中国科学:信息科学 在审文章 onlinelatex

SCIENTIA SINICA Informationis

论文

类增量学习中基于语义强化的可持续对抗攻击

刘心晶[1], 王珠珠[2*], 孟宪佳[2*], 刘泰峰[1], 马卓然[1], 杨易龙[3], 马卓[1]

1. 西安电子科技大学网络与信息安全学院, 西安 710126
2. 西北大学信息科学与技术学院, 西安 710127
3. 西安电子科技大学人工智能学院, 西安 710126

* 通信作者. E-mail: z.z.wang@foxmail.com, xianjiam@nwu.edu.cn

国家重点研发计划 (批准号:2023YFB2904000)、国家自然科学基金 (批准号:62573346、62302361、U23A20307、U23A20306、62501444、U25A20429)、中国博士后科学基金 (批准号:GZB20250406)、陕西省创新能力支撑计划-科技创新团队项目 (批准号:2025RS-CXTD-010)、中央高校基本科研业务费专项资金 (批准号:QTZX26096)、“111” 计划 (B16037) 的资助项目

摘要 目前, 许多广泛应用的神经网络模型并非基于静态数据训练, 而是需要适应不断变化的新增数据. 类增量学习 (CIL) 应运而生, 通过逐步更新模型以应对新旧数据的分布差异, 防止模型发生灾难性遗忘. 然而, 随着新类别的加入, 模型的知识域发生剧烈变化, 导致对抗扰动随着增量学习的进行而失效. 为解决类增量学习场景下对抗样本易失效的问题, 本文提出了一种基于语义强化的有目标对抗样本生成方法. 核心思想是以 “强化对抗扰动的目标类语义” 为优化目标, 分三步提升对抗样本在类增量学习过程中的鲁棒性:(1) 候选样本生成与语义筛选: 利用扩散模型生成目标类样本, 并结合初始模型的预测熵对样本进行筛选, 从中选取具有强目标类语义的样本; (2) 浅层语义特征提取: 基于模型浅层特征在类增量学习过程中的稳定性, 提取目标类样本的浅层语义信息. (3) 基于语义约束的对抗样本生成: 设计一组语义约束函数, 使对抗样本在特征空间中同时 “靠近目标类、远离非目标类”, 从而生成鲁棒的对抗样本. 本文综合实验表明, 在类别规模提升至 10 倍时, 本文方法的平均攻击成功率相比基准方法提升了 39.38%.

关键词 对抗攻击, 增量学习, 有目标对抗, 鲁棒性, 语义强化

1 引言

1.1 研究背景

在实际应用中, 许多广泛部署的神经网络模型并非基于静态数据训练, 而是需要持续应对随着时间推移不断增加的新数据. 为使模型能够适应这种动态变化, 类增量学习 (Class Incremental Learning, CIL) 应运而生, 并被广泛应用. CIL 的核心思想在于使模型能够逐步在新数据上进行更新, 这些新数据的分布和标签通常与旧数据存在差异, 从而重点解决灾难性遗忘问题, 确保模型在面对新旧数据时保持良好的性能 [1,2]. 然而, 在类增量学习的过程中, 随着新类别的加入, 模型的决策边界发生剧烈的变化. 这种决策边界的漂移对对抗攻击产生了显著影响 [3,4], 尤其是有目标对抗攻击. 在这类攻击中, 攻击者通过精心设计的扰动来欺骗模型, 使其产生特

引用格式: 刘心晶, 王珠珠, 孟宪佳, 等. 类增量学习中基于语义强化的可持续对抗攻击. 中国科学: 信息科学, 在审文章
Xinjing Liu, Zhuzhu Wang, Xianjia Meng, et al. Sustainable Adversarial Attacks Based on Semantic Enhancement in Class-Incremental Learning. Sci Sin Inform, for review

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0474

定的错误预测. 然而, 在 CIL 过程中, 随着新类别的不断加入, 模型的知识域会发生显著漂移 [2,5~7]. 这种知识域的漂移对对抗攻击产生了重要影响, 尤其是有目标的对抗攻击. 在此类攻击中, 攻击者通过精心设计的扰动诱导模型产生特定的错误预测 [3,4,8,9]. 然而, 在类增量学习场景下, 随着新类别样本的持续加入, 模型的知识域不断变化, 导致先前有效的对抗扰动逐渐失效, 从而显著降低对抗攻击的效果, 表现出明显的不可持续性. 因此, 研究有目标对抗样本在类增量学习过程中的脆弱性, 不仅是提升对抗攻击鲁棒性的一项重要挑战, 也为优化对抗防御策略提供了新的研究方向.

1.2 相关工作

现有的有目标对抗攻击研究提出了多种方法, 这些方法通常可分为两大类, 基于梯度的方法和基于生成模型的方法, 以下将对这两种主要的对抗攻击方法进行详细阐述.

1.2.1 基于梯度的方法

基于梯度的对抗攻击方法通常需要根据模型输出的 logits 来计算损失函数, 并迭代更新对抗样本. 为了使对抗样本对模型的微调等变化具有鲁棒性, 此类方法通常在一个替代模型上计算梯度, 以增强对抗的可迁移性. Goodfellow 等人 [4] 提出的快速梯度符号法 (FGSM, Fast Gradient Sign Method) 在输入梯度方向上对图像施加单步扰动. 而 Dong 等人 [3] 提出动量迭代的 FGSM 方法 (MIFGSM, Momentum Iterative FGSM), 通过引入动量项来改进 FGSM, 使得扰动在每次迭代中积累历史梯度信息, 从而增强对抗攻击的有效性和稳定性. Byun 等人 [10] 所提出的 CFM 方法在两种对立的竞争噪声下生成对抗扰动, 在该方法中, 结合不同噪声源的扰动, 使得生成的对抗样本在多个模型间具有更强的迁移能力, 从而提高攻击的迁移性. Ge 等人 [11] 所提出的 CleanSheet 方法通过利用从目标模型的干净训练数据中提取的鲁棒特征来诱导误分类, 从而构建了一个通用的对抗扰动. Xu 等人 [12] 所提出的 UnivIntruder 方法通过 CLIP [13] 构建对齐的替代模型来模拟目标模型的行为, 从而产生针对性与通用性具备的对抗扰动.

1.2.2 基于生成模型的方法

基于生成模型的攻击方法通过训练生成模型来创造对抗样本, 具体地, 这些方法训练一个生成器模型, 学习如何在特定条件下生成有目标对抗样本. 通常而言, 基于生成模型的对抗噪声具有更强的语义特性, 因此在不同的模型间具备更强的可迁移性. Fang 等人 [14] 采用了一个单类条件生成器, 并通过 CLIP [13] 的文本嵌入进行引导, 从而实现高效的多目标对抗攻击, 其迁移性显著高于仅使用标签条件的方法. Wang 等人 [15] 所提的 TTAA 方法引入了双重判别器, 从 logit 空间和特征空间中捕获信息, 生成具有较强迁移性的对抗样本. Sun 等人 [16] 提出的 GAKer 方法在生成对抗样本时, 通过注入潜在感染的表示到输入的中间特征中, 使得对抗样本在特征空间中更接近目标类别. Guo 等人 [17] 的 AdvDiffVLM 方法通过修改扩散模型反向生成过程中的得分, 生成包含自然对抗目标语义的对抗样本. 而 Li 等人 [18] 提出了一种即插即用模块 AIM, 利用目标对象中的语义信息来提升对抗样本的迁移性.

1.3 动机与贡献

本文的主要贡献总结如下.

- 针对 CIL 场景中对抗样本易失效的问题, 本文提出了一种基于语义强化的有目标对抗样本生成方法. 该方法首先筛选具有强语义的目标类样本, 再执行目标类语义嵌入, 以提升对抗样本在 CIL 过程应对知识域漂移时的鲁棒性.

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0474

• 提出基于预测熵的样本生成—筛选算法以增强目标类语义稳定性: 先利用扩散模型生成目标类样本, 再据模型预测熵筛选语义更强的样本, 并结合数据增强以抑制, 语义波动、提升泛化能力, 同时, 利用 CIL 过程中更为稳定的浅层特征, 设计目标类语义提取与嵌入机制: 将目标类浅层表征注入对抗样本生成模型, 并引入语义约束, 使生成样本在表征空间中贴近目标类、远离非目标类, 从而确保语义稳定性与鲁棒性.

• 通过综合实验, 验证所提方法在 CIL 中的有效性与鲁棒性. 通过在 CIFAR-100 和 ImageNet-100 数据集上的综合实验验证了方法在 CIL 中的鲁棒性, 在类别规模提升至原来的 9 倍的条件下, 所提方法显著提高对抗样本的持续性, 攻击成功率提升了 39.38%.

本文的其余部分组织如下. 第 2 章对增量学习中的有目标对抗攻击问题进行描述, 第 3 章介绍了提出基于预测熵的样本筛选与增强算法和基于模型浅层特征提取与嵌入的对抗样本生成算法. 第 4 章对当前主流 CIL 设置开展对抗攻击实验, 验证所提方法的有效性和鲁棒性, 并探讨了防御方法. 最后, 第 5 章对本文的工作进行了总结和讨论.

2 问题描述

本文旨在设计一种对类增量学习鲁棒的对抗样本生成方法. 接下来详细介绍增量学习场景, 并基于此场景定义攻击者的能力与目标.

2.1 场景描述

给定一个初始目标模型 $f_1(x)$, 该模型是在一组初始数据集 $D_1 = (X_1, Y_1)$ 上进行训练获得的. 增量学习的全部数据集为 $D = D_1, D_2, ..., D_m$, 其中 m 代表增量学习次数, 其中每次增量学习的子数据集为 $D_i = (X_i, Y_i)$. 每次在进行增量学习时, 按如下目标更新模型 $f_{i-1}(x)$ 为 $f_i(x)$:

$$\begin{aligned} &\arg\min_{\theta} \mathcal{L}(f_i^{\theta}(X_i), Y_i), \\ \text{s.t.} \quad &\arg\max_{\theta} \quad \mathbb{E}[f_i^{\theta}(X_j) = Y_j], \quad \text{for } j \leqslant i, \end{aligned} \tag{1}$$

其中 θ 代表模型的参数,X_i 和 Y_i 包含 D_i 中的所有样本和标签. 需要注意的是, 旧类别数据 (X_j, Y_j) 可能会与新类别数据 (X_i, Y_i) 融合后交替出现, 也可能不交替出现, 这取决于训练过程是否重放旧类别样本 [1, 19]. 需注意, 为了符号表示的简洁性, 在后文中隐藏模型参数 θ.

