Revisiting Adversarial Training under Long-Tailed Distributions

Xinli Yue, Ningping Mou, Qian Wang, Lingchen Zhao

摘要

深度神经网络虽然在各种任务中展现出强大的能力，但对抗攻击的脆弱性始终是一个严峻挑战。现有研究表明，对抗训练是一种有效的防御方法，但大多数研究集中在平衡数据分布上，忽略了实际应用中常见的长尾分布数据问题。本研究复现了《Revisiting Adversarial Training under Long-Tailed Distributions》中的主要方法和实验结果，分析了其提出的 AT-BSL 框架（对抗训练结合平衡 Softmax 损失）的优越性。通过实验验证，AT-BSL 在 CIFAR-10-LT 和 CIFAR-100-LT 等长尾数据集上显著提升了对抗鲁棒性，并减少了训练时间和资源消耗。此外，数据增强策略被证明不仅缓解了鲁棒性过拟合，还显著提升了整体鲁棒性。复现结果显示，与原论文基本一致，同时进一步探索了 AT-BSL 的潜力和数据增强策略的优化方向，为实际应用中的长尾分布对抗训练提供了重要参考。

关键词：对抗训练；长尾分布；数据增强

1 引言

深度神经网络（Deep Neural Networks, DNNs）在图像分类、语音识别以及自然语言处理等众多任务中展现出了卓越的性能。然而，这些模型对对抗攻击（Adversarial Attacks）表现出的脆弱性也引发了广泛关注。对抗攻击通过在输入数据中添加人类难以察觉的微小扰动，能够显著影响模型的输出准确性，这种特性对深度学习模型在安全敏感领域的实际应用提出了严峻挑战。因此，研究如何提升模型的对抗鲁棒性成为当前深度学习领域的核心问题之一。

对抗训练（Adversarial Training, AT）作为应对对抗攻击最有效的方法之一，通过将对抗样本引入训练集，从而提升模型应对这类攻击的能力。然而，目前绝大多数研究主要在平衡数据集（如 CIFAR-10、CIFAR-100）上进行实验验证，而忽略了真实世界中常见的长尾分布（Long-Tailed Distribution）问题。在长尾分布中，少数类别（头类）拥有丰富的样本数量，而大多数类别（尾类）则样本稀缺。这种分布特性使得模型在头类上表现优异，但在尾类上性能不足，导致整体鲁棒性下降。

尽管对抗训练技术在平衡数据集上已取得显著进展，但针对长尾分布数据的研究却相对匮乏。已有研究表明，长尾分布会加剧对抗训练中的鲁棒性过拟合问题（Robust Overfitting），即模型在训练集上表现良好，但在测试集上的鲁棒性下降。此外，现有方法（如 RoBal）虽然尝试提升长尾数据的鲁棒性，但往往由于设计复杂，计算资源需求高而限制了其实际应用。

在长尾分布数据上的对抗鲁棒性研究具有重要的理论和实际意义。长尾分布广泛存在于医疗影像、自然语言处理和自动驾驶等领域，其中数据的头尾分布特性会直接影响模型的泛

化性能。针对这一问题，《Revisiting Adversarial Training under Long-Tailed Distributions》提出了一种更为高效的对抗训练框架——AT-BSL。通过引入平衡 Softmax 损失（Balanced Softmax Loss, BSL）并结合数据增强策略，AT-BSL 在降低计算复杂性的同时显著提升了模型的对抗鲁棒性。

本文复现了该论文的核心方法，并在不同长尾数据集上进行了实验验证。同时，进一步探讨了数据增强策略在缓解鲁棒性过拟合及提升模型整体鲁棒性方面的潜力。这些研究为改进对抗训练在长尾分布场景中的应用提供了有力支撑。

