

基于子空间的可学习初始化单步对抗训练方法

胡 军^{1,2}, 占程曦^{1,2}

¹重庆邮电大学计算机科学与技术学院 重庆 中国 400065

²重庆邮电大学网络空间大数据智能计算教育部重点实验室 重庆 中国 400065

摘要 快速对抗训练因其被证明既高效又鲁棒而受到了广泛关注,但是快速对抗训练方法存在严重的灾难性过拟合问题,导致快速对抗训练的模型无法抵御攻击性较强的对抗样本。子空间对抗训练使用网络模型子空间的低维特性解决了快速对抗训练的灾难性过拟合问题。然而,现有子空间对抗训练方法使用基于随机采样的初始化方式,使得样本初始化值并不总是很好地指向样本损失增大的方向,导致方法的鲁棒性不强。为此,本文提出了一种基于子空间的可学习初始化单步对抗训练方法。具体地,该方法在使用动态线性降维法采集目标网络关键参数子空间的基础上,另外训练了一个三层架构的简单卷积网络,利用干净样本在目标网络子空间中的梯度信息指导卷积神经网络更新参数,进而为每个干净样本生成与其自身梯度相关的初始化值,从而有效克服现有方法的不足。实验结果表明,该方法在 CIFAR-10 和 CIFAR-100 数据集上分别针对 PGD-50 和 AA 评估方法实现了 53.39% 和 25.82% 的鲁棒精度,且该方法训练的最优模型与最终模型间仅存在 0.03%~0.4% 的鲁棒精度差异。同时,该方法在不同扰动步长组合、不同采样方法组合以及不同目标网络均表现出对基线方法的鲁棒精度提升。此外,该方法相较于标准多步对抗训练 PGD-10-AT 仅使用 21% 的训练时间,具有显著的计算优势,且相较于其他单步对抗训练也表现出一定的鲁棒性优势。总之,该方法使用较少的时间成本,实现了更好的鲁棒性和稳定性。

关键词 深度神经网络; 对抗样本; 对抗防御; 单步对抗训练; 子空间

中图分类号 TN915.08; TP309.2 DOI 号 10.19363/J.cnki.cn10-1380/tn.2025.09.09

Single-Step Adversarial Training with Learnable Initialization Based on Subspace

HU Jun^{1,2}, ZHAN Chengxi^{1,2}

¹School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

²Key Laboratory of Cyberspace Big Data Intelligent Security, Ministry of Education, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Abstract Fast Adversarial Training (FAT) has garnered considerable interest owing to its demonstrated efficiency and robustness. However, many fast adversarial training methodologies are susceptible to severe catastrophic overfitting, leading to models trained via fast adversarial training being vulnerable to more aggressive adversarial samples. Subspace Adversarial Training (Sub-AT) effectively mitigates the issue of catastrophic overfitting inherent in FAT by leveraging the intrinsic low-dimensional properties of the network model subspace. Nonetheless, prevailing approaches to Sub-AT rely on random sampling-based initialization, which may not consistently align the initialization values with the direction of increasing sample loss, thereby resulting in diminished robustness. To this end, this paper proposes a learnable initializing single-step adversarial training method based on subspace. Specifically, the proposed method, on the foundation of utilizing dynamic linear dimensionality reduction to gather the critical parameter subspace of the target network, additionally trains a simple convolutional network with a three-layer architecture. It employs the gradient information of clean samples within the target network subspace to guide the convolutional neural network in updating parameters, thereby generating an initialization value related to the gradient of each clean sample, effectively overcoming the limitations of existing methods. Experimental results demonstrate that the proposed method achieves robust accuracies of 53.39% and 25.82% for PGD-50 and AA evaluation methods on the CIFAR-10 and CIFAR-100 datasets, respectively, and the difference in robustness accuracy between the optimal model and the final model trained by the proposed method ranges from 0.03% to 0.4%. Moreover, it exhibits robustness improvements over the baseline method across different disturbance step combinations, sampling method combinations, and various target networks. Furthermore, it consumes merely 21% of the training duration in contrast to the conventional multi-step adversarial training PGD-10-AT, thereby yielding a notable computational advantage. Additionally, it demonstrates certain robustness benefits over alternative single-step adversarial training techniques. In summary, the proposed method attains superior robustness and stability while requiring reduced time expenditure.

通讯作者: 胡军, 博士, 教授, Email: hujun@cqupt.edu.cn

本课题得到国家重点实验室稳定支持基金(No. JBS252800290)资助。

收稿日期: 2023-12-18; 修改日期: 2024-03-07; 定稿日期: 2025-08-20

Key words deep neural networks; adversarial examples; adversarial defense; single-step adversarial training; subspace

1 引言

随着计算机计算能力的提升和数据量的爆发式增长, 深度神经网络近年来在计算机视觉、自然语言处理、语音识别和医疗诊断等领域表现出了强大的优势。然而, 深度神经网络在实际应用中可能遭到对抗样本的恶意攻击, 即添加了难以察觉的扰动所生成的对抗样本很容易欺骗深度神经网络^[1], 导致其输出错误的结果。因此, 对抗样本的生成和防御目前已经成为深度学习中的一个热点研究问题。

近年来, 研究人员先后提出了多种攻击方法以生成对抗样本。Szegedy 等人^[1]发现深度神经网络易受到对抗样本的干扰, 并使用 L-BFGS (Limited-memory Broyden Fletcher Goldfarb Shanno) 方法生成对抗样本。Goodfellow 等人^[2]提出了快速梯度符号法 (Fast Gradient Sign Method, FGSM) 来生成对抗样本。Madry 等人^[3]提出了多次迭代计算梯度的投影梯度下降 (Project Gradient Descent, PGD) 生成对抗样本。Croce 等人^[5]提出了两种无参数攻击方法, 即具有对数比差的自动 PGD 攻击 (Adaptive Projected Gradient Descent with Distance-based Label Regularization, APGD-DLR) 和具有交叉熵的自动 PGD 攻击 (Adaptive Projected Gradient Descent with Cross-Entropy, APGD-CE), 前者在进行对抗攻击时采用了一个神经网络模型作为强化学习的策略网络辅助调整生成对抗样本的参数, 而后者使用交叉熵损失函数和投影梯度下降算法生成对抗样本。Hu 等人^[6]提出了一种基于增强空间动量的对抗攻击方法 (Enhanced Spatial Momentum Iterative Fast Gradient Sign Method, ESMI-FGSM), 该方法通过在空间动量的基础上引入邻域采样方法生成对抗样本。

为了提高模型对对抗样本的鲁棒性, 研究人员先后提出了多种方法进行对抗防御^[7]。这些防御方法主要包括数据预处理、对抗样本检测、网络鲁棒性强化三类。其中, 网络鲁棒性强化中的对抗训练 (Adversarial Training, AT)^[3]被认为是最有效的一种防御策略, 其采用模型梯度信息生成对抗扰动, 将生成的对抗扰动添加到干净样本中以生成对抗样本, 将得到的对抗样本和干净样本以一定的比例混合为网络模型的训练数据集, 最后用混合数据集训练目标网络模型得到鲁棒的网络模型。对抗训练可以表述为一个极小极大问题^[3], 其中内部最大化问题聚焦于如何最大化分类损失以生成对抗样本, 外部最

小化问题聚焦于如何最小化对抗样本的损失以获得鲁棒模型。一般来说, 内部最大化问题的求解程度越高, 对抗训练过程中造成的分类损失越大, 经过训练降低对抗样本损失后模型能抵御对抗攻击越强。因此, 内部最大化问题的求解对于对抗训练至关重要^[13]。