2.2 攻击者能力与目标

本文假设一个攻击者仅能访问初始训练集 $D_1 = (X_1, Y_1)$, 以及在初始数据集 D_1 上训练的初始模型 $f_1(x)$. 攻击者对 $f_1(x)$ 的访问能力为白盒, 在本文中, $f_i(x)$ 也称为目标模型, 其中 $f_1(x)$ 为初始目标模型. 目标模型的 CIL 训练策略和增量学习的子数据集均对攻击者不可见. 在这一场景下, 攻击者的目标是基于任何源样本 X_s, 生成对抗样本 X_{adv}, 来欺骗任意一个 CIL 更新轮次的目标模型 $f_i(x)$. 具体地, 攻击者需要在无法访问任何新类别的数据集以及更新后的模型 $f_i(x)$ 的条件下, 构建一个通用的对抗样本生成方法 $\Phi(\cdot)$, 能够基于源样本生成对抗样本: $X_{adv} \leftarrow \Phi(X_s)$, 从而使 $f_i(x)$ 错误地将 X_{adv} 分类为特定的目标类 y_t[1)]. 为实现该目标, 攻击者通过以下过程来优化对抗样本生成方法中的参数, 最终制作对抗样本:

$$\begin{aligned} &w^* = \arg\min_{w} \mathbb{E}_{x_s \in X_s}[\mathcal{L}(f_i(\Phi^w(x_s)), y_t)], \forall i \geqslant 0 \\ &\text{s.t.} \quad ||x_{adv} - x_s||_{\infty} \leqslant \epsilon, \quad x_s \in X_s \end{aligned} \tag{2}$$

1) 本文旨在探讨有目标对抗场景, 不包含无目标对抗攻击, 这是由于无目标对抗可通过向源样本加入足够大的对抗噪声实现.

https://www.sciengine.com/doi/10.1360/SSI-2025-0474

Accepted

其中, w 表示对抗样本生成方法 $\Phi(\cdot)$ 中的可训练参数, X_s 为源样本, 其标签不包含目标类 y_t, 模型 $f_i(x)$ 是在一系列数据集 $\{D_1, D_2, ..., D_i\}$ 上逐步训练得到的, $\mathcal{L}(\cdot)$ 是用于计算模型预测与目标标签之间差异的损失函数. 对抗样本与源样本的距离受到 l_∞ 范数约束, 阈值为 ϵ, 目的是确保对抗样本在输入空间中的变化小, 保持对抗样本的隐蔽性. 需注意, 由于攻击者仅能访问 $f_1(x)$, 无法直接优化公式 2, 因此需借助 $f_1(x)$ 进行求解.

此处定义对抗样本的可持续性, 其表示为对抗样本在不同新增类别数目下的对抗攻击成功率.

定义 1 对抗样本的可持续性: 给定一组对抗样本 $\{x_1, x_2, \ldots, x_n\}$, $f_i(x)$ 为第 i 次增量学习获得的模型, D_i 为第 i 轮的增量学习子数据集, x_{adv} 为生成的对抗样本, y_t 为目标类标签. 对抗样本的可持续性度量定义为以下函数:

$$S_i = \mathbb{E}_{x \in X_{test}}[\mathbb{I}(f_i(x_{adv}) = y_t)] \tag{3}$$

其中, X_{test} 表示未知的测试数据集, $\mathbb{I}$ 是指示函数, 当 $f_i(x_{adv}) = y_t$ 时值为 1, 否则为 0. S_i 表示在第 i 轮训练后, 对抗样本成功使模型发生特定错误预测的概率, 即对抗样本在该轮增量学习模型上的攻击成功率. S_i 随着训练轮次的增加, 表示对抗样本在不同 CIL 训练阶段的可持续性.

3 方案设计

本文提出一种基于语义强化的有目标对抗样本生成方法, 增强对抗样本在增量学习场景下的可持续性. 本章将首先通过实验中的关键观察引出研究动机, 再阐述方案详细设计.

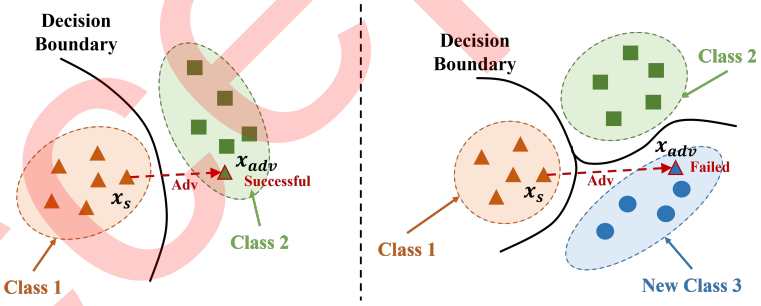


图 1 (网络版彩图) CIL 前后模型知识域变化示意图. 左侧图表示模型增量学习前攻击, 右侧图表示增量学习后攻击. 新增类别为 Class 3, 红色虚线箭头连接了源样本 x_s 与其对应的对抗样本 x_{adv}.

Figure 1 (Color online) Illustration of model knowledge domain changes before and after CIL. The left picture shows the attack against the model before CIL, and the right picture shows the attack after CIL. The red dashed arrow connects source samples x_s and adversarial examples x_{adv}.

3.1 核心思想

传统的对抗攻击通常基于目标模型梯度, 寻找最小的扰动, 以使受扰后的对抗样本越过目标模型的决策边界 [3,4], 从源类别所在的源域迁移至目标类别所在的目标域. 尽管已有对抗扰动迁移性增强的研究, 如 Clean-Sheet [18] 和 Univintruter [12] 等, 其旨在生成具有与目标类样本语义相似的对抗样本, 但仅能够在基于新任务微调的模型上维持鲁棒性, 无法适应 CIL 中模型知识域的剧烈变化. 参考文献 [2], 在图 1 中对 CIL 前后模型的知识域变化进行了抽象描述. 当模型在新类别数据上进行更新后, 原目标域发生漂移, 而新类别的知识域占据部分原目标域. 这导致在样本中添加相同的对抗扰动无法使 Class 1 样本的被分类为目标类 Class 2, 而是落入 Class 3 的知识域中. 换而言之, 目标模型决策边界的显著变化使原对抗样本所加入的噪声方向和大小不适用于更新后知识域显著漂移后的模型 [2], 导致对抗样本失效.

https://www.sciengine.com/doi/10.1360/SSI-2025-0474

本文从实验角度进一步对该现象进行说明. 图 2 展示了当前几种主流的对抗样本在面对 CIL 更新 [19] 时的对抗样本对目标类的 GradCAM[2], 横轴的数字代表了模型学习的类别数, 初始模型学习类别数为 10, 每个轮次的增量类别数为 10. 对抗攻击包括基于梯度的方法 [3] 和基于生成模型的方法[11,12,14,16,18], 图中分别展示了这两类中典型方法的结果. 每个 GradCAM 图像上方的数字衡量了其与初始增量 (10 类) 时的图像的归一化欧氏距离. 可观察到, 随着学习类别数量的增加, 现有对抗攻击的有效性逐渐减弱, 对目标类产生贡献的区域也随着训练过程推移而逐渐缩小, 欧式距离逐渐增大. 即, 针对初始目标模型生成的对抗样本语义产生了漂移. 综合以上讨论, 以往的对抗攻击失效原因在于, 对抗样本仅基于初始目标模型梯度生成目标类语义, 这种语义具有较强的过拟合特性. 随着 CIL 过程的推进, 模型知识域的变化使这类对抗样本退化为普通噪声. 因此, 本文的核心思想是强化对抗样本的目标类语义, 减小对抗样本在初始模型上的过拟合, 从而生成在 CIL 过程中具有鲁棒性的对抗样本. 以下对本文方法进行详细介绍.

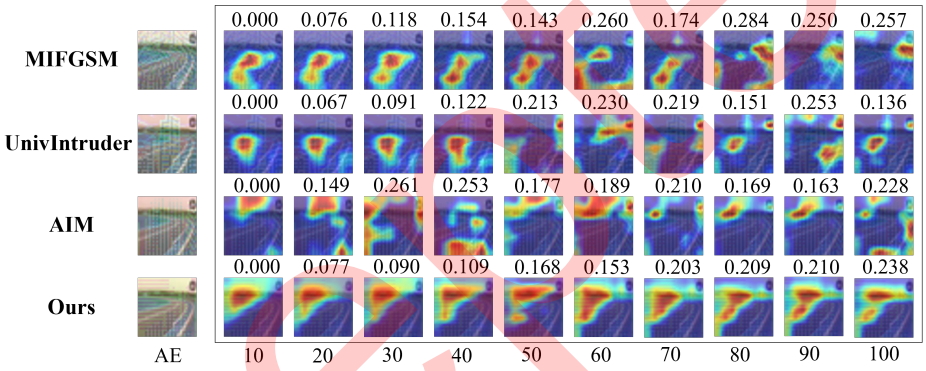


图 2 (网络版彩图) CIL 过程中, 不同攻击产生的对抗样本的 GradCAM.

Figure 2 (Color online) GradCAM of different adversarial examples during the CIL process.

3.2 方法总述

本文提出基于语义强化的对抗样本生成方法, 该方法进一步分为两部分主要算法, 样本生成与筛选 (Sample Generation & Selection) 和语义提取与对抗样本生成 (Semantics Extraction & AE Generation), 如图 3 所示. 其中, 第一部分的目的是获得对目标模型而言具有强语义的目标类样本; 第二部分则基于筛选出的强目标类语义样本, 提取稳定的目标类语义, 并生成鲁棒的对抗样本. 接下来本文对每个模块算法进行详细介绍.

