|  |  |
| --- | --- |
| **文章信息** | **方法** |
| **阅读程度**：  方法精读  **文章标题：**  Regularizer to Mitigate Gradient Masking Effect during Single-Step Adversarial Training  **中文标题：**  缓解单步对抗性训练中的梯度掩蔽效应  **发表于：**  CVPR2019  **作者：**  Vivek B S, Arya Baburaj and R. Venkatesh Babu  **单位：**  Video Analytics Lab, Department of Computational and Data Sciences Indian Institute of Science, Bangalore, India  （又是这个实验室，挺niubility的跟GAMA NuAT NCAT这两个阿三是同一个实验室） | 关于梯度隐蔽现象：    梯度隐蔽现象指的是当前对抗训练过程中使用的攻击策略，随着训练的推进难以继续生成有效的对抗样本而导致训练停滞不前的现象。  左侧三图是FGSM对抗训练过程的展示：下左图展示的是在训练到第20轮时，对此时的模型参数使用PGD，FGSM，R-FGSM三种算法生成对抗样本，分别统计三种对抗样本经过网络预测后的平均损失值，发现虽然PGD最高，但是三者总体相差不多，说明此事FGSM算法尚且能够生成有效的对抗样本。右下图展示的是在训练到100轮时的情况，显然此事PGD算法产生的对抗样本的损失值远大于FGSM算法与R-FGSM算法的损失值，此事FGSM及R-FGSM算法均无法产生有效的对抗样本。上图展示的是随着训练过程的推进，R-FGSM与FGSM产生的对抗样本经过网络预测后的logits值的欧式距离，显然在约50轮的时候这个值陡增，此时出现了梯度隐蔽，样本点叠加一个很小的随机噪声产生的新样本，原样本与新样本经过FGSM后生成对抗样本，两个对抗样本经过预测后的logits值差异很大。**通过这个现象我们可以捕捉到梯度隐蔽的出现。**  右侧三图展示的是PGD对抗训练过程：显然PGD对抗训练过程不存在上述问题。  所以作者提出的策略就是引入一个正则化项去限制logits欧式距离陡增这一现象的发生：    作者还强调，仅使用R-FGSM或同时使用R-FGSM与FGSM无法达到类似的效果，无法取得真正的鲁棒性。 |
| **摘要及介绍** | **实验** |
| 神经网络容易受到敌对样本的影响：样本具有难以察觉的噪声，用来操纵网络的预测。为了学习鲁棒模型，引入了一种称为对抗性训练的训练程序。（介绍对抗训练）  在对抗性训练过程中，模型用包含对抗性样本的小批量进行训练。为了扩大对大型数据集和网络的对抗性训练的规模，在训练时使用了快速和简单的方法（如FGSM：快速梯度符号法）。  研究表明，使用单步对抗训练方法（即使用FGSM等非迭代方法生成的对抗样本）训练的模型不是鲁棒的，相反，它们学会了通过掩盖梯度来产生较弱的对手。  在这项工作中，我们提出了一个正则化术语的训练损失，以减轻梯度掩蔽在单步对抗训练的影响。  当FGSM和R-FGSM（在计算FGSM样本之前添加小随机噪声的对手）产生的对抗样本的logits值之间的距离变得很大时，所提出的正则化项导致训练损失增加。  所提出的单步对抗训练比计算成本昂贵的最先进的PGD对抗训练方法更快，而且也取得了相同的效果。 | 确实是简单有效且解释性好的办法 |