对抗训练根据生成对抗样本的步数可以分为两类, 即多步对抗训练和快速对抗训练。一般多步对抗训练如 PGD 对抗训练 PGD-AT^[3]和 TRADES^[14]等能产生较强对抗扰动。但是, 防御者需要对送入分类器的每个小批次训练数据进行数据增强, 在生成对抗样本时需要大量前向和反向传播计算, 导致多步对抗训练计算成本较高。快速对抗训练为了提高训练效率, 通常使用大步长的一步或少步梯度计算的快速梯度符号下降法 (FGSM) 取代小步长多步迭代的投影梯度下降法 (PGD) 生成对抗样本。其中, 单步对抗训练仅仅使用一步对抗攻击生成对抗样本, 能最大限度地降低训练耗时, 提高训练效率。例如, Shafahi 等人^[15]使用具有冗余批次和累积扰动的单步梯度进行对抗训练提出了 Free-AT。Wong 等人^[4]在使用 FGSM 生成对抗样本前为样本加入了均匀分布中随机采样的噪声提出了 Fast-AT (也称 FGSM-RS)。Kim 等人^[16]沿着对抗方向验证内部区间并寻找合适的步长进行对抗训练提出了 Single-AT。Sriramanan 等人^[17]引入一个松弛项来寻找更合适的梯度方向进行对抗样本的生成提出了 GAT。这些单步对抗训练方法已被证明既高效又具有一定的鲁棒性, 但普遍存在灾难性过拟合的问题, 即针对投射梯度下降 (PGD) 攻击的鲁棒性精度在训练过程中突然下降到 0%。即便部分方法克服了灾难性过拟合, 其鲁棒精度也难以令人满意。

为了解决灾难性过拟合问题, Li 等人于 2021 年提出的动态线性降维法 DLDR^[18], 对模型的参数进行筛选, 使模型位于维度很小的子空间中更新参数, 即在子空间中训练时对较少的主要参数进行更新也能达到常规训练中更新全部参数接近的训练效果。同年, Li 等人首次将子空间训练与对抗训练相结合, 以动态线性降维 DLDR 为基础提出了 Sub-AT^[19], 以解决单步对抗训练中的灾难性过拟合问题。该方法在对抗训练模型出现灾难性过拟合前使用 DLDR 算法对模型参数进行采样得到模型子空间, 并在子空间中进行抗训练优化, 通过子空间约束梯度增长速度, 防止灾难性过拟合。但是, 该方法中采用基于随

机性的简单样本初始化, 其初始化值来源于均匀分布, 限制了子空间对抗训练的性能。因此, 本文对 Sub-AT 进行了改进, 使用生成网络生成对抗样本的初始化值以生成更强的对抗样本, 并使用这些对抗样本对目标模型子空间进行对抗训练, 以更新模型子空间网络参数。根据上述思想, 本文提出了一种基于子空间的可学习初始化单步对抗训练方法, 并通过实验对该方法的有效性进行了验证。

2 相关工作

本节介绍对抗训练方法及其相关变体, 其中包括对抗训练、快速对抗训练和子空间对抗训练方法。

2.1 对抗训练

对抗训练^[20](Adversarial Training, AT)的核心思想在于将对抗样本加入模型训练集, 与干净样本一同训练网络模型, 拟合数据的分布, 覆盖深度神经网络模型的分类盲区, 从而提升模型的鲁棒性。

对抗训练可以被表述为一个极小极大问题, 给定一个参数为 w 的目标网络 $f(\cdot, w)$, 一个包括干净样本 x 及其对应标签 y 的数据分布 D , 损失函数 $\mathcal{L}(f(x, w), y)$, 以及威胁约束 Δ , 对抗训练的目标函数可定义如下:

$$\min_w \mathbb{E}_{(x, y) \sim D} [\max_{\delta \in \Delta} \mathcal{L}(f(x + \delta; w), y)] \quad (1)$$

GoodFellow 等人^[2]在提出 FGSM 攻击方法的同时, 提出将对抗样本作为输入数据用以训练分类器, 即 FGSM-AT。FGSM 对抗样本生成公式如下:

$$x_{\text{adv}} = x + \varepsilon \text{sgn}(\nabla_x \mathcal{L}(f(x, y; w), y)) \quad (2)$$

FGSM-AT 使用基于梯度的单步攻击 FGSM 生成对抗样本, 在生成对抗样本时仅需计算一步梯度信息, 因此该方法有着实现简单、训练代价低廉的优点, 缺点是无法为模型提供足够的鲁棒性, 经过 FGSM-AT 训练的模型无法抵挡多步对抗攻击。

相较于一步攻击生成训练数据的 FGSM-AT, PGD 对抗训练以 PGD 攻击的大量计算为代价, 通过多次小步攻击迭代产生对抗样本。PGD 对抗训练通过对对抗样本的多次优化, 使对抗训练损失达到最小, 实现模型的鲁棒。PGD 使用了一步对抗扰动的随机处理, 并使用大量的迭代次数来获得较强的对抗样本。多步对抗训练通常比单步对抗训练提供更好的鲁棒性保证, 但在生成对抗样本迭代计算梯度时, 面临较高的计算和时间成本。

2.2 快速对抗训练

快速对抗训练(Fast Adversarial Training, FAT)使用一步或少步对抗攻击生成对抗样本。相较于多步

对抗训练, 简单而少步的梯度计算能够使快速对抗训练方法节约大量的训练时间, 但对于梯度信息的不充分计算也导致了对抗样本往往不能达到较强的对抗性, 这也直接影响了对抗训练的效果。基于此, 研究人员先后提出了一系列的 FAT 方法以提高对抗样本质量, 从而提高快速对抗训练的性能。例如, Shafahi 等人^[15]提出的 Free-AT 利用单步梯度的冗余批次和累积扰动实现了显著的模型鲁棒性。Wong 等人^[4]提出的 Fast-AT 方法使用 FGSM-AT 与随机初始化相结合来优化对抗样本的生成过程, 其可以取得与基于 PGD 的对抗训练^[3]媲美的性能。具体来说, Fast-AT 为干净样本附加随机初始化, 并对带有随机初始化的干净样本执行 FGSM 攻击生成对抗样本, 其内部最大化问题求解的目标函数如下:

$$\delta^* = \prod_{[-\varepsilon, \varepsilon]^d} [\eta + \alpha \text{sgn}(\nabla_x \mathcal{L}(f(x + \eta; w), y))] \quad (3)$$

其中, δ^* 为对抗扰动, $\eta \in U(-\varepsilon, \varepsilon)$ 为采样自均匀分布的随机噪声, α 为扰动步长。Fast-AT 在生成对抗样本时, 首先对干净样本 x 附加随机采样的初始化值 η 。之后对附加 η 的样本进行一次 FGSM 攻击得到初始扰动。为了防止扰动量过大, 最后对附加 η 的初始扰动执行投影操作, 得到对抗扰动 δ^* 。

Wong 等人^[4]首次提出灾难性过拟合现象, 该现象在快速对抗训练以及多步对抗训练中均有可能发生。例如, 在快速对抗训练中由于 FGSM 大步长的特性, 众多对抗样本都获得了一个较大的扰动量, 部分样本由于超出了最大扰动预算而被投影至最大扰动处, 导致了大部分对抗样本在最大扰动附近聚集, 模型训练时大量优化了这些在最大扰动附近的样本, 从而造成模型分类边界的高度扭曲, 最终导致该模型对训练时使用的攻击方法具有非常高的防御能力, 而在面对其他攻击方法时几乎无法防御^[16]。