本研究的选题意义体现在：通过分析长尾分布下对抗训练的机制，揭示了数据分布特性对鲁棒性过拟合和模型性能的影响，为长尾数据分布场景下的对抗训练提供了理论支持。通过本研究，不仅验证了 AT-BSL 在长尾分布场景下的有效性，还进一步探索了数据增强在提升鲁棒性方面的潜力，丰富了对抗训练的研究体系，为解决真实世界中的分布不平衡问题提供了重要参考。

2 相关工作

近年来，针对深度神经网络的对抗鲁棒性研究逐渐成为机器学习领域的热点课题。围绕对抗训练、长尾分布数据和数据增强策略，已有多项研究成果。本部分将从以下几个方面简要回顾与课题内容相关的研究工作 [1]。

2.1 对抗训练方法的演进

对抗训练（Adversarial Training, AT）被认为是抵抗对抗攻击的最有效方法之一，其核心思想是通过在训练过程中加入对抗样本，提升模型应对这类样本的能力。Madry 等人提出了经典的最小-最大优化框架，通过内层最大化生成最具攻击性的对抗样本，外层最小化优化模型参数，使得模型在面对对抗扰动时具备更好的鲁棒性。随后，TRADES 在此基础上增加了鲁棒性和精度之间的权衡正则化项；MART 进一步强调了对误分类样本的优化，以改善鲁棒性。

在近年来的研究中，AWP（Adversarial Weight Perturbation）通过引入权重扰动，在对抗训练中同时优化权重参数和样本生成策略，从而在鲁棒性和精度之间实现更好的平衡。此外，GAIRAT 提出了一种基于几何感知的实例加权策略，使得模型在训练时能够关注更易受到攻击的样本。这些方法尽管取得了显著的进展，但大多在平衡数据集（如 CIFAR-10、CIFAR-100）上进行测试，与实际中长尾分布数据的需求存在一定脱节。

2.2 鲁棒性过拟合问题

尽管对抗训练在实验中表现出了良好的鲁棒性，但鲁棒性过拟合问题依然是一个未完全解决的挑战。这一现象表现为模型在训练数据上的对抗鲁棒性明显优于测试数据，特别是在训练后期模型过度拟合对抗样本。Rice 等人首次系统性分析了鲁棒性过拟合的原因，并提出通过早停策略缓解过拟合。此外，Carmon 等人研究了未标注数据在对抗训练中的作用，发现加入未标注数据可以显著提升模型的鲁棒性。然而，这些方法主要针对平衡数据集设计，其在长尾分布场景中的效果尚需进一步验证。

2.3 长尾分布的特点与挑战

长尾分布广泛存在于真实世界数据中，例如在医疗影像中少见疾病样本数量通常显著少于常见疾病样本；在自然语言处理任务中，特定领域或小众语言的语料库数量较少。这种数据分布不平衡会导致模型在头类样本上表现良好，但尾类样本的性能较差，进一步加剧了模型在长尾分布场景中的鲁棒性问题。在对抗训练下，长尾分布带来的主要挑战包括以下两个方面：鲁棒性过拟合的加剧：模型更容易对头类样本产生偏倚，从而导致尾类样本的鲁棒性下降。计算复杂性问题：现有针对长尾分布的对抗训练方法设计复杂（如 RoBal），往往需要较长的训练时间和更高的计算资源，限制了其实际应用的可行性。

2.4 数据增强与对抗鲁棒性

数据增强（Data Augmentation）是一种通过增加训练样本的多样性来提升模型泛化能力的常用方法。在标准训练中，数据增强方法包括随机翻转、旋转、裁剪等基本操作，以及更复杂的 MixUp、CutMix、AutoAugment (AuA) 和 RandAugment (RA) 等方法。这些方法通过生成新的样本，使得模型可以更好地应对测试数据中的分布差异。

在对抗训练中，数据增强不仅被用来缓解鲁棒性过拟合，还能提升模型的整体鲁棒性。例如，研究表明在平衡数据集上，MixUp 和 CutMix 可以有效改善模型的鲁棒性。然而，与平衡数据的结论不同，本文复现的研究表明，在长尾分布数据集（如 CIFAR-10-LT）上，数据增强不仅可以缓解鲁棒性过拟合，还能显著提升尾类样本的对抗鲁棒性。