3.3 样本生成与筛选

首先, 攻击者需获取一组具有强目标类语义的样本. 尽管公共数据集或图像网站也能够提供目标类样本, 但此类样本中可能含有其他类别语义, 例如, 当目标类为‘自行车’时, 样本可能还包括‘公路’. 这将导致目标类样本语义混淆, 使提取出的语义不稳定. 为了减小语义混淆, 本方法首先基于扩散模型 [20,21] 生成目标类样本. 参考现有文献 [22], 假设目标类为‘Cat’, 首先制作一组提示词用于描述目标类, 然后利用扩散模型根据提示词生成候选目标类样本集合. 然而, 仅通过生成模型得到的样本仍可能存在语义强度差异, 因此, 攻击者需进一步筛选具有强语义的目标类样本. 从持续学习的角度看, 对抗扰动需要在模型多次更新过程中保持强目标类语义, 从而抵抗模型的灾难性遗忘 [23]. 根据模型灾难性遗忘相关理论, 样本被模型遗忘的速率与其学习速率呈负相关关系. 具体而言, 在训练阶段较早被模型正确分类的样本通常对应更加稳定的语义特征, 在后续模型更新过程中更难被遗忘. 为构造在 CIL 过程中可持续的对抗样本, 应优先选择遗忘速率较低的样本, 即学习

2) GradCAM 是一种常用的工具, 用于分析对目标类贡献最大的区域.

https://www.sciengine.com/doi/10.1360/SSI-2025-0474

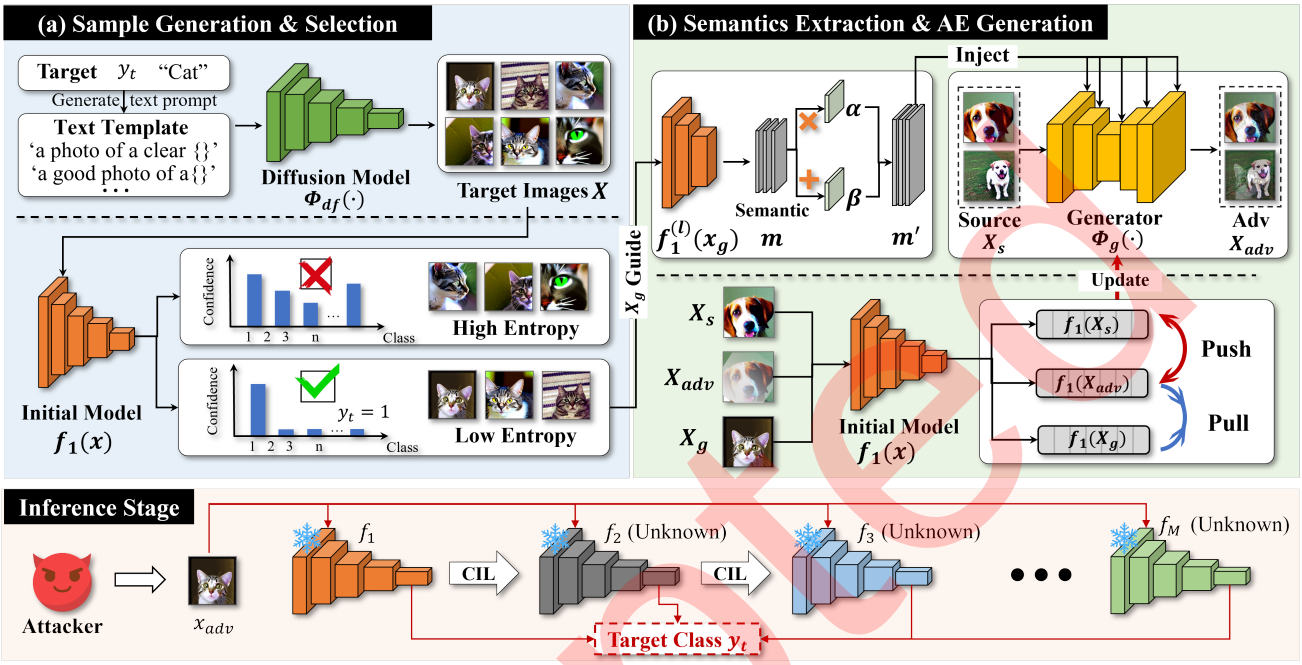


图 3 (网络版彩图) 算法示意图. (a) 样本生成与筛选, 旨在找到具有强语义的目标类样本; (b) 语义提取与对抗样本生成, 旨在提取稳定的目标类语义, 并生成鲁棒的对抗样本.

Figure 3 (Color online) Overview of the proposed method: (a) Target Class Sample Generation and Selection, which aims to identify samples with strong semantics of the target class; (b) Target-Class Semantic Extraction and Adversarial Example Generation, which focuses on subtly embedding the target class semantics to generate adversarial examples.

速率较高的样本作为目标类语义来源. 然而, 攻击者通常无法直接观察或控制持续学习模型的训练过程, 只能获得初始模型的白盒访问权限, 无法直接确定初始训练集中学习速率高的样本. 为解决该局限性, 本文提出一种基于预测熵的样本筛选方法, 利用初始模型上的预测分布来近似估计样本的学习速率, 以近似地筛选学习速率高的强语义样本.

假设分类模型为 $f(x)$, 训练数据为 D, 训练过程可以表示为对 m 个数据子集的逐步优化, 其中 $D = \{D_1, D_2, .., D_m\}$, 其中 $D_i = (X_i, Y_i)$ 代表训练数据和标签, 每组 (X_i, Y_i) 可能存在不同的分布. 当 $f(x)$ 在 D_i 上进行更新时, 优化目标为最小化预测向量与训练标签的熵: $\min \mathcal{L}[f(X_i), Y_i)]$. 其中 $\mathcal{L}$ 通常为交叉熵损失函数. 该优化过程使模型对训练样本逐渐形成确定性的类别预测. 在 CIL 中, $f_1(x)$ 则近似于在第一组数据 D_1 上训练的早期模型. 因此, 若一组样本 X 在模型 $f_1(x)$ 上的预测结果接近真实标签, 则可以认为 X 具有较高的学习率. 对于任意样本 $x_i \in X_i$, 若其具有强语义, 模型在训练后往往能够以较高置信度预测其真实标签, 即预测概率分布 $P(f(x_i)|x_i)$ 在标签类别 y_i 上显著高于其他类别, 对应较小的预测熵. 由此将预测熵作为样本语义强度的近似指标.

本文的样本生成与筛选方法见算法 1. 首先, 根据目标类别 y_t 通过提示词生成函数 $\Phi_p(\cdot)$ 构造一组目标类文本提示词 T, 并利用预训练扩散模型 $\Phi_{df}(\cdot)$ 生成候选样本集合 X. 随后, 算法利用初始目标模型 $f_1(x)$ 对生成样本进行语义一致性验证. 对于每个候选样本 x_i, 首先计算模型预测概率向量 $p_i = P(f_1(x_i))$, 并仅保留被模型正确分类为目标类别 y_t 的样本. 对于这些样本进一步计算预测熵 $\mathbb{H}(p_i)$, 以衡量模型预测的不确定性. 最后, 根据预测熵对样本进行排序, 并选取预测熵最低的 r 比例样本作为最终的目标类样本集合 X_g. 其中,$r \in (0, 1]$ 为超参数, 用于控制筛选样本的比例. 需注意, 该算法中 $\Phi_p(\cdot)$ 的输入 y_t 表示目标类对应的文本. 第 12 行中, 超参数 $r \in (0, 1]$, 表示从样本集合 $\hat{X}$ 中选取预测熵最低的 r 比例的样本.

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0474

算法 1 样本生成与筛选算法

输入: 预训练的扩散模型 $\Phi_{df}(\cdot)$, 提示词生成函数 $\Phi_p(\cdot)$, 初始目标模型 $f_1(x)$, 目标类 y_t, 超参数 r;

生成一组目标类提示词 $T = \Phi_p(y_t)$;
生成目标类样本 $X = \Phi_{df}(T)$;
初始化集合 $\hat{X} = \emptyset$;
for $x_i \in X$ **do**
    计算 $p_i = P(f_1(x_i))$;
    if $\arg\max_k p_{i,k} = y_t$ **then**
        $\mathbb{H}(p_i) = -\sum_{k=1}^{C} p_{i,k} \log(p_{i,k})$;
        $\hat{X} = \hat{X} \cup x_i$;
    end if
end for
$X_g = \{x_i \in \hat{X} | \mathbb{H}(p_i) \leqslant \mathbb{H}_r(\hat{X})\}$;

输出: 目标类样本 X_g.

3.4 语义提取

本模块将筛选后得到的强语义样本作为引导样本, 攻击者进一步提取其中的目标类语义特征作为生成对抗样本的优化方向. 现有对抗方法中的语义提取模块通常侧重于学习通用语义表示. 例如, 一些方法利用预训练视觉模型提取高层语义特征 [12,14], 或通过频域分析挖掘类别相关属性 [24,25]. 这些方法的核心假设是, 不同结构的模型能够提取相似的目标类语义表示, 从而使生成的对抗样本具有较强的跨模型迁移能力. 然而, 在 CIL 场景中, 模型参数在持续更新过程中会逐渐发生漂移, 从而导致模型知识域不断变化. 如章节 3.1所述, 随着新类别不断引入, 模型高层语义空间会逐渐偏移, 原有类别的语义表示可能退化为一般噪声. 因此, 依赖高层语义表示的对抗扰动往往难以在 CIL 更新后的模型上保持持续攻击效果. 从深度网络表征学习的理论角度分析, 深层神经网络通常呈现明显的层级语义结构. 已有研究表明, 网络浅层主要提取通用视觉特征, 如边缘、颜色和纹理等低级结构信息, 而深层则逐渐编码更加抽象的类别相关语义模式. 而浅层特征在不同任务之间具有更强的可迁移性和稳定性, 而深层特征则更加依赖具体任务分布 [26].

在持续学习场景中, 这一现象进一步被强化. 为了在不破坏已有知识的情况下快速适应新类别, 许多持续学习策略倾向于保持浅层参数相对稳定, 而主要更新深层参数以适应新的类别分布 [8,27,28]. 因此, 相比深层语义表示, 浅层语义在多次模型更新过程中具有更高的稳定性. 基于这一关键特征, 本文利用初始模型 $f_1(x)$ 的浅层网络作为语义提取器, 从强语义样本中提取具有更高稳定性的目标类语义表示, 并将其作为对抗噪声生成的优化方向. 具体而言, 对于输入样本 x, 其浅层语义表示可以表示为:

$$m = f^{(l)}(x), \quad \text{where} \quad l \in \{1, 2, \ldots, L\}, \quad L \ll \text{total layers} \tag{4}$$

其中 $f^{(l)}(\cdot)$ 表示模型第 l 层的特征映射, m 为提取出的目标类语义表示. 由于 l 远小于模型总层数, 所得到的特征主要反映输入样本的稳定低层语义结构, 从而能够在 CIL 过程中保持更强的可持续性.