Andriushchenko 等人^[21]认为使用随机初始化在一定条件下仍无法抵御多步对抗攻击(如 PGD 等), 导致灾难性过拟合, 因而提出了 FGSM-GA。该方法使用一个正则器来加强对抗样本生成的优化以提高模型的鲁棒性。Kim 等人^[16]表示, Fast-AT 采用具有扰动量最大的对抗样本, 而不是对抗方向的对抗样本, 因此 Fast-AT 仍有可能导致灾难性过拟合。他们因此提出了 Single-AT 来确定适当的步长, 以产生强大的对抗本来提高模型的鲁棒性, 同时一定程度上缓解灾难性过拟合。Sriramanan 等人^[17]提出了 GAT 对抗训练, 通过使用分类损失的松弛项, 提高了生成的对抗样本的质量, 该方法可以防止灾难性过拟合并实现先进的防御性能, 但由于引入了计算

成本昂贵的组件, 其训练效率十分有限。

2.3 子空间对抗训练

子空间对抗训练(Subspace Adversarial Training, Sub-AT)的基本思想是训练少量网络模型的核心参数以达到训练全部参数相近的效果, 从而可以在保证性能的同时节约时间成本^[18]。并且, 将子空间训练用于对抗训练, 还可以将 AT 的梯度下降限制在低维子空间而不是整个参数空间中, 参数空间维度的降低可以隐式地抑制梯度的增长速度, 达到避免灾难性过拟合的目的^[19]。

DLDR 算法^[18]采用模型参数子空间训练的思想, 在目标模型的训练过程中对模型参数进行一定次数采样, 形成离散的参数采样点, 并以此描述参数轨迹特征。该算法通过筛选模型主要参数, 找到近似覆盖参数轨迹的低维子空间, 其子空间采样步骤如下:

- (1) 首先对对抗训练模型检查点 $\{w_1, w_2, \dots, w_t\}$ 进行采样, 将模型参数对齐为参数长度为 N 的向量 w_i ;
- (2) 其次将对齐后的参数矩阵 $[w_1, w_2, \dots, w_t]$ 进行奇异向量分解(SVD), 得到维度为 d 的子空间 $[u_1, u_2, \dots, u_d]$ 的正交基。

一般而言, 深度神经网络的参数数量远大于 d 维子空间的参数数量, 即该算法实现了对目标模型主要参数的提取, 以及对目标网络参数空间的降维。

子空间对抗训练 Sub-AT^[19]为一个即插即用的训练算法, 其可以与多种对抗训练方法结合。本文主要讨论快速对抗训练算法, 因此在后文的理论分析与实验中, 均使用 Sub-AT 与快速对抗训练结合的 Sub-Fast-AT 作为讨论对象和对比方法。Sub-Fast-AT 在生成对抗样本时首先对目标模型进行单步对抗训练, 即仅使用附加随机采样噪声的一步 FGSM 攻击生成对抗样本对目标模型进行对抗训练, 并在该模型发生灾难性过拟合现象前使用 DLDR 算法采样目标模型参数子空间, 以提取网络学习到的有利于鲁棒性和自然精度的特征。然后, 将模型参数倒退到初始化状态, 并通过将梯度投影到子空间中约束对抗训练的优化。Sub-AT 的对抗样本生成函数和 min-max 优化函数如下:

$$\delta = \prod_{[-\epsilon, \epsilon]^d} [\eta + \alpha \text{sgn}(\nabla_x L(f(x + \eta; w_s), y))] \quad (4)$$

$$x_{\text{adv}} = x + \delta \quad (5)$$

$$\min_{w_s} \max_{(x, y) \sim D} \mathbb{E} L(f(x_{\text{adv}}; w_s), y) \quad (6)$$

其中, x 表示干净样本, y 为真实标签, η 为随机采样自均匀分布 $U_{[-\epsilon, \epsilon]}$ 的噪声, x_{adv} 为对抗样本, w_s 为目标网络参数子空间。

3 基于子空间的可学习初始化单步对抗训练方法

本节首先对子空间单步对抗训练中存在的问题进行分析, 然后提出一种基于子空间的可学习初始化单步对抗训练方法。

3.1 子空间单步对抗训练的问题分析

在已有的子空间对抗训练方法中, 生成对抗样本时采取了朴素的随机初始化的方式。如图 1 所示是子空间单步对抗训练 Sub-AT 在训练过程中每个迭代轮次的损失值, 可以发现各个轮次的损失存在较大的波动。在给定分类模型时, 每个样本的优化方向是确定的。然而, 随机初始化将干净样本移动到任意位置, 这可能导致以下两种情况。

- (1) 部分样本向损失减小的方向负优化, 而制作对抗样本的目标是增大样本损失从而提高对抗样本扰动性, 为样本附加减小损失的初始化值不利于提高样本的扰动性;
- (2) 部分样本向损失增大的方向正优化, 但与最佳优化方向存在偏差, 样本的扰动性仅由正交于分类边界的扰动分量提供。

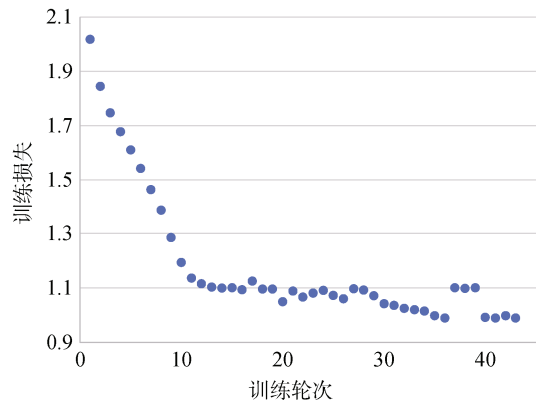


图 1 Sub-Fast-AT 训练中的损失变化

Figure 1 Loss variation in Sub-Fast-AT training

如上所述, 求解 min-max 目标函数的内部最大化问题时, 使用随机初始化方式导致生成质量不稳定的对抗样本, 进而限制了对抗训练的性能。为了进一步提高目标网络模型的鲁棒性, 有必要改进随机初始化带来的不确定影响。

3.2 方法改进

受到初始化值生成网络^[26]的启发, 我们在保留 Sub-AT 子空间采样步骤以及训练目标网络核心参数的同时, 引入了用于生成样本依赖的初始化值的简单架构生成网络, 以稳定初始化方向, 从而增大对抗样本损失, 优化函数的内部最大化求解过程。

根据上述思想, 我们提出了一种基于子空间的可学习初始化单步对抗训练方法(Learnable Subspace Adversarial Training, L-Sub-AT), 该方法包含以下三个训练步骤:

(1) 使用 DLDR 采样法对目标网络进行子空间参数采样。首先, 使用耗时少、计算量少的 FGSM 攻击生成对抗样本, 对目标模型进行对抗训练, 并使用 DLDR 算法在每个训练轮次中均匀地采样两次子空间参数。采样结束后, 将目标模型参数倒退到初始状态;

(2) 使用生成网络协助生成对抗样本。对干净样本 x 进行一步 FGSM 攻击, 获得 x 在目标网络低维子空间 $\{w_s\}$ 中的梯度信息。其次, 干净样本与梯度信息在通道维度拼接后作为生成网络的输入值(以 CIFAR-10 数据集为例, 输入样本的尺寸为 $32 \times 32 \times 6$), 生成网络将输入的图像梯度信息送入卷积层, 并输出依赖于输入样本梯度信息的初始化值 η 。将初始化值 η 与干净样本 x 结合, 得到增广样本 $x' = x + \eta$ 。使用 FGSM 对增广样本 x' 进行单步梯度攻击, 得到 x' 在目标网络低维子空间 $\{w_s\}$ 中的梯度信息。最后, 将梯度信息投影至 $[-\varepsilon, \varepsilon]$ 即可得到对抗扰动 δ , 对抗扰动 δ 与干净样本 x 相加, 即为对抗样本 x_{adv} 。其详细步骤如图 2 所示;