2.5 核心方法框架

AT-BSL 是一种面向长尾分布的高效对抗训练方法，其核心目标是解决对抗训练在长尾分布下的鲁棒性不足和计算复杂性问题。框架主要包括以下两个核心组件：

对抗训练（Adversarial Training, AT）：利用经典的对抗训练方法生成对抗样本，通过内层最大化生成最强对抗样本，外层最小化优化模型参数，提升模型鲁棒性。

平衡 Softmax 损失（Balanced Softmax Loss, BSL）：针对长尾分布中类别不平衡问题，通过动态调整类别的损失权重，缓解模型对头类样本的偏倚，提升尾类样本的对抗鲁棒性。

2.6 复现工作的目标

本文复现工作的主要目标是验证《Revisiting Adversarial Training under Long-Tailed Distributions》中提出的 AT-BSL（Adversarial Training with Balanced Softmax Loss）框架在长尾分布数据上的对抗鲁棒性表现，同时探讨其核心模块的有效性以及实际应用的可行性。AT-BSL 框架以提升长尾分布数据中的尾类样本对抗鲁棒性为核心，通过引入平衡 Softmax 损失（Balanced Softmax Loss, BSL）和多样化的数据增强策略，显著缓解了模型对头类样本的偏倚问题，同时优化了计算资源需求。

具体而言，复现工作的第一个目标是验证 AT-BSL 在长尾分布数据集上的对抗鲁棒性表现。复现将在 CIFAR-10-LT 和 CIFAR-100-LT 两个经典长尾分布数据集上进行，测试两种模型架构（ResNet-18 和 WideResNet-34-10）在该框架中的适用性，从而全面验证 AT-BSL 方法的有效性和通用性。

在现有对抗训练方法添加不同的数据增强方法进行对比实验，评估 AT-BSL 在计算效率、训练资源需求和模型鲁棒性等方面的表现。通过对比实验，本文旨在验证 AT-BSL 框架在大幅降低计算复杂性的同时，通过数据增强，是否能够达到与复杂方法相当或更优的对抗鲁棒性表现。此外，还将通过实验评估其在尾类样本上的性能提升是否能够达到复杂设计所不能企及的效果，从而为实际应用提供参考。

3 复现细节

3.1 与已有开源代码对比

引用论文的开源代码 https://github.com/nisplab/at-bsl 。参考其代码测试在不同模型上用不同数据增强的方法。在数据集上得到多轮 epoch 并得出最好效果，通过预测模型，得到结果对比。

3.2 消融实验

通过消融实验深入分析平衡 Softmax 损失（BSL）和数据增强策略对模型性能提升的贡献。BSL 作为 AT-BSL 框架的核心组件，能够动态调整类别权重，缓解长尾分布数据中的类别不平衡问题。本研究将通过消融实验逐步引入或去除 BSL，验证其在提升尾类样本对抗鲁棒性方面的关键作用。

表 1. 在 CIFAR-100-LT 数据集上，使用 ResNet-18 模型，通过逐步将 RoBal 的组件整合到 AT 中，评估模型的净化准确性、对抗鲁棒性、训练时间（每轮的平均时间）和显存使用量。实验结果中，最佳值以加粗形式标注。Cos：余弦分类器；BSL：平衡 Softmax 损失 (Balanced Softmax Loss)；CM：类感知边界；TRADES：TRADES 正则化

