|  |  |
| --- | --- |
| **文章信息** | **方法** |
| **阅读程度**：  粗读  **文章标题：**  Perceptual adversarial robustness Defense against unseen threat models  **中文标题：**  针对未见威胁模型的感知对抗鲁棒性防御  **发表于：**  ICLR2021  **作者：**  Cassidy Laidlaw  **单位：**  University of Maryland | 1. **神经直觉威胁模型**   就是对传统威胁模型的一种扩充，这里引用的是LPIPS距离去作为约束生成对抗样本。前面的工作对抗样本的生成基本上都是基于p范数或者其他的一些范数，当噪声大小在这些约束的范围内时，认为对抗性噪声对人类是不可见的。作者认为LPIPS约束是一种更可靠的约束，去衡量什么样的噪声是不可见的，即对抗样本与原样本之间满足LPIPS约束，认为此时对抗性噪声是不可见的。  关于LPIPS约束其实很简单，将两个样本输入神经网络之间，对比神经网络各层输出特征图的输出之间的欧氏距离，当欧式距离小于epsilon时，认为两个样本之间没有显著差异。这里作者认为，成功地将两个样本之间的视觉差异，通过神经网络进行了合理的量化。（本质上说，2范数是直接对比样本之间的二范数距离来衡量样本之间的差异，LPIPS约束是衡量两个样本经过神经网络后所有特征图之间的二范数距离来衡量样本之间的差异）。   1. **感知对抗攻击**   有了直觉威胁模型作为约束，只需要生成满足此约束的对抗样本，即为感知对抗攻击，作者的主要贡献也在于解决了如何生成对抗样本这个问题。  第一种生成方法PPGD（感知PGD），通常来说PGD攻击是求梯度来修改样本从而最大化损失函数，之后将获得的对抗样本映射到p范数的约束空间中，PPGD中第一步梯度上升容易实现，第二步映射难以实现（因为本文用的是LPIPS约束，不是p范数）。这里作者使用泰勒公式去求解此问题。  第二种方式LPA（拉格朗日感知攻击，lambda像是拉格朗日系数？）是参照cw攻击，将约束条件纳入损失函数之中自行训练：，最大化此损失函数同时获得使交叉熵损失增大，同时满足LPIPS约束的对抗样本。   1. **感知对抗训练**   就用感知对抗攻击进行对抗训练，这里有两个版本，第一种情况，计算LPIPS约束的网络与主网络是同一个网络，此时每次主网络训练后同时更新神经知觉威胁模型网络。第二种使用一个固定的网络进行LPIPS约束的计算 |
| **摘要** |  |
| 对抗性鲁棒性的一个关键挑战是缺乏对人类感知的精确数学特征，用于定义人眼无法察觉的对抗性攻击。  目前大多数攻击和防御试图通过考虑限制性对抗性威胁模型来避免这个问题，如L2或L∞距离、空间扰动等。（这里的意思是，在此之前，人类考虑所谓的人眼无法察觉的扰动是通过限制噪声的大小来实现的，作者本文要提出一种新的威胁模型NPTM，来描述何为人眼不可察觉）  然而，对任何这些限制性威胁模型都很健壮的模型对其他威胁模型仍然很脆弱，也就是说，它们对不可预见的攻击的泛化性很差。（揭示本文的主要研究方向）。  此外，即使一个模型对几个限制性威胁模型的联合是鲁棒的，它仍然容易受到其他威胁模型中没有包含的难以察觉的对抗性例子的影响。  为了解决这些问题，我们提出了针对所有难以察觉的对抗性实例集合的对抗性训练。  由于这个集合在没有人的情况下难以计算，所以我们使用深度神经网络来近似它。  我们称这种威胁模型为神经感知威胁模型（NPTM），它包括对自然图像的有界神经感知距离的对抗性例子（基于神经网络去近似真实感知距离）。  通过广泛的知觉研究，我们发现神经知觉距离与人类对对抗性例子的可感知性的判断有很好的相关性，验证了我们的威胁模型。  在NPTM条件下，我们开发了新的知觉对抗性攻击和防御。  由于NPTM非常广泛，我们发现针对感知攻击的感知对抗训练（PAT）对许多其他类型的对抗攻击具有鲁棒性。  我们在CIFAR-10和ImageNet-100上测试PAT，对抗五种不同的对抗性攻击： L2、L∞、空间、上着色和JPEG。  我们发现PAT在对抗这五种攻击的联合情况下实现了最先进的鲁棒性——比其他最佳攻击的精度提高了一倍多——而没有对其中任何一种攻击进行训练。也就是说，PAT可以很好地推广到不可预见的扰动类型。这在不能假定特定威胁模型的敏感应用程序中是至关重要的，而据我们所知，PAT是第一个具有这种特性的对抗性训练防御。 |  |