3.5 对抗样本生成

本模块将提取出的目标类语义注入基于卷积结构的对抗样本生成器中, 将源样本转化为对抗样本. 生成过程的核心目标是在保持样本视觉自然性的同时, 将目标类别语义稳定地嵌入源样本. 神经网络的分类学习中, 样本类别的判定通常依赖于特征空间中的类内紧致性与类间可分性, 许多研究通过同时减小同类样本距离并增大异类样本距离来学习具有判别性的嵌入空间 [29]. 考虑到 CIL 过程中的模型遗忘对类间距影响加剧, 在语义引导的对抗样本生成过程中, 仅推动样本靠近目标类别语义无法保证其稳定落入目标类别决策区域, 还需要同时降低其与非目标类别语义的相似度, 以扩大类别间语义间隔. 近期针对有目标对抗攻击的研究亦表明, 通过强化目标类别特征并抑制原类别或其他类别特征, 可显著提升语义迁移能力与攻击稳定性 [12]. 基于上述理

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0474

算法 2 对抗样本生成算法

输入: 初始 CIL 模型 $f_1(\cdot)$, 生成器模型 $\Phi(w,\alpha,\beta)$, 目标类样本 X_g, 删除 y_t 类样本后的初始训练集 D_0', 噪声预算 ϵ, 步长 η, 层数 l.

初始化 w, α, β;
for each batch $(X,Y) \subset D_0'$ **do**
  for $x_g \in X_g$ **do**
    初始模型浅层特征提取 $m = f_1^{(l)}(x_g)$;
    计算生成模型第 0 层 $h_0' = f_0(X)$ 户
    for each $i \in \{1,2,\ldots,L-1\}$ **do**
      $m' = (1+\alpha_i)m + \beta_i$;
      $h_i = f_i^{bn}(h_{i-1}')$
      $h_i' = \mathrm{Interp}(m', w_i, h_i)$;
    end for
    $X_{adv} = \mathrm{Interp}(m, w_L, h_{L-1}'')$;
    $\mathcal{L} = \mathcal{L}_{CE}[f_1(X_{adv}), f_1(X_g)] - \mathcal{L}_{CE}[f_1(X_{adv}), X]$;
    $\alpha \leftarrow \alpha - \eta\nabla_\alpha\mathcal{L}$;
    $\beta \leftarrow \beta - \eta\nabla_\beta\mathcal{L}$;
    $w \leftarrow w - \eta\nabla_w\mathcal{L}$;
  end for
end for
给定非目标类的源样本 X_s, 计算 $X_{adv} = \Phi(X_s, X_g, w, \alpha, \beta)$;

输出: 对抗样本 X_{adv}.

论基础, 本文定义了一组语义生成约束条件: 1) 目标语义一致性约束, 即对抗样本在特征空间中应与目标类别语义保持较高相似度; 2) 非目标语义分离约束, 即对抗样本需与非目标类别语义保持充分距离, 其示意见图 3 中绿色区域. 前者确保生成样本的语义表征稳定向目标类别迁移, 而后者通过扩大类别间语义间隔避免语义混淆, 从而降低知识域漂移带来的对抗语义失效风险.

已有研究充分探讨了语义引导的图像生成模型 [18,30,31], 其能够将语义信息嵌入生成器的每个 BatchNorm 层, 使生成器在不同层的特征空间中保持目标类语义的一致性. 基于上述思想, 本文构建了一种语义引导的对抗样本生成算法（见算法 2）. 该算法通过一个卷积生成器 $\Phi(\cdot)$ 在多层特征空间中逐步注入目标类别语义, 从而实现源样本语义向目标类别语义的隐蔽迁移. 具体地, 对于目标类别样本 x_g, 首先利用初始模型 $f_1(\cdot)$ 提取其浅层语义特征 m, 并将其作为目标语义原型注入生成器的中间层特征中. 随后, 生成器在每个特征层通过语义嵌入模块对中间特征进行调制, 其中第 10、12 行的 $\mathrm{Interp}(\cdot)$ 表示特征插值函数, 用于融合目标语义特征与当前层特征, 而 $f_i^{bn}(\cdot)$ 表示生成模型第 i 个 BatchNorm 层的特征变换操作. 将生成器模型表示为 $x' \leftarrow \Phi_g(x)$, 参考已有研究的生成器模型结构 [18], , 本文在生成器的中间特征层引入一组仿射变换参数以控制语义注入过程. 具体而言, 每一层特征图通过尺度参数 α_i 与位移参数 β_i 进行调整, 从而实现对目标语义的逐层嵌入, 这两组参数不断更新, 以使生成的对抗样本在模型各个层的特征空间中接近于目标类. 其计算方式如下:

$$\begin{aligned} m' &= (1+\alpha_i)m + \beta_i, \\ i &= 1,2,...,N \end{aligned} \tag{5}$$

其中, α_i 和 β_i 分别表示第 i 层的尺度参数和位移参数, 计算方法为: $\alpha_i = \mathrm{Conv}(x_g^i), \beta_i = \mathrm{Conv}(x_g^i)$, N 表示语义注入模块的总数目, 其与生成器的 BatchNorm 层数相同, x_g^i 是目标类的一个样本, 作为目标类语义的引导样本, 通常, 此样本能够在公共数据集获取或通过搜索引擎获取, w_i 是权重矩阵, h_i 表示模型第 i 层的嵌入值矩阵. 经过多轮优化参数 w, α, β 迭代更新生成器后, 其能够将目标类别语义隐蔽地嵌入输入样本, 最终得到具有稳定目标语义的对抗样本 X_{adv}.

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0474

4 实验

4.1 实验设置

数据集和模型. 本文考虑了之前工作中广泛采用的两个数据集: CIFAR-100 [32] 和 ImageNet-100 [33], 每个数据集由 100 个类组成. 基于当前的类增量学习研究 [2], 这两个数据集被分为 10 组, 每组包含 10 个类, 用于模拟类增量学习. 其中, 初始模型 f_1 基于前 10 个类的样本进行训练. 目标类是从初始模型训练时的 10 个类中随机选择的, 这 10 个类是攻击者能够访问的类. 在模型设置方面, 本文选择类增量学习中的典型模型 [2], ResNet-32 和 ResNet-50, 二者均具有相同的卷积层结构. 其中, ResNet-32 模型用于 CIFAR-100 数据集的训练, ResNet-50 模型用于 ImageNet-100 数据集的训练, 以保证最佳的模型性能.

CIL 方法和基准方法. 在 CIL 领域, 围绕模型更新策略已产生一系列研究, 依据文献 [2, 34], 现有 CIL 方法被分为 5 种类型: 1) 微调 (Fine-tuning) 方法, 作为通用的 CIL 基准策略, 通常通过调整模型后几层参数, 在新类别样本上进行训练; 2) 回放 (Replay) 方法 [19], 通过存储旧类样本, 并将之加入新类数据进行交替训练, 缓解灾难性遗忘; 3) 知识蒸馏方法 (KD, Knowledge Distillation), 通常使用旧模型的软目标进行正则化, 例如 iCaRL [8] 与 PodNet [35]; 4) 动态网络方法, 通过扩展模型架构或调整参数以适应新类数据, 如 DER 方法扩展新主干结构 [28], Foster 方法通过蒸馏算法压缩模型参数 [36]; 5) 模型校正方法, 通过对模型进行后期校准, 减少对新类别样本的预测偏差, 如 BiC [37] 在验证集上训练线性校正层, WA [38] 在优化步骤后归一化分类器权重. 所有基准方法的默认噪声预算设置为 $\epsilon = 32/255$. 为充分探究本文所提方案的有效性, 本文在所有 5 种类型的 CIL 方法上均进行了测试实验, 包括 Fine-tuning [27], Replay [19], iCaRL [8], Foster [36] 和 WA [38] 方法[3].

考虑到当前对 CIL 的对抗样本生成方法的缺失, 本文将可迁移对抗样本生成方法 [39] 作为基准方法. 可迁移对抗样本样本方法旨在使对抗样本在模型结构、参数发生变化后仍然能够产生对抗效果. 二者攻击的目标模型均与源模型具有不同的参数. 参考章节 1.2, 可迁移的对抗样本方法可进一步分为: 1) 基于迭代的方法, 本文选取 MIFGSM [3]、CleanSheet [11] 和 UnivIntruder [12] 方法作为基准方法; 2) 基于生成模型的方法, 本文将 AIM [18]、GAKer [16] 和 CGNC [14] 作为基准方法.

评估指标. 本文从两方面考虑性能评估指标. 根据定义 1, 考虑每轮 CIL 更新后的攻击成功率 (ASR, Attack Success Rate), 该指标为对抗攻击性能的典型指标, ASR_i 表示在第 i 个任务更新后, 模型 $f_i(\cdot)$ 将对抗样本分类到目标类中的比例, 能够衡量对抗样本随着更新轮次的可持续性. 另一方面, 为评估攻击在整个 CIL 过程中的可持续性, 本文定义平均攻击成功率 (SASR, Sustainable Attack Success Rate), 该指标计算了在整个增量学习过程中, 所有目标类别上的平均攻击成功率, 计算方法如下:

$$\mathrm{SASR} = \frac{1}{M \times C} \sum_{i=1}^{M} \sum_{j=1}^{C} \mathbb{I}[f_i(x_{adv}^j) = y_t] \tag{6}$$

其中 M 表示增量次数,C 表示测试集中用于评估攻击性能的对抗样本数量,$\mathbb{I}(\cdot)$ 是一个指示函数, 正确分类时为 1, 否则为 0, $x \in X_{test}$, $C = |X_{test}|$.

