中国科学 : 信息科学 2025 年 第 55 卷 第 8 期: 1986–2001
SCIENTIA SINICA Informationis
论文

基于贝叶斯能量对抗后训练的黑盒对抗防御方法

刁云峰[1], 姜凯超[1], 郭丹[1,2*], 梁振宇[3,4*], 时增林[1], 钱振兴[5], 汪萌[1,2]

1. 合肥工业大学计算机与信息学院, 合肥 230009
2. 合肥综合性国家科学中心人工智能研究院, 合肥 230094
3. 国防科技大学电子对抗学院, 合肥 230038
4. 合肥综合性国家科学中心信息安全研究中心, 合肥 230031
5. 复旦大学计算机科学技术学院, 上海 200433
* 通信作者. E-mail: guodan@hfut.edu.cn, liangzy21@nudt.edu.cn

收稿日期: 2024–11–02; 修回日期: 2025–01–26; 接受日期: 2025–04–01; 网络出版日期: 2025–08–12

国家自然科学基金 (批准号: 62302139, 62272144, 72188101, 62020106007)、安徽省科技重大专项 (批准号: 202203a05020011)、安徽省杰出青年科学基金 (批准号: 2408085J040)、 中央高校基本科研业务费专项资金 (批准号: JZ2023HGTA0202, JZ2023HGQA0101)、国家级大学生创新训练项目 (批准号: 202410359032) 和合肥综合性国家科学中心资助项目

摘要 深度神经网络在视觉分类任务上表现出卓越的性能, 但其安全性也面临着重大挑战, 特别是分类器的输出结果容易受到对抗攻击的恶意操纵. 为应对此问题, 对抗训练作为一种有效的防御机制得到了快速发展. 然而, 现有对抗训练方法大多依赖于白盒防御策略, 即需要访问模型的结构参数并对模型进行重新训练, 这在许多实际应用场景中并不切实际, 尤其是对于大规模预训练模型的鲁棒性增强. 此外, 重新训练模型在提升鲁棒性的同时往往会以牺牲模型精度为代价, 使得这些模型在正常分类任务及其下游任务中难以胜任. 为了解决上述问题, 本文提出了一种新的黑盒防御方法, 称为贝叶斯能量对抗后训练. 在数据层面, 该方法从能量的角度出发, 对对抗样本和干净样本的联合数据分布进行建模; 在模型层面, 则通过贝叶斯视角考虑附加模型参数的完整后验分布, 实现了对数据和模型的全贝叶斯对待. 作为一种后训练黑盒防御方法, 该方法通过冻结预训练模型并附加一个小规模的贝叶斯组件, 将原始模型转化为具有弹性恢复能力的鲁棒性模型, 而无需重新训练或访问原始模型参数. 大量的实验结果表明, 本文提出的黑盒防御方法能够在不降低原始模型精度的前提下, 有效抵御基于梯度的白盒和黑盒攻击, 其性能优于现有的白盒防御方法.

关键词 对抗样本, 深度学习, 对抗防御, 贝叶斯神经网络, 能量模型

1 引言

深度学习技术在计算机视觉领域, 特别是视觉分类和检测等方面, 得到了广泛的应用, 并深入渗透至国民经济的多个关键领域, 包括人脸识别与公共安全、自动驾驶、医疗诊断等. 尽管如此, 当前的深度学习技术仍被视为一种 “黑箱” 算法, 其可解释性和安全性存在显著缺陷, 使得在实际应用中难

引用格式: 刁云峰, 姜凯超, 郭丹, 等. 基于贝叶斯能量对抗后训练的黑盒对抗防御方法. 中国科学: 信息科学, 2025, 55: 1986–2001, doi: 10.1360/SSI-2024-0326
Diao Y F, Jiang K C, Guo D, et al. Post-train black-box defense through energy-based Bayesian adversarial training. Sci Sin Inform, 2025, 55: 1986–2001, doi: 10.1360/SSI-2024-0326

以确保其安全性和可控性. 特别是在医疗、自动驾驶、监控、金融等关乎人生命财产安全的关键领域, 当基于深度学习的算法在面临恶意攻击或遇到缺陷数据时, 其脆弱的决策能力可能对人们生命健康和财产安全构成严重威胁. 我国《人工智能安全治理框架》[1] 强调, 在大力发展人工智能的同时, 必须高度重视可能带来的安全风险和挑战. 习近平在中共中央政治局第九次集体学习时也强调, 要加强人工智能发展的潜在风险研判和防范, 维护人民利益和国家安全, 确保人工智能安全、可靠、可控 [2]. 由此可见, 研究安全、可靠、可控的深度学习技术已成为国家层面的重要战略布局.

对抗样本是指在原始输入样本中加入精心设计的微小扰动, 导致机器学习模型产生错误的输出 [3]. 基于深度学习的视觉分类模型因为极易受到对抗样本的威胁而导致其安全问题备受关注 [4]. 目前, 对抗样本已被证实可应用于各种视觉分类任务中. 当基于深度学习的视觉分类技术被应用于人脸识别与公共安全、医疗诊断、自动驾驶、视频监控等关乎人类生命与财产安全的场景时, 深度学习模型其脆弱的决策能力在遭受恶意攻击时将严重威胁人们的生命健康与财产安全. 因此, 探索对抗样本的防御机制并增强视觉分类器的对抗性鲁棒性, 具有非常重要的研究意义.

对抗训练 (adversarial training, AT) [5] 是目前最有效且广泛应用的对抗防御方法之一, 其原理是在模型训练过程中引入对抗样本, 以增强模型对潜在对抗攻击的鲁棒性. 然而, 对抗训练要求使用者完全掌握模型的结构和参数, 并对标准预训练模型进行重新训练, 因此通常被视为是一种白盒防御方法. 白盒防御虽然理论上可以提升模型的鲁棒性, 但在一些现实场景下并不适用. 首先, 出于商业或安全原因, 模型所有者可能不愿共享模型信息. 其次, 为应对实际应用中的复杂场景, 现实场景下的视觉分类器通常包含亿级乃至千亿级参数, 并在大规模数据集上训练. 鉴于对抗训练相较于标准训练需要更高的计算成本和时间消耗, 对如此庞大的模型在大规模数据集上重新进行对抗训练对开发者和终端用户而言均不切实际. 第三, 从终端用户的角度来看, 大型模型通常在大规模数据集上预训练后共享, 用户可直接对预训练模型进行微调以适应下游任务. 因此保留预训练模型的完整性, 不仅能避免重新训练带来的高昂成本, 还能防止模型在应用于下游任务时性能下降. 最后, 对抗训练在改善模型鲁棒性的同时, 通常会以降低对良性样本的分类准确率为代价 [6,7], 而标准训练的预训练模型对良性样本具有良好的特征表示能力. 因此, 保留预训练模型的完整性也有助于缓解模型鲁棒性与准确率之间的权衡问题. 鉴于此, 本文致力于开发一种无需了解模型架构或参数, 且无需重新训练的黑盒防御框架.

贝叶斯防御方法 [8] 通过学习模型的参数分布来集成不同的决策边界, 从而抵御不同种类的对抗攻击. 然而, 目前的贝叶斯防御方法存在两个方面的弊端. (1) 在数据层面, 尽管现有的贝叶斯防御方法可以对模型的参数分布进行建模, 但难以估计完整的对抗样本分布, 忽略了数据流形建模对模型精度和鲁棒性的影响. (2) 在模型层面, 尽管贝叶斯防御方法在理论上可以学习模型的参数分布和决策边界分布, 但是由于现实场景中的分类器往往包含着数十亿甚至百亿千亿规模的模型参数量, 贝叶斯神经网络难以在如此高维的空间下直接采样模型的参数分布. 针对上述问题, 本文创新性地提出了一种基于后训练策略的黑盒对抗防御方法, 称为贝叶斯能量对抗后训练. 贝叶斯能量对抗后训练通过对干净样本分布、对抗样本分布和模型的参数分布进行联合概率分布建模, 实现了对于数据和模型的全贝叶斯对待. 该方法采用后训练贝叶斯策略进行优化, 在冻结的预训练模型后面附加一个小规模的贝叶斯模型单元, 并仅对该单元进行鲁棒优化. 这一设计不仅保留了原始模型的完整性, 还大幅降低了资源开销. 贝叶斯能量对抗后训练的整体框架如图 1 所示, 其主要分为基于能量的分布建模以及贝叶斯边界修正两部分. 对于基于能量的分布建模部分, 与传统对抗训练方法将对抗样本视为 “点估计” 问题 (图 1(a)) 不同, 本文方法从能量视角出发, 将判别式分类器解释为生成式鲁棒性分类器, 因此可以利用能量函数参数化建模数据分布和对抗样本的联合概率分布 $p(\tilde{\mathbf{x}}, \mathbf{x}, y)$, 实现了干净样本与对抗样本一对多的映射关系. 同时, 通过在优化过程中分配给对抗样本低能量值, 使其落入干净样本附近的高密度概率区域, 从而有效缓解了模型分类精度与鲁棒性之间的权衡问题 (图 1(d)). 对于贝叶斯边界修正部分, 不同的鲁棒性分类器在应对不同类型的对抗样本时, 通常表现出不同的鲁棒性. 因此, 考虑分类器的后验概率分布 $p(\theta|\tilde{\mathbf{x}}, \mathbf{x}, y)$, 理论上可以涵盖所有可能的分类决策边界, 以提高模型的鲁棒性. 这使