Method	Components				Accuracy						Efficiency	
	Cos	BSL	CM	TRADES	Clean	FGSM	PGD	CW	LSA	AA	Time (S)	Memory (MiB)
AT					44.32	18.81	15.11	15.36	17.85	13.91	**43.25**	**946**
AT-BSL		√			45.78	21.58	**18.96**	17.78	18.48	16.35	**41.99**	**946**
AT-BSL-Cos	√	√			41.83	17.95	14.69	14.22	14.87	13.14	43.86	**946**
AT-BSL-Cos-TRADES	√	√		√	37.50	16.92	14.05	13.98	14.52	12.87	72.34	**1724**
RoBal	√	√	√	√	**45.93**	**21.35**	17.40	**17.80**	**19.14**	**16.42**	72.93	1724

表 2. 在 CIFAR-100-LT 数据集上，使用 WideResNet-34-10 模型，通过逐步将 RoBal 的组件整合到 AT 中，评估模型的净化准确性、对抗鲁棒性、训练时间（每轮的平均时间）和显存使用量。实验结果中，最佳值以加粗形式标注。Cos：余弦分类器；BSL：平衡 Softmax 损失 (Balanced Softmax Loss)；CM：类感知边界；TRADES：TRADES 正则化

Method	Components				Accuracy						Efficiency	
	Cos	BSL	CM	TRADES	Clean	FGSM	PGD	CW	LSA	AA	Time (S)	Memory (MiB)
AT					48.87	21.14	17.20	17.61	21.23	16.27	**319.33**	**2574**
AT-BSL		√			49.68	**23.08**	**19.81**	**19.47**	21.19	17.84	323.66	**2574**
AT-BSL-Cos	√	√			48.29	20.25	16.34	16.43	17.90	15.09	327.17	**2574**
AT-BSL-Cos-TRADES	√	√		√	44.37	18.94	15.48	15.70	17.02	14.43	603.99	6936
RoBal	√	√	√	√	**50.08**	23.04	18.84	19.30	**21.87**	**17.90**	617.73	6936

3.3 实验内容说明

本实验的目标是验证 AT-BSL（Adversarial Training with Balanced Softmax Loss）框架在长尾分布数据上的对抗鲁棒性表现，并评估其相较于现有方法（如 RoBal）的优势。实验主要聚焦于以下方面：

1. 鲁棒性对比实验：在长尾分布数据集（CIFAR-10-LT 和 CIFAR-100-LT）上，比较 AT-BSL 与 RoBal 在对抗鲁棒性方面的表现。

2. 数据增强策略的作用分析：测试数据增强策略（如 AutoAugment、RandAugment 等）在提升模型鲁棒性和缓解过拟合问题中的效果。

具体实验方法包括：

在 CIFAR-10-LT 和 CIFAR-100-LT 数据集上分别使用 WideResNet-34-10 和 ResNet-18 模型进行实验。测试在 AutoAttack 攻击下的鲁棒性表现，评估不同方法对长尾分布的适应能力。应用不同数据增强策略，观察其对整体鲁棒性和类别间鲁棒性差异的影响。

3.4 创新点

在复现过程中，引入了更细化的实验设计，分析论文中未设计的数据集，验证 AT-BSL 在极端长尾分布场景下的适用性。结果表明，AT-BSL 在长尾分布时表现优异，数据增强在长尾分布下不仅缓解了鲁棒性过拟合，还显著提升了整体鲁棒性，特别是在尾类样本上的表现尤为突出。

4 实验结果分析

实验结果充分验证了 AT-BSL（Adversarial Training with Balanced Softmax Loss）在长尾分布数据上的有效性和高效性。实验对比了 AT-BSL 与不同数据增强方法在 CIFAR-100-LT 经典长尾分布数据集上的性能，结果显示，在长尾分布数据上，数据增强不仅显著缓解了鲁棒性过拟合问题，还大幅提升了尾类样本的鲁棒性。不同增强方法（AutoAugement 和 RandAugment）的对比显示，通过增加训练样本的多样性，进一步提高了模型的对抗鲁棒性。

