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，

# 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.