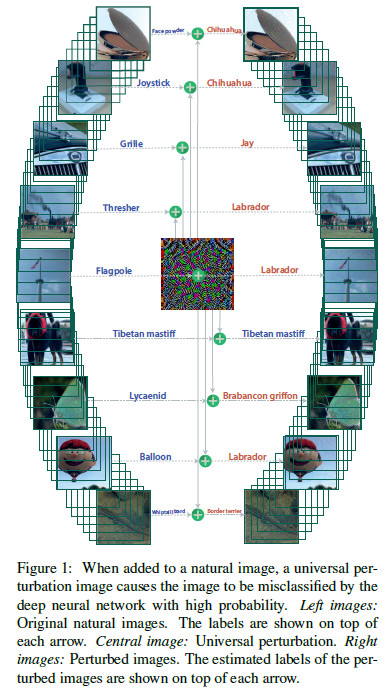
Universal adversarial perturbations

# 代码地址：

<https://github.com/LTS4/universal>

# 核心思想：

本文提出一种 universal对抗扰动，universal是指同一个扰动加入到不同的图片中，能够使图片被分类模型误分类，而不管图片到底是什么。示意图：

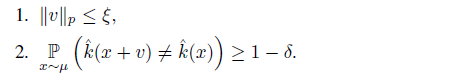


# 形式化的定义：

对于d维数据分布 ，里面的每一个样本，存在一个分类器。是一个扰动，v满足：

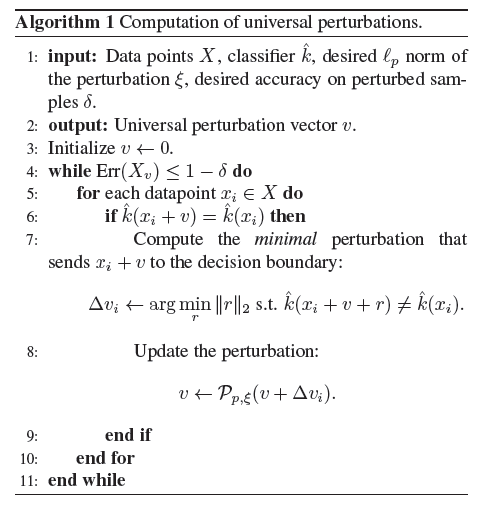


同时，还有一个范数的约束：



换句说，需要找到一个对抗扰动v，这个扰动可以加到所有的样本点上，而且会以的概率让对抗样本被分类错误。

作者提出了一种算法来寻找这种对抗扰动：



其中：



表示把v限制在以 为半径的球上。

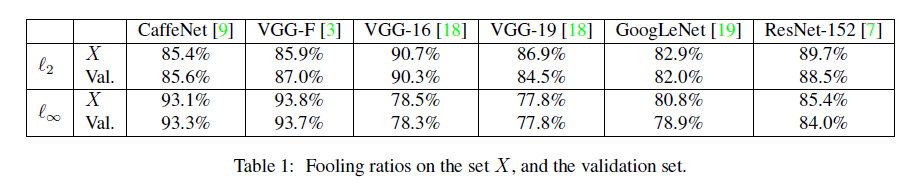
这个算法的思想是：

从u里面采样出一个样本集 里面有m个图片，然后迭代地寻找能够让m个样本 比例的样本被分类错误的对抗扰动。

一开始v=0,没有什么扰动，然后对于每个样本,看它加上扰动v后，会不会分类错误，如果分类错误，则下一个样本，否则寻找一个微小的扰动，使得x+v+ 被分类错误。持续这个过程，直到满足错误率。

