|  |  |
| --- | --- |
| **文章信息** | **方法** |
| **阅读程度**：  方法精读  **文章标题：**  Subspace Adversarial Training  **中文标题：**  子空间对抗训练  **发表于：**  CVPR2022  **作者：**  Tao Li  **单位：**  Department of Automation, Shanghai Jiao Tong University | **作者首先探究了灾难性过拟合：**  当灾难性过拟合发生时，训练集对抗样本的正确率上升，验证集对抗样本的正确率降为0，训练集自然样本的正确率出现抖动，作者认为这是因为在灾难性过拟合发生时，网络专注于学习对抗性信息而忽略了自然特征，因此训练集对抗样本正确率上升，自然样本正确率出现抖动。此外作者还观测到，在灾难性过拟合发生时，学习过程中回传给模型的梯度的大小出现剧烈抖动，因此基于这一现象，作者提出子空间对抗训练，通过将梯度信息映射到一个子空间中来防止梯度信息的波动。  （**本文提出的子空间对抗训练方法数学解释性挺好的，后续也可以拿来使用**）  **Phase 1：**获得子空间的一组基，即确定在哪个子空间中进行对抗训练：  首先进行一次预训练，在灾难性过拟合发生之前的训练批次中按照一定的策略采样t组参数值，并且均拉长，形成形状为(N,t)的矩阵，其中N代表参数的个数。  第五行WTW形成一个(t,t)的矩阵，可以求得W矩阵的特征值，只选取最大的d个特征值及其特征向量，这d个特征向量就形成了d维子空间的一组基。（数学上暂时还没搞清楚，但大概是这么回事）。  **Phase 2：**用W1进行初始化，保证参数在步骤一提取到的子空间，对提取到的梯度g，使用PPTg，将梯度信息映射到子空间中（子空间中的两个向量相加自然还在子空间中）。 |
| **摘要及介绍** | **实验** |
| 单步对抗性训练（AT）由于被证明既有效又鲁棒而受到了广泛的关注。然而，存在一个严重的灾难性过拟合问题，即在训练过程中，对投影梯度下降（PGD）攻击的鲁棒精度突然下降到0%。（大背景及存在的问题）  在本文中，我们从一个新的优化角度来解决这个问题，并首先揭示了每个样本的快速增长梯度与过拟合之间的密切联系，这也可以用于理解多步AT中的鲁棒过拟合。（两种过拟合，单步AT的灾难性过拟合与多步AT的灾难性过拟合）  为了控制梯度的增长，我们提出了一种新的AT方法，子空间对抗训练（Sub-AT），它将AT约束在一个仔细提取的子空间中。（本文提出的方法）  它成功地解决了这两种过拟合问题，并显著提高了其鲁棒性。在子空间中，我们还允许具有较大步长、较大半径的单步AT，进一步提高了鲁棒性性能。  因此，我们实现了最先进的单步在性能。**在没有任何正则化项的情况下**，我们的单步AT对CIFAR-10上8/255强半径的PGD- 50攻击可以达到51%以上的鲁棒精度，达到与标准多步PGD-10 AT的竞争性能，具有巨大的计算优势。 | 对比实验基本上是与基础PGD进行的对比 |