Adversarial Examples Are Not Bugs, They Are Features

这篇文章研究对抗样本这种现象：对抗样本是数据分布的固有特征，并对此进行研究。

1

为了研究这种现象，文章将输入数据的特征区分，特征可以理解为图片的某个维度像素值这样，分为：
鲁棒特征，扰动此特征不会改变预测结果
不鲁棒特征，相反
对抗样本是不鲁棒特征的存在引起的。

文章实验：
1:只包含鲁棒特征的数据集可以训练一个良好的鲁棒模型，这个实验说明对抗样本这种现象不是网络训练过程产生的，是数据集固有的特性，figure1(a)上面
2:使用扰动后的对抗样本训练一个高精度的模型，此模型在原始没有加扰动的数据集上还可以正确分类，这个实验说明不鲁棒的特征是和预测分类有关的，翻转这些特征值可以引起分类错误，figure1(b)
3:只使用不鲁棒特征进行训练，也可以训练一个好的模型，但是不鲁棒，figure1(a)下面

2

模型定义，2分类模型(x,y)，x是从D分布上选取的数据，y是正负1.
首先将x做一个映射f，从原始数据空间上映射到实数空间，再将实数空间映射到均值为0方差为1的分布上。f可以理解为是图片某个维度的像素值对应到均值为0方差为1的分布上之后的数据。

ρ-useful特征，对于一个特征f，如果f的值和分类标签正相关，那么称特征是有用的，即此特征有助于标签正确分类
γ-robust useful特征，f是有用的，并且在f上加扰动之后，无论怎么扰动，此特征和标签还是正相关，那么称此特征是鲁棒有用的，即此特征被扰动了之后还是有助于标签正确分类的
Useful non-robust特征，有用，但是扰动之后有可能会无助于标签正确分类。这些特征帮助模型进行标签分类，但是也引起了对抗样本。

分类模型，线性模型，c=sgn(w*f+b)

标准训练：使模型预测结果尽可能得与y(1/-1)一致，即y*c尽可能得大，那么就是使-y*c尽可能得小

鲁棒训练：使数据在扰动下的最大损失尽可能得小

3 找到鲁棒与不鲁棒的特征，将输入的特征分开，并构造了两种数据集：
鲁棒数据集：从原始数据集中删除了不鲁棒的特征
标签错误数据集，模型训练时每个标签都是错误的，同时轻微扰动不鲁棒特征，即使用对抗样本进行训练，但是这样的模型可以在标准数据集上（没有改动的数据集上）预测良好
对应文章开头的两种实验

3.1 区分鲁棒与不鲁棒特征
从原始数据分布构建鲁棒特征分布：
如果某个特征对于分类是有用的，那么我们保留这个特征不变，否则，改变这个特征，让这个特征在新的鲁棒分布下期望为0，即对于预测没有影响(5)
同时由于改变候选的特征有多种，选择一种与原始的最接近的(6)

结果如图2
左边是三种数据分布
原始数据
鲁棒的数据，即去除了不鲁棒的特征
标签错误数据集，即加了扰动使分类错误
右边是结果
标准数据集训练高准确率低鲁棒率
标准数据集再进行鲁棒训练高准确率高鲁棒率
鲁棒数据集训练高准确率高鲁棒率
标签错误数据集还是可以正确分类，高准确率低鲁棒率

3.2 只用不鲁棒特征是可以训练数据集的
构造不鲁棒特征数据集步骤：
1:得到一个对抗样本x_{adv}（7），对抗样本的分类t要么是随机的要么是与原始分类y有关但是不同的
2:去除x_{adv}中与y有关的鲁棒特征
去除鲁棒特征文章提供了两种方案
1:如果此对抗样本的分类是随机选取的，那么就保存不鲁棒特征，其他置为0，那么在训练中只有不鲁棒特征对分类起作用的是不鲁棒特征（8）
2:如果分类是与原始分类有关的，就让不鲁棒特征在分类中起作用，让鲁棒特征在原始分类中起反作用

使用这样的数据集进行训练，在标准数据集上的预测结果还可以，表1.

3.3 不鲁棒特征可以引起对抗样本的转移性
使用不鲁棒数据集在不同模型上进行训练，模型预测精度越好，即使用不鲁棒特征越充分，越容易收到对抗转移攻击，说明不鲁棒特征可以引起对抗样本转移性

4 提出研究此问题的框架

（10）数据分布是高斯分布，预测结果是正负1
（11）模型是使预测结果尽可能得与正确结果相同
（12）鲁棒训练是在扰动范围下使最坏的点不那么坏

定理一大概想说明在存在对抗样本情况下的损失比不存在对抗样本情况下的损失是大这么多的，其中与方差的迹有关，方差的迹又与每个特征的不鲁棒性有关。如果存在某个特征非常不鲁棒，那么迹就很大，对抗样本存在时的损失就会比不存在时的损失大很多

定理二大概想说明方差随着扰动范围变大而变大

定理三说模型梯度和分类向量之间的夹角更小，式子表示夹角的cos值，越大说明夹角越小，此定理可能想说明鲁棒模型更平滑一些

图4，在不同的扰动范围下数据分布以及模型学习到的分类边界，数据的中心点μ不变，但是方差一致变大，特征x1就是不鲁棒的特征，x2就是鲁棒的特征，改变x1容易引起分类错误，x2则相反。