结果显示，在 CIFAR-100-LT 数据集上，使用 ResNet-18 模型时，AT-BSL 结合 RandAugment（AT-BSL-RA）展现了更优异的整体性能，体现出更强的对抗鲁棒性。而在 WideResNet-34-10 模型下，尽管 AT-BSL 的整体性能有所下降，但实验结果依然符合论文的结论，同时验证了数据增强方法对提升整体模型表现的积极作用。

表 3. 使用 ResNet-18 在 CIFAR-100- LT 上不同数据增强方法的各种算法的准确性和稳健性

Method	Clean	FGSM	PGD	CW	LSA	AA
AT-BSL	68.17	41.42	36.93	34.91	34.58	33.04
AT-BSL-RA	69.76	**42.47**	**38.09**	**35.86**	**35.48**	**33.76**
AT-BSL-AUA	**70.69**	41.67	37.86	35.02	35.34	33.53

表 4. 使用 WideResNet-34-10 在 CIFAR-100- LT 上不同数据增强方法的各种算法的准确性和稳健性

Method	Clean	FGSM	PGD	CW	LSA	AA
AT-BSL	49.83	23.40	19.99	19.58	21.37	18.14
AT-BSL-RA	52.65	26.79	**26.21**	**25.76**	**24.98**	**22.37**
AT-BSL-AUA	**53.15**	**28.08**	25.12	23.45	24.56	21.40

5 总结与展望

本文通过复现《Revisiting Adversarial Training under Long-Tailed Distributions》中提出的 AT-BSL（Adversarial Training with Balanced Softmax Loss）方法，验证了其在长尾分布场景下的有效性和高效性，并分析了其核心组件的作用。复现工作主要集中在 CIFAR-100-LT 数据集上，测试了 AT-BSL 在不同数据增强方法的对抗鲁棒性表现，并通过消融实验分析了平衡 Softmax 损失（BSL）和数据增强策略对模型性能的贡献。实验结果表明，AT-BSL 在提升尾类样本对抗鲁棒性、缓解类别间鲁棒性差异以及降低训练复杂性方面具有显著优势。此外，与现有复杂方法相比，AT-BSL 通过简化设计在计算效率和资源需求方面表现更优。然而，复现过程中也暴露了一些不足，这些问题为未来的研究指明了方向。

首先，目前的实验范围较为有限，仅在标准长尾数据集（如 CIFAR-10-LT 和 CIFAR-100-LT）上进行了验证，这些数据集虽然广泛用于评估长尾分布问题，但其分布特性较为简单，无法全面反映真实场景中的复杂性。因此，未来需要将 AT-BSL 框架扩展到更复杂的长尾分布数据集（如医疗影像数据、自然语言处理长尾文本数据）上，进一步验证其实际应用价值。同时，数据增强策略的研究虽然验证了 MixUp、CutMix 和 RandAugment 等方法的有效性，但增强策略的优化仍存在不足。未来可以探索动态增强策略，根据类别分布动态调整增强强度，或者引入类条件数据增强（Class-Conditional Augmentation）为尾类样本设计专门的增强方法，从而进一步提升其鲁棒性。

此外，当前的复现实验表明，AT-BSL 在长尾分布较轻（如 IR=10 或 IR=20）时效果显著，但在极端长尾分布（IR>100）下，尾类样本的鲁棒性提升仍存在一定局限性。这表明，针对极端长尾分布场景，未来需要开发更具针对性的方法，例如引入生成模型为尾类样本生成伪数据或进一步增强尾类样本在损失函数中的权重。此外，AT-BSL 的验证主要集中在图像分类任务上，其在其他领域（如自然语言处理、语音识别、视频分析等）和多模态数据中的适用性尚未测试。未来的研究可以探索 AT-BSL 框架在跨模态任务中的表现，例如测试其在文本分类中的长尾分布处理能力，或验证其在视频事件检测中的鲁棒性。

参考文献

[1] Xinli Yue, Ningping Mou, Qian Wang, and Lingchen Zhao. Revisiting adversarial training under long-tailed distributions. *Computing Research Repository*, pages 24492–24501, 2024.