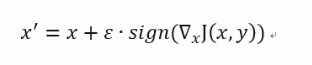
**1.FGSM的原理**

FGSM的全称是Fast Gradient Sign Method(快速梯度下降法，**在白盒环境下**，通过求出模型对**输入的导数**，然后用**符号函数得到其具体的梯度方向**，接着乘以一个步长，得到的“扰动”加在原来的输入  上就得到了在FGSM攻击下的样本。

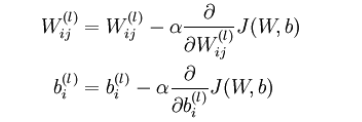
　　FGSM的攻击表达如下：



　　　　ε 是一个调节系数（一个很小的扰动，其他的文章中应该是符号η），sign() 是一个符号函数，代表的意思也很简单，就是取一个值的符号，当值大于 0 时取 1，当值等于 0 时取 0，当值小于 0 时取 -1，∇ 表示求 x 的梯度，可以理解为偏导，J 是训练模型的损失函数。

那么为什么这样做有攻击效果呢？

就结果而言，攻击成功就是模型分类错误，就模型而言，就是加了扰动的样本使得模型的loss增大。而**所有基于梯度的攻击方法都是基于让loss增大这一点来做的。**可以仔细回忆一下，在神经网络的反向传播当中，我们在训练过程时就是沿着梯度方向来更新更新w，b的值。这样做可以使得网络往loss减小的方向收敛。



那么现在我们既然是要使得loss增大，而模型的网络系数又固定不变，唯一可以改变的就是输入，因此我们就利用loss对输入求导从而“更新”这个输入。(当然，肯定有人问，神经网络在训练的时候是多次更新参数，这个为什么仅仅更新一次呢？主要因为我们希望产生对抗样本的速度更快，毕竟名字里就有“fast”，当然了，多次迭代的攻击也有，后来的PGD（又叫I-FGSM)以及MIM都是更新很多次，虽然攻击的效果很好，但是速度就慢很多了）

**原论文解读和公式推导**

主要参考链接[【论文笔记】（FGSM）Explaining and Harnessing Adversarial Examples - 李斯赛特 - 博客园 (cnblogs.com)](https://www.cnblogs.com/setdong/p/16300720.html)

有问题，待定