Adversarial Training and Robustness for Multiple Perturbations

Abstract
现在做对抗样本防御的工作都是针对某一种特定的攻击，如对于p范数如$\ell_1, \ell_2, \ell_\infty$或者空间变换如移动、旋转的攻击。在模型针对一种攻击做对抗训练时，一般不会增加模型对于其他攻击的鲁棒性。本文解释了这种情况出现的原因，并且对模型进行对抗多种攻击组合的鲁棒性训练，并且提出了一种有效且强力的针对1范数进行攻击的算法。

1 Introduction

Can we achieve adversarial robustness to different types of perturbations simultaneously?
我们能对模型进行同时对抗多种不同攻击的鲁棒性训练吗?

Mutually Exclusive Perturbations (MEPs)
ℓ_∞ *and* ℓ_1, ℓ_∞ *and rotation/translation*
互斥扰动组合，即模型对于一部分扰动的抵抗能力增强会导致对于其他部分扰动的抵抗能力减弱。

提出两种对于多攻击的训练方法
Average strategy：对于一个对抗样本x，将所有攻击类别中最好的对抗样本都挑出来进行分析、训练
Max strategy：对于一个对抗样本x，在所有攻击类别的对抗样本中只挑出来攻击最成功的（分类损失最大的）进行分析、训练

SLIDE
文章提出的一种针对1范数进行攻击的算法。

图1展示了对于max策略进行实验的结果。

gradient-masking
梯度遮盖，即对于无穷范数攻击进行鲁棒性训练的模型，会让人觉得针对1范数、2范数也是鲁棒的。为什么会这样，图3给了个大概解释。对于使用梯度攻击的PGD，由于沿轴是平滑的，所以找不到对抗样本；但是对抗样本是存在且很好找的，多使用随机初始点或者使用梯度无关的攻击。

Affine adversary
即多种攻击的线性组合，攻击程度提升了，以前是某个范数上的攻击，现在可以用时进行多个范数的攻击（加上空间的变换），后续部分会给出组合攻击的例子与证明。

2 Theoretical Limits to Multi-perturbation Robustness

2.1 对抗风险定义，定义在多个攻击域上的攻击成功率

$l(f(x), y)$
损失函数，如果$f(x) \neq y$所产生的损失

$R_{adv}(f; S) := E_{(x,y)\sim D}[\max_{r\in S} l(f(x + r), y)]$
分类器f在数据集D与扰动集合S上所产生的最大损失的期望
即对于每一个样本，找到它在攻击域（ϵ ball）中最坏的点，将所有样本的最坏点统计求期望。
R_{adv}^{max}
对一个样本点，找到所有攻击域中最坏的点
R_{adv}^{avg}
对一个样本点，找到每个攻击域中最坏的点，并取平均

2.2 实验所用模型，二分类模型
在文章中，所有证明的假设是基于此模型的，此模型输入为d+1维度，第0维度是对无穷扰动鲁棒，其余维度不鲁棒。
由于此模型是一个特例模型，那么从此模型分析出来的对抗风险是一般模型一个下界，鲁棒性是一般模型的一个上界。

2.3 防御无穷范数攻击和1范数攻击是互斥的，证明见附录
即模型对于两种攻击总共能防御的上界是1/2
证明解析：证明是用特例证明的，即证明了鲁棒的上界
扰动r_∞, r_1如文章中所定义，是一个特殊的扰动，与模型假设配合，假设此扰动可以改变模型预测的结果。即加在第0维度的1范数扰动和加在后续维度的无穷范数扰动可以改变模型预测的结果。
定义两种x分类状态p_{+-}, p_{-+}，前者第0维度导致分类为正后续维度导致分类为负，后者相反，扰动加到这两种状态分布的x上时会产生有效的结果。
对于无穷范数扰动，加上扰动分类不变，分两种情况：
a:原始分类为1，因为无穷范数扰动加在除了第0维度的后续维度上，那么不改变第0维度对结果的贡献，所以要第0维度为正；又因为无穷范数扰动会改变所改变的维度对于结果的贡献，那么后续维度要为负。所以此时的概率就是$Pr[y = +1] * p_{+-}$
b:同理，分类为-1的时候为概率如文章所示。
对于1范数也是类似的。
对于x，加上两个范数扰动之后不改变分类结果的概率和为1.即一个增加另一个肯定减少。
又由于这是一个特例，所以是鲁棒性的上界。

2.4 防御无穷范数攻击和空间变换攻击是接近互斥的
即模型对于两种攻击总共能防御的上界接近1/2

2.5 组合攻击
如果模型对于多种攻击域都是鲁棒的，对于其线性组合的扰动是不是鲁棒的？详细见附录
a:在线性模型中，对一种p范数组合鲁棒的模型对其他组合都鲁棒
定义一些中间变量v，v是现有扰动对于结果的改变亮，最大最小为max,min
由于是线性的，那么两个攻击域的线性组合的最大最小是两个点攻击中的最大最小端点
那么加上仿射扰动改变结果的概率，分为两种情况：
原来是正，加上扰动为负了，那么此种情况就是
$\omega^T r + \omega^T x + b < 0 \,|\, y = +1$
$\omega^T r + h(x) < 0 \,|\, y = +1$
$\omega^T r < -h(x) \,|\, y = +1$
同理另一种情况是
$\omega^T r > h(x) \,|\, y = -1$（此处我认为可能是$\omega^T r > -h(x) \,|\, y = -1$，不过无所谓了证明和它没关系）
那么由于不等号的性质吗，就可以换成扰动对于结果影响的最大最小值，即将仿射扰动通过最大最小转换成了两部分扰动的集合。证明了两部分相等。
b:在非线性模型中，对一种p范数组合鲁棒的模型对其他组合不一定鲁棒
c:在线性模型中，对一种无穷范数和空间攻击鲁棒的模型对其他组合不一定鲁棒

3 New Attacks and Adversarial Training Schemes

训练中的风险定义，即在训练中找到使分类最不准确的点、损失最大的点
$$\hat{R}_{adv}(f;S) = \sum_{i=1}^{m} \max_{r \in S} L(f(x^{(i)} + r), y^{(i)})$$
然后优化这个风险，最小化这个风险，即
$$\min \hat{R}_{adv}(f;S) = \min \sum_{i=1}^{m} \max_{r \in S} L(f(x^{(i)} + r), y^{(i)})$$
对于$x^{(i)} + r$，可以使用一种强力的攻击算法找到，记为$A(x)$

对于两种处理多种攻击域对抗样本策略，所以有两种训练方式
max找到所有攻击域中最坏的点
avg将每个攻击域中最坏的点都取出来求平均

SILDE攻击算法
为什么要有这个算法，因为PGD对于1范数的更新太慢，每次只能优化更新一个变化最快的维度，即只能沿着一个维度下降。
此算法一次更新多个维度，设置一个更新的百分比q，每次更新q比例的维度，更新速度指数级提升。

4 Experiments

表格1、2、3以及4、5.