Towards Evaluating the Robustness of Neural Networks

# 论文代码：

http://nicholas.carlini.com/code/nn robust attacks.

# 论文motivation:

蒸馏网络说自己短小强撼，可以为目标网络模型提供很强的鲁棒性，号称自己能够把已经出现的攻击的成功率从95%锐减到0.5%。作者不信这个邪，于是基于L0,L2,L无穷范数提出三种新的攻击方法，尝试攻击蒸馏网络，而且还攻击成功了。而且作者认为，基于他们提出的三种攻击而生成的对抗样本具有很强的迁移性，新的防御模型应该要抵抗这些攻击。

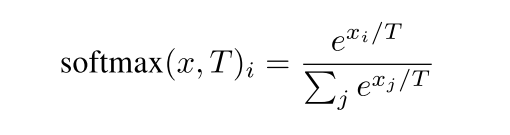
作者还尝试使用不同的目标函数去生成对抗样本，实验发现不同的目标函数会大大影响攻击效果。===》白盒攻击，

# 蒸馏网络:Distillation

蒸馏网络一开始是提出来用于导师训练中把一个很大的导师模型降低为一个更小的蒸馏模型。蒸馏方法首先会以标准的方式训练得到一个导师模型，然后使用导师模型计算训练集里面所有样本的soft label,这里的soft label是指网络输出的各个类别的概率值（也就是softmax之后的输出结果），然后把整个softmax给的概率向量作为样本的标签变化整个训练集。这么做的好处是可以使得学生模型尽快的学习到导师的映射能力，相比hard label来说，soft label至少会包含更多的类别信息。

Defensive distillation 是指使用这种套路增强网络鲁棒性的一种防御方法。Defensive distillation与原始的distillation有两点不同：在defensive distillation中，学生模型和导师模型规模一样；defensive distillation会使用一个更大的蒸馏温度T，进而隐藏模型额外信息。

softmax函数是神经网络的最后一层，defensive distillation修改softmax 函数：



引入了一个温度参数T。显然有softmax(x,T)=softmax(x/T,1)。引入T的作用在于，当T很大的时候，T会把x之间的差异抹除掉，使得softmax(x,T)的输出接近于一个随机分布；而当T很小的时候，T会把x中最大的那个强调出来，其它分量给滤除掉。引入T后，各个分量之间的偏序关系还是保持不变的，即argmax softmax(x,T)=argmax softmax(x)。

举个例子：x=[1,2,3,4,5]

softmax(x) =[0.011656230956039607, 0.03168492079612427, 0.0861285444362687, 0.23412165725273662, 0.6364086465588308]

softmax(x/10)= [0.1621203478685732, 0.1791706936926544, 0.19801424004056153, 0.21883957945767904, 0.2418551389405318]

softmax(x\*10)=[4.248161380306793e-18, 9.357198133414646e-14, 2.06106004620887e-09, 4.539786860886242e-05, 0.9999546000702375]

softmax(x)中各个量的差异很大，最后一个是最大的。

softmax(x/10)后，各个差异不大了。

softmax(x\*10)后，各个量差异特别大，把最大的那个量的强调出来。

但是，这些结果中各个量的偏序关系是与x一致。

Defensive distillation 的步骤有四个：

1). 首先训练一个导师网络，最后一层softmax设置温度参数T。

2). 计算训练集里面的各个样本的soft label.把样本的y替换成soft label。

3). 训练学生网络，softmax也使用温度参数T。首先在训练学生网络的时候，受soft label的影响，输入到softmax之前的logit就被放大了T倍，

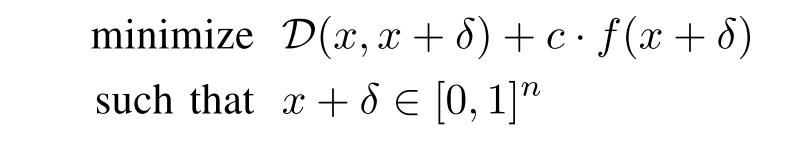
4). 学生网络在测试阶段，softmax中把温度设置为1。

最重要是这个学生网络中softmax把温度T设置为1。

测试的时候，温度T设置为1,相比于训练阶段，相似的x最后输入到softmax的值会扩大T倍，参考到上面的softmax(x\*10)这个例子，这种扩大T倍会把softmax中正确标签的输出概率强制为1，而其他标签强制为0。这就是一个hard softmax了，很难获取梯度。🡺一个想法：：：**是不是，只需要hard softmax就可以了呢？也不用复杂的蒸馏了。**

而logit就只会“扩大”T倍，这其实就是softmax(x,T)=softmax(x/T,1)的意思。测试阶段，温度设置为1，但是x还是x。

# 作者提出的攻击方法



其中f()函数是一个很特别的函数：

C是分类函数

# 借鉴之处：

* 把自己提出的新攻击方法当作检验模型鲁棒性的baseline.
* 论文写的特别好，将背景技术啥的介绍的很详细，给出自己的理解。