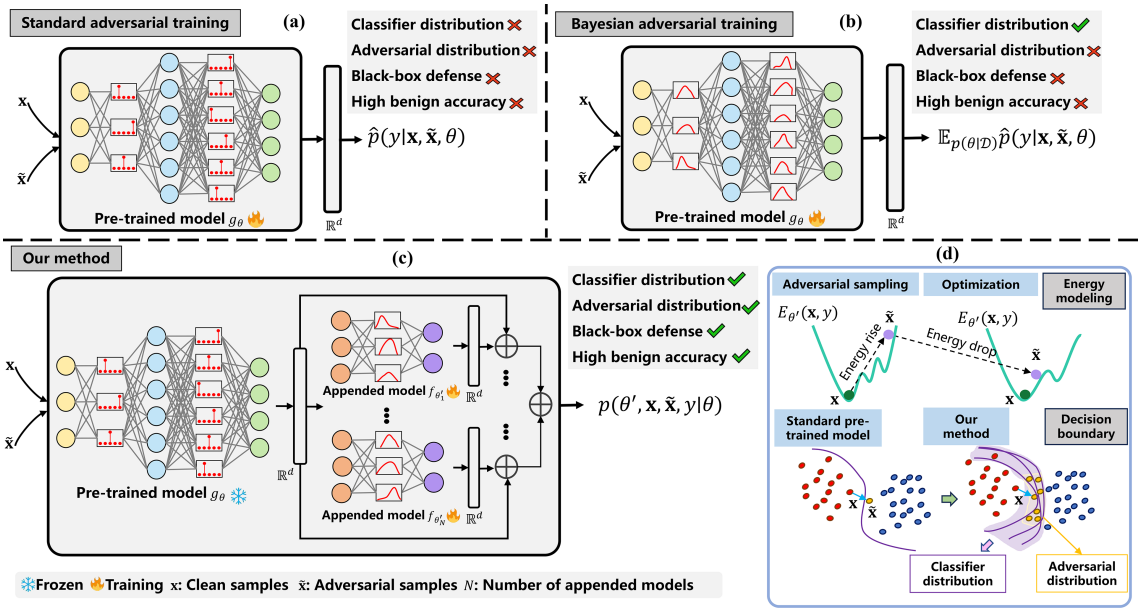

图 1　(网络版彩图) 贝叶斯能量对抗后训练的框架示意图. (a) 标准对抗训练; (b) 贝叶斯对抗训练; (c) 本文方法; (d) 贝叶斯能量对抗后训练的优化过程.

Figure 1　(Color online) Structure of the post-train energy-based Bayesian adversarial training. (a) Standard adversarial training; (b) Bayesian adversarial training; (c) our proposed method; (d) processes of post-train Bayesian energy adversarial training.

得联合概率分布 $p(\tilde{\mathbf{x}},\mathbf{x},y)$ 进一步泛化为 $p(\tilde{\mathbf{x}},\mathbf{x},y,\theta)$. 然而, 传统的贝叶斯对抗训练方法 (图 1(b)) 需要对模型重新进行贝叶斯优化, 在一些需要黑盒防御的实际场景下并不适用; 此外, 现实场景中的分类器往往包含着数十亿甚至百亿千亿规模的模型参数量, 贝叶斯神经网络难以在如此高维的空间下直接采样模型的参数分布. 因此, 本文设计了一种基于 "贝叶斯后训练" 的黑盒防御策略. 该方法在冻结的标准预训练模型 g_θ 后面添加一个参数为 θ' 的微型贝叶斯模型单元 $f_{\theta'}$, 并仅对 $f_{\theta'}$ 进行鲁棒性优化, 因此 $p(\tilde{\mathbf{x}},\mathbf{x},y,\theta)$ 在黑盒设置下可以重新表示成 $p(\tilde{\mathbf{x}},\mathbf{x},y,\theta'|\theta)$, 如图 1(c) 所示. 相比于白盒防御, 基于贝叶斯能量对抗后训练的黑盒防御方法不仅保留了预训练模型的完整性, 还显著降低了资源开销, 在实际应用中的灵活性更高. 为评估模型的鲁棒性表现, 本文在 3 个数据集、4 种基线模型上评估了共计 13 种对抗攻击方法的防御表现. 大量实验结果表明, 该方法在面对基于梯度的白盒对抗攻击和黑盒对抗攻击时, 能够在保持模型原始精度的同时显著提升鲁棒性, 其鲁棒性甚至优于现有的白盒防御方法. 特别是在大扰动强度的极端条件下, 本方法相比于现有的对抗训练方法有显著的鲁棒性优势. 此外, 贝叶斯能量对抗后训练方法不仅对基于梯度的攻击具有防御效果, 还能有效应对其他类型的对抗攻击, 包括自动集成攻击、基于优化的攻击以及基于输入变换的攻击方法等.

本文的第 2 节介绍了相关工作; 第 3 节详细阐述了本文介绍的基于能量模型的后训练贝叶斯对抗训练方法; 第 4 节为实验设置、实验结果及对其的分析与讨论; 第 5 节是对本文的总结和对未来工作的展望.

2　相关工作

对抗训练. 自视觉分类器在面对对抗样本时的脆弱性被揭示以来 [4,9~11], 众多对抗防御策略相继被研发和应用, 以有效应对这一安全挑战. 这些方法涵盖输入去噪 [12]、梯度正则化 [13,14]、防御蒸馏 [15] 以及对抗训练 [5,16,17] 等. 近年来, 对抗训练被证实为最有效的防御策略之一 [18], 其通过引入对抗样本, 显著增强了模型的鲁棒性和泛化能力, 进而在面对多种攻击时展现出优异的防御效果. 早

期, Madry 等 [5] 将考虑对抗因素的模型训练视为一个鞍点问题. 然而, 随着对抗攻击技术的不断发展 [17, 19, 20], 众多新的对抗训练策略应运而生, 旨在进一步优化传统方法. 在针对多种对抗攻击的防御研究中, Hammoudeh 等 [21] 提出了一种具有理论可证明鲁棒性的策略, 以增强模型在面对多样化攻击时的稳健性. Bortolussi 等 [8] 则通过分析贝叶斯神经网络的对抗攻击几何结构, 揭示了其在过参数化极限下的鲁棒性与脆弱性特征. Zhao 等 [22] 通过分析脆弱模型在攻击下的表现, 确定易受攻击的样本的类型, 并在对抗训练过程中为这些易受攻击样本赋予更高的权重, 从而使模型更关注这些样本, 提升整体的鲁棒性.

Mirza 等 [16] 将鲁棒分类器重新解释为能量模型, 并通过分析能量景观, 揭示了对抗训练中目标攻击与非目标攻击的不同影响. Kim 等 [23] 提出了一种相位移对抗训练方法, 旨在通过频率转换提升高频信息的学习效果. 尽管上述方法在提升鲁棒性方面取得了显著成效, 但它们在训练过程中往往忽视了鲁棒性重新训练所带来的巨大开销. 更重要的是, 许多方法以牺牲干净样本的分类准确率为代价, 来换取对抗样本的分类准确率. 干净样本的分类准确率作为大多数视觉分类任务的核心指标, 其重要性不言而喻. 因此, 如何在两者之间取得平衡, 成为了当前研究的热点和亟待解决的问题. Zhang 等 [24] 通过限制干净样本与对抗样本之间的输出概率差异, 提出了一种在良性准确率与对抗鲁棒性之间进行权衡的方法 (TRADES). Diao 等 [25] 则研究了对抗样本与良性样本在数据流形上的关系, 以控制两者间的准确率与鲁棒性差异. Co 等 [26] 结合雅可比 (Jacobi) 正则化与模型集成, 提出了 Jacobians 集成方法, 有效改善了模型在普适对抗扰动下的鲁棒性与准确率之间的权衡. 此外, Sitawarin 等 [27] 分析了随机变换防御策略的鲁棒性与准确率之间的折中, 揭示了防御效率与分类精度之间的平衡问题. 尽管上述方法在这一权衡上取得了一定进展, 但仍未完全解决准确率与鲁棒性之间的权衡问题 [24, 25, 27].

黑盒防御. 现有的对抗训练方法大多依赖于白盒防御策略, 即需要对模型的全部参数细节进行深入的掌握, 并需要重新训练模型, 这在许多实际应用场景中并不切实际. 黑盒防御策略无需访问模型内部参数, 同时展现出对不同类型数据的广泛适应性. 最近已有少部分学者开始关注黑盒防御问题. Chen 等 [28] 针对黑盒分数查询攻击, 提出了一种通过混淆深度神经网络的输出 logits 来误导攻击者的黑盒防御方法 AAA. AAA 虽然能够抵御黑盒分数查询攻击, 但无法防御白盒攻击. Zhang 等 [29] 结合了去噪平滑与零阶优化的方法, 对认证黑盒防御进行了研究, 尽管该方法对黑盒防御的鲁棒性可以提供理论保证, 但其鲁棒性的提升并不显著, 且缺乏在大规模数据集和大规模神经网络上的验证.

3　基于贝叶斯能量对抗后训练的黑盒对抗防御方法

3.1　威胁模型

在黑盒对抗防御设定下, 攻击者和防御者的能力定义如下.

攻击者能力. 在白盒威胁模型中, 攻击者完全了解目标模型的结构、参数和训练数据. 在黑盒威胁模型中, 攻击者无法直接访问目标模型的内部信息, 而只能通过目标模型的输入和输出关系来实施攻击.

防御者能力. 白盒防御允许防御者访问模型结构和参数, 并可重新训练以提升鲁棒性. 黑盒防御是指防御者不访问模型的结构和参数, 且不对模型进行重新训练的防御方法. 虽然黑盒防御在现实中的应用场景更加广泛, 但由于资源受限, 其设计相比白盒防御方法面临更大的挑战. 本文重点研究黑盒防御方法, 旨在不改变预训练模型的前提下, 有效抵御白盒和黑盒攻击.