(3) 使用第 2 步中制作的对抗样本 x_{adv} 进行对抗训练, 优化子空间模型参数。

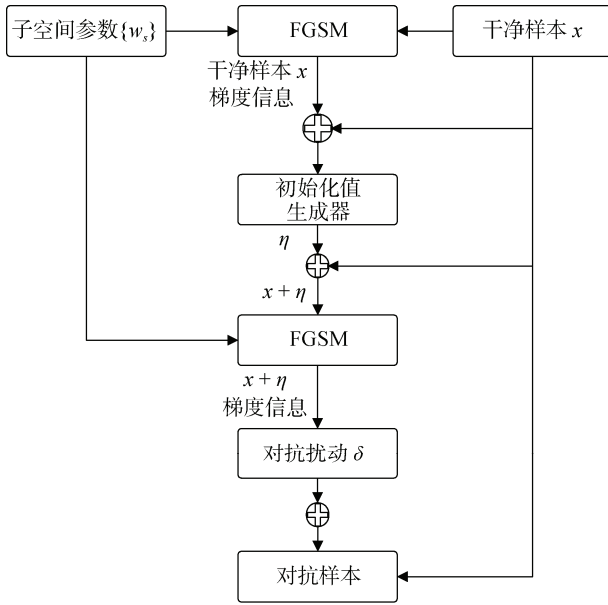


图 2 L-Sub-AT 对抗样本生成流程图

Figure 2 L-Sub-AT adversarial example generation flow

对于目标网络, 我们采用典型的图像分类网络

的架构, 定义为 $y = f(x; w)$, 而对于采样后的目标网络子空间, 定义为 $y = f(x; w_s)$, 其中 x 代表输入图像, y 代表预测标签, w 代表目标网络的参数, w_s 代表目标网络子空间的参数。对于生成网络, 其输入是干净样本图像及其符号梯度。公式如下:

$$s_x = \text{sgn}(\nabla_x L(f(x; w_s), y)) \quad (7)$$

我们使用 FGSM 攻击快速地制作对抗样本, 并对目标网络进行一次对抗训练, 由于在一定的迭代轮次之后目标网络会发生灾难性过拟合, 我们参照 Sub-AT 的做法, 在模型发生灾难性过拟合之前对模型参数进行两次采样, 以确保采样的子空间参数是有效且学习到尽可能多的对抗性的。

初始化值 η_g 用于制作对抗样本, 生成过程可以表示为

$$\eta_g = \varepsilon g(x, s_x; \theta) \quad (8)$$

其中, $g(\cdot; \theta)$ 代表参数为 θ 的生成网络, ε 表示将生成网络输出的值投影到 $[-\varepsilon, \varepsilon]$ 。

我们将干净样本图像与来自目标网络子空间的梯度信息相拼接, 得到生成网络的输入。生成网络产生一个依赖于样本的对抗初始化值。为了尽可能降低额外的训练开销, 生成网络仅包括三层。第一层包括一个卷积层, 有 64 个大小为 $3 \times 3 \times 6$ 的滤波器, 以及一个批量归一化层。第二层为 ResBlock 残差块, 有 64 个大小为 $3 \times 3 \times 64$ 的滤波器。第三层包括一个卷积层, 有 64 个大小为 $3 \times 3 \times 3$ 的滤波器, 以及一个批量归一化层。生成网络采用 ReLU^[23] 作为激活函数。

内部最大化问题近似解的求解如下:

$$S = \text{sgn}(\nabla_x L(f(x_j + \eta_g; w_s), y)) \quad (9)$$

$$\delta_g(\theta) = \prod_{[-\varepsilon, \varepsilon]^d} (\eta_g + \alpha S) \quad (10)$$

其中, η_g 表示每张干净样本的对抗初始化值, 它受干净样本 x 在子空间中的梯度信息 s_x 影响, 能为训练过程中的 FGSM 扰动提供一些信息指导。

根据上述对梯度、初始化值以及扰动的定义, 我们可以通过寻找更好的生成网络参数 θ 来进一步最大化损失, 即 $\max_{\theta} L(f(x + \delta_g(\theta); w_s), y)$, 并通过寻找更好的目标网络子空间参数 w_s 来进一步最小化损失。因此, 联合优化的目标函数定义为

$$\min_{w_s} \max_{\theta} \mathbb{E}_{(x, y) \sim D} L(f(x + \delta_g(\theta); w_s), y) \quad (11)$$

训练过程中, 生成网络与目标网络子空间进行博弈。前者最大化损失以生成更有效的对抗初始化, 以帮助生成更有效的对抗样本。而后者则最小化损

失以更新参数, 使模型获得对对抗样本的鲁棒性。生成网络可以学习干净样本在目标网络子空间中的分布, 从而在目标网络子空间的样本空间中找到对于该样本有效的初始化方向, 并且可以根据目标模型在不同训练阶段的鲁棒性动态地生成对抗初始化, 与均匀分布的随机初始化相比更具有稳定性。

基于子空间的可学习初始化对抗训练 L-Sub-AT 如算法 1 所示。该算法由三阶段组成, 第一阶段(第 1 步~第 3 步)采取一定的训练方法对目标模型进行对抗训练(如 Fast-AT、GAT 等), 采样该模型的训练轨迹, 并通过参数轨迹提取该目标模型的参数子空间。第二阶段(第 7 步~第 12 步)以固定频率训练生成网络, 目标网络每迭代 h 个轮次, 更新一次生成网络参数。第三阶段(第 13 步~第 16 步)在提取的子空间中, 制作对抗样本并执行对目标网络子空间的对抗训练。

算法 1 基于子空间的可学习初始化对抗训练 L-Sub-AT

输入: 子空间维数 d , DLDR 采样次数 t , 学习率 α , 批量大小 n , 训练数据 $\{x_i, y_i\}$, 数据集大小 M , 参数为 w 的目标网络 $f(\cdot, w)$, 参数为 θ 的生成网络 $g(\cdot, \theta)$ 和用于表示生成网络训练间隔的超参数 h , 对抗训练扰动步长 ε , 初始化值步长 $\mu = \varepsilon$ 。

输出: 训练后的目标网络参数 w_k 。

1) 使用 FGSM 或其他对抗攻击方法生成对抗样本训练目标网络模型, 采样参数轨迹:

$$\{w_1^s, w_2^s, \dots, w_t^s\}$$

2) 对模型参数轨迹执行 DLDR 动态线性降维法, 获得目标模型参数子空间:

$$P = [u_1, u_2, \dots, u_d]$$

3) 将模型参数回滚到未经过训练的初始状态:

$$w_0 = w_1^s$$

4) FOR $j=1$ TO M DO

5) 采样小批量数据 $\{(x_i, y_i)\}_{i=1}^n$;

6) $s_{x_i} = \text{sgn}(\nabla_{x_i} L(f(x_i; w_j), y_i))$

7) IF $j \text{ MOD } h = 0$ THEN

8) $\eta_g = \mu g(x_j, s_{x_j}; \theta)$

9) $S_j = \text{sgn}(\nabla_x L(f(x_j + \eta_g; w_j), y_j))$

10) $\delta = \prod_{[-\varepsilon, \varepsilon]^d} (\eta_g + \varepsilon S_j)$

11) $\theta \leftarrow \theta + \nabla_x L(f(x_j + \delta; \theta), y_j)$

12) END IF

13) 生成对抗样本 $\{x_i^{\text{adv}}\}_{i=1}^n$

14) $g_k^{\text{adv}} \leftarrow \frac{1}{n} \sum_{i=1}^n \nabla_w L(f(x_i^{\text{adv}} + \delta, w_k), y)$

