|  |  |
| --- | --- |
| **文章信息** | **方法** |
| **阅读程度**：  精读  **文章标题：**  Toward Efficient Robust Training against Union of ℓp Threat Models  **中文标题：**  对联合ℓp威胁模型进行有效的鲁棒性训练  **发表于：**  NIPS2022  **作者：**  Gaurang Sriramanan  **单位：**  University of Maryland, College Park  *这篇文章的数据部分有点问题，作者在测试l1 l2 l∞范数单独进行攻击时的分类正确率，又测了最坏情况下的分类正确率（三种攻击有一种成功即认为攻击成功），但是后者的数值大于前者，这理论上是不可能的。* | 4.1 首先作者进行了一个简单验证试验：  使用FGSM方法生成的对抗样本记为R-FGSM，C代表使用了线性增长的攻击强度，发现，使用C也会出现灾难性过拟合，只不过来得晚一点，作者**认为FGSM作为单步攻击是不行的，但是C的使用是有效果的**。然后换了个攻击方法，叫核攻击：，通过最大化这个损失函数生成对抗样本，最后一项是对抗样本与干净样本的距离度量，用核范数进行距离度量。发现使用此单步攻击，也会出现灾难性过拟合，但是当引入C后，绿色的线未发生灾难性过拟合。  4.1.2 作者介绍了如何在单步时利用梯度信息，基于1范数寻找对抗样本：  这里是像素点关于损失函数的导数，是像素点的变化量，导数乘以变化量就是损失函数值的变化量（共有d个像素点），累加起来就是总体的变化量。如果不考虑条件(b)的约束，那么只需要将拉满即可，即如果是正的，增大最多为,即如果是负的，减小最多为。但实际上由于(b)的存在，不能直接取最大的变化量，因此将的绝对值从大到小排列，然后从绝对值最大的对应的像素点开始分配扰动（或）当分类的扰动总的1范数已经达到，剩下的像素点不再添加扰动。（说实话我感觉这种扰动分配方法真的不一定是损失函数增大最快的方法）  具体训练时，作者依据预先给定的频率依次使用基于1范数的攻击和基于无穷范数的攻击对一个minibatch生成对抗样本，考虑到更改不同攻击方式后获得的梯度回传会破坏另一种范数的鲁棒性，因此使用动量式更新：。 |
| **摘要及介绍** | **实验** |
| 深度神经网络受到被称为对抗性攻击的精心设计的扰动的严重脆弱性，这导致了各种训练技术的发展，以产生稳健的模型。虽然现有方法的主要焦点是针对解决在单一威胁模型下实现的最坏情况下的性能，但安全关键系统必须同时对多个威胁模型具有鲁棒性。  现有的方法，在这种威胁模型（如ℓ∞、ℓ2、ℓ1）的联合下处理最坏情况的性能，要么利用对抗性训练方法，需要在实践中计算代价昂贵的多步攻击，要么依赖对对单一威胁模型稳健的预训练模型的微调。  在这项工作中，我们表明，通过仔细选择目标函数用于健壮的训练，可以实现类似的，甚至改善最坏情况的性能在一个威胁模型只利用单步攻击在训练期间，从而实现显著减少训练所需的计算资源。  此外，先前的工作表明，对抗ℓ1威胁模型的对抗性训练相对困难，即使是多步反向训练的模型也容易出现梯度掩蔽和灾难性过拟合。  然而，我们提出的方法——当专门应用于ℓ1威胁模型时——使我们能够获得第一个仅通过单步对抗性攻击训练的ℓ1鲁棒模型。最后，为了证明我们的方法的优点，我们利用一组现代攻击评估来更好地估计在考虑的威胁模型联合下最坏情况的性能。 | 实验也很简单，就是基于三个数据集开展的防御效果展示，有一说一，防御效果只是与其他方法持平，但是速度很快。我觉得这工作意义不大。 |