3.2　后训练黑盒对抗防御

标准对抗训练可以被看作是内部损失最大化和外部损失最小化的 min-max 组合优化问题 [5]. 给定数据 $\mathbf{x} \in \mathbf{X}$, 其对应标签 $y \in \mathbf{y}$ 和分类模型 g_θ, 内部最大化的目标是通过最大化损失来找到一个攻

击性最强的对抗样本, 然后再通过外部最小化对抗样本的平均损失. 具体优化过程为

$$\min_{\theta} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\max_{\|\tilde{\mathbf{x}}-\mathbf{x}\|\in\Omega} \mathbf{L}(\tilde{\mathbf{x}}, y; \theta) \right], \tag{1}$$

其中, $\mathcal{D}$ 表示训练数据; $\mathbf{L}$ 表示损失函数, 对于分类问题通常使用交叉熵损失; $\tilde{\mathbf{x}}$ 表示 $\mathbf{x}$ 对应的对抗样本, Ω 表示扰动空间. 由于式 (1) 需要已知模型全部细节并且更新模型参数 θ 进行重新训练, 因此可以被视为是一种白盒防御方法. 由于重新训练的开销巨大, 这在很多场景下具有局限性. 不仅如此, 由于对抗样本和干净样本存在分布上的差异 [30], 模型的重新鲁棒训练会导致模型隐空间特征偏离自然流形空间, 导致白盒对抗训练后的模型精度下降 [25]. 为解决上述问题, 本文提出了一种新的后训练黑盒防御方法, 在冻结的预训练模型的后面附加一个可学习的微小模型 $f_{\theta'}$, 并只对 $f_{\theta'}$ 进行鲁棒性优化. 相比于原模型的输出 $g_\theta(x)$, 附加模型的输出可以通过跳跃连接表示为

$$\text{logits} = f_{\theta'}(g_\theta(\mathbf{x})) + g_\theta(\mathbf{x}). \tag{2}$$

因此, 可以将式 (1) 修改为只针对 $f_{\theta'}$ 进行鲁棒性优化:

$$\min_{\theta'} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\max_{\|\tilde{\mathbf{x}}-\mathbf{x}\|\in\Omega} \mathbf{L}(\tilde{\mathbf{x}}, y, \theta; \theta') \right]. \tag{3}$$

后训练黑盒对抗防御方法通过冻结预训练模型, 并仅利用预训练模型的输出结果对微型附加模型进行参数优化, 从而在最大程度上保证了预训练模型的完整性, 并有效加速了训练过程.

3.3　能量视角下的对抗样本分布建模

早期实验结果表明 (见 4.6 小节), 直接将 min-max 对抗训练方法应用于后训练黑盒防御框架中, 对分类器的鲁棒性提升有限. 原因在于, 标准的预训练模型的隐空间特征是从干净样本中直接学习得到的, 因此容易受到对抗样本的影响. 针对此问题, 黑盒防御设置必须有针对性地设计更有效的鲁棒性优化方法. 与标准对抗训练 [5] 仅在训练过程中考虑特定攻击 (如 PGD [5] 攻击) 下的对抗样本不同, 理想的鲁棒性模型应能在训练过程中观测到所有潜在的攻击方法生成的不同对抗样本, 即观测到对抗样本的完整分布. 此外, 鉴于对抗训练方法常导致模型精度与准确率之间的权衡问题, 理想的鲁棒性模型还应能同时正确分类对抗样本和干净样本. 因此, 本文提出对干净样本、对抗样本和分类标签的联合概率分布进行建模:

$$p_{\theta'}(\mathbf{x}, \tilde{\mathbf{x}}, y) = p_{\theta'}(y|\tilde{\mathbf{x}}, \mathbf{x})\, p_{\theta'}(\tilde{\mathbf{x}}, \mathbf{x}) = p_{\theta'}(y|\tilde{\mathbf{x}}, \mathbf{x})\, p_{\theta'}(\tilde{\mathbf{x}}|\mathbf{x})\, p_{\theta'}(\mathbf{x}), \tag{4}$$

其中 $p_{\theta'}(y|\tilde{\mathbf{x}}, \mathbf{x})$ 可以通过判别式分类器计算分类损失得到. 对于对抗样本和干净样本的联合概率分布建模 $p_{\theta'}(\tilde{\mathbf{x}}, \mathbf{x})$, 可以进一步利用贝叶斯公式拆解为 $p_{\theta'}(\tilde{\mathbf{x}}|\mathbf{x})\, p_{\theta'}(\mathbf{x})$, 其中 $p_{\theta'}(\tilde{\mathbf{x}}|\mathbf{x})$ 表示完整的对抗样本分布. 从能量模型的观点, 干净样本的分布 $p_{\theta'}(\mathbf{x})$ 可以直接用能量函数参数化表示 [31]:

$$p_{\theta'}(\mathbf{x}) = \frac{\exp(-E_{\theta'}(\mathbf{x}))}{Z(\theta')} = \frac{\sum_{y\in\mathbf{y}} \exp(f_{\theta'}(\mathbf{x})[y])}{Z(\theta')}, \tag{5}$$

其中, $E_{\theta'}(\cdot) : \mathbb{R}^D \to \mathbb{R}$ 表示能量函数. 一般而言, 能量模型通过式 (5) 中的指数函数对能量函数进行参数化建模, 将每个数据点映射为对应的标量, 从而描述连续密度函数的分布. $g_{\theta'}$ 表示参数为 θ' 的附加模型, $g_{\theta'}(\mathbf{x})[y]$ 则表示附加模型输出的分类标签 y 对应的分数. $Z(\theta') = \int_{\mathbf{x}} \exp(-E_\theta(\mathbf{x}))$ 是一个归一化常数, 确保模型输出满足概率分布的要求. 在训练过程中, 能量函数 $E_{\theta'}(\mathbf{x})$ 会将低能量值分配给高概率密度区域, 例如训练数据, 而将高能量值分配给低概率密度区域, 例如异常样本或不太常见的样本. 对于能量模型优化, 一个常见的做法是对模型参数 θ' 进行最大似然估计, 其对数似然的梯度可以表示为

$$\frac{\partial \log p_{\theta'}(\mathbf{x})}{\partial \theta'} = \mathbb{E}_{p_{\theta'}(\mathbf{x}')} \left[\frac{\partial E_{\theta'}(\mathbf{x}')}{\partial \theta'} \right] - \frac{\partial E_{\theta'}(\mathbf{x})}{\partial \theta'}. \tag{6}$$

由于 $Z(\theta')$ 难以计算解析解, 因此对于式 (6), 通常采用蒙特卡洛 (Monte Carlo) 采样的方式近似求解:

$$\frac{\partial \log p_{\theta'}(\mathbf{x})}{\partial \theta'} \approx \frac{\partial}{\partial \theta'}\left[\frac{1}{L_1}\sum_{i=1}^{L_1} E_{\theta'}\left(\mathbf{x}_i^+\right) - \frac{1}{L_2}\sum_{i=1}^{L_2} E_{\theta'}\left(\mathbf{x}_i^-\right)\right], \tag{7}$$

其中, $\left\{\mathbf{x}_i^+\right\}_{i=1}^{L_1}$ 表示一个批次内的所有训练样本, $\left\{\mathbf{x}_i^-\right\}_{i=1}^{L_2}$ 表示从 $p_{\theta'}(\mathbf{x})$ 中采样获得的一个批次内的样本, 其满足独立同分布的设定. 具体地, $\mathbf{x}^-$ 可以通过随机梯度郎之万动力学 (stochastic gradient Langevin dynamics, SGLD) [32] 迭代采样获得:

$$\mathbf{x}_{t+1}^- = \mathbf{x}_t^- + \frac{c^2}{2}\frac{\partial \log p_{\theta'}\left(\mathbf{x}_t^-\right)}{\partial \mathbf{x}_t^-} + c\epsilon, c > 0, \epsilon \in \mathbf{N}(0, \mathbf{I}), \tag{8}$$

其中 c 是采样中的步长, ϵ 服从标准的高斯 (Gauss) 分布 N, I 是一个单位矩阵. 参考式 (5), 对抗样本的分布 $p_{\theta'}(\tilde{\mathbf{x}}|\mathbf{x})$ 也可以用能量模型参数化表示:

$$p_{\theta'}(\tilde{\mathbf{x}}|\mathbf{x}) = \frac{\exp\left(-E_{\theta'}(\tilde{\mathbf{x}}|\mathbf{x})\right)}{\tilde{Z}_{\theta'}} = \frac{\sum_{y\in\mathbf{y}}\exp\left(f_{\theta'}(\tilde{\mathbf{x}})[y]\right)}{\tilde{Z}_{\theta'}}. \tag{9}$$

针对式 (9), 本文参考对抗训练框架采取两阶段的优化策略, 其优化过程如图 1 所示. 在对抗样本的采样阶段, 本文旨在建模完整的对抗样本分布. 尽管对抗样本的分布无法直接观测, 但从能量的角度来看, 对抗样本大多是在原始数据分布的边缘或之外生成的, 位于数据分布的低密度区域, 因此具有较高的能量. 受此能量观点的启发, 本文采样具有高能量值的对抗样本, 这些样本旨在生成远离干净样本分布且更具侵略性的样本, 其采样过程可以用能量函数表示为

$$\tilde{\mathbf{x}}_{t+1} = \tilde{\mathbf{x}}_t - \frac{c^2}{2}\frac{\partial \log p_{\theta'}\left(\tilde{\mathbf{x}}_t, y\right)}{\partial \tilde{\mathbf{x}}_t} + c\epsilon, c > 0, \epsilon \in \mathbf{N}(0, \mathbf{I}). \tag{10}$$