15) $w_{k+1} \leftarrow w_k - \alpha P(P^T g_k^{\text{adv}})$

16) END FOR

其中第 12 步生成对抗样本步骤如算法 2 所示。

算法 2 L-Sub-AT 对抗样本生成

输入: 批量大小 n , 训练数据 $\{x_i, y_i\}$, 样本梯度 s_i , 参数为 w 的目标网络 $f(\cdot, w)$, 参数为 θ 的生成网络 $g(\cdot, \theta)$, 对抗训练扰动步长 ε , 初始化值步长 $\mu = \varepsilon$ 。

输出: 对抗样本 x_i^{adv} 。

1) FOR $i=1$ TO n DO

2) $\eta_g = \mu g(x_i, s_i; \theta)$

3) $S_i^* = \text{sgn}(\nabla_x L(f(x_i + \eta_g; w_i), y_i))$

4) $\delta = \prod_{[-\varepsilon, \varepsilon]^d} (\eta_g + \varepsilon S_i^*)$

5) $x_i^{\text{adv}} = x_i + \eta_g + \delta$

6) END FOR

该算法使用两次 FGSM 单步攻击生成对抗样本, 但其生成步骤与两步迭代对抗训练 PGD-2-AT 不同。具体而言, PGD-2-AT 的两步 FGSM 攻击在目标网络的参数下进行, 第一步 FGSM 攻击生成的梯度受到固定步长、预定义投影范围、随机初始化和符号函数的约束, 限制了对抗样本的对抗性。相比之下, 该算法使用生成网络与子空间生成对抗样本的初始化值, 并且没有步长和投影的约束, 子空间为目标网络的低维核心参数, 去除了冗余参数的影响, 得到的梯度信息更准确, 计算的对抗扰动更强, 这一点在第 4 节中进行了实验验证。L-Sub-AT 训练框架如图 3 所示。输入为干净样本, 输出为鲁棒网络模型参数。

4 实验结果分析

本节对基于子空间的可学习初始化对抗训练方法进行了实验分析, 具体包括优化对抗样本的有效性, 不同扰动步长下的稳定性, 以及不同数据集和模型上的泛化性。此外, 我们与基线方法还进行了对抗样本损失的对比。

4.1 实验环境

4.1.1 数据集

实验采用数据集 CIFAR-10 和 CIFAR-100, 并按照 9:1 的比例划分为训练集和验证集。

4.1.2 训练参数与评估方法

对于单步对抗训练防御方法, 我们采用了典型

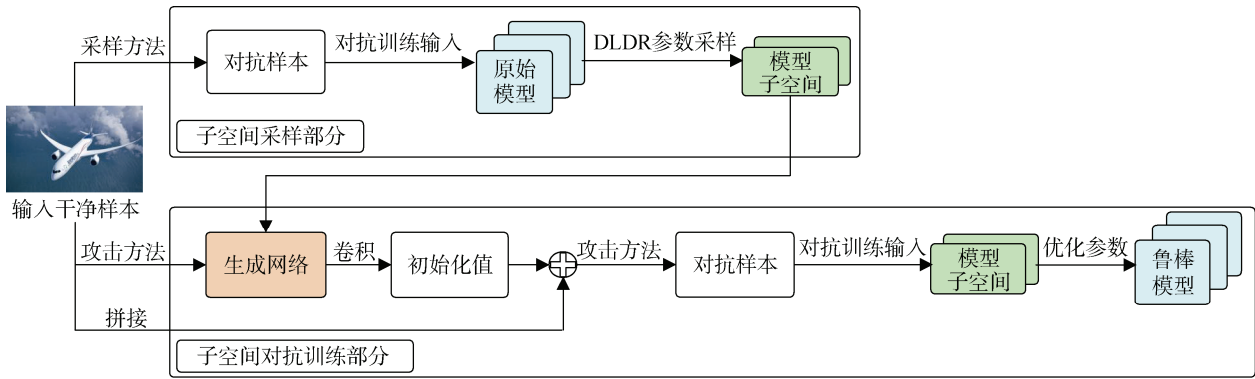


图3 L-Sub-AT 训练框架
Figure 3 L-Sub-AT training framework

的扰动半径为 $\varepsilon = 8/255$ 的 L_∞ 范数, 并使用 Wong 等人^[4]的推荐步长 $\alpha = 1.25\varepsilon$ 。对于多步单抗训练方法, 我们遵循 Rice 等人^[22]的设定, 采用步长 $\alpha = 2/255$ 进行训练。对于评估方法, 我们采用了 PGD-20^[5], PGD-50^[4]和 AutoAttack^[5]三种强力攻击方法。其中, PGD-20 与 PGD-50 分别使用小步长迭代 20 次与 50 次产生足够强力的对抗样本。AutoAttack 是一种组合攻击方法, 其中包含了两种低参数的白盒攻击 APGD-CE 和 APGD-DLR, 以及两种黑盒攻击 FAB 和 Square Attack。其中 Square Attack 为基于随机查询的黑盒攻击方法, 并非基于梯度进行攻击, 因此 AutoAttack 的攻击成功率可以作为验证防御方法是否依赖梯度掩蔽进行防御的衡量指标。

4.1.3 模型训练

在实验中, 我们采用 PreAct-ResNet-18^[27]和 WideResNet-28-10^[24]作为训练模型。在与基线方法进行对比时, 我们分别考虑三种学习率计划: (1)循环学习率, 在前半段 epoch 中将学习率从 0 线性增大到最大学习率, 在后半段 epoch 中将学习率从最大学习率线性下降至 0; (2)分段学习率, 将学习率初始化为 0.1 训练模型, 并在训练 epoch 为 100 和 150 分别衰减为当前学习率的 1/10; (3)恒定学习率, 将学习率置为 1。训练 200 epoch 并进行子空间采样, 以观察采用的方法是否存在灾难性过拟合。在采样的子空间中训练 40 个 epoch, 保证足够的收敛性。采用的批次大小为 128, 使用 SGD 优化器, 动量为 0.9, 权重衰减为 $10e^{-4}$, 生成网络的训练参数 h 为 20, 即目标网络每训练 20 个轮次, 生成网络训练 1 次。另外, 应用了常规的像素填充、随机裁剪和水平翻转作为数据增强手段。

4.1.4 实验结果评估

本文的主要目标是在尽可能减少自然精度损失的情况下提高鲁棒精度, 因此我们在训练期间同时

考虑了鲁棒精度与自然精度两项指标。实验中, 我们随机选取数据集中 10% 的样本作为验证集。评测鲁棒精度时使用评估方法对验证集样本进行对抗攻击生成对抗样本, 进而使用模型对这些对抗样本进行分类, 分类正确率即为鲁棒精度, 其反映了模型在存在异常输入的环境下正确分类的能力。评测自然精度时则不对验证集样本使用对抗攻击, 仅对干净样本进行分类并记录分类正确率, 该指标反映了模型在面对正常输入时的正确分类能力。

此外, 我们还考察了最优模型性能和最终模型性能。具体地, 将最优模型鲁棒精度记录为最优值, 将训练 100 次后的模型鲁棒精度记录为最终值, 并统计最优值与最终值间的差值, 用于衡量模型训练的稳定性。