其他设置. 本文方法使用文献 [22] 中经典的预训练扩散模型 Stable Diffusion v2 Model 生成目标类样本, 并使用其默认推理设置进行对抗样本生成: 噪声调度器设置为去噪扩散隐式模型 DDIM (Denoising Diffusion Implicit Models), 扩散步数设置为 50, 无分类器引导系数为 7.5, 初始随机噪声服从标准高斯分布. 在随机性设置方面, 本文对每个攻击样本进行 5 次独立随机采样, 并报告其平均结果以减小随机波动的影响. 对样本筛选的设置采取生成样本中预测熵最低的 10%, 即 $r = 0.1$. 为提取目标类语义, 对两种目标模型, 均提取第 1 层卷积层语义, 即 $l = 1$. 训练过程采用 Adam 优化器 [40], 学习率设置为 $2e^{-4}$, 将动量衰减因子设置为 0.5 和 0.999,

3) 所述 CIL 方法见 https://github.com/LAMDA-CL/PyCIL.

https://www.sciengine.com/doi/10.1360/SSI-2025-0474

表 1 不同攻击方法和 CIL 方法在 10 个目标类上的平均 SASR. 所有 CIL 方法均在 CIFAR-100 和 ImageNet-100 数据集上训练. Clean Acc 表示模型在十个任务上的平均准确率.

Table 1 Average SASR across ten target classes for various attacks and CIL methods. All CIL methods are trained using the CIFAR-100 and ImageNet-100 datasets. Clean Acc denotes the average accuracy of models trained across ten tasks.

Dataset	Attack	FT	RP	MEMO	DER	Foster	WA	BiC	iCaRL	PodNet	AVG.
CIFAR-100	Clean Acc	26.23	58.38	71.69	70.90	64.89	65.54	60.39	62.85	55.65	59.61
	MIFGSM	7.29	13.41	21.71	36.85	17.93	16.51	18.95	17.94	44.17	21.64
	GAKer	0.04	0.02	0.01	0.00	0.00	0.04	0.00	0.00	5.43	0.62
	AIM	**9.94**	30.96	0.05	0.04	48.47	56.20	0.04	28.15	0.02	19.32
	CGNC	0.57	0.12	0.00	0.00	0.48	0.30	0.00	0.00	0.08	0.17
	CleanSheet	1.23	2.31	3.80	1.16	2.02	2.71	4.02	6.00	7.88	3.46
	UnivIntruder	4.85	31.23	34.82	40.23	31.23	41.82	33.62	32.43	42.20	32.49
	Ours	**5.56**	**50.82**	**52.13**	**60.06**	**54.25**	**81.37**	**58.62**	**57.69**	**82.91**	**55.94**
ImageNet-100	Clean Acc	24.91	58.75	71.69	74.14	68.40	67.39	61.58	57.89	67.28	61.34
	MIFGSM	**9.80**	9.36	16.05	26.23	22.46	15.05	9.48	11.08	68.02	20.84
	GAKer	0.14	1.08	0.33	0.17	0.22	0.00	0.00	0.00	0.17	0.23
	AIM	0.23	4.46	2.26	4.34	1.77	1.09	1.60	1.60	3.60	2.33
	CGNC	0.07	0.11	0.05	0.06	0.05	0.05	0.06	0.01	0.06	0.06
	CleanSheet	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.47	0.00	0.05
	UnivIntruder	0.34	4.17	1.63	3.36	2.82	2.44	0.75	0.93	2.53	2.11
	Ours	**7.70**	**29.47**	**33.57**	**74.10**	**38.09**	**43.39**	**24.84**	**34.80**	**76.69**	**40.30**

训练中批数据大小为 128, 优化进行了 50 个迭代轮次. 此外, 根据可迁移黑盒对抗攻击的标准设置 [12, 41~43], 设置噪声预算 ϵ 为 16/255.

4.2 攻击性能

首先将本文所提方法与不同 CIL 方法下的几种基准攻击方法进行比较, 使用 CIFAR-100 和 ImageNet-100 数据集上所有 CIL 任务的平均 SASR 来评估攻击性能. 结果如表 1所示, 表中数据是在整个 CIL 过程中测试的 SASR 结果, 即, 模型完成了 100 个类别的增量学习, 每次增加类别数为 10 个, 下划线标记了每种 CIL 设置下的最佳结果. 为简洁表示, FT 代表 Fine-tuning 方法, RP 代表 Replay 方法. 从表中结果能够看出, 在几乎所有 CIL 场景下, 本文所提方案均优于基准方法. 在 CIFAR-100 上, 本文方法在所有 CIL 设置中平均高于所有基准方案 42.99%, 在 ImageNet-100 数据集上同样提升了 35.76%. 因此, 本文方法在两个数据集上达到了 39.38% 的平均性能提升, 显著提升了对抗攻击的可持续性. 尽管本文方法在 FT (Fine-tuning) 设置中未取得最优性能, 但考虑到该设置下的平均模型精度 (Clean Acc) 仅为 25.57%, 可推断模型在增量学习中产生了较严重的灾难性遗忘, 模型无法有效进行增量学习, 因此在 FT 场景下的对抗样本性能下降可忽略不计. 进一步地, 分析基于生成模型 (GAKer、AIM、CGNC 和本文方法) 和基于迭代 (MIFGSM、CleanSheet 和 UnivIntruder) 的对抗方法的性能. AIM 考虑了将目标类语义注入对抗样本, 因此相比于 GAKer 和 CGNC 方法具有更强的可持续性. 与可迁移对抗样本生成方法, 例如 AIM、UnivIntruder 方法中通用的目标类语义嵌入相比, 本文方法中的对抗样本语义依据初始模型预测熵进行了筛选, 并经过模型浅层特征提取得到了强化, 因而在 CIL 的知识域漂移中具有更强的鲁棒性, 取得了更优的性能. 对于迭代攻击, 本文方案也显著优于 MIFGSM、CleanSheet 和 UnivIntruder. 由于 MIFGSM 中的对抗噪声需依据初始模型 $f_1(\cdot)$ 的梯度生成, 其易在 $f_1(\cdot)$ 上过拟合. 而 CleanSheet、UnivIntruder 与 AIM 类似, 三者均侧重于更为通用的目标类语义, 而随着 CIL 过程中模型的知识域漂移, 二者的性能也会随之下降. 总体而言, 本文方案的表现优于所有基准方法,

https://www.sciengine.com/doi/10.1360/SSI-2025-0474

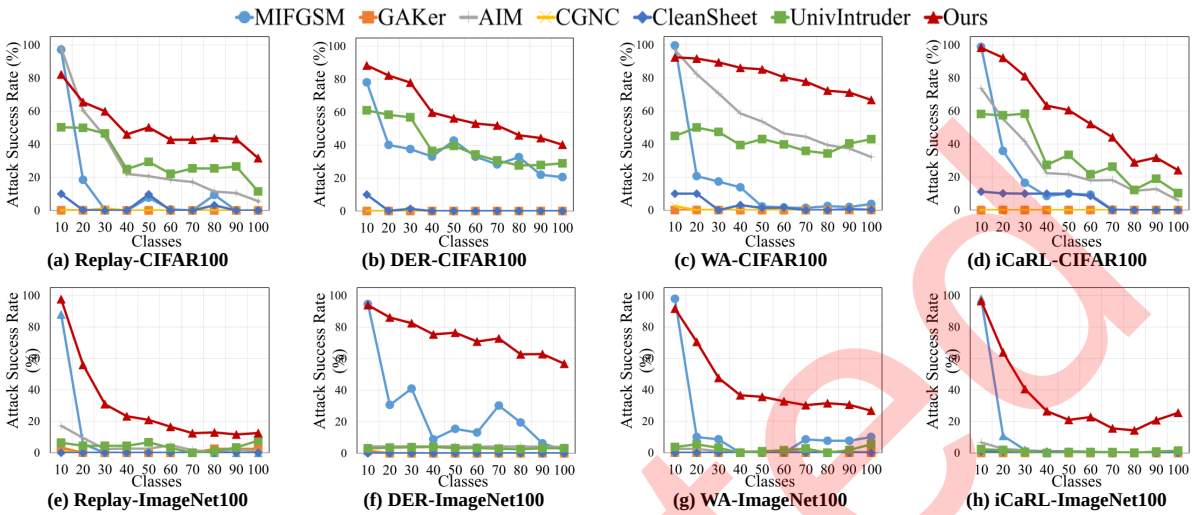


图 4 (网络版彩图) 不同 CIL 方法下基准攻击与提出攻击的 ASR 曲线. 子图 (a)-(d) 展示了在 CIFAR-100 数据集上, 目标类为 skyscraper 时, 4 种 CIL 方法的结果. 子图 (e)-(h) 展示了在 ImageNet-100 数据集上, 目标类为 candy store 时, 4 种 CIL 方法的结果.

Figure 4 (Online Color) ASR curves for both baseline attacks and the proposed attack across various CIL methods. Each subfigure illustrates the ASR across incremental tasks. Subfigures (a)–(d) present results for the CIFAR-100 dataset with 'skyscraper' as the target class, evaluated on four representative CIL methods. Subfigures (e)–(h) show the corresponding results for the ImageNet-100 dataset using 'candy store' as the target class, evaluated on the same four CIL methods.

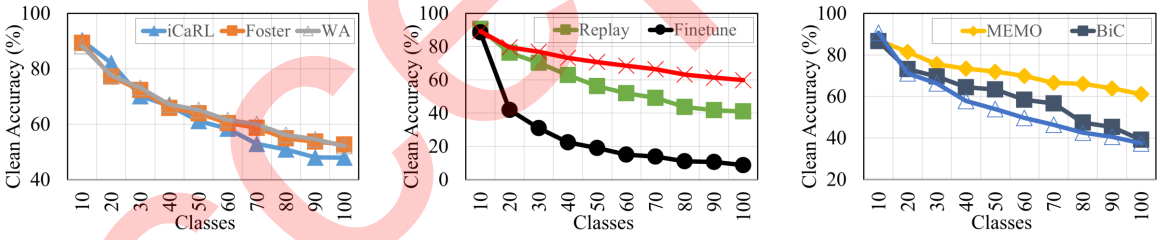


图 5 (网络版彩图) CIFAR-100 数据集上, 不同 CIL 方法的 Clean Acc, 增量类别数为 10.

Figure 5 (Online Color) Clean Accuracy of different CIL methods on CIFAR-100, with 10 new classes each.

这突显了本文所提语义强化方法的有效性, 该模块在防止过拟合的同时确保了对抗性样本的可持续性.