不同于对抗训练最小化预测分类概率 $\log p_{\theta'}(y|\tilde{\mathbf{x}})$ (即最大化分类损失) 来生成对抗样本, 式 (10) 中最小化对抗样本的联合概率分布 $\log p_{\theta'}(\tilde{\mathbf{x}}, y)$, 使生成的对抗样本位于概率密度低的区域, 同时添加高斯噪声 $c\epsilon$ 以增加生成的对抗样本的多样性. 在模型更新阶段, 其优化目标是最小化对抗样本与干净样本之间的能量差距, 旨在将对抗样本拉回到其对应的干净样本周围, 从而降低其能量值. 因此优化过程中需分配给对抗样本低能量值, 使对抗样本位于真实样本周围的高概率密度区域:

$$\frac{\partial \log p_{\theta'}(\tilde{\mathbf{x}}|\mathbf{x})}{\partial \theta'} \approx \frac{\partial}{\partial \theta'}\left[\frac{1}{L_1}\sum_{i=1}^{L_1} E_{\theta'}\left(\mathbf{x}_i^+\right) - \frac{1}{L_2}\sum_{i=1}^{L_2} E_{\theta'}\left(\tilde{\mathbf{x}}_i|\mathbf{x}_i^+\right)\right]. \tag{11}$$

综上, 对于联合概率分布 $p_{\theta'}(\mathbf{x}, \tilde{\mathbf{x}}, y)$ 的优化可以采用极大似然估计的方式:

$$\log p_{\theta'}(\mathbf{x}, \tilde{\mathbf{x}}, y) = \log p_{\theta'}(y|\tilde{\mathbf{x}}, \mathbf{x}) + \log p_{\theta'}(\tilde{\mathbf{x}}|\mathbf{x}) + \log p_{\theta'}(\mathbf{x}). \tag{12}$$

3.4　贝叶斯边界修正

尽管式 (12) 可以考虑完整的对抗样本分布, 但是对于模型参数的学习仍然是一个 '点估计' 问题, 即旨在学习一组固定的模型参数来匹配样本输入到输出的映射. 基于 '点估计' 的鲁棒性优化在白盒防御设置下虽然是可行的, 但在后训练黑盒防御设置下, 由于附加模型的参数量有限, 只针对附加模型参数的 '点估计' 优化难以使模型习得高度鲁棒的分类边界. 在贝叶斯学习框架下, 一个可以正确分类 $\mathbf{x}$ 的模型, 有无限多种方法绘制其分类边界. 同理, 对于可以正确分类对抗样本 $\tilde{\mathbf{x}}$ 的鲁棒性模型, 同样也应有无穷多种方式绘制其分类边界, 且不同分类边界间的鲁棒性表现也存在差异. 因此, 考虑多个鲁棒分类模型边界的集合, 即分类边界的分布, 能够赋予模型更强的鲁棒性. 这归因于不同类型的对抗样本到分类边界的距离存在差异, 进而导致生成对抗样本的威胁模型也呈现出多样性. 一个包含多种

算法 1 贝叶斯能量对抗后训练.

输入: 训练样本 $\mathbf{x}$; 训练迭代的次数 N_{tra}; 从模型 θ 中采样的轮数 M_1; 对抗样本分布的采样轮数 M_2; 从模型 θ' 中采样的轮数 M_3; 参数为 $\{\theta'_1,\ldots,\theta'_N\}$ 的附加模型 $f_{\theta'}$; 附加模型的数量 N;
初始化: 随机初始化附加模型参数 $\{\theta'_1,\ldots,\theta'_N\}$;
主迭代: $N_{tra}\leftarrow 0$, $N\leftarrow 0$;
for $t=1$ to M_1 **do**
  根据式 (8) 采样, $\mathbf{x}_t^-\leftarrow\mathbf{x}_{t-1}^-$;
end
根据式 (7) 计算 $h_1=\frac{\partial\log p_{\theta'}(\mathbf{x})}{\partial\theta'}$;
for $t=1$ to M_2 **do**
  根据式 (10) 采样, $\tilde{\mathbf{x}}_t\leftarrow\tilde{\mathbf{x}}_{t-1}$;
end
根据式 (11) 计算 $h_2=\frac{\partial\log p_{\theta'}(\tilde{\mathbf{x}}|\mathbf{x})}{\partial\theta'}$;
使用交叉熵损失计算 $h_3=\frac{\partial\log p_{\theta'}(y|\mathbf{x},\tilde{\mathbf{x}})}{\partial\theta'}$;
$h_{\theta'}=h_1+h_2+h_3$;
for $t=1$ to M_3 **do**
  最小化 $h_{\theta'}$, 更新参数 θ'_N;
end
return $\theta'_1,\ldots,\theta'_N$;
输出: 附加模型权重 $\{\theta'_1,\ldots,\theta'_N\}$;

分类边界的集合, 能够提供多样化的类间距离和局部边界平滑性, 从而有效抵御多种类型的攻击 [31]. 为学习分类边界的分布, 本节从贝叶斯视角重新考虑附加模型的参数分布学习过程, 因此式 (4) 可以扩展为

$$p(\theta',\mathbf{x},\tilde{\mathbf{x}},y|\theta)=p(\mathbf{x},\tilde{\mathbf{x}},y|\theta')p(\theta'|\theta),\tag{13}$$

其中, $p(\mathbf{x},\tilde{\mathbf{x}},y|\theta')$ 等价于式 (4) 的求解. 通过这种方式, 在黑盒对抗防御的过程中可以实现对干净样本、对抗样本和分类器的全贝叶斯对待. 具体而言, 可以采用贝叶斯模型平均的方式来优化式 (13):

$$\begin{aligned}p\left(y'|\mathbf{x}',\mathbf{x},\tilde{\mathbf{x}},y\right)&=E_{\theta'\sim p(\theta')}\left[p\left(y'|\mathbf{x}',\mathbf{x},\tilde{\mathbf{x}},y,\theta,\theta'\right)\right]\\&\approx\tfrac{1}{N}\textstyle\sum_{i=1}^{N}p\left(y'|\mathbf{x}',\theta'_i,\theta\right),\theta'\sim p\left(\theta'|\mathbf{x},\tilde{\mathbf{x}},y,\theta\right),\end{aligned}\tag{14}$$

其中 $\mathbf{x}'$ 是一个新样本, y' 是 $\mathbf{x}'$ 的预测标签, $p\left(\theta\right)$ 是模型先验概率, N 是贝叶斯附加模型的数量, θ 是冻结的预训练模型, 可以用以下迭代采样的方法对附加模型进行训练:

$$\begin{aligned}&\{\mathbf{x},\tilde{\mathbf{x}},y\}_t\mid\theta,\theta'_{t-1}\sim p\left(\mathbf{x},\tilde{\mathbf{x}},y\mid\theta,\theta'_{t-1}\right),\\&\theta'_t\mid\{\mathbf{x},\tilde{\mathbf{x}},y\}_t,\theta\sim p\left(\theta'\mid\{\mathbf{x},\tilde{\mathbf{x}},y\}_t,\theta\right),\end{aligned}\tag{15}$$

其中 t 是训练迭代次数. 尽管附加模型 $f_{\theta'}$ 可以考虑任意的模型结构, 但本文在实验中发现 $f_{\theta'}$ 采取简单的两层全连接层就可以获得良好的防御效果, 因此本文将 $f_{\theta'}$ 都统一设置为两层全连接模型. 算法 1 所示为算法的整体流程图.

4　实验结果

4.1　实验设置

本文采用 SGAHMC (stochastic gradient adaptive Hamiltonian Monte Carlo) 算法作为优化器, 具体超参设置参考文献 [33]. 为了验证本方法在图像数据上的有效性, 本文选取了 3 个常用的图像数据集, 包括 CIFAR-10 [34], CIFAR-100 [34] 和 ImageNet [35]. 鉴于该方法采取了贝叶斯防御策略, 本研究选

表 1　不同对抗训练方法在 **EOT-PGD** [36] 攻击扰动下的鲁棒性比较 (%). 粗体表示最优结果.

Table 1　Robustness (%) comparison with various Bayesian defenses under different EOT-PGD [36] attack budget. The best results are in bold.

Data	Defense	0	0.035	0.055	0.07
CIFAR-10	None	93.6	0	0	0
	PGD-AT [5]	80.3	31.1	15.5	10.3
	Adv-BNN [36]	79.7	37.7	16.3	8.1
	IG-BNN [37]	83.6	50.2	26.8	16.9
	Ours	**93.8**	**83.3**	**63.9**	**50.8**

择了最先进的贝叶斯防御方法 Adv-BNN [36] 和 IG-BNN [37] 作为基准. 在 CIFAR-10 数据集上, 本研究采用 VGG-16 网络作为目标网络, 以保持与比较方法原文中的默认设置一致. 与现有仅适用于小型神经网络的贝叶斯防御方法 [36,37] 不同, 本方法训练时可以避免占用大量内存, 故可训练更宽更深的神经网络, 如 WideResNets [38]. 鉴于大多数对抗训练工作使用 WideResNets 网络, 本研究选择了常用的 WRN28-10 作为基线网络, 与其他对抗训练方法 [5,36,37] 进行对比实验. 对于黑盒防御, 本文与之前的黑盒防御方法 AAA [28] 方法进行比较. 与 AAA 只能防御基于分数的黑盒攻击不同的是, 本文方法可以同时防御白盒和黑盒攻击. 除非特别提及, 所有参与比较的对抗训练方法在训练中所采样的对抗样本的扰动幅值都为 8/255, 迭代次数为 10 次.

本文的实验平台采用主频为 2039.813 GHz 的 AMD EPYC 7542 32-Core CPU 和 NVIDIA GeForce RTX 3090 GPU, 并搭载 24 GB 运行内存. 实验中使用开源的机器学习框架 PyTorch 对提出的方法进行实现.

