|  |  |
| --- | --- |
| **文章信息** | **内容** |
| **阅读程度**：  精读  **文章标题：**  The Enemy of My Enemy is My Friend: Exploring Inverse Adversaries for Improving Adversarial Training  **中文标题：**  敌人的敌人就是我的朋友：探索反向对抗样本来优化对抗训练  **发表于：**  CVPR2023  **作者：**  Junhao Dong  Seyed-Mohsen Moosavi-Dezfooli  Xiaohua Xie\*  中山大学谢晓华教授的团队  第二作者是DeepFool的作者，我比较有印象（国际合作呜呜呜）  **单位：**  中山大学  伦敦帝国学院  **本质上还是对损失函数的优化。**  **（将自然样本与对抗样本的对齐改为自然样本与反向对抗样本的对齐，并基于反向对抗样本会带来额外计算开销这一事实提出了三种在对抗训练中引入反向对抗样本的策略（实例级生成，类级通用，一次性））**  没有想到2023年的今天，还能看到仅通过改进损失函数来优化对抗训练的论文，还是我太菜了呜呜呜。 | 1. Introduction   一个小实验：将样本按照自然样本分类精度进行降序排列，即分类精度高的类是TOP 50分类精度低的类是BOTTOM 50.结果显示对于正常训练来说差别不大，但对于对抗训练而言，Top50%的精度下降速度显然慢于BOTTOM50%，这说明在训练时那些本来就被错误分类的自然样本应当被特殊关照。    三个贡献：**1.** 通过分析将本来就错分的自然样本与其对抗样本的类别概率向量进行对齐的不必要性与有害性（细想一下确实是这样），我们提出了一种新的对抗训练框架基于反向对抗样本，（反向对抗样本是通过梯度下降将自然样本（真实类别为c）推入类别c的高置信度区域），通过对齐对抗样本与反向对抗样本，来引导网络对对抗样本与高置信度区域中的反向对抗样本输出相近的概率分布向量。**2.**基于所提出的反向对抗训练（IAT）范式，我们进一步设计了一种特定于类的通用反向对抗样本生成策略（对一个类生成一个通用反向对抗性噪声，而不是对每一个样本生成一个特定的反向对抗性噪声），以减轻隐藏在决策表面上的类级不平衡。我们还提出了一种一次性的反向对抗样本生成策略，以降低计算成本，而性能损失可以忽略不计（整个训练过程中只生成一次反向对抗样本）。**3.**大量的实验证明了与最先进的方法相比IAT有效性，作者还在使用额外的合成数据（基于DDPM生成）及大模型的条件下与其他方法进行对比。此外，我们有效地实现了在自然精度和对抗性鲁棒性之间更好的权衡。我们的方法还可以与单步对抗性训练方法相结合，作为一个即插即用的组件，以低成本提高鲁棒性。（即对于多步对抗训练而言算是一个新的损失函数），对于单步对抗训练而言是一个即插即用的零件。   1. Background   主要介绍的是trades的正则化方法（对齐自然样本与对抗样本的kl散度）   1. Method   首先是反向对抗样本的生成方式（对每一个样本都生成自己的反向对抗样本）：    对抗训练损失函数：    （贡献1，一种新的对抗训练策略）  首先是反向对抗样本的生成方式（对每一类都生成通用的反向对抗噪声）    动量训练策略：      一次性策略：  在t<T时用f（x）即自然样本获得p(t)，在t=T时进行一次反向对抗样本生成，之后t>T时就是用这一批反向对抗样本（全程只生成了一次反向对抗样本）  （贡献2，通用反向对抗噪声）  总结来说：作者提了三种框架，公式三四是对每个实例生成反向对抗样本，公式67是对每个类生成通用反向对抗样本，这两种反向对抗样本还是实时生成的，最后一种一次性策略只生成一次反向对抗样本，因此相对于普通对抗训练只有极小的开销。  对一个minibatch进行如下操作，首先得到所有样本的对抗样本  之后根据每个样本的真实类别，对其添加对应类别的通用反向对抗性噪声得到反向对抗样本。  同时会计算C（类别数）个损失函数（第8行）用于在此batch遍历结束后更新通用反向对抗性噪声（即C个反向对抗性噪声每个batch更新一次）  最后用公式7作为损失函数优化网络进行训练即可。   1. Experiments   实验一自然是一组在多个数据集上与其他对抗训练策略进行的对比实验：  实验二统计了训练时间，来量化IAT带来的计算量开销  **实验三统计了在更大的噪声下几种方法的鲁棒性：（我猜HAT方法是专注于解决大噪声鲁棒性问题的一种方法）**  **提供了一种说明对抗训练策略优势性的角度**    **实验四统计了在使用额外数据时对抗训练策略的优势性**  **提供了一种说明对抗训练策略优势性的角度**  **（小伙子很鸡贼）**    **实验五统计了在使用此方法对单步对抗训练带来的提升**  **提供了一种说明对抗训练策略优势性的角度**    实验三四五是真的鸡贼，先限定一个范围（大噪声）（添加数据）（单步对抗训练），然后与此范围内的sota进行比较、  之后是一些消融实验 |
| **摘要** |
| 尽管当前的深度学习技术已经产生了相近的性能在各类计算机视觉任务上，然而他们仍然是脆弱的，在面对对抗样本时。（标准开场白）  对抗训练和它的各种变体已经被证实是最有效的防御策略。这些方法通常通过缩小自然样本与对抗样本输出概率分布向量的差异。（已有方法与本方法的主要差别）  然而，这当网络对自然样本本来就分类错误时，这可能对网络训练是有害的。  我们提出了一种新的对抗训练策略来鼓励网络对对抗样本和反向对抗样本产生相似的输出。反向对抗样本的生成与对抗样本相反，通过梯度下降来修改自然样本像素值来鼓励网络对修改后的样本做出正确的分类结果。  在各种数据集与网络结构上的广泛实验证明了我们的训练策略取得了SOTA，在鲁棒性分类精度与自然样本分类精度上。此外，使用通用反向对抗样本，我们进一步以很小的计算量代价提高了单步对抗训练技术。 |