4.2 对抗训练防御性能对比

我们首先对比了多种快速对抗训练方法, 包括 Single-AT^[16]、Free-AT^[15]、GAT^[17]、FGSM-SDI^[26]、具有随机初始化的 FGSM 对抗训练 Fast-AT^[4]以及使用正则器辅助训练的 GradAlign^[25]。对于上述防御方法, 我们使用分段学习率进行对抗训练。相关研究表明, 精心设计的循环学习率计划有助于克服灾难性过拟合问题, 因此我们对具有随机初始化的单步对抗训练 Fast-AT 和使用昂贵正则器进行单步对抗训练的 GradAlign 两种防御方法设立了循环学习率对照组。评估模型鲁棒性采用 PGD-50^[15]攻击, 其训练轮次设置为 100 次。评估指标为 PGD-50 攻击下, 鲁棒精度和自然精度的最优值、最终值以及它们的差值。不同对抗训练方法在 CIFAR-10 和 CIFAR-100 数据集上的结果如表 1 和表 2 所示。

可以发现, 使用分段学习率的 Fast-AT 和 GradAlign 发生了灾难性过拟合现象。使用循环学习率的 GradAlign 和使用分段学习率的 Single-AT、GAT 在最优值和最终值之间存在较大的差值, 因此有可

表 1 CIFAR-10 数据集单步 AT 方法防御效果比较
Table 1 Comparison of defense effects of single-step AT methods on CIFAR-10

训练方法	鲁棒精度			自然精度		
	最优值	最终值	差值	最优值	最终值	差值
Fast-AT (循环)	45.94	45.79	0.15	82.46	83.66	1.20
GradAlign (循环)	48.10	46.93	1.17	80.45	81.16	0.71
Fast-AT (分段)	39.95	0.00	3.95	79.31	89.95	10.64
GradAlign (分段)	42.19	0.02	42.17	70.98	88.90	17.92
Single-AT	35.52	32.57	2.95	83.58	86.93	3.35
Free-AT	47.40	47.12	0.28	79.59	80.20	0.61
GAT	50.13	41.46	8.67	82.42	84.45	2.03
FGSM-SDI	48.06	47.92	0.14	81.29	81.30	0.01
Fast-Sub-AT	47.92	47.58	0.34	82.09	82.47	0.38
Ours	51.25	51.20	0.05	81.51	81.64	0.13

表 2 CIFAR-100 数据集单步 AT 方法防御效果比较
Table 2 Comparison of defense effects of single-step AT methods on CIFAR-100

训练方法	鲁棒精度			自然精度		
	最优值	最终值	差值	最优值	最终值	差值
Fast-AT (循环)	16.54	0.00	16.54	34.45	47.31	12.86
GradAlign (循环)	24.22	24.07	0.15	50.72	51.84	1.12
Fast-AT (分段)	17.84	0.00	17.84	41.74	46.83	5.09
GradAlign (分段)	23.80	15.74	8.06	49.37	54.02	4.65
Single-AT	16.16	0.95	15.21	56.02	59.31	3.29
Free-AT	23.60	23.05	0.55	51.08	51.81	0.73
GAT	23.12	20.39	2.73	58.38	57.18	1.2
FGSM-SDI	21.91	0.00	21.91	43.11	39.29	3.82
Fast-Sub-AT	24.79	24.29	0.5	56.38	56.52	0.14
Ours	25.82	25.42	0.4	56.29	56.40	0.17

最能在后续的训练中发生训练效果不稳定的现象。我们的方法在 CIFAR-10 和 CIFAR100 数据集上都取得了最好的鲁棒性,并在最优值与最终值之间实现了较小的差值,这表明根据我们的方法可以训练出具有更好的鲁棒性和稳定性的模型。

4.3 不同扰动步长防御性能对比

我们进一步使用目标网络模型 ResNet-18 以及数据集 CIFAR-10, 更换不同扰动步长以模拟不同攻击强度需求下的场景,使用 PGD-50 和 AutoAttack (AA)作为评估手段,观察鲁棒精度以及最优值与最终值的差值。一般地,使用越大的扰动步长将导致对抗样本获得更多的扰动量,网络模型也将因此获得

更强的防御能力。然而,对抗样本的研究意义在于如何在人眼正常识别的条件下误导深度神经网络错误分类。Goodfellow 等人^[2]发现使用低于 8/255 作为对抗扰动步长时,对抗样本对深度神经网络模型的误导能力十分有限,在使用超过 16/255 的对抗扰动步长时,图像上的改动将变得十分明显,此时虽然可以误导深度神经网络错误分类,但也会因明显的改动被人眼轻易识别。因此我们遵循 FGSM^[2]中的扰动步长推荐值 $\epsilon=8/255$,并逐渐增大到 $1.5\epsilon=12/255$ 和 $2\epsilon=16/255$ 。评估指标为 PGD-50 攻击下,鲁棒精度和自然精度的最优值、最终值以及它们的差值。实验结果如表 3 所示,我们的方法在评估下的鲁棒性能相比 Fast-AT 以及基线方法均有所提升,甚至在设置大扰动步长时超越了 PGD-10 多步对抗训练的性能。并且,我们的方法将鲁棒性最优值与最终值间的差值控制在了 0.3%以内,说明我们的方法为子空间快速对抗训练提供了一定的稳定作用。评估指标为 AutoAttack 攻击下鲁棒精度和自然精度的最优值、最终值以及它们的差值。如表 4 所示,我们的方法在最佳模型检查点处与 PGD-10-AT 的鲁棒精度仍有差距,但在最终模型检查点处,PGD-10-AT 的鲁棒精度与

表 3 不同扰动步长下对 PGD-50 的防御效果比较
Table 3 Comparison of defense effects against PGD-50 with different perturbation steps

训练方法 (步长)	子空间 采样步长	最优值	最终值	差值
Fast-AT	/	39.95	0.00	39.95
Sub-Fast-AT (8/255)	8/255	48.32	47.58	0.74
Ours(8/255)	8/255	51.25	51.20	0.05
Ours(12/255)	8/255	51.86	51.76	0.1
Ours(16/255)	8/255	53.03	52.74	0.29
Ours(16/255)	12/255	53.39	53.30	0.09
PGD-10-AT	/	50.79	42.51	8.28

表 4 不同扰动步长下对 AA 的防御效果比较
Table 4 Comparison of defense effects against AA with different perturbation steps

训练方法 (步长)	子空间 采样步长	最优值	最终值	差值
Fast-AT	/	37.55	0.00	37.55
Sub-Fast-AT (8/255)	8/255	44.04	43.99	0.05
Ours(8/255)	8/255	45.81	45.76	0.05
Ours(12/255)	8/255	45.89	45.86	0.03
Ours(16/255)	8/255	46.11	46.00	0.11
Ours(16/255)	12/255	46.13	45.94	0.19
PGD-10-AT	/	47.29	41.08	6.21

优值相比下降了约 6.21%，这说明 PGD-10-AT 在训练后期会出现训练不稳定的现象，引起训练后期模型性能的下降。相比之下我们的方法在训练中的最优值和最终值的差值不超过 0.2%，这说明我们的方法具有较好的训练稳定性。

我们将本文提出的方法采样子空间，并使用 FGSM 生成对抗样本进行对抗训练，以及 PGD-50 和 AutoAttack 作为评估方法。两种方法在 4 组子空间采样扰动步长 ε_1 和对抗训练扰动步长 ε_2 参数设置下的结果如图 4 和图 5 所示，我们的方法在不同扰动步长下相较基线方法均有所提升，说明了在更大扰动步长下，我们的方法依然能求得内部最大化问题的更优解。