4.2　白盒攻击评估

与贝叶斯防御方法的比较. 由于本文采用了后训练贝叶斯的防御策略, 因此首先与贝叶斯对抗防御方法 Adv-BNN [36] 和 IG-BNN [37] 进行比较. 本文参考之前贝叶斯防御工作的默认评估设置, 将 EOT (expectation-over-transformation) 算法与 PGD 攻击结合, 组成一种攻击性更强的对抗攻击方法 EOT-PGD. 具体而言, 在每一次迭代中, EOT-PGD 攻击通过平均多个随机变换后样本的梯度来估计期望梯度. 参考之前贝叶斯防御的设置, 本文采用 VGG16 作为骨干网络, 扰动边界设置为 $\epsilon \in [0, 0.07, 0.005]$, 实验结果如表 1 所示. 首先, 针对添加随机变换的 EOT-PGD 攻击, 基于贝叶斯防御的方法的防御表现普遍要比对抗训练 PGD-AT 要高. 其次, 本研究提出的方法在保持模型精度不变的前提下, 针对不同大小的扰动设置, 均超越了当前最先进的贝叶斯防御方法, 实现了最优的鲁棒性表现. 具体而言, 即使在扰动大小为 0.07 的极端情况下, 本研究方法仍然实现了 50.8% 的鲁棒性, 相较于之前最先进的贝叶斯防御方法 IG-BNN, 提高了 33.9%, 与对抗训练方法 PGD-AT 相比, 则提升了 40.5%.

与对抗训练方法的比较. 为了进一步证明本文所提出防御方法的鲁棒性, 本文将其与多个具有先进性能的对抗训练方法进行比较, 包括常用的 PGD-AT [5], TRADES [24], MART [19], 以及最近提出的 LAS-AT [20] 和 AWP [17]. 此外, 本文还将本方法与 FAT [39] 和 LBGAT [40] 进行比较, FAT 和 LBGAT 在保持高准确度的同时也能实现强大的鲁棒性. 所有对抗训练方法均使用 WRN34-10 网络. 在表 2 中展示了这些方法在 CIFAR-10 [34] 和 CIFAR-100 [34] 两个通用数据集上, 面对 PGD [5] 和 FGSM [41] 白盒攻击时的鲁棒性表现. 与现有对抗训练方法相比, 本文所提出的方法无论是在干净样本的分类精度以及对基于梯度的对抗攻击防御能力方面均表现出显著优势. 为进一步验证本文方法在更强攻击强度下的防御表现, 图 2 展示了不同防御方法在面对 EOT-PGD 攻击, 扰动边界设置为 $\epsilon \in [0, 0.07, 0.005]$ 时的防御表现. 随着攻击强度的增加, LAS-AT 和 AWP 的鲁棒性在扰动阈值为 0.07 时均降至 25% 以下, 而其他基线方法都降低到了 20% 以下. 相比之下, 尽管本文方法的防御性能随着攻击扰动的增大

表 2　与不同对抗训练方法的鲁棒性 (%) 比较 (贝叶斯能量对抗后训练的实验重复进行了三次).

Table 2　Comparative robustness (%) analysis with different adversarial training methods (the experiments of our proposed method were repeated 3 times).

Method	CIFAR-10				CIFAR-100			
	Clean	FGSM	PGD-20	PGD-50	Clean	FGSM	PGD-20	PGD-50
Standard training	**96.1**	49.5	0	0	**80.7**	14.4	0	0
PGD-AT	85.2	56.1	55.1	54.9	60.9	32.1	31.7	31.5
TRADES	85.7	64.7	56.1	55.9	58.6	31.5	28.7	26.6
MART	84.2	67.5	58.6	58.1	60.8	27.6	26.4	25.8
LAS-AT	87.7	67.2	60.2	59.8	64.9	36.9	36.4	36.1
AWP	85.6	62.9	58.1	57.9	60.4	34.6	33.9	33.7
FAT	88.0	65.9	49.9	48.8	–	–	–	–
LBGAT	88.2	57.8	54.7	54.3	60.6	35.5	34.8	34.6
Ours	95.1 ± 0.9	$\mathbf{94.1 \pm 0.8}$	$\mathbf{93.9 \pm 1.0}$	$\mathbf{93.7 \pm 1.0}$	80.2 ± 0.5	$\mathbf{54.8 \pm 1.2}$	$\mathbf{54.7 \pm 1.2}$	$\mathbf{54.5 \pm 1.3}$

表 3　本方法与文献 [44] 的鲁棒性 (%) 比较.

Table 3　Robustness (%) comparison between the proposed method and [44].

Data & model	Norm	Extra data	Clean	Robustness
CIFAR-10				
[44] (WRN28-10)	l_∞	80M TI	89.5	64.1
[44] (WRN70-16)	l_∞	80M TI	91.1	67.2
Ours (WRN28-10)	l_∞	n/a	$\mathbf{95.1 \pm 0.9}$	$\mathbf{93.7 \pm 0.9}$
[44] (WRN70-16)	l_2	80M TI	94.7	82.2
Ours (WRN28-10)	l_2	n/a	$\mathbf{95.1 \pm 0.9}$	$\mathbf{93.6 \pm 0.8}$
CIFAR-100				
[44] (WRN70-16)	l_∞	80M TI	69.2	39.0
Ours (WRN28-10)	l_∞	n/a	$\mathbf{80.2 \pm 0.5}$	$\mathbf{54.7 \pm 1.0}$

而有所下降, 但在 $\epsilon = 0.07$ 的极端情况下, 其鲁棒性仍然高达 48.6%, 是 LAS-AT 的两倍. 这一结果验证了本文方法在极端扰动条件下的优越防御性能.

为了评估本文方法在大规模数据集上的性能, 本文方法进一步在 ImageNet 数据集 [35] 上进行了比较实验. 鉴于对抗训练在完整 ImageNet 数据集上需要耗费巨大的计算资源 (如在 128 张 Nvidia V100 GPU 上需耗时 38 小时 [42]), 本研究选取了 ImageNet-1000 中前 100 个类别样本的子集, 并对训练样本进行了 64×64 的下采样处理. 模型结构选择 ResNet-18, 其在 ImageNet 子集上标准训练的准确率为 67.60%. 选择的威胁模型包括 FGSM, PGD, EOT-PGD 以及 APGD [43]. 实验结果如图 3 所示, 展示了不同防御方法在干净样本上的准确率、抵御不同威胁模型的鲁棒表现以及训练耗时. 本文方法在准确率比标准训练结果只降低 2.7% 的前提下, 对不同攻击方法的防御性能全面优于对比的基线方法. 此外在计算耗时方面, 本文所提出的对抗后训练方法在单卡 RTX 3090 GPU 上只需要 10 个小时即可训练完成, 比标准训练用时少 2 个小时, 仅为 LAS-AT 方法训练时间的 1/14. 综合 3 个数据集的实验结果表明, 本文方法在保证分类准确率的前提下, 显著提升了对梯度对抗攻击的鲁棒性, 从而验证了针对干净样本、对抗样本及分类器联合分布建模的有效性.

已有研究 [44~46] 表明, 使用额外的训练数据和更大规模的模型进行对抗训练可以显著提升模型的鲁棒泛化能力. 为此, 本文选择与 Gowal 等的工作 [44] 进行比较. 文献 [44] 采用了更深更宽的 WRN70-16 网络骨干结构, 并额外使用了 50 万张来自 80M-Ti Million Tiny Images (80M TI) 数据集 [47] 中的

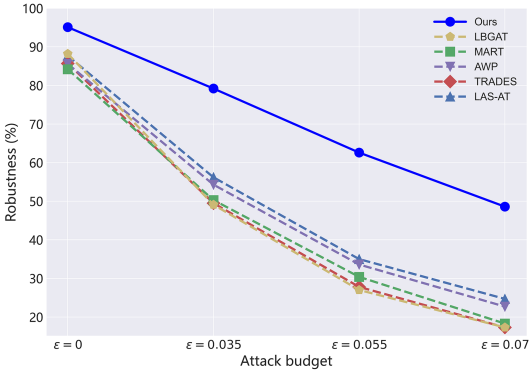

图 2　(网络版彩图) CIFAR-10 数据集下基于不同 EOT-PGD [36] 攻击扰动的对抗训练方法鲁棒性 (%) 对比.

Figure 2　(Color online) Robustness (%) comparison with various AT-based methods under different EOT-PGD [36] attack budget on the CIFAR-10 dataset.

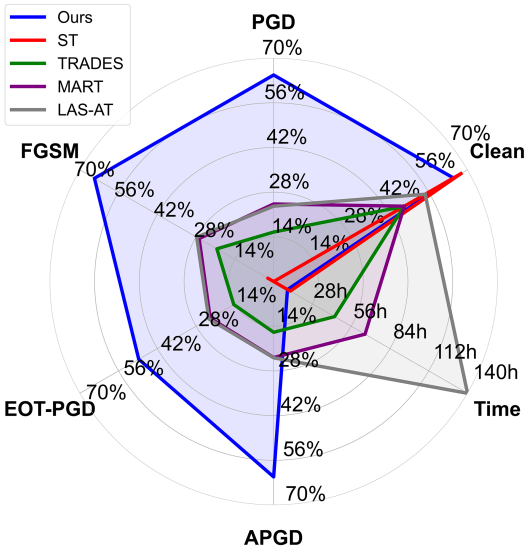

图 3　(网络版彩图) 使用 ResNet-18 在 ImageNet 子集上进行鲁棒性比较.

Figure 3　(Color online) Robustness comparison using ResNet-18 on the ImageNet subset.