为进一步探讨在 CIL 训练过程中各类对抗攻击的性能, 本文记录了在 2 种数据集上使用不同的 CIL 方法更新时, 所有对抗攻击方法在不同更新轮次时的 ASR, 结果如图 4所示. 图中红色的折线代表本文方法, CIL 训练每次引入 10 个新类别. 可以观察到, 本文方法在整个 CIL 过程中均显著高于所有基准方法, 在类别增加时, ASR 会产生缓慢下降, 且慢于绝大多数基准方案. 从 ASR 稳定性角度而言, 本文方法的 ASR 折线波动相比于基准方法小, 表明本方法在不同的更新轮次上均具有较好的攻击稳定性, 同时也验证了本方法的目标类语义强化方法的优越性, 该语义强化方法保证了稳定的目标类语义, 因此受 CIL 过程的知识域漂移影响较小. 此外, 有部分 CIL 方法上的攻击会产生 ASR 骤降的情况, 如 Replay 和 iCaRL, 这归因于 CIL 方法依赖于固定长度缓存来存储先前的数据或模型参数, 以减轻灾难性遗忘 [2]. 一旦缓存已满, 模型保留早期知识的能力就会减弱, 导致 ASR 显著下降, 因此, 随着 CIL 方法保留早期知识能力增强, 本方法的攻击性能持续性也会增强. 作为辅助说明, 本文在图 5展示了在 CIFAR-100 数据集上进行 CIL 时, 对初始的 10 类数据的分类精度. 能够观察到, 各类 CIL 方法在学习新类别的过程中, 对旧数据分类精度也在缓慢下降.

不同目标类时的对抗样本鲁棒性. 本小节探讨当设置不同的目标类别时, 本文方法的 SASR, 结果如图 6 所示, 其中红色线标注了本文方法在不同 CIL 设置中的 SASR 平均值, 每种 CIL 训练总类别数为 100, 每次更新 10

https://www.sciengine.com/doi/10.1360/SSI-2025-0474

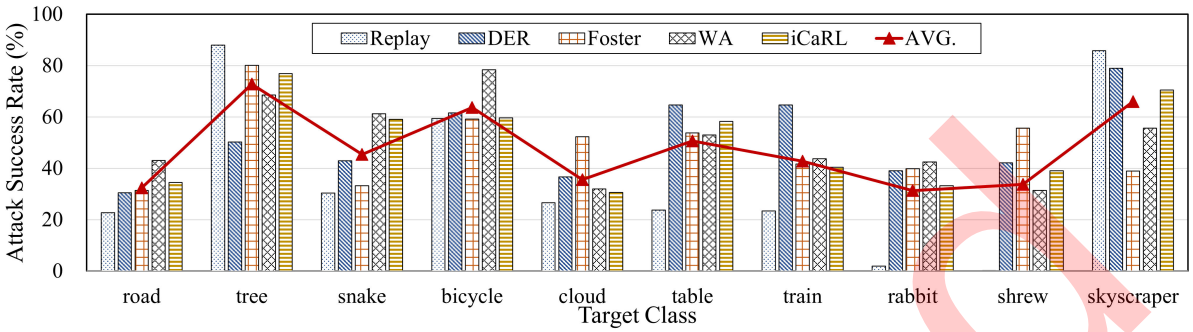


图 6 (网络版彩图) 不同目标类别下的平均 SASR 对比. 模型在 CIFAR-100 数据集上基于 5 种 CIL 方法训练. 结果表明所提方法在多个目标类别上均能维持较高的 SASR.

Figure 6 (Online Color) SASR across different ten attack target classes and five CIL methods, evaluated using models trained on the CIFAR-100 dataset. The results demonstrate the effectiveness of the proposed method in maintaining high ASR across diverse target classes.

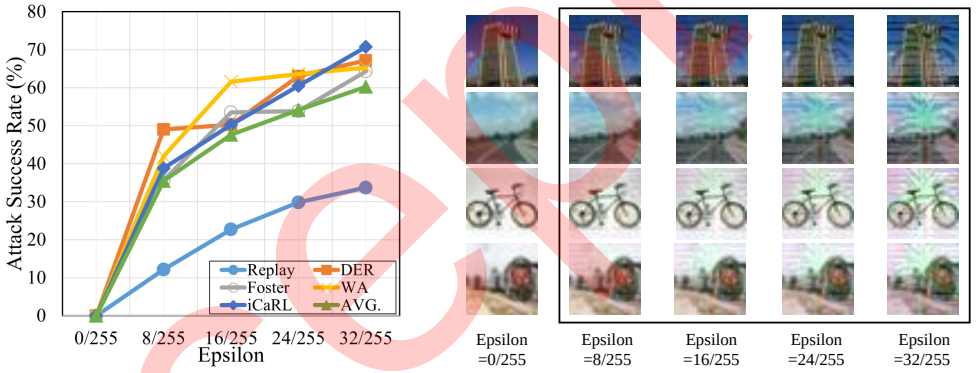


图 7 (网络版彩图) 不同噪声预算时的 SASR 和攻击样本展示. 数据集 CIFAR-100, 目标类 'tree'.

Figure 7 (Online Color) SASR and visualization of adversarial exampes under different nnoise budgets ϵ. Dataset is CIFAR-100 and the target class is 'tree'.

个类. 图中结果表明方法在不同目标类设置下均有效, 平均 SASR 均在 35% 以上. 同时可以观察到, 在部分类别, 如类别 1、3 和 9 上, 能够达到较高的平均 SASR, 而有些类别, 如 0 和 7 相对较低. 该现象与已有研究中普遍观察到的 '类别鲁棒性差异' [44] 一致, 尽管诸如 CIFAR-100 等数据集的训练样本类别数量均衡, 不同目标类别的对抗样本仍然存在显著差异. 当目标类的类内一致性更高, 即与其他类间隔更大时, 攻击在 CIL 过程中具有更高的成功率. 这与图中观察一致, 例如, 'road' 常包含车辆、背景与场景纹理等多源语义, 其类别表征更易被上下文因素主导, 相比 'tree' 和 'bicycle' 等主体更明确的类别, 更难形成可被持续诱导的稳定目标类语义. 此外, 在 CIL 过程中, 模型的灾难性遗忘也会呈现类别不均衡的特征, 不同目标类的可达性差异化, 从而一定程度上影响 SASR 的稳定性. 例如, iCaRL 通过回放样本与蒸馏等机制减轻遗忘, 往往能保持较高的旧类分类精度与更稳定的表示空间, 从而使不同目标类上的平均 SASR 整体更高且波动更小. 考虑到类别语义强度的差异性, 本文建议攻击者尽量选取图像样本中主体较为明确、目标模型分类精度高的类别作为对抗攻击的目标类别.

4.3 扰动可见性

本小节探究不同扰动约束下, 对抗样本的隐蔽性和攻击性能. 本文使用 l_∞ 范数约束对抗扰动, 噪声预算为 ϵ. 图 7中记录并展示了不同 ϵ 时, 在不同 CIL 设置上的 ASR 和攻击样本,ϵ 分别设置为 0/255、8/255、16/255、

https://www.sciengine.com/doi/10.1360/SSI-2025-0474

24/255 和 32/255, 其中 0/255 表示无对抗扰动. 从图中可看出, 随着扰动预算增大,ASR 也逐渐均匀上升. 每当 ϵ 增加 8/255, SASR 通常能够增加约 10%. 即使当噪声预算 ϵ 进一步降低到 8/255 时, 大部分 CIL 上的 ASR 仍然维持在 20% 以上. 这表明即使在更严格的扰动设置下, 本文方法的仍然具有较强的攻击性能. 随着 ϵ 的增大, 样本中目标类 'tree' 的语义也同步增强, 对抗扰动在视觉上更为显著. 因此, 考虑到攻击性能与隐蔽性的平衡, 本文建议设置 $\epsilon = 16/255$.

增量类别规模. 为进一步探讨更大规模增量学习场景下方法的有效性, 本实验随机抽样 ImageNet 数据集的额外 100 个类别作为新的增量类, 构建 200 类的增量学习设置, 并在不同的 CIL 方法下评估攻击成功率, 结果如图 8 所示. 能够看出, 随着增量类别规模显著增加, 不同 CIL 方法下的攻击成功率均呈现缓慢下降趋势, SASR 为 40.189. 该现象源于大规模类增量下, 目标模型对早期类别的遗忘显著加剧, 而不同 CIL 方法对旧类语义的保持能力将直接影响攻击性能. 由于显著的模型灾难性遗忘, 现有研究通常仅在增量类别规模为 100 的设置下探讨攻击方法性能, 以在保持模型对目标类识别的能力的条件下进行更具可比性的评估. 这一趋势主要源于大规模类增量设置下目标模型对早期类别的遗忘程度显著加剧, 导致目标类的判别边界与语义表征难以保持稳定. 在此情况下, 模型自身对目标类的分类能力已明显退化, 因此在此条件下对抗攻击成功率可比性较弱.

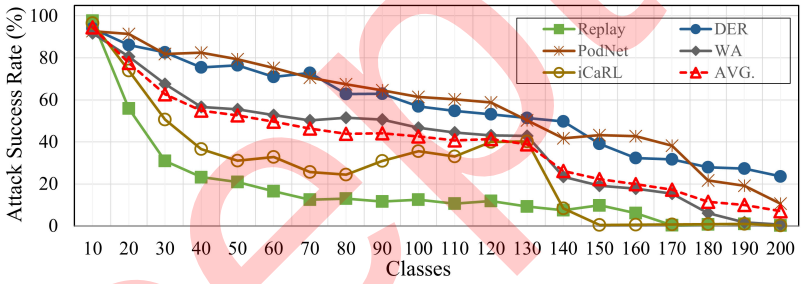


图 8 (网络版彩图) 当增量类别为 200 时, 不同 CIL 方法下的攻击成功率. 数据集 CIFAR-100, 目标类 'tree'.

Figure 8 (Online Color) Attack success rate across different CIL methods at 200 incremental classes on CIFAR-100 (target class: tree).

无目标攻击性能. 本实验进一步探讨在噪声预算 $\epsilon = 16/255$ 和 $\epsilon = 32/255$ 的条件下, 本文方法的攻击性能. 对于无目标攻击, 需使对抗语义尽可能远离源类. 基于此, 对本文算法进行调整: 生成源类提示词并筛选样本, 然后使模型浅层提取出的对抗样本语义远离源类别, 即, 仅保留损失函数的第二项:$\mathcal{L} = -\mathcal{L}_{CE}[f_1(X_{adv}), X]$. 本节实验保持相同的 10 个初始类别和增量类别设置, 每次增量训练类别数为 10, 并记录当增量类别增加时的攻击成功率, 结果如表 2. 可以看出, 在不同的噪声预算下, 无目标攻击的成功率均均超过 90%. 这是由于无目标对抗攻击仅需使预测类别不等于原类别, 其可行解范围远大于有目标攻击.