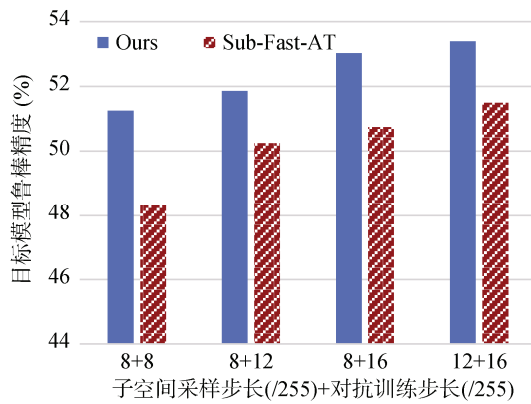


图 4 CIFAR-10 上与基线方法在不同步长下的鲁棒精度对比(PGD-50)

Figure 4 Robust accuracy comparison with baseline method at different step sizes on CIFAR-10 (PGD-50)

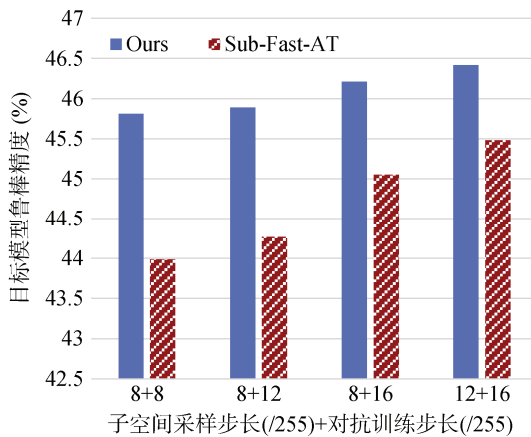


图 5 CIFAR-10 上与基线方法在不同步长下的鲁棒精度对比(AA)

Figure 5 Robust accuracy comparison with baseline method at different step sizes on CIFAR-10 (AA)

4.4 不同子空间采样策略效果对比

进一步，我们更换子空间采样策略，在 CIFAR-10

上与基线方法进行了实验对比。具体地，我们采用 GAT 攻击方法生成对抗样本，用于执行采样阶段的模型训练，并从这些模型的训练轨迹中进行 DLDR 动态线性降维法采样子空间，采用扰动半径为 $8/255$ 的 L_∞ 扰动。为了观察不同的采样方法与对抗训练方法组合对子空间单步对抗训练的影响，我们设立了仅使用 GAT 进行单步对抗训练作为空白对照组，子空间采样阶段使用 GAT 同时对抗训练阶段使用 FGSM 组，子空间采样阶段和对抗训练阶段均使用 GAT 组。评估指标为 PGD-50 攻击下，鲁棒精度和自然精度的最优值、最终值以及它们的差值。对比结果如表 5 所示，我们的方法在最优模型检查点以及最终模型检查点均实现了对基线方法的鲁棒性提升。值得注意的是，自然精度和鲁棒精度之间存在权衡，更好的鲁棒精度总是伴随着较低的自然精度，虽然我们的方法略微降低了自然精度，但在可接受的范围之内，且我们的方法更多地提升了鲁棒精度。

表 5 GAT 子空间采样策略下的鲁棒性对比
Table 5 Robustness comparison with GAT subspace sampling strategy

训练方法	鲁棒精度			自然精度		
	最优值	最终值	差值	最优值	最终值	差值
GAT	50.13	41.46	8.67	82.42	84.45	2.03
Sub-GAT-AT	51.15	50.80	0.35	81.61	81.76	0.15
Ours(GAT)	53.80	53.59	0.21	81.23	81.44	0.21
Sub-Fast-AT	49.96	49.21	0.75	82.01	82.03	0.02
Ours(FGSM)	52.33	51.86	0.47	81.53	81.57	0.04

4.5 不同目标网络防御效果对比

最后，我们将目标模型更换为 WideResNet-28-10 架构，在 CIFAR-10 上与基线方法进行了对比实验。其中，采用扰动半径为 $8/255$ 的 L_∞ 扰动。评估指标为 PGD-20 攻击下，鲁棒精度和自然精度的最优值、最终值以及它们的差值。

如表 6 所示，我们的方法在 CIFAR-10 数据集上相较于基线方法最优鲁棒精度提升了 2.17%，最终鲁棒精度提升了 1.92%，同时最优自然精度仅下降了 0.63%，最终自然精度仅下降了 0.7%。但是我们的方法在鲁棒精度上出现了更高的鲁棒精度差值。针对如上问题，我们设立了额外的对照组“Ours*”，并将初始化值步长设置为 $\mu = 0.75\varepsilon$ 进行实验。由实验结果可知，我们降低了初始化值的幅度，因此损失了一定的鲁棒精度，同时也得到了最低的鲁棒精度差值，这意味着降低总扰动量有利于我们的方法稳定训练。综合以上，我们的 L-Sub-AT 方法可以应用于

其他网络模型并提升其鲁棒性。

表 6 WideResNet-28-10 目标网络模型下鲁棒性对比
Table 6 Robustness Comparison on WideRes-Net-28-10 Network

训练方法	鲁棒精度			自然精度		
	最优值	最终值	差值	最优值	最终值	差值
Fast-AT	43.66	0.00	43.66	84.60	89.74	5.14
Sub-Fast-AT	48.24	47.52	0.72	86.86	86.96	0.10
Ours	50.41	49.44	0.97	86.23	86.26	0.03
Ours*	49.84	49.30	0.54	86.58	86.44	0.14

我们对比了各种方法在 CIFAR-10 数据集上的鲁棒精度以及训练时间以衡量每个方法的训练效果及成本, 如图 6 所示。我们的方法在相同的实验环境下, 对比多步对抗训练 PGD-10-AT 优势明显, 并可以在更短的时间内使用更少的计算资源, 同时获得更好的防御性能。尽管我们的方法在训练时间上略高于 FGSM-SDI 和 Sub-AT, 但对比这些方法具有明显的鲁棒性能优势。总之, 我们的方法通过牺牲少部分训练时间获得了更好的鲁棒性, 实现了模型鲁棒性和训练时间成本的平衡。

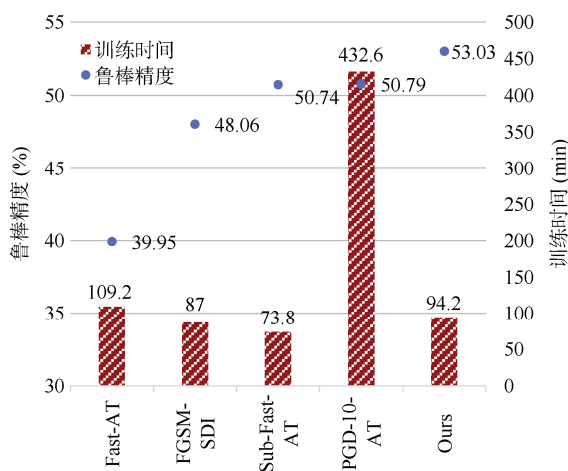


图 6 CIFAR-10 下模型鲁棒精度及训练时间对比
Figure 6 Comparison of robust accuracy and training time on CIFAR-10

4.6 不同子空间初始化方法的训练损失对比

我们针对 2.3 节提出的子空间对抗训练现存的问题进行了一组对比实验, 我们将提出的方法在训练 0~40 个轮次的训练损失与 Sub-Fast-AT 进行了对比。如图 7 所示, 我们的方法生成的对抗样本在训练后期能达到更高的损失值, 更高的损失值意味着引入生成网络后, 制作的样本依赖的初始化值增强了对抗样本的对抗性, 即我们的方法能够更好地求解