图像用于训练, 在多个鲁棒性基准测试上均取得了显著的性能提升. 本文遵循 Gowal 等 [44] 所提出的评估方案, 采用 PGD-40 攻击, 在 l_∞ 范数下, 将扰动大小设置为 8/255, 同时在 l_2 范数约束下, 将扰动大小设置为 128/255, 以此为基础进行了全面评估, 实验结果如表 3 所示. 在不引入额外数据、不改变模型骨干结构且无需重新训练的前提下, 本方法与文献 [44] 相比不仅维持了较高的良性分类准确率, 还在 PGD 白盒攻击测试中展现出了更高的鲁棒性能.

4.3　黑盒攻击评估

本节旨在评估黑盒攻击场景下的防御性能. 为此, 本节选取黑盒防御方法 AAA [28] 和对抗训练方法 AT 进行比较. 鉴于 AAA 是一种专门针对黑盒分数查询攻击的防御方法, 因此本节采用常用的黑盒分数查询攻击方法 Bandits [48] 进行评估. 在黑盒攻击设置上, 本文使用 l_∞ 和 l_2 威胁模型进行防御评估. 由于 Chen 等 [28] 和 Ilyas 等 [48] 没有报告 CIFAR-10 数据集上 AT 和 AAA 对 l_2 Bandits 的结果, 所以在表 4 中仅报告了本方法在 l_2 攻击下的评估结果. 由于黑盒防御方法不需要对预训练模型进行重新训练, 因此 AAA 和本研究方法都可以保持比 AT 方法更高的准确率. 尽管如此, 本研究方法的准确率仍然比 AAA 高出 1.3%. 更重要的是, 与先前的黑盒防御方法 AAA 相比, 本方法在查询次数为 100 次和 2500 次的攻击设置下, 分别实现了 14.2% 和 16.3% 的鲁棒性提升.

4.4　更具挑战性的攻击评估

4.4.1　AutoAttack 评估

本节采用 AutoAttack [43] 框架进行鲁棒性评估. AutoAttack 是一种极为强大的自动集成攻击框架, 融合了 3 种白盒攻击策略 APGD-CE, APGD-DLR, FAB [49] 和一种黑盒攻击 Square [50]. 早期的实验结果表明, 直接使用标准训练的预训练模型进行后训练黑盒防御, 难以有效抵御 AutoAttack 攻击. 本文推测, 这一现象归因于预训练模型的隐空间特征主要基于干净样本进行训练, 因此预训练模型本身易受对抗样本的攻击. 尽管本文提出的方法在理论上能够考虑对抗样本的完整分布, 但由于实际采

表 4　在 CIFAR-10 数据集上针对 Bandits 攻击的鲁棒性 (%) 比较 (查询次数 = 100/2500).

Table 4　Robustness (%) under Bandits attack on CIFAR-10 (@query = 100/2500).

Method	Norm	Clean	Bandits	
			@100	@2500
Standard training	$l_\infty = 8/255$	96.2	69.9	41.0
AT	$l_\infty = 8/255$	87.0	83.6	76.3
AAA	$l_\infty = 8/255$	94.8	80.9	78.4
Ours	$l_\infty = 8/255$	96.1	**95.1**	**94.7**
Standard training	l_2	**96.2**	1.3	0
Ours	l_2	96.1	**93.1**	**92.8**

表 5　在 CIFAR-10 数据集上使用 JEM 作为预训练模型的鲁棒性 (%).

Table 5　Robustness (%) on CIFAR-10 using JEM as the pre-trained model.

Method	Clean	APGD-CE	APGD-DLR	FAB	SQUARE	Auto-Attack
JEM [31]	**92.9**	6.6	12.5	10.9	18.7	5.5
JEM+Ours	89.3	**39.9**	**32.9**	**75.8**	**39.8**	**29.5**

表 6　在 CIFAR-10 数据集上抵御不同攻击方法的防御表现 (%).

Table 6　Robustness (%) on CIFAR-10 against different attack strategies.

Method	Clean	C&W [52]	DeepFool [51]	MIG [11]	Admix [53]
JEM [31]	**92.9**	61.3	11.4	20.9	19.1
JEM+Ours	89.3	**62.5**	**21.4**	**30.2**	**30.4**

样次数有限, 难以涵盖干净样本对应的所有对抗样本. 由于标准预训练过程中学习的特征分布存在固有脆弱性, AutoAttack 凭借其广泛的攻击策略, 总能发现后训练黑盒防御可能忽视的对抗子空间.

为验证上述假设, 该实验将干净数据上训练的预训练模型替换为具有鲁棒性的预训练模型, 旨在使预训练模型的隐空间特征同样具备鲁棒性. 鉴于本文采用了基于能量模型的黑盒防御框架, 本节采用联合能量模型 JEM [31] 作为黑盒防御框架的预训练模型. 作为一种生成式分类器, JEM 从能量的观点对数据分布 $p_\theta(\mathbf{x}, y)$ 进行联合概率建模, 相较于判别式分类器, 展现出更强的鲁棒性. 从表 5 可以看出, 与原始 JEM 模型相比, 经过黑盒防御的 JEM 模型在 AutoAttack 的每种攻击方法下均表现出显著的鲁棒性提升, 从而在面对 AutoAttack 的整体攻击时, 模型的鲁棒性得到大幅提升.

4.4.2　针对不同攻击类型的鲁棒性评估

贝叶斯能量对抗后训练方法不仅能够有效抵御基于梯度的对抗攻击, 还能对其他类型的对抗攻击表现出一定的防御能力. 为了验证这一点, 本节采用了多种不同类型的攻击方法进行鲁棒性评估, 包括基于优化的对抗攻击方法 Deepfool [51] 和 CW [52], 基于数据增强的对抗攻击方法 Admix [53] 和基于动量积分梯度的对抗攻击方法 MIG [11]. 本实验与 AutoAttack 评估保持实验设置一致, 均采用能量模型 JEM 作为黑盒防御的预训练模型以增强预训练模型的防御能力. 实验结果如表 6 所示, 采用本文提出的方法对 JEM 模型进行黑盒防御后, 其在应对多种不同类型的攻击方法时表现出显著的防御能力提升, 尤其是在针对原预训练模型效果不佳的攻击场景中, 鲁棒性显著增强, 其中针对 Deepfool, MIG 和 Admix 的防御性能都提升了 10% 左右.

4.5　梯度混淆评估

已有研究表明 [54], 部分依赖梯度混淆的防御策略容易被自适应对抗攻击轻易绕过, 因此并不构成

表 7　消融实验分析.

Table 7　Ablation experiments.

JEBM	NUM	PT	Clean (%)	PGD (%)	EOTPGD (%)	APGD (%)
×	0	×	93.6	0	0	0
×	1	✓	**93.9**	19.3	16.6	19.2
✓	1	✓	**93.9**	48.8	43.8	46.1
✓	3	✓	**93.9**	92.8	81.3	90.0
✓	5	✓	93.8	**93.2**	**83.3**	**90.3**
✓	7	✓	93.8	**93.2**	**83.3**	90.0

真正的安全防御. 然而, 与梯度混淆策略仅隐藏单一决策边界梯度不同的是, 基于后训练策略的贝叶斯能量对抗训练方法的核心优势在于其能够建模干净样本与对抗样本的联合数据分布, 并将单一的决策边界扩展为决策边界分布. 因此, 本文所提方法的鲁棒性并非源自梯度混淆, 即其梯度本质上保持完整, 未出现被打断、随机化或梯度爆炸/消失等问题.

为深入阐释本文方法鲁棒性并非来源于梯度混淆策略, 本文采用文献 [54] 所提出的梯度混淆评估协议进行实验验证. 首先, 基于梯度混淆的防御策略通常对迭代攻击的鲁棒性高于单步攻击. 为此, 我们在 CIFAR-10 数据集上, 利用 VGG16 模型架构, 对比了本文方法在面对单步攻击 FGSM [41] 与迭代攻击 EOT-PGD-20 时的鲁棒性. 结果显示, 本文方法对 FGSM 的鲁棒性为 93.2%, 而对 EOT-PGD-20 的鲁棒性为 83.3%, 表明本文方法对迭代攻击的鲁棒性低于单步攻击, 这与梯度混淆现象不符. 其次, 基于梯度混淆的防御方法对于扰动范围大的攻击的鲁棒性通常高于小扰动范围攻击的鲁棒性, 而这与本文实验结果表 1 相悖. 最后, 基于梯度混淆的防御方法通常无法抵御自适应对抗攻击方法, 如 FAB 和 EOT 攻击. 本文执行了 EOT-PGD 和 FAB 攻击, 其实验结果分别如表 1 和 5 所示. 实验结果表明, 本文方法对自适应攻击具有鲁棒性, 从而进一步证实本研究方法鲁棒性不依赖于梯度混淆现象.

4.6　消融实验分析

4.6.1　联合能量建模与贝叶斯修正策略的消融分析

本方法对于鲁棒性的提升主要归因于干净样本和对抗样本的联合能量建模和贝叶斯修正策略. 其中, 贝叶斯修正策略的鲁棒性表现与附加模型的数量密切相关. 因此, 本节分别评估了联合能量建模 (joint energy-based modeling, JEBM)、附加模型数量 (number, NUM) 和后训练方法 (post-train, PT) 在本研究方法中的贡献. 本节选用 CIFAR-10 数据集和 VGG16 预训练模型, 以 PGD-20 攻击作为鲁棒性评估的攻击方法, 表 7 中展示了本方法在不同消融分析下的良性准确率和对抗鲁棒性, 其中 NUM = 0 表示标准预训练模型的实验结果. 从表 7 中可观察到, 后训练黑盒防御策略并未如传统白盒防御一样在干净样本的分类准确率上造成较大损失, 甚至小幅超过了标准训练模型的良性准确率 (93.6%). 其次, 通过对干净样本与对抗样本的联合分布进行建模, 黑盒防御的鲁棒性从 18.8% 显著提升至 48.8%. 进一步地, 同时考虑对抗样本与数据的联合分布以及模型参数的联合分布可以进一步提高模型的鲁棒性. 尽管贝叶斯神经网络理论上需要采样大量模型进行推理, 但在实际使用中发现采样 3 个以上的附加模型即可实现非常高的鲁棒性表现, 在附加模型数量等于 5 时, 模型的鲁棒性表现最佳.

