|  |  |
| --- | --- |
| **文章信息** | **方法** |
| **阅读程度**：  粗度了解  **文章标题：**  Adversarial Training and Robustness for Multiple Perturbations  **中文标题：**  对抗训练和对多种扰动的鲁棒性  **发表于：**  NIPS2019  **作者：**  Florian Tramèr  **单位：**  斯坦福 | 方法很简单，进行了一些理论分析  方法:作者给出了四种攻击，1范数，2范数与无穷范数，并且指出了如何利用梯度去实现这三种范数的攻击： For the ∞-norm, the steepest descent direction is sign(g), and for 2, it is g/g的欧式距离 For the 1-norm, the steepest descent direction is the unit vector e （按这样添加扰动，可以获得无穷范数，2范数，1范数为1的扰动）。但是对于1范数而言，最速扰动仅对一个像素值添加了扰动，且扰动大小是1的时候还超出了像素范围（对一个像素点添加的最大扰动为1-x或-x），因此作者提出了这样的一范数扰动：先从梯度矩阵中找出绝对值最大的一部分梯度，其余梯度值都置为0（即只更新变化速度最快的像素点），然后就是求符号函数，并且用一范数做归一化，保证扰动值的1范数为1，然后乘以就得到了1范数为的扰动。  训练过程是这样，第一种MAX，在每一步对一个样本用所有方法生成对抗样本，选择使损失函数最大的对抗样本进行对抗训练，第二种AVG，就是每种对抗样本都用，然后求平均  关于理论分析部分，作者定义了一个Mutually Exclusive Perturbations (MEPs)概念，，即某两种攻击的平均错误率大于1/C,其中C为分类问题的类别数，即认为两种扰动为互斥扰动。然后理论分析了无穷范数与1范数扰动是互斥扰动，无穷范数扰动与空间变换为互斥扰动。  最后还设计了一种仿射攻击，简单来说就是对两种攻击进行加权平均。 |
| **摘要及介绍** | **实验** |
| 针对对抗性例子的防御，如对抗性训练，通常是针对单一的扰动类型（例如，小的`∞-noise）而定制的。（背景）  对于其他的扰动，这些防御并不能提供保证，有时甚至会增加模型的脆弱性。（问题）  我们的目的是理解这种鲁棒性权衡背后的原因，并训练同时对多种扰动类型具有鲁棒性的模型。（目的）  我们证明了对不同类型的lp有界扰动和空间扰动的鲁棒性的权衡，必须存在在一个自然和简单的统计环境中。我们通过在MNIST和CIFAR10上展示类似的稳健性权衡，证实了我们的正式分析。  我们提出了新的多扰动对抗训练方案，以及对l1范数的有效攻击，并利用这些方案表明，针对多重攻击训练的模型无法实现与单独攻击训练的模型竞争的鲁棒性。  特别是，我们发现使用一阶`∞、`1和`2攻击MNIST的对抗性训练仅达到50%的鲁棒精度，部分原因是梯度掩蔽。  最后，我们提出了在扰动类型之间进行线性插值的仿射攻击，并进一步降低了反向训练模型的精度。 |  |