对抗训练目标函数中的内部最大化问题。

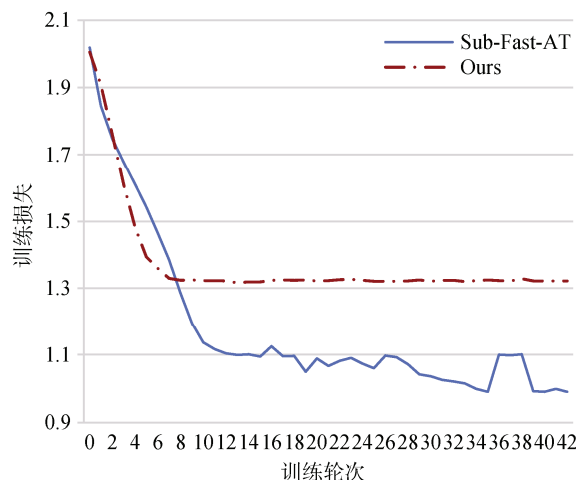


图 7 与基线方法的训练损失对比
Figure 7 Comparison of training loss with baseline method

此外, 我们的方法收敛速度更快, 并且生成的对抗样本在训练后期更加稳定, 损失波动较小, 有助于最终目标模型获得稳定的鲁棒性。究其原因, 主要是生成网络通过干净样本在目标模型中的梯度信息生成初始化值, 样本依赖的初始化值可以在一定程度上帮助稳定样本初始化值的方向, 减少基于随机性的初始化带来的不稳定影响。

5 结论

在本文中, 我们提出了一种基于子空间的可学习初始化单步对抗训练方法。具体来说, 我们在使用 DLDR 动态线性降维法采样子空间(即提取目标模型关键参数用于后续的参数更新)后, 采用生成网络以干净样本及其来自子空间的梯度信息为条件生成更有效的初始化。在训练阶段, 生成网络和目标网络子空间交替优化并进行博弈, 前者学习产生一个动态的依赖样本的初始化, 以根据当前目标网络生成更强的对抗样本。后者采用生成的对抗样本进行训练以更新目标模型关键参数, 以降低对抗样本损失。与基线方法 Sub-AT 相比, 我们提出的方法专注于优化对抗样本强度, 进而优化子空间对抗训练的效果。得益于依赖样本的初始化值, 我们的方法可以在多种场景下实现最优模型与最终模型之间较小的性能差值, 即在一定程度上稳定对抗训练效果。与近年来的单步对抗训练方法以及 PGD-10 多步对抗训练相比, 我们的方法更具有性能和成本优势。同时, 由于子空间的低维特性, 我们的方法允许在使用更大的扰动步长进行对抗训练而不发生过拟合。此外, 我们的方

法在不同数据集、不同的参数采样策略、不同目标模型网络的场景下的性能提升也具有一定的泛化性优势。

参考文献

- [1] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[C]. *International Conference on Learning Representations*, 2014: 1-10.
- [2] Goodfellow I, Shlens J, Szegedy C, et al. Explaining and harnessing adversarial examples[C]. *International Conference on Learning Representations*, 2015: 1-11.
- [3] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[C]. *International Conference on Learning Representations*, 2021:1-28.
- [4] Wong E, Rice L, And Kolter J Z. Fast is better than free: Revisiting adversarial training[C]. *International Conference on Learning Representations*, 2020: 1-17.
- [5] Croce F, Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks[C]. *The 37th International Conference on Machine Learning*, 2020: 2206-2216.
- [6] Hu, J, Wei, G H, Xia, S Y, et al. Adversarial attack method based on enhanced spatial momentum[J]. *International Journal of Machine Learning and Cybernetics*, 2025, 16: 789-802.
- [7] Guo M H, Yang Y Z, Xu R, et al. When NAS meets robustness: In search of robust architectures against adversarial attacks[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 628-637.
- [8] Zhang H, Wang J. Defense against adversarial attacks using feature scattering-based adversarial training[C]. *The 33rd Conference on Neural Information Processing Systems*, 2019: 1-5.
- [9] Pang T, Du C, Zhu J. Robust deep learning via reverse cross-entropy training and thresholding Test[C]. *International Conference on Learning Representations*, 2017: 1-20.
- [10] Metzen J H, Genewein T, Fischer V, et al. On detecting adversarial perturbations[C]. *International Conference on Learning Representations*, 2017: 1-12.
- [11] Carlini N, Wagner D. Towards evaluating the robustness of neural networks[C]. *2017 IEEE Symposium on Security and Privacy*, 2017: 39-57.
- [12] Shaham U, Yamada Y, Negahban S. Understanding adversarial training: Increasing local stability of supervised models through robust optimization[J]. *Neurocomputing*, 2018, 307: 195-204.
- [13] Wang Y, Ma X J, Bailey J, et al. On the convergence and robustness of adversarial training[C]. *The 36th International Conference on Machine Learning*, 2019: 6586-6595.
- [14] Zhang H Y, Yu Y D, Jiao J T, et al. Theoretically principled trade-off between robustness and accuracy[C]. *The 36th International Conference on Machine Learning*, 2019: 7472-7482.
- [15] Shafahi A, Najibi M, Ghiasi M A, et al. Adversarial training for free![C]. *The 33rd Conference on Neural Information Processing Systems*, 2019: 1-16.
- [16] Kim H, Lee W, Lee J. Understanding catastrophic overfitting in single-step adversarial training[C]. *The AAAI Conference on Artificial Intelligence*, 2021, 35(9): 8119-8127.
- [17] Sriramanan G, Addepalli S, Baburaj A. Guided adversarial attack for evaluating and enhancing adversarial defenses[C]. *The 34th Conference on Neural Information Processing Systems*, 2020, 33: 20297-20308.
- [18] Li T, Tan L, Huang Z H, et al. Low dimensional trajectory hypothesis is true: DNNs can be trained in tiny subspaces[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(3): 3411-3420.
- [19] Li T, Wu Y W, Chen S Z, et al. Subspace adversarial training[C]. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 13409-13418.
- [20] Matyasko A. Towards deep neural networks robust to adversarial examples[D]. Nanyang Technological University, 2020.
- [21] Ibrahim H, Pik Kong N S, Ng T F. Simple adaptive median filter for the removal of impulse noise from highly corrupted images[J]. *IEEE Transactions on Consumer Electronics*, 2008, 54(4): 1920-1927.
- [22] Rice L, Wong E, Kolter J Z. Overfitting in adversarially robust deep learning[C]. *The 37th International Conference on Machine Learning*, 2020: 8093-8104.
- [23] Agarap A F. Deep learning using rectified linear units (ReLU)[C]. *International Conference on Learning Representations*, 2018: 1-18.
- [24] Zagoruyko S, Komodakis N. Wide residual networks[C]. *The British Machine Vision Conference*, 2016: 1-12.
- [25] Andriushchenko M, Flammarion N. Understanding and improving fast adversarial training[J]. *Advances in Neural Information Processing Systems*, 2020, 33: 16048-16059.
- [26] Jia X J, Zhang Y, Wu B Y, et al. Boosting fast adversarial training with learnable adversarial initialization[J]. *IEEE Transactions on Image Processing*, 2022, 31: 4417-4430.
- [27] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. *IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770-778.



胡军 于 2010 年在西安电子科技大学大学模式识别与智能系统专业获得博士学位。现任重庆邮电大学计算机科学与技术学院教授。中国计算机学会杰出会员。研究领域为多粒度认知计算、人工智能安全和图分析与挖掘等。Email: hujun@cqupt.edu.cn



占程曦 于 2017 年在江西财经大学软件工程专业获得学士学位。现在重庆邮电大学计算机技术专业攻读硕士学位。研究领域为人工智能安全。Email: s210231243@stu.cqupt.edu.cn