|  |  |
| --- | --- |
| **文章信息** | **方法** |
| **阅读程度**：  方法精读  **文章标题：**  Towards Efficient and Effective Adversarial Training  **中文标题：**  高效且有效的对抗训练  **发表于：**  NIPS2021  **作者：**  Gaurang Sriramanan, Sravanti Addepalli  **单位：**  Video Analytics Lab, Department of Computational and Data Sciences  Indian Institute of Science, Bangalore, India | 首先回顾一下，矩阵的核范数指的是矩阵的奇异值的和，矩阵的F范数指的是矩阵的每个元素值的平方和开根号。  **这是本文生成对抗样本的方法**，最大化第一项自然就是为了生成对抗样本，第二项是对抗样本与干净样本分类结果之间的差值的核范数，f(X)是一个N\*C的矩阵，N是batch\_size，C是分类类别数，作者说使用核范数作为正则化有如下好处：第一最大化范数，自然是在最大化自然样本与对抗样本分类结果之间的差异，这对于对抗样本的生成是一个正面因素，第二，**核范数的因素增大了损失函数面的局部光滑性，有助于避免梯度隐蔽（不懂），**第三作者认为使用核范数生成的对抗样本具有更好的多样性，作者为此还进行了一个对比实验（将核范数变为F范数生成对抗样本，证明核范数生成的对抗样本多样性更好）。  而最训练过程要最小化的损失函数如下，其中第一项的损失函数使用自然样本计算，第二项是最小化对抗样本与干净样本分类结果之间的差值的核范数，作者认为1.核范数是F范数的上界，因此最小化核范数也是最小化分类结果之间的差异 2.核范数可以近似矩阵的秩，而矩阵的每列代表的是不同的类，每行代表的是不同的样本，可以促进一个批次样本之间的相互帮助（不懂，也没说怎么帮助）  一些其他的改进，作者认为单步NuAT还不行，改进为NuAT2，即第一步最大化使用核范数，第二步最大化只使用交叉熵损失函数。  作者还提出了NuAT-H，即通过观测损失值的波动，自行在NuAT与NuAT2之间进行灵活切换。  此外还有 Model Weight Averaging技术 |
| **摘要及介绍** | **实验** |
| 深度神经网络对对抗性攻击的脆弱性激发了人们对提高其鲁棒性的极大兴趣。  然而，目前最先进的对抗性防御涉及到在训练过程中使用10步的对手，这使得它们在计算上无法应用于大规模数据集。  虽然最近的单步防御显示出了有希望的方向，但它们的鲁棒性与多步训练方法不同。  在这项工作中，我们通过在网络预测上引入一种新的核范数正则化器来加强在数据样本附近的函数平滑。  虽然之前的工作独立地考虑每个数据样本，但所提出的正则化器在训练小批中使用对抗样本的联合统计来增强攻击生成和训练期间的优化，在有效的防御中获得最先进的结果。(**核范数的主要目的是在训练的一个小批中使用对抗样本的联合统计来增强攻击效果，而不同于已有的工作仅仅对每个数据样本独立地生成对抗样本**)(**这一点对比损失好像也能做到！**)  我们通过在训练迭代中加入网络权值的指数平均来获得进一步的收益。（其他的小trick）。  最后，我们引入了一种混合训练方法，该方法结合了两步变体的有效性和单步防御的高效性。  与trades和PGD-AT等多步骤防御相比，我们证明了更好的结果，但计算成本显著更低。 |  |