表 2 CIFAR-100 数据集上不同 ϵ 设置下随类别增量变化的无目标对抗攻击成功率 (ASR)

Table 2 Untargeted adversarial attack success rate (ASR) on CIFAR-100 under different ϵ with increasing classes.

ϵ	M=1	M=2	M=3	M=4	M=5	M=6	M=7	M=8	M=9	M=10	SASR
16/255	83.67	84.16	86.89	92.41	90.67	95.34	94.73	96.72	98.77	99.72	**92.46**
32/255	85.02	88.66	92.42	96.77	95.65	98.88	96.47	99.21	99.15	99.89	**95.90**

4.4 攻击成本分析

本小节对本文方法中两个主要高开销环节的计算成本进行系统分析, 包括: (1) 利用扩散模型生成并筛选候选目标样本; (2) 迭代的对抗样本训练. 实验中批次大小设置为 256, 训练轮数为 50, 数据集为 CIFAR-100. 表 3 汇总了两个模块在不同硬件平台上的运行时间、计算量 (TFLOPs) 及对应的云端成本. 可以看出, 样本

https://www.sciengine.com/doi/10.1360/SSI-2025-0474

生成与筛选模块 (模块 (a)) 的计算开销较低. 该模块仅使用预训练扩散模型进行前向推理, 无需额外训练, 同时可复用预先计算的文本嵌入, 因此整体运行时间均低于 0.12 小时, 对应计算量不足 1 TFLOPs, 云端成本可忽略. 主要计算开销集中在对抗样本训练模块 (模块 (b)). 在 RTX 4060 (8GB) 上, 该模块运行约 4.3 小时, 在 UCloud 3080Ti (12GB) 上约 3.2 小时, 总计算量约 34.1 TFLOPs, 对应云端成本约 0.05 美元. 综合两个模块, 单次完整攻击流程的总体成本低于 0.06 美元, 证明了本文方法在实际部署场景下具有较低的运行成本.

表 3 样本生成与筛选模块 (模块 (a)) 与对抗样本训练模块 (模块 (b)) 的运行时间、计算量和云端成本, 数据集为 CIFAR-100.

Table 3 Runtime, floating-point operations per second, and cost of the sample generation and selection module (module (a)) and the training module (module (b)) on CIFAR-100.

Module	Ucloud 1080Ti (12GB) runtime (hours)	RTX 4060 (8GB) runtime (hours)	Floating-point operations per second (TFLOPs)	Cost (dollars)
(a)	0.083	0.112	0.888	0.001
(b)	3.200	4.300	34.100	0.050
Total	3.283	4.412	34.988	0.051

4.5 消融实验

本小节对实验中关键设置对性能的影响进行探讨, 包括: 1) CIL 中不同类别增量设置下的攻击性能; 2) 攻击时不同筛选比例参数 r 设置下的攻击性能.

类别增量. 改变 CIL 过程中初始训练和更新的类别数, 测试并记录 ASR, 结果如图 9 所示, Init 代表初始训练的类别数, Inc 表示增量学习的类别数, 测试数据集为 CIFAR-100. 可以看出, 本文方法在大部分 CIL 设置下均能保持较高的攻击性能, 随着类别数增加, 方法的 ASR 在大多数情况下高于 40%, 并且在类别数在 70 以内时, 平均 ASR 超过 50%. 尽管随着类别增加, ASR 存在下降, 但下降趋势较稳定且缓慢. 此外, 与图 4中的结论类似, 重放 Replay 与基于蒸馏的 iCaRL 方法上的 ASR 略低于其他 CIL 设置, 这归因于模型在这两种方法上略低的精度.

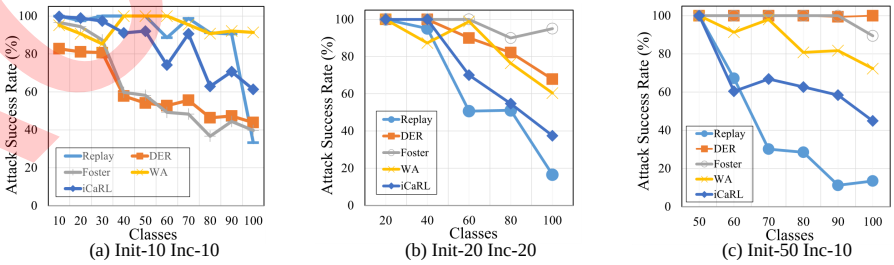


图 9 (网络版彩图) CIFAR-100 数据集下不同 CIL 设置的 ASR 曲线: (a) 共 10 个任务, 每个任务 10 个类; (b) 共 5 个任务, 每个任务 20 个类; (c) 共 6 个任务, 初始任务 50 个类, 增量任务每个 10 个类.

Figure 9 (Online Color) ASR curves under different CIL Settings on CIFAR-100: (a) 10 tasks with 10 new classes each, (b) 5 tasks with 20 new classes each, (c) 6 tasks with 50 classes in the initial task and 10 classes per incremental task.

筛选比例参数. 筛选比例参数 r 是筛选强目标类语义的关键, 为探究该参数对攻击性能的影响, 测试 $r \in (0, 1.0]$ 时, 不同 CIL 设置上的 SASR, 并将结果记录于表 4, 其中实验其余设置与表 1中保持一致, 注意, 数据集分为 10 个任务, 每个任务 10 个类. 可以看出, 当 r 越小时, ASR 越高. 即样本的预测熵越低, 代表初始模型对所生成目标类样本的置信度越高, 反映了所筛选出的样本具有越强的目标类语义. 当 $r = 1.0$, 代表不做任何语义筛选, 该设置下的引导样本 X_g 可能存在较多语义混淆的样本, 使攻击效果下降. 随着 r 值增大, 方法在各类

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0474

表 4 CIFAR-100 数据集下, 筛选比例参数 r 不同时, 攻击不同 CIL 模型的 SASR.
Table 4 Selection ratio parameter r vs. SASR under different CIL methods. Dataset: CIFAR-100.

CIL	$r = 0.05$	$r = 0.1$	$r = 0.2$	$r = 0.4$	$r = 0.6$	$r = 1.0$
Replay	88.13	89.97	87.04	84.40	78.06	61.97
DER	57.79	60.25	58.13	53.49	50.39	38.22
Foster	60.90	61.48	60.24	58.54	51.45	40.77
WA	94.17	94.07	93.72	91.24	85.18	76.04
iCaRL	82.19	83.89	81.87	78.95	72.73	62.10

CIL 设置上的 ASR 均缓慢下降. 与最优设置 $r = 0.1$ 相比, 当 $r = 1.0$ 时, SASR 在各类 CIL 上平均下降了 22.11%. 而当 r 减小至 0.05 时, X_g 样本显著较少, 无法覆盖目标类的特征空间, 导致用于生成模型 $\Phi_g(\cdot)$ 训练的引导样本较少, 攻击效果有十分微小 (仅 1.30%). 综上, 本文建议设置较小的 r 值, 并设置 $r = 0.1$ 为默认值.

类别增量顺序. 本小节继续探究当类别增量顺序变更时的攻击性能. 固定 CIFAR-100 数据集的 10 个初始类别, 随机化 90 个增量类别以测试本文对抗样本在 10 个初始类上的 SASR, 结果见表 5. 其中, 噪声采用默认值 16/255, 每次增量任务中的类别个数为 10. 根据表中结果, 变更在不同 CIL 方法的增量顺序后 (Shuffled), 本文方法的 SASR 均与原顺序结果 (Original) 保持在相近水平.SASR 的绝对差值分别为 5.13% (Replay)、6.78% (DER)、2.53% (PodNet)、2.04% (WA) 以及 3.73% (iCaRL), 平均波动幅度约为 4.0%. 因此, 尽管不同 CIL 方法在类别增量顺序变更后会表现出一定的性能起伏, 但整体波动幅度较小, 证明了本文攻击方法对类别增量训练顺序不敏感.

表 5 CIFAR-100 数据集下, 增量类别训练顺序变更前 (Original) 与变更后 (Shuffled) 的 SASR.
Table 5 Comparison of SASR under original and shuffled class-incremental training orders on CIFAR-100.

CIL Method	Replay	DER	PodNet	WA	iCaRL
Original	50.82	60.06	82.91	81.37	57.69
Shuffled	45.69	66.84	85.44	79.33	61.42

扩展数据集性能. 为了评估方法在不同数据分布下的泛化能力, 本小节将实验扩展至 CUB-200 和 Food-101 两个数据集. 与相关研究 [45] 保持一致,CUB-200 采用 100 个初始类别并进行 10 轮、每轮 10 类的增量学习; Food-101 则使用 11 个初始类别并以每轮 10 类进行增量扩展. 两种数据集均采用 ResNet-18 作为增量学习模型, 并在 FT、RP、DER、iCaRL 和 BiC 等典型 CIL 方法下进行评估. 表 6 给出了不同方法下的模型性能及对应攻击成功率. 在类别分布较均衡的 CUB-200 上,CIL 方法的平均准确率较高, 部分接近 50%, 此时本文方法在不同 CIL 算法上均获得较高 (平均 33.42%) 的平均攻击成功率, 整体与模型精度保持一致变化趋势.

表 6 不同类增量学习方法在 CUB-200 与 Food-101 数据集上的性能对比.
Table 6 Comparison of performance of different CIL methods on CUB-200 and Food-101.

Indicator	CUB-200 (Init-100 Inc-10)					Food-101 (Init-11 Inc-10)				
	FT	**RP**	**DER**	**iCaRL**	**BiC**	**FT**	**RP**	**DER**	**iCaRL**	**BiC**
Clean Acc	12.98	28.09	50.90	25.44	49.70	5.86	11.63	18.70	12.46	12.39
SASR	17.98	22.47	57.52	37.89	34.15	7.96	17.66	20.04	15.43	9.20

相比之下,Food-101 由于具有明显的长尾分布特性, 其增量模型整体性能显著下降, 所有 CIL 方法的平均精度低于 20%. 尽管如此, 本文方法仍能够在多种算法上维持稳定的攻击效果, 其 SASR 仍接近对应模型精度

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0474

水平. 综上, 当 CIL 模型分类精度较高时, 本方法能够实现接近模型精度水平的攻击成功率; 而在 CIL 中更具挑战性的长尾数据分布下, 仍然具备一定攻击有效性, 证明所提出方法在不同数据集上均具有良好的泛化能力.

