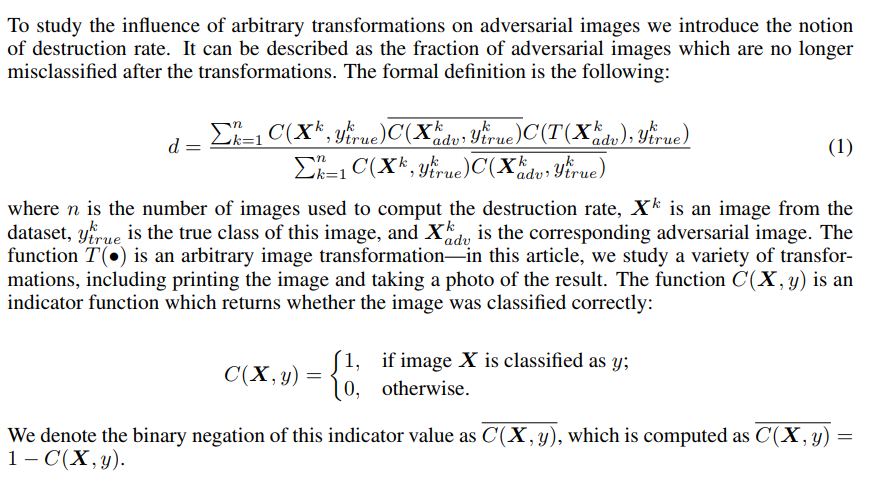
1. 使用ImageNet训练一个Inception V3网络。
2. 寻找验证集里面特定样本的对抗样本，（使用不同的对抗方法，以及不同的）
3. 把原始样本和对抗样本打印出来，并用手机拍这些打印出来的图片，图片有二维码定位用。
4. 使用Inception V3模型识别手机拍出来的图片。



可以看到，Photo Transformation还是对对抗样本的对抗性有一定的损失作用的。其中FGSM的影响相对比较小，因为FGSM改动大。而迭代的都影响大，这是因为它们改动小，然后这些扰动又被传感器的精度问题给破坏了。



这个表表示，Photo Transform对对抗性的破坏能力。Destrction rate的定义如下：



表示经过Photo Transformation后，有多少“有效”对抗样本（指对抗样本让分类器分类错误，而原样本可以被正确分类）又不会被错误分类的比例。

# 其它变换对对抗样本的影响

作者探究了不同的子变换对对抗样本的对抗性的影响。这个是通过直接改数字图像来完成的。