鉴于多数对抗攻击策略均依赖于计算分类器对输入数据的梯度, 故损失梯度的特性成为理解模型鲁棒性的关键. 因此, 为深入探究本方法提升模型鲁棒性的内在机制, 本节对模型的损失梯度进行了可视化分析. 对于确定性模型, 其梯度是在单一分类边界的模型上计算得出的; 而本方法则通过多个模型上的梯度平均来计算期望损失梯度. 先前理论研究指出, 在数据量足够大的前提下, 具有无限宽度的贝叶斯神经网络的期望损失梯度趋于 0 [8]. 鉴于本方法的附加模型可视为小型贝叶斯神经网络, 其或可同样受益于这一特性. 为验证此假设, 本文从 CIFAR-10, CIFAR-100 和 ImageNet 子集 3 个数据集

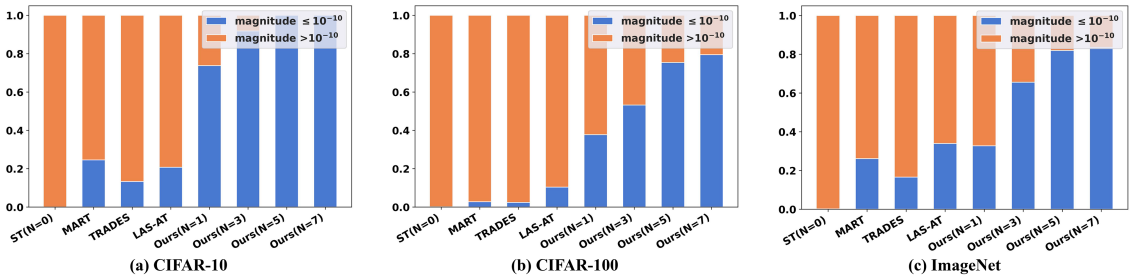

图 4　(网络版彩图) 贝叶斯能量对抗后训练的期望损失梯度单元大小在 10^{-10} 以上和以下的可视化表示. N 表示附加模型的个数, $N = 0$ 表示标准训练.

Figure 4　(Color online) Visualization of the expected loss gradient unit magnitude (PCM) of Bayesian energy adversarial post-training for values above and below 10^{-10}. N is the number of appended models. $N = 0$ is standard training.

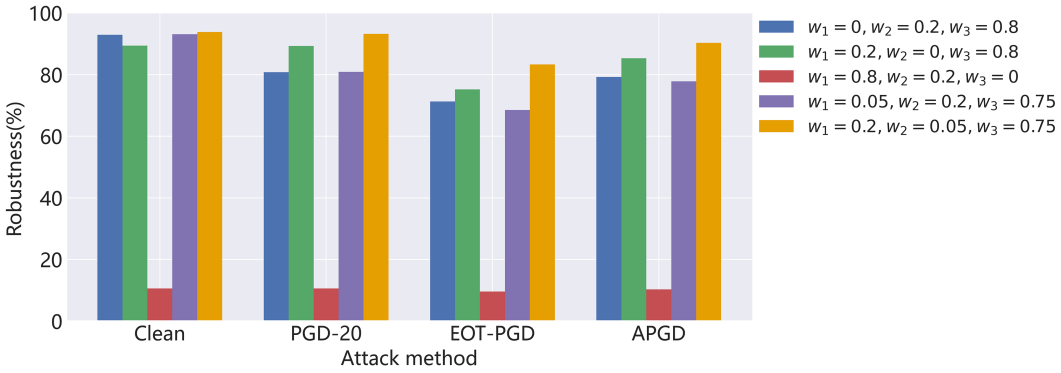

图 5　(网络版彩图) 使用 VGG16 在 CIFAR-10 数据集上贝叶斯能量对抗后训练的攻击梯度权重消融可视化.

Figure 5　(Color online) Attack gradient weight ablation of post-train Bayesian energy-based adversarial training on CIFAR-10 with VGG16.

中各随机抽取了 500 张图片并计算其期望损失梯度. 图 4 呈现了 3 个数据集期望梯度分量接近零的比例. 其中 ST 表示标准训练模型, N 表示附加模型的个数. 如图 4 所示, 对抗训练和本文方法都会减少梯度分量以提高鲁棒性. 随着附加模型数量 (N) 的增加, 贝叶斯能量对抗后训练模型的梯度分量逐渐接近于零, 表明期望损失梯度趋于消失. 当附加模型数量为 5 或 7 时, 模型的鲁棒性最高, 同时损失梯度在 3 个数据集上都已近乎完全消失, 这进一步验证了本研究方法鲁棒性的来源. 考虑到附加模型数量为 5 或 7 时的防御性能接近, 为节省计算资源, 在本文的实验设置中, 默认附加模型数量为 5.

4.6.2　攻击梯度权重的消融分析

如算法 1 所示, 在贝叶斯能量对抗后训练的优化过程中含有 h_1, h_2, h_3 这 3 个不同的梯度优化目标, 分别对应干净样本分布的能量建模梯度, 对抗样本分布的能量建模梯度, 以及对抗样本和干净样本的分类损失梯度. 在实际实验中, 我们对算法 1 中的梯度 h_1, h_2, h_3 的组合进行加权, 总体的优化梯度可以加权表示为 $h_{\theta'} = w_1h_1 + w_2h_2 + w_3h_3$. 本节对 3 种梯度进行不同设置的消融实验以分析不同梯度权重对于模型精度和鲁棒性的影响. 具体的消融设置如下: (1) 不考虑干净样本分布建模梯度 ($w_1 = 0$); (2) 不考虑对抗样本分布建模梯度 ($w_2 = 0$); (3) 不考虑分类损失梯度 ($w_3 = 0$); (4) 和 (5) 同时考虑 3 种梯度, 但对于 w_1, w_2 的侧重不同. 图 5 显示了在不同梯度权重设置下的消融实验结果. 从实验结果可以看出, 分类损失对于结果的影响最大 (设置 (3)), 这是因为分类损失优化模型的分类边界使得模型能够正确区分对抗样本与干净样本, 因此在优化过程中需要增大分类损失梯度 h_3 的权重. 此外, 在考虑分类损失的前提下, 相比只考虑干净样本分布建模 (设置 (1)) 和对抗样本分布建模 (设置 (2)), 同时考虑干净样本和对抗样本的联合分布建模在干净样本的分类表现上取得了最好的

效果, 而侧重干净样本的建模 (设置 (5)) 鲁棒性优于侧重对抗样本的建模 (设置 (4)), 实现最佳鲁棒性. 由此可见 3 种梯度共同作用于贝叶斯附加模型的参数优化过程, 使得模型在鲁棒性和分类准确率之间取得良好的平衡. 实验结果显示当 $w_1 = 0.2, w_2 = 0.05, w_3 = 0.75$ 时模型精度和鲁棒性最佳, 因此该权重作为本文默认的梯度权重设置.

5　总结与展望

本文提出了一种新的后训练黑盒防御策略, 该策略无需重新训练即可将标准预训练模型转化为一种具有弹性恢复能力的鲁棒性模型, 同时保持预训练模型精度不受影响. 本研究方法实现了对数据分布、对抗样本分布及模型参数分布的全贝叶斯对待, 显著增强了后训练黑盒防御方法的鲁棒性能. 大量的实验结果和分析表明, 本方法无需深入了解预训练模型的参数细节和重新训练预训练模型, 仅凭后训练微小的贝叶斯附加模型, 即可显著提升模型鲁棒性, 并在良性准确率与鲁棒性之间实现了良好的平衡. 面对基于梯度估计的白盒与黑盒攻击, 本方法均展现出卓越的鲁棒性, 同时避免了高昂的内存占用与计算成本.

尽管本文的研究主要聚焦于图像分类模型, 但所提出的黑盒防御技术作为一种通用方法, 同样具备扩展到其他任务类型上的潜力, 比如视频理解和目标检测. 此外, 尽管借助更鲁棒的预训练模型, 本文方法能在一定程度上抵御 AutoAttack 攻击, 但其鲁棒性相较于最先进的白盒方法仍有所不足. 如何进一步提升后训练黑盒防御方法在 AutoAttack 攻击下的表现, 仍是当前研究面临的一项挑战. 这两方面将在我们后续的研究工作中进行验证与探索.

参考文献

1 全国网络安全标准化技术委员会. 人工智能安全治理框架. 2024. https://www.cac.gov.cn/2024-09/09/c_1727567886199789.htm

2 新华网. 习近平: 推动我国新一代人工智能健康发展. 2018. http://www.xinhuanet.com/politics/leaders/2018-10/31/c_1123643321.htm

3 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

4 Xia M F, Ye Z P, Zhao W, et al. Adversarial attack and interpretability of the deep neural network from the geometric perspective. Sci Sin Inform, 2021, 51: 1411–1437 [夏萌霏, 叶子鹏, 赵旺, 等. 几何视角下深度神经网络的对抗攻击与可解释性研究进展. 中国科学: 信息科学, 2021, 51: 1411–1437]

5 Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks. In: Proceedings of ICLR, 2018

6 Wang H, Zhang A, Zheng S, et al. Removing batch normalization boosts adversarial training. In: Proceedings of ICML, 2022. 23433–23445

7 Wei X, Zhao S, Li B. Revisiting the trade-off between accuracy and robustness via weight distribution of filters. IEEE Trans Pattern Anal Mach Intell, 2024, 46: 8870–8882

8 Bortolussi L, Carbone G, Laurenti L, et al. On the robustness of Bayesian neural networks to adversarial attacks. IEEE Trans Neural Netw Learn Syst, 2025, 36: 6679–6692

9 Gao R J, Guo Q, Yu H K, et al. Adversarial attack method against image classification based on haze perturbation. Sci Sin Inform, 2023, 53: 309–324 [高瑞均, 郭青, 余洪凯, 等. 基于雾扰动的图像分类对抗性攻击方法. 中国科学: 信息科学, 2023, 53: 309–324]

10 Yuan H, Chu Q, Zhu F, et al. AutoMA: towards automatic model augmentation for transferable adversarial attacks. IEEE Trans Multimedia, 2023, 25: 203–213

11 Ma W, Li Y, Jia X, et al. Transferable adversarial attack for both vision transformers and convolutional networks via momentum integrated gradients. In: Proceedings of ICCV, 2023. 4630–4639

12 Li J, Xu D, Qin Y, et al. A feature guided denoising network for adversarial defense. In: Proceedings of ICUS, 2022. 393–398

13 Li Q, Hu Q, Lin C, et al. Revisiting gradient regularization: inject robust saliency-aware weight bias for adversarial defense. IEEE Trans Inform Forensic Secur, 2023, 18: 5936–5949

14 Jia X, Zhang Y, Wei X, et al. Prior-guided adversarial initialization for fast adversarial training. In: Proceedings of ECCV, 2022. 567–584

15 Huang B, Chen M, Wang Y, et al. Boosting accuracy and robustness of student models via adaptive adversarial distillation. In: Proceedings of CVPR, 2023. 24668–24677

16 Mirza M H, Briglia M R, Beadini S, et al. Shedding more light on robust classifiers under the lens of energy-based models. In: Proceedings of ECCV, 2024

17 Wu D, Xia S T, Wang Y. Adversarial weight perturbation helps robust generalization. In: Proceedings of NeurIPS, 2020. 33: 2958–2969

18 Sengphanith R, Marez D, Berk J N, et al. Evaluating the efficacy of different adversarial training strategies. In: Proceedings of Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications V, 2023. 525–532

19 Wang Y, Zou D, Yi J, et al. Improving adversarial robustness requires revisiting misclassified examples. In: Proceedings of ICLR, 2019

20 Jia X, Zhang Y, Wu B, et al. LAS-AT: adversarial training with learnable attack strategy. In: Proceedings of CVPR, 2022. 13398–13408

21 Hammoudeh Z, Lowd D. Provable robustness against a union of L_0 adversarial attacks. In: Proceedings of AAAI, 2024. 38: 21134–21142

22 Zhao P, Yuan H, Chu Q, et al. Delving deeper into vulnerable samples in adversarial training. In: Proceedings of ICASSP, 2024. 4490–4494

23 Kim Y, Kim S, Seo I, et al. Phase-shifted adversarial training. In: Proceedings of UAI, 2023. 1068–1077

24 Zhang H, Yu Y, Jiao J, et al. Theoretically principled trade-off between robustness and accuracy. In: Proceedings of ICML, 2019. 7472–7482

25 Diao Y, Wang H, Shao T, et al. Understanding the vulnerability of skeleton-based human activity recognition via black-box attack. Pattern Recognition, 2024, 153: 110564

26 Co K T, Martinez-Rego D, Hau Z, et al. Jacobian ensembles improve robustness trade-offs to adversarial attacks. In: Proceedings of IJCNN, 2022. 680–691

27 Sitawarin C, Golan-Strieb Z J, Wagner D. Demystifying the adversarial robustness of random transformation defenses. In: Proceedings of ICML, 2022. 20232–20252

28 Chen S, Huang Z, Tao Q, et al. Adversarial attack on attackers: post-process to mitigate black-box score-based query attacks. In: Proceedings of NeurIPS, 2022. 35: 14929–14943

29 Zhang Y, Yao Y, Jia J, et al. How to robustify black-box ML models? A zeroth-order optimization perspective. In: Proceedings of ICLR, 2022

30 Stutz D, Hein M, Schiele B. Disentangling adversarial robustness and generalization. In: Proceedings of CVPR, 2019. 6976–6987

31 Grathwohl W, Wang K C, Jacobsen J H, et al. Your classifier is secretly an energy based model and you should treat it like one. In: Proceedings of ICLR, 2020

32 Welling M, Teh Y W. Bayesian learning via stochastic gradient langevin dynamics. In: Proceedings of ICML, 2011. 681–688

33 Springenberg J T, Klein A, Falkner S, et al. Bayesian optimization with robust Bayesian neural networks. In: Proceedings of NeurIPS, 2016. 29

34 Krizhevsky A. Learning multiple layers of features from tiny images. In: Handbook of Systemic Autoimmune Diseases. 2009. https://api.semanticscholar.org/CorpusID:18268744

35 Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: Proceedings of CVPR, 2009. 248–255

36 Liu X, Li Y, Wu C, et al. Adv-BNN: improved adversarial defense through robust Bayesian neural network. In: Proceedings of ICLR, 2019

37 Doan B G, Abbasnejad E M, Shi J Q, et al. Bayesian learning with information gain provably bounds risk for a robust adversarial defense. In: Proceedings of ICML, 2022. 5309–5323

38 Zagoruyko S. Wide residual networks. In: Proceedings of British Machine Vision Conference, 2016

39 Zhang J, Xu X, Han B, et al. Attacks which do not kill training make adversarial learning stronger. In: Proceedings of ICML, 2020. 11278–11287

40 Cui J, Liu S, Wang L, et al. Learnable boundary guided adversarial training. In: Proceedings of ICCV, 2021. 15721–15730

41 Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proceedings of ICLR, 2015

42 Xie C, Wu Y, Maaten L V D, et al. Feature denoising for improving adversarial robustness. In: Proceedings of CVPR, 2019. 501–509
43 Croce F, Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: Proceedings of ICML, 2020. 2206–2216
44 Gowal S, Qin C, Uesato J, et al. Uncovering the limits of adversarial training against norm-bounded adversarial examples. 2020. ArXiv:2010.03593
45 Rebuffi S A, Gowal S, Calian D A, et al. Data augmentation can improve robustness. In: Proceedings of NeurIPS, 2021. 34: 29935–29948
46 Wang Z, Pang T, Du C, et al. Better diffusion models further improve adversarial training. In: Proceedings of ICML, 2023. 36246–36263
47 Torralba A, Fergus R, Freeman W T. 80 million tiny images: a large data set for nonparametric object and scene recognition. IEEE Trans Pattern Anal Mach Intell, 2008, 30: 1958–1970
48 Ilyas A, Engstrom L, Madry A. Prior convictions: black-box adversarial attacks with bandits and priors. In: Proceedings of ICLR, 2019
49 Croce F, Hein M. Minimally distorted adversarial examples with a fast adaptive boundary attack. In: Proceedings of ICML, 2020. 2196–2205
50 Andriushchenko M, Croce F, Flammarion N, et al. Square attack: a query-efficient black-box adversarial attack via random search. In: Proceedings of ECCV, 2020. 484–501
51 Moosavi-Dezfooli S M, Fawzi A, Frossard P. Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of CVPR, 2016. 2574–2582
52 Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proceedings of SP, 2017. 39–57
53 Wang X, He X, Wang J, et al. Admix: enhancing the transferability of adversarial attacks. In: Proceedings of ICCV, 2021. 16158–16167
54 Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. In: Proceedings of ICML, 2018. 274–283

Post-train black-box defense through energy-based Bayesian adversarial training

Yunfeng DIAO[1], Kaichao JIANG[1], Dan GUO[1,2*], Zhenyu LIANG[3,4*], Zenglin SHI[1], Zhenxing QIAN[5] & Meng WANG[1,2]

1. *School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China*
2. *Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230094, China*
3. *Electronic Engineering Institute, National University of Defense Technology, Hefei 230038, China*
4. *Information Security Research Center, Hefei Comprehensive National Science Center, Hefei 230031, China*
5. *School of Computer Science, Fudan University, Shanghai 200433, China*
* Corresponding author. E-mail: guodan@hfut.edu.cn, liangzy21@nudt.edu.cn

Abstract Deep neural networks have demonstrated outstanding performance in vision classifications. However, their security may be compromised by several factors, including their universal vulnerability to adversarial attacks. This has facilitated research on adversarial training (AT), the most widely used defense mechanism. However, existing AT-based methods are typically regarded as the white-box setting, considering that they require access to model parameters and re-training the victim under modified training regimes. White-box defenses are often impractical in real-world scenarios, such as re-training large-scale foundation models, mainly due to limited computational resources. Moreover, AT-based methods tend to improve robustness at the cost of sacrificing clean accuracy, which makes these models less effective for standard classification tasks and downstream applications. To address these challenges, we propose a new black-box defense framework called "Post-train Bayesian Energy Adversarial Training," a fully Bayesian treatment over the data, adversaries, and classifier. Our new post-train strategy is achieved by fixing the pre-trained model and appending tiny Bayesian components behind it, which transforms it into a resilient one without the need for re-training or accessing the model knowledge. The extensive results demonstrate that our proposed black-box defense outperforms existing white-box defense in terms of defending against both gradient-based white-box and black-box attacks without sacrificing clean accuracy.

Keywords adversarial examples, deep learning, adversarial defense, Bayesian neural network, energy-based model