5 结论

本文面向类增量学习场景中对抗样本难以持续的问题, 提出了一种基于语义强化的有目标对抗样本生成方法, 具体地, 设计了基于预测熵的强语义目标类样本筛选、基于浅层特征的目标类语义提取和对抗样本生成方法, 并通过系统实验验证了所提方法在对抗样本持续性方面的有效性. 在性能评估方面, 本文在 9 种典型的 CIL 设置上对所提方法进行了全面实验. 结果表明, 当类别规模扩展至原来的 10 倍时, 本方法的平均攻击成功率比现有基准方法提升了 39.38%, 显著增强了对抗样本在类增量学习中的可持续性. 此外, 对不同攻击目标、增量学习设置及关键参数的实验结果进一步证明了本文方法在隐蔽性和鲁棒性方面的优势. 尽管本文方法在经典图像分类模型上取得了良好效果, 但针对其他类型模型的攻击效果仍有待深入探索. 例如, 未来可进一步研究如何在 Transformer 等大型模型的增量学习机制下, 优化样本语义增强算法, 在保证攻击隐蔽性的同时, 实现良好的攻击性能.

参考文献

1 Wang L, Zhang X, Su H, et al. A comprehensive survey of continual learning: Theory, method and application. In: Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2024. 5362-5383

2 Zhou D W, Wang Q W, Qi Z H, et al. Class-incremental learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2024

3 Dong Y, Liao F, Pang T, et al. Boosting adversarial attacks with momentum. In: Proceedings of Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 9185–9193

4 Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014

5 Masana M, Liu X, Twardowski B, et al. Class-incremental learning: survey and performance evaluation on image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2022, 45: 5513–5533

6 Li L, Tan Y, Yang S, et al. Adaptive decision boundary for few-shot class-incremental learning. In: Proceedings of Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2025. 18359–18367

7 Pan Q, Wang L, Zhang J, et al. Lunch: Adaptive balancing of continual learning via hyperparameter uncertainty. Science China Information Sciences, 2026, 69

8 Rebuffi S A, Kolesnikov A, Sperl G, et al. icarl: Incremental classifier and representation learning. In: Proceedings of Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), 2017. 2001–2010

9 Sun K, Li M, Lin Z. Pareto adversarial robustness: Balancing spatial robustness and sensitivity-based robustness. Science China Information Sciences, 2025, 68

10 Byun J, Kwon M J, Cho S, et al. Introducing competition to boost the transferability of targeted adversarial examples through clean feature mixup. In: Proceedings of Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023. 24648–24657

11 Ge Y, Wang Q, Huang H, et al. Hijacking attacks against neural network by analyzing training data. In: Proceedings of 33rd USENIX Security Symposium (USENIX Security 24), 2024. 6867–6884

12 Xu B, Dai X, Tang D, et al. One surrogate to fool them all: Universal, transferable, and targeted adversarial attacks with clip. In: Proceedings of Proceedings of The ACM Conference on Computer and Communications Security (CCS), 2025. 1-18

13 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of International conference on machine learning (PMLR), 2021. 8748–8763

14 Fang H, Kong J, Chen B, et al. Clip-guided generative networks for transferable targeted adversarial attacks. In: Proceedings of European Conference on Computer Vision (ECCV), 2024. 1-19

15 Wang Z, Yang H, Feng Y, et al. Towards transferable targeted adversarial examples. In: Proceedings of Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), 2023. 20534–20543

16 Sun Y, Yuan S, Wang X, et al. Any target can be offense: Adversarial example generation via generalized latent infection. In: Proceedings of European Conference on Computer Vision (ECCV), 2024. 383–398

17 Guo Q, Pang S, Jia X, et al. Efficient generation of targeted and transferable adversarial examples for vision-language models via diffusion models. IEEE Transactions on Information Forensics and Security (TIFS), 2024

18 Li T, Ma X, Jiang Y G. Aim: Additional image guided generation of transferable adversarial attacks. In: Proceedings of Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2025. 4941–4949

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0474

19 Ratcliff R. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. Psychological review, 1990, 97: 285
20 Saharia C, Chan W, Chang H, et al. Palette: Image-to-image diffusion models. In: Proceedings of ACM SIGGRAPH 2022 Conference Proceedings, 2022. 1-10
21 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Advances in neural information processing systems (NeurIPS), 2020, 33: 6840–6851
22 Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models. In: Proceedings of Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), 2022. 10684–10695
23 Jagielski M, Thakkar O, Tramèr F, et al. Measuring forgetting of memorized training examples. In: Proceedings of The Eleventh International Conference on Learning Representations (ICLR) poster, 2023. 1-22
24 Yang H, Jeong J, Yoon K J. Facl-attack: frequency-aware contrastive learning for transferable adversarial attacks. In: Proceedings of AAAI Conference on Artificial Intelligence (AAAI), 2024. 6494-6502
25 Wang Y, Wu Y, Wu S, et al. Boosting the transferability of adversarial attacks with frequency-aware perturbation. IEEE Transactions on Information Forensics and Security (TIFS), 2024, 19: 6293–6304
26 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of 3rd International Conference on Learning Representations (ICLR), 2014. 1-14
27 Yosinski J, Clune J, Bengio Y, et al. How transferable are features in deep neural networks? In: Proceedings of Proceedings of the 28th International Conference on Neural Information Processing Systems (NeurIPS), 2014. 3320–3328
28 Buzzega P, Boschini M, Porrello A, et al. Dark experience for general continual learning: a strong, simple baseline. Advances in neural information processing systems (NeurIPS), 2020, 33: 15920–15930
29 Saadabadi M S E, Dabouei A, Malakshan S R, et al. Hyperspherical classification with dynamic label-to-prototype assignment. In: Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. 17333-17342
30 Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2021, 43: 4217–4228
31 Park T, Liu M Y, Wang T C, et al. Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 1-10
32 Krizhevsky A, Hinton G, et al. Learning multiple layers of features from tiny images., 2009
33 Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database. In: Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009. 248-255
34 Van de Ven G M, Tuytelaars T, Tolias A S. Three types of incremental learning. Nature Machine Intelligence, 2022, 4: 1185–1197
35 Douillard A, Cord M, Ollion C, et al. Podnet: Pooled outputs distillation for small-tasks incremental learning. In: Proceedings of 16th European Conference on Computer Vision (ECCV), 2020. 86–102
36 Wang F Y, Zhou D W, Ye H J, et al. Foster: Feature boosting and compression for class-incremental learning. In: Proceedings of 18th European conference on computer vision (ECCV), 2022. 398–414
37 Wu Y, Chen Y, Wang L, et al. Large scale incremental learning. In: Proceedings of Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019. 374–382
38 Zhao B, Xiao X, Gan G, et al. Maintaining discrimination and fairness in class incremental learning. In: Proceedings of Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020. 13208–13217
39 Gu J, Jia X, de Jorge P, et al. A survey on transferability of adversarial examples across deep neural networks. Transactions on Machine Learning Research, 2024
40 Diederik K. Adam: A method for stochastic optimization. 3rd International Conference for Learning Representations (ICLR), 2014
41 Naseer M M, Khan S H, Khan M H, et al. Cross-domain transferability of adversarial perturbations. Advances in Neural Information Processing Systems (NeurIPS), 2019, 32
42 Naseer M, Khan S, Hayat M, et al. On generating transferable targeted perturbations. In: Proceedings of Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 7708–7717
43 Chen P, Yang J, Lin J, et al. A practical clean-label backdoor attack with limited information in vertical federated learning. In: Proceedings of 2023 IEEE International Conference on Data Mining (ICDM), 2023. 41–50
44 Tian Q, Kuang K, Jiang K, et al. Analysis and applications of class-wise robustness in adversarial training. In: Proceedings of Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021. 1561–1570
45 Zhou D W, Ye H J, Ma L, et al. Few-shot class-incremental learning by sampling multi-phase tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45: 12816–12831

Sustainable Adversarial Attacks Based on Semantic Enhancement in Class-Incremental Learning

https://www.sciengine.com/doi/10.1360/SSI-2025-0474

Accepted

Xinjing Liu[1], Zhuzhu Wang[2*], Xianjia Meng[2*], Taifeng Liu[1], Zhuoran Ma[1], Yilong Yang[3] & Zhuo Ma[1]

1. *School of Cyber Engineering, Xidian University, Xi'an 710126, China*
2. *School of Information Science & Technology, Northwestern University, Xi'an 710127, China*
3. *School of Artificial Intelligence, Xidian University, Xi'an 710126, China*
* Corresponding author. E-mail: z.z.wang@foxmail.com, xianjiam@nwu.edu.cn

Abstract Many deployed neural networks are not trained once on a static corpus but must continually adapt to newly arriving data. Class-Incremental Learning (CIL) updates a model stage by stage to mitigate catastrophic forgetting; however, as new classes are introduced, the model' s representation space undergoes substantial drift, causing adversarial perturbations —especially targeted ones—to degrade rapidly. Existing approaches rarely maintain stable target semantics under such domain shift. To address this, we propose a semantics-reinforced targeted adversarial example generation method for CIL. The core idea is to explicitly strengthen the target-class semantics of the perturbation and thereby improve robustness across increments. The framework proceeds in three steps: (i) Candidate generation and semantic screening: a diffusion model synthesizes target-class candidates, which are then filtered by the initial model' s predictive entropy to retain samples with strong target semantics; (ii) Shallow-semantic feature extraction: leveraging the relative stability of shallow features during CIL, we extract target-class shallow semantics and inject them into the generation process; (iii) Semantic-constraint optimization: a set of semantic constraints guides the adversarial examples to be close to the target class and far from non-target classes in feature space, improving robustness against representation drift. Comprehensive experiments demonstrate that, when the number of classes increases by a factor of ten, the proposed method improves the sustainable attack success rate by 39.38%.

Keywords adversarial attack, incremental learning, targeted adversarial attack, robustness, semantic enhancement

Accepted

https://www.sciengine.com/doi/10.1360/SSI-2025-0474