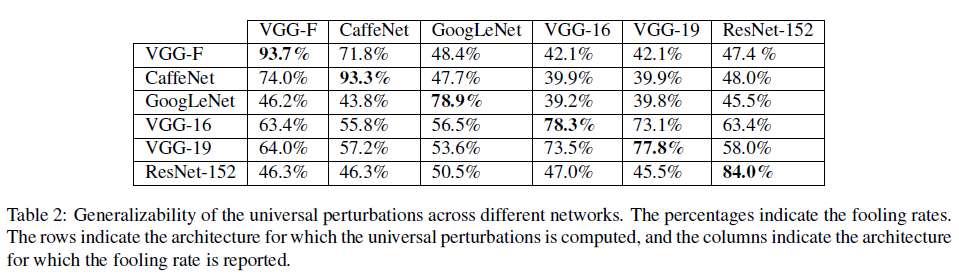
# 对抗效果：

基于一个数据集，寻找到的universal 对抗扰动，可以使得新样本也分类错误：

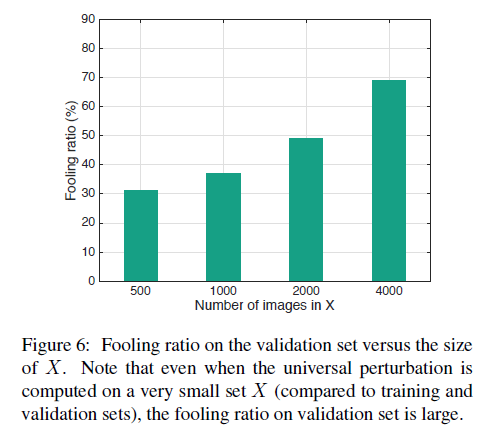


其中X是训练集，Val是验证集。

不同模型寻找到universal扰动，也可以使其它模型分类错误。



X训练集样本量的影响：



作者使用的是ImageNet，改数据集有1000类。可以看出当X包含4000个样本时，每个类平均4个样本，攻击成功率可以达到70%左右。而x包含500个样本，平均每个类0.5个样本，对抗成功率也有30%。

# Universal pertubations存在性解释：

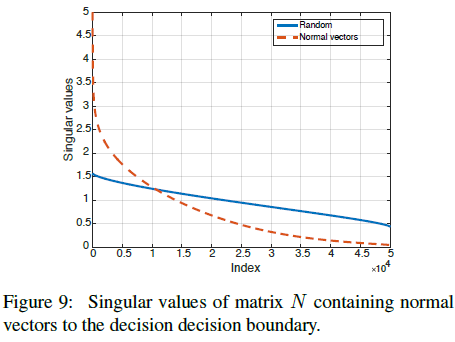
对于验证集里面的每个样本x，我们寻找它的对抗扰动r(x)= 

这种r(x)其实是看作是x处分类模型决策界的法向量。

作者提取n个样本处的这种法向量，形成矩阵N:



通过对N进行SVG分解，作者发现N的奇异值有一些特别大，而另外一些特别小：



这种现象意味着，这些法向量其实可以存在冗余的，换句话说这些法向量所在决策界存在着冗余性和相关性。基于SVG分解的向量构建的对抗扰动其中也能取得38%的对抗准确性。这就说明了，神经网络学习得到的决策界，在高维空间是存在相似的相关性的。通过样本子集X 可以获得m个样本的决策界相关性，这种相关性在其它不同的样本周围的决策界上依然存在。

Universal Pertubation则是以最大化成功率的使用这些法向量构建扰动，因而它也会学习到决策界的相关性。

# 启发与未来工作点

* Universal pertubations的存在，使得我们可以通过预训练找好这种通用扰动，然后在实际预测的时候直接把这些扰动加到正常样本上来实施攻击。在这种攻击模式下，攻击者在预测阶段是可以做到**实时**的，可以不用像传统方法那样花大量的时间针对测试阶段遇到的样本计算对抗扰动。 在实时对抗领域，universal pertubation会有很大的空间。
* 应该也可以使用GAN来产生这种universal pertubation。Generator接受噪声，产生universal pertubations，然后把这种扰动加到所有的正常样本上，使得鉴别器（这个鉴别器可以事先训练，GAN训练过程中固定住它）把x+v识别错，同时在损失函数中加入对v的范数约束。
* 有目标的pertubations.文中提到的方法是无targeted的，只要universal pertubation把原始样本被分类器识别错了就行。这里的targeted可以是： 1. 只针对特定的类别，做干扰的，例如universal pertubations可以加入所有图片上，但只会把含有狗的图片攻击成功。 2. Universal pertubations加入所有图片后，会把所有图片的狗的得分都大化。