IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

发布日期 xxxx 00, 0000，当前版本日期 xxxx 00, 0000。数字对象标识符
10.1109/ACCESS.2024.0429000

利用对抗样本的脆弱性进行校正

FUMIYA MORIMOTO[1], RYUTO MORITA[1], SATOSHI ONO [1]（IEEE会员）
[1]鹿儿岛大学理工学研究科信息科学与生物医学工程系，鹿儿岛县鹿儿岛市郡元1-21-40，邮编890-0065。
通讯作者：SATOSHI ONO（电子邮件：ono@ibe.kagoshima-u.ac.jp）。本研究部分获得日本学术振兴会科研费（项目编号JP 22K12196）和日本科学技术振兴机构A-STEP（项目编号JPMJTM20T0）的资助。

arXiv:2601.00270v1 [cs.CR] 1 Jan 2026

摘要 基于深度神经网络的分类器在处理对抗样本时容易出错。对抗样本是经过微小扰动、无法被人类察觉的输入数据，会给安全相关应用带来重大风险。因此，人们开展了大量研究，以开发能够减轻其威胁的防御机制。现有方法大多侧重于依据输入样本特征区分对抗样本，强调对抗样本检测，却未解决攻击发生前如何正确分类样本的问题。有些任务只需拒绝检测到的对抗样本，而另一些任务则必须识别原始输入的正确类别，例如自动驾驶中的交通标志识别。本研究旨在提出一种校正对抗样本的方法，以估计其原始输入的正确标签。我们的方法通过重新攻击对抗样本，使其越过决策边界，从而准确预测标签，有效解决使用白盒攻击方法生成的、感知上难以察觉的对抗样本的校正问题。然而，对于黑盒攻击在远离决策边界处生成的对抗样本，或被定向攻击错误分类到低置信度类别的对抗样本，如何有效校正仍然是一项挑战。所提方法采用直接的方法，仅将对抗样本作为输入，因此能够应对多种攻击，同时无需调整参数或进行预训练。结果表明，所提方法能够稳定地校正由各种攻击方法生成的对抗样本，包括定向攻击和黑盒攻击。此外，与传统校正方法和输入变换方法相比，该方法在应对各种攻击时表现得更加稳定。

索引词 深度神经网络、对抗样本、对抗防御、人工智能安全、标签校正

I. 引言

近期研究表明，基于深度神经网络（DNN）的分类器容易误识别对抗样本（AE）。对抗样本是经过特殊微扰、幅度很小且人眼无法察觉的输入数据 [1]。这一漏洞给自动驾驶中的交通标志识别 [2]–[5] 和基于图像的个人身份验证 [6]–[8] 等安全关键任务带来了严重问题。鉴于对抗样本可能在现实世界应用中被利用，解决其脆弱性对于确保应用系统的安全至关重要。将深度神经网络直接集成到各种系统所带来的风险，促使研究人员探索如何利用深度神经网络开发能够防御对抗样本的方法。

例如，输入变换 [9], [10] 是一种旨在通过图像变换等预处理来降低对抗样本影响的方法。然而，由于相同的变换会应用于所有输入，良性样本同样会受到该变换的扭曲，从而降低分类准确率。此外，该方法还需要根据深度神经网络的输入数据类型（如图像和音频）进行预处理方法开发，以及针对具体任务进行参数调整。

与此同时，已有检测方法 [11] 能够根据输入特征区分对抗样本，并保持良性样本的识别准确率。然而，这些方法只是检测并丢弃对抗样本，却无法识别攻击前图像的正确类别。尽管对于许多任务而言，简单地拒绝被检测为对抗样本的输入可能已经足够，但对于要求在攻击前识别输入的任务来说，这会成为一个关键问题。例如，当停车标志遭到攻击以混淆自动驾驶汽车时，仅检测并丢弃该对抗样本攻击是不够的，因为汽车将无法在正确的位置停车（图1(a)）。此外，

检测到对抗样本后停车同样很危险，因为汽车可能会停在不该停车的地方，例如限速标志遭到攻击时。因此，需要进行一些后处理，而不是拒绝检测到的对抗样本，以便利用它们识别停车标志。

从对抗样本中估计正确分类标签（即输入样本在遭受攻击前的标签）这一任务被定义为校正 [12]。如图 1(b) 中自动驾驶的示例所示，将校正与对抗样本检测器结合使用，即使停车标志遭到篡改，也能使自动驾驶车辆在停车标志前正确停车。除了此类关键任务外，检测对抗样本并准确识别其正确标签，对一般分类问题也大有裨益。大量关于输入变换的研究，以及一些关于融合标签校正的检测器的研究，都表明对抗样本校正的重要性日益提升 [13], [14]。

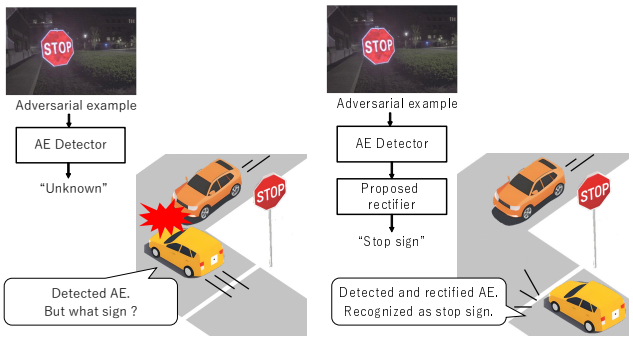
因此，本研究旨在提出一种从对抗样本中推断正确标签的校正方法。所提方法着眼于对抗样本的脆弱性，并通过重新攻击来纠正错误分类结果，使其正确归类为原始输入对应的类别。近期的攻击方法能够生成扰动极小、几乎难以察觉的对抗样本，其具体表现取决于目标深度神经网络模型的特性。这意味着生成的对抗样本位于特征空间中的分类边界附近，因此，添加扰动很可能改变其分类结果。能够改变对抗样本分类结果的微小扰动称为对抗样本的脆弱性 [11]。鉴于这种脆弱性，使用所提方法重新攻击对抗样本，可以有效地使错误分类结果与正确标签一致。

由于防御者通常能够访问所防御模型的内部信息，我们的方法采用白盒攻击方法对对抗样本进行重新攻击。通过计算能够降低误分类类别置信度的梯度，该方法可以高效地重新攻击对抗样本，以纠正其类别。由于我们的方法假设所有输入均为对抗样本，因此可以不受限制地对其进行重新攻击，并持续调整，直至分类结果发生变化。这减轻了对

领域特定预训练的需求，这对所提方法大有裨益。

本研究的主要目标是探究通过重新攻击，将黑盒攻击生成的对抗样本纠正为其正确标签的可行性。近年来，深度神经网络的实际应用发展迅速，各类图像和音频处理服务已作为 API 提供。同样，以生成式人工智能为代表的各类云服务商业人工智能（AI）系统也在迅速增加。这些 AI 系统中有许多采用深度神经网络模型，但其源代码和内部架构信息并未公开，因此在实施对抗攻击时必须使用黑盒攻击技术。因此，对纠正遭受黑盒攻击的对抗样本的方法的需求日益增长。

尽管预计所提方法将通过使用基于梯度的扰动再次攻击对抗样本来改变



(a) 仅使用对抗样本检测器　(b) 使用检测器和所提校正器

图 1. 所提方法在自动驾驶交通标志识别中的应用示例 [4]。

对抗样本的分类结果，但阐明该方法准确估计原始输入类别的方式至关重要。例如，如果使用相同的攻击方法生成并纠正对抗样本，我们的方法预计能够轻易地从对抗样本中辨别出原始输入的正确类别标签。然而，如果采用不同于最初所用攻击方案的方法进行再次攻击，该方法并不总能确定原始输入类别。此外，尝试纠正使用黑盒攻击生成的、缺乏来自目标深度神经网络的梯度信息的对抗样本时，也将面临重大挑战。

本研究的另一目标是考察所提方法能否校正通过定向攻击生成的对抗样本；这类攻击会将原始类别误分类为置信度较低的类别。对抗攻击通常分为两类：非定向攻击（将输入误分类为不同于原始分类结果的标签）和定向攻击（故意将输入误分类为指定的目标标签）。非定向攻击旨在降低正确预测结果的置信度，而定向攻击则试图提高目标预测结果的置信度。定向攻击所诱导类别的置信度越低，对抗样本中的扰动就越大，从而使准确校正类别更加困难。

我们通过涵盖图像和语音识别任务的实验（涉及至多七种攻击方法），验证了所提方法能够稳定地将对抗样本校正回正确标签，且不受数据类型和防御模型的影响。

本研究的贡献总结如下：

- **无训练、与数据和检测器无关的校正器：**通过设计为独立于对抗样本检测器运行，所提方法在通用性、灵活性和效率方面具有优势。与图像平滑等输入变换预处理不同，后者需要针对每种输入数据和任务进行专门设计、实现和调整，而我们的方法可普遍应用，不受

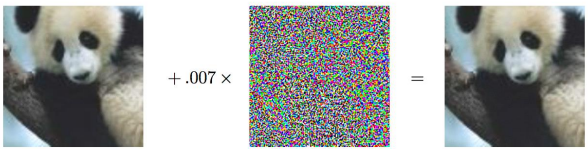


图2. Goodfellow 等人展示的对抗样本示例 **[15]**。

输入数据类型的限制。此外，由检测器执行对抗样本检测，因此所提校正器无需预先训练或调整。这是因为校正器在假设输入为对抗样本的情况下运行，因此无需预设攻击强度或迭代次数即可进行重新攻击；只需持续执行，直至模型识别出的标签发生变化。此外，我们的校正器可与任何检测器无缝集成，从而便于在现有系统中开发适用于特定任务或数据类型的新检测器和应用。

- **所提方法鲁棒性的实验验证：**首先，对该方法在黑盒攻击（包括基于分数和基于决策的攻击）以及定向攻击下的有效性进行概念分析。随后，通过实验证明该方法应对这些挑战的能力。此外，将所提方法抵御上述攻击的效果与一种独立于检测器运行的现有校正器进行比较 [12]。该先前方法采用可解释人工智能（XAI）技术消除关注区域，从而实现校正，其性能优于输入变换方法。输出比较表明，所提方法性能更优，也凸显了其相较于输入变换方法的优势。最后，通过实验验证了我们的方法在音频和图像模态中的适用性。

本文后续章节安排如下：第二节回顾对抗攻击与防御方面的先前研究。第三节阐述所提方法的基本概念，即通过重新攻击来校正对抗样本。第四节介绍实验设置与结果，评估该方法在图像和音频模态下应对各种攻击和数据集时的有效性，并详细比较我们的方法与现有方法，展示该方法抵御定向攻击的能力以及与检测器的交互情况。最后，第五节总结全文并展望未来研究方向。

II. 相关工作

A. 预备知识

在对抗攻击中，对抗样本由攻击者有意生成，正如 Goodfellow 等人所示的示例（图2） [15]。这里，$\boldsymbol{x}'$ 是通过向输入$\boldsymbol{x}$添加微小扰动 $\boldsymbol{\delta}$ 而生成的对抗样本，其定义如下：

$$\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{\delta}, \quad s.t.\ C(\boldsymbol{x}') \neq C(\boldsymbol{x})$$

其中$C(\cdot)$表示分类器的分类结果。扰动 $\boldsymbol{\delta}$ 定义为L_p 范数低于 ε ($\|\boldsymbol{\delta}\|_p < \varepsilon$)。

我们将对手知识分为两类：白盒和黑盒。在白盒场景下，对手掌握目标模型的梯度和参数的完整信息。相比之下，在黑盒场景下，对手不了解模型，无法获取其不同层级的内部信息。白盒攻击（例如利用模型梯度信息的基于梯度的攻击 [15]–[20],）威力强大。然而，在黑盒场景下，可以实施两类攻击：基于分数的攻击 [21] 和基于决策的攻击[22], [23]。前者利用预测结果及其置信度，而后者仅依据预测结果。

通常，对手会实施非定向攻击，将输入样本错误分类为不同于其原始分类结果的标签（非定向攻击场景），或分类为某个预先指定的标签（定向攻击场景）。通常，非定向攻击比定向攻击更容易实现。这是因为非定向攻击会降低正确预测的置信度，而定向攻击旨在提高目标预测的置信度。

B. 对抗攻击

1）基于梯度的攻击

Goodfellow 等人提出了快速梯度符号法（FGSM） [15], , 该方法根据模型的梯度生成对抗样本。它通过沿梯度方向迈出一步来增大梯度损失，执行无需迭代的一步攻击。Kurakin 等人提出了基本迭代法（BIM） [16], , 也称为 I-FGSM，通过以较小的步长迭代应用 FGSM，提升了 FGSM 的攻击性能。Madry 等人改进了 BIM，并提出了投影梯度下降（PGD）方法 [17]。与从原始输入开始的 BIM 不同，PGD 从随机点初始化，并持续执行随机攻击。尽管存在这一差异，这两种方法通常被视为相同。

Moosavi-Dezfooli 等人提出了 DeepFool 方法 [18] , 该方法旨在以最小的扰动实现成功攻击。与 FGSM 及类似方法需要手动设置扰动参数不同，DeepFool 在扰动距离最小时将决策边界视为线性。在这种情况下，通过使用泰勒展开对决策边界进行线性化，并沿正交向量寻找对抗样本，从而得出该正交向量。

Carlini 等人提出了一种名为 CW 的攻击方法，将对抗样本的生成视为一个优化问题，旨在最小化未受攻击的输入与对抗样本之间的差异 [19]。CW 能够实现最小扰动，并展现出较高的攻击成功率。

Papernot 等人提出了基于最大雅可比显著图的攻击（JSMA） [20]。该方法计算雅可比矩阵，并据此生成对抗显著图。随后，贪心算法会选出对抗显著图中数值最高的像素并对其施加扰动。重复这些步骤，直至达到扰动像素数量的上限，最终生成一个对抗样本。

2）基于分数的攻击

Narodytska 等人提出了一种名为局部搜索（LS）的基于分数的攻击方法 [21]，，该方法通过降低原始标签的预测概率来生成对抗样本。它通过贪心局部搜索，从原始输入生成仅有少量像素扰动的局部邻近图像。随后，选择原始标签预测概率最低的图像。重复这些步骤，直到扰动图像的预测标签与原始图像的标签不同。

3）基于决策的攻击

Brendel 等人提出了一种名为边界攻击的基于决策的攻击（DBA），其假设条件是仅提供预测标签。该方法沿决策边界逐步接近原始输入，同时保持图像分类错误 [22]，以此最小化扰动量。

Chen 等人提出了 HopSkipJumpAttack（HSJA）方法，该方法使用蒙特卡罗方法近似局部决策边界，从而估计决策边界梯度 [23]。HSJA 从对抗样本附近的区域估计与决策边界表面正交的方向，并结合二分搜索来最小化扰动量。

C. 对抗防御

采用深度神经网络模型的现实系统面临着恶意实体发起对抗攻击的风险，这些实体以造成损害为动机。因此，人们开发了各种对抗防御方法，以保护系统免受此类攻击。鉴于许多现实环境中存在不可预测的随机性，评估系统抵御对抗样本的鲁棒性，就是对其在最坏情况下韧性的一种测试。

采用深度神经网络的系统所用的对抗防御方法通常分为三类：对抗训练 [15]，[17], [24], [25], 输入变换 [9], [10], [26]–[28],和检测方法 [11], [29]–[34]。

其中，对抗训练是最常见的方法，它将对抗样本纳入训练数据，从而增强系统抵御对抗样本的能力。然而，尽管这些方法有效，却往往会降低良性样本的准确率，并增加计算开销。

输入变换通过预处理输入数据，为减轻对抗样本的影响提供了一种途径。在此类任务中例如图像分类，R&P 变换 [9]和 JPEG 变换 [10] 会用于修改输入图像。然而，对所有输入统一应用变换可能会扭曲良性样本，从而降低分类准确率。此外，预处理方法必须适应深度神经网络的输入数据类型，包括图像、音频和文本。

与另外两种方法相比，检测方法能够准确识别良性样本，但在需要精确分类原始输入的任务中面临挑战，例如自动驾驶中的交通标志识别 [35]。

D. 利用对抗样本脆弱性进行对抗防御

以攻为守（A^2*D*）是一种针对对抗样本脆弱性的防御方法。如果对抗样本在特征空间中靠近决策边界，再次对其发起攻击就能轻易使其越过边界，从而改变分类结果 [11]。对抗样本脆弱性的衡量方法是：对可能具有对抗性的输入数据再次发起攻击，采用迭代搜索攻击，并评估再次攻击成本，即改变识别结果所需的迭代次数。通过比较对抗样本与良性样本（根据待攻击的案例分别准备为训练样本）改变类别所需攻击成本的差异，可以识别对抗样本。基本迭代法 [16]、JSMA [20], 和 DBA [22]用于迭代攻击，而 k 近邻（k-NN）算法或标准分数（Z 分数）则用于根据攻击成本区分对抗样本。

Attackdist 方法是一种基于两个主要假设的检测方法。首先，攻击算法生成的对抗扰动必须接近最优解。其次，最优解位于决策边界附近 [34]。如果对抗样本被再次攻击，其扰动应显著小于良性样本中的扰动。该方法利用对抗扰动的 Lp 范数进行检测。

另一种检测方法基于两个基本原理利用 CW攻击方法。首先，CW攻击造成的扰动会通过迭代逐渐减小，从而使其接近决策边界。其次，对于已经受到攻击的图像，CW再次攻击产生的扰动远小于对原始图像进行攻击时产生的扰动 [33]。该方法通过实施额外的 CW再次攻击，并根据迭代次数进行判别，来识别由 CW攻击生成且难以检测的对抗样本。

尽管上述方法通过重新攻击输入图像展现出较强的对抗样本检测能力，但它们仅专注于对抗样本检测，并未考虑识别正确的原始输入类别。

E. 近期对抗防御方法

在对抗防御领域最先进的研究中，本节介绍一些近期出现的、与我们的方法相似的思路。

Salman 等人提出了非对抗方法，该方法利用对抗样本的反向梯度 [36]。他们的方法通过生成能够最小化损失、而非施加对抗扰动的非对抗样本（un-AE），提升了模型的性能以及对受损图像的鲁棒性。

最近，Chen 等人提出了对抗视觉提示（AVP）方法，通过视觉提示增强对抗鲁棒性 [37]。AVP 通过设计提示，在测试期间预先校正对抗样本的分类结果，从而提高鲁棒性。

Wang 等人提出了 FeConDefense 方法 [38]，该方法建立在 Mao 等人关于反向攻击方法 [39]的先前研究之上。这两种技术利用结合对比损失和特征一致性损失的伪损失梯度，将反向扰动引入对抗样本，从而恢复自然图像。

我们所提方法的理念与上述方法一致，这些方法利用对抗样本的梯度，沿对抗扰动的反向引入扰动。然而，尽管当前最先进的方法能够提高模型鲁棒性，但它们与我们的研究目标存在显著差异。例如，Un-AEs 通过应对域偏移来增强模型对受损图像的鲁棒性，却并未直接提高对抗鲁棒性。与此同时，AVP 对良性样本和对抗样本应用相同的提示，导致前者的分类准确率大幅下降。类似地，反向攻击会显著降低良性样本的分类准确率，而关于 FeConDefense 所获得的分类准确率的细节则十分有限。此外，AVP 和 FeConDefense 仅对基于梯度的攻击有效，而这类攻击是训练防御模型所必需的；它们也未评估防御模型针对各种对抗攻击的性能。另外，这些方法被归类为输入变换方法，而我们的方法作为检测器的后处理技术，可以避免良性样本分类准确率下降。

近年来，自然语言和语音处理领域针对对抗样本检测开展了广泛研究。自然语言处理（NLP）中的方法，例如频率引导词语替换（FGWS） [40], TextFirewall [41], 词级差分反应 [42],以及基于数据和模型不确定性的对手检测 [43] ，还有语音处理中的 Acoustic-decoy [44] 和 FraudWhisler [45] ，都只专注于检测对抗样本。然而，当前的方法还包括用于校正检测到的对抗样本标签的机制。例如，随机替换与投票（RS&V） [13] 通过替换对抗样本中的同义词生成多个相似句子，并根据这些句子分类结果的一致性进行检测，同时也能校正其标签。反应式扰动散焦（RAPID） [14] 提出了一种方法，将检测器与聚焦于校正器的扰动相结合，并在后处理阶段使用伪语义过滤来识别和校正对抗样本标签。该方法需要为检测器训练一个神经网络，并使用校正器来响应。不过，这些方法仅适用于自然语言模态，并且依赖特定检测器才能发挥作用，这使它们有别于可适配任何检测器的方法。

F. 对抗样本纠正

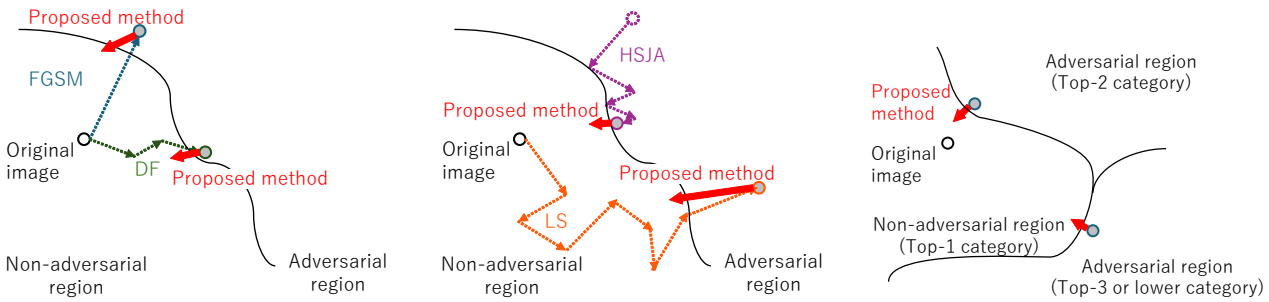
很少有研究探讨能够与任意检测器集成的校正器，例如所提方法；Kao 等人提出的方法是少数示例之一 [12]。他们提出了一种纠正方法，用于解决现有防御方法的局限性。他们的研究首次关注各种后处理技术的需求，以揭示检测器检测到的对抗样本的正确分类结果。他们通过修改或移除利用可解释人工智能估计出的已识别感兴趣区域，探讨了校正对抗样本并将其恢复到正确状态的可行性。该方法不需要输入排除或计算密集型处理，因此缓解了当前防御方法所存在的局限性。此外，与作为纠正检测到的对抗样本的方法而实施的四种基线方法（用于去噪的自编码器、JPEG 压缩、全图像高斯模糊和图像中的随机像素替换）相比，该方法表现更优。然而，其校正成功率在很大程度上取决于用于生成对抗样本的攻击方法。此外，该方法的一个局限性是，它只能校正针对能够利用可解释人工智能方法的深度神经网络生成的对抗样本。

第 II-E 节介绍的一些最新自然语言处理深度神经网络方法主要侧重于检测对抗样本，但也能对其进行校正。RS&V [13],例如，通过同义词替换施加扰动，以检测和校正对抗样本。它会针对一个输入生成多个不同的扰动模式，将这些受扰动的输入送入深度神经网络模型，然后对输出标签进行多数投票。如果多数标签与原始输入的标签一致，则判定该输入为良性；如果多数标签发生变化，则将该输入识别为对抗样本，并将多数标签作为校正后的结果。

RAPID 是一种防御方法，由基于学习的检测器和基于扰动去聚焦的语义纠正器组成。该纠正器会对检测到的对抗样本施加安全扰动，以抵消对抗扰动，并生成与原始输入含义相近的句子。具体而言，它根据词语重要性和分类概率进行同义词替换，确保语义保持不变。

FGWS [40]采用同义词替换来检测对抗样本并预测正确标签，与 RAPID 类似。该方法根据词频特征施加扰动。

上述方法专为自然语言处理中的深度神经网络设计。此外，RAPID 和 FGWS 等方法需要领域专用数据和模型训练，而我们提出的方法无需进行此类准备，且可以轻松地与各种检测器结合。



(a) 校正由基于梯度的攻击生成的对抗样本，例如 FGSM [15]和 DF [18]。(b) 校正由决策攻击生成的对抗样本，例如 LS [21] 和 HSJA [23]。(c) 校正由定向攻击生成的对抗样本，例如 FGSM [15]和 DF [18]。
t

图3. 所提方法的概念性解释。

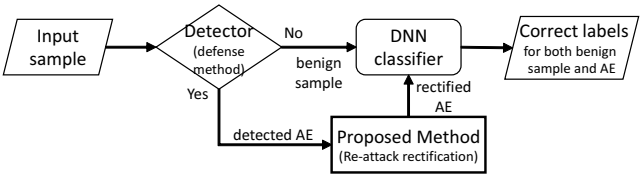


图4. 所提方法与对抗样本检测方法之间的关系。

III. 所提对抗样本纠正方法

A. 核心思想

所提出的对抗样本校正方法的核心，是通过重新攻击防御方法识别出的对抗样本，估计良性样本的正确标签。由于白盒攻击方法依据损失函数的梯度生成对抗样本，因此对抗样本位于对抗区域中、靠近决策边界的位置。值得注意的是，与生成对抗样本的标准对抗攻击不同，重新攻击无需获知正确类别或原始输入。通过施加与生成对抗样本时方向相反的扰动，可以使这些样本越过决策边界，回到非对抗区域。因此，我们的校正方法利用目标防御模型得出的损失函数梯度，推动对抗样本朝着降低误判类别置信度的方向移动，从而预测其正确标签。

所提方法专门用于独立校正对抗样本，因此能够与任意对抗样本检测器集成。该方法假设其输入仅为检测器识别出的对抗样本，不包含良性样本。因此，所提方法可以持续重新攻击，直至对抗样本的标签发生改变。这一设计无需进行特定任务的预训练或参数调整，这是所提方法的一项显著优势。

所提方法利用重新攻击来纠正对抗样本，可广泛应用于各种任务和数据模态。尽管平滑或噪声添加/移除等替代方法也能纠正对抗样本，但仍需要针对深度神经网络的输入数据类型（包括图像、声音和视频）定制特定的预处理方法，然而，我们的方法不需要此类预处理或调整，也不受深度神经网络输入数据类型的限制。此外，该方法适用于存在对抗攻击方法的任何深度神经网络。

B. 特征空间中的概念性解释

图3展示了所提方法针对各种攻击策略的工作原理。快速梯度符号法生成的对抗样本只需对输入的梯度进行一次计算，因此会产生较大的扰动（图3(a)）；相比之下，基本迭代法和DeepFool等方法通过迭代过程生成扰动较小的对抗样本。我们的方法利用白盒攻击方法进行重新攻击，使对抗样本朝着降低被误分类类别置信度的方向移动，从而将其转变为非对抗样本。使用与生成对抗样本时相同的白盒攻击方法进行再次攻击，有可能成功校正标签。此外，即使重新攻击所用的方法与通过白盒攻击方法生成对抗样本时采用的原始攻击策略不同，所提方法仍能利用梯度推动对抗样本朝着降低错误识别类别置信度的方向移动，从而恢复其正确类别。

不利用目标防御模型损失函数梯度的基于分数的黑盒攻击（如局部搜索）会在距离决策边界相对较远的位置生成对抗样本，因此难以对其进行校正，如图3(b)所示。然而，由于我们的方法假设所有输入都是对抗样本，因此可以不断重新攻击，直到对抗样本的分类结果发生变化。这样，该方法便能持续重新攻击局部搜索生成的对抗样本，直至其跨越决策边界，甚至有可能将远离边界的对抗样本校正回其原始类别。

基于决策的黑盒攻击（如 HSJA）通常从对抗区域附近开始搜索，并朝原始输入方向进行二分搜索，如图3(b)所示。因此，HSJA 生成的对抗样本最终会接近

对抗区域与非对抗区域之间的边界，这与白盒攻击生成的对抗样本类似。因此，即使生成对抗样本时没有使用损失梯度，我们的再攻击方法仍能有效纠正其标签。

相反，对于通过定向攻击将其错误分类到置信度显著较低的类别中的对抗样本，由于扰动较大，校正难度会不断增加。尽管如此，我们的方法预计仍能有效校正这些对抗样本，因为它会持续再次攻击，直到对抗样本的分类结果发生变化。成功校正的关键，尤其是针对旨在被识别为置信度最低的前三个类别或更低类别的对抗样本时，取决于错误分类类别区域与正确类别（Top-1）区域之间的接近程度。如图3(c)所示，如果存在连接错误分类类别与正确类别的充分边界区域，则可利用我们的方法成功校正标签。鉴于深度神经网络输入的高维性及其涵盖的众多类别，所提方法被认为能有效应对定向攻击，因为预计许多类别都与非对抗区域相邻。

C. 理论基础

本节将从理论角度探讨所提方法如何利用白盒攻击方法校正对抗样本。对抗攻击方法通常通过求解以下优化问题，生成最小的对抗扰动 $\boldsymbol{\delta}$，以改变分类结果：

$$\text{minimize } \|\boldsymbol{\delta}\|_p, \quad s.t.\ C(\boldsymbol{x}+\boldsymbol{\delta}) \neq C(\boldsymbol{x})$$

优化后的扰动 $\boldsymbol{\delta}^*$在对抗区域内取值且尽可能小，因此非常接近原始分类区域（非对抗区域）的决策边界。因此，如果对抗样本 $\boldsymbol{x}+\boldsymbol{\delta}^*$ 稍微向原始输入的方向移动，就会被分类为原始类别，即，

$$C\left(\boldsymbol{x}+(1-\mu)\boldsymbol{\delta}^*\right) = C(\boldsymbol{x}), \quad \mu \gtrsim 0$$

其中，$\mu \gtrsim 0$表示 μ 近似等于 0，但大于 0。

在此，我们考虑对上述包含理想扰动 $\boldsymbol{\delta}^*$ 的对抗样本进行重新攻击。具体而言，

$$\text{minimize } \|\boldsymbol{\delta}'\|_p, \quad s.t.\ C(\boldsymbol{x}+\boldsymbol{\delta}^*+\boldsymbol{\delta}') \neq C(\boldsymbol{x}+\boldsymbol{\delta}^*)$$

上述问题的近似解 $\boldsymbol{\delta}'^*$ 可视为 $-\mu\boldsymbol{\delta}^*$。

$$\|-\mu\boldsymbol{\delta}^*\|_p \approx 0, \quad C(\boldsymbol{x}+\boldsymbol{\delta}^*-\mu\boldsymbol{\delta}^*) = C(\boldsymbol{x}) \neq C(\boldsymbol{x}+\boldsymbol{\delta}^*)$$

因此，要纠正对抗样本，必须在原始输入$\boldsymbol{x}$ 仍未知的条件下，确定方向 $-\boldsymbol{\delta}^*$ 和幅度 μ。

这里，我们假设$\boldsymbol{x}+\boldsymbol{\delta}^*$ 足够接近$C(\boldsymbol{x})$这一原始类别的决策边界，并且对于$C(\boldsymbol{x})$以外的类别，其逻辑值（即 Softmax 前输出）和$C(\boldsymbol{x}+\boldsymbol{\delta}^*)$的逻辑值都足够低。这些假设使我们能够将分类视为一个二分类问题。此外，假设整个模型可以视为线性模型，且采用不含正则化项的二元交叉熵损失，则可以通过计算在$\boldsymbol{x}+\boldsymbol{\delta}^*$处使$C(\boldsymbol{x}+\boldsymbol{\delta}^*)$的逻辑值降低的方向，得到 $-\boldsymbol{\delta}^*$ 的方向。

$$-\boldsymbol{\delta}^* = \lambda\nabla L(\boldsymbol{\theta}, \boldsymbol{x}+\boldsymbol{\delta}^*, C(\boldsymbol{x}+\boldsymbol{\delta}^*)),\ \lambda > 0$$

其中，$L(\boldsymbol{\theta}, \boldsymbol{x}+\boldsymbol{\delta}^*, C(\boldsymbol{x}+\boldsymbol{\delta}^*))$表示参数为 $\boldsymbol{\theta}$ 的二元分类器对输入$\boldsymbol{x}+\boldsymbol{\delta}^*$ 及其目标$C(\boldsymbol{x}+\boldsymbol{\delta}^*)$的损失函数。这意味着，可以通过使用快速梯度符号法或基本迭代法等白盒攻击重新攻击对抗样本，从而纠正该对抗样本，并估计原始输入$\boldsymbol{x}$的类别$C(\boldsymbol{x})$。

请注意，只要 $-\boldsymbol{\delta}^*$ 的方向指向$C(\boldsymbol{x})$所在区域，就无需精确估计其方向。然而，较大的 μ、其他类别相对于$C(\boldsymbol{x})$和$C(\boldsymbol{x}+\boldsymbol{\delta}^*)$具有显著的对数几率，以及复杂的决策边界等因素，都会加大这些假设的偏差。对抗样本在特征空间中偏离这些假设越远，就越难以准确估计 $-\boldsymbol{\delta}^*$的方向，从而使对抗样本的纠正更加困难。

D. 处理流程

图4展示了所提对抗样本纠正方法与传统防御方法之间的相互作用。所提方法会对通过现有检测方法识别出的对抗样本再次发起攻击，以推断其原始图像所属的正确类别。该方法不受限制，兼容多种对抗攻击方法，包括白盒攻击和黑盒攻击。不过，鉴于防御者通常拥有访问受保护深度神经网络内部信息的权限，本研究重点关注白盒攻击。所提方法无需预先调整扰动量，但根据输入样本和目标防御模型计算扰动方向至关重要。此外，如III-C节所述，使用能够计算损失函数梯度的白盒攻击进行重新攻击，可以最有效地估计最佳方向。

E. 再攻击方法

尽管所提方法可以利用任何攻击方法重新攻击对抗样本，但鉴于我们能够获取有关深度神经网络模型的内部信息，本文选择使用快速梯度符号法 [15], 基本迭代法 [16], 和 DeepFool [18] 方法重新攻击对抗样本。所提方法可以采用任何白盒攻击方法。为了证明即使简单方法也足以实现有效校正，我们选用了快速梯度符号法、基本迭代法和深度欺骗这三种最简单且最具代表性的白盒算法。

算法1使用快速梯度符号法重新攻击

输入： 检测到的对抗样本$\boldsymbol{x}_a$，扰动大小ϵ，迭代次数s

输出： 重新攻击的对抗样本$\boldsymbol{x}'_a$

$\epsilon_s \leftarrow \epsilon/s$
$\boldsymbol{g} \leftarrow \nabla_{\boldsymbol{x}_a} L(\theta, \boldsymbol{x}_a, y_a)$
$\epsilon \leftarrow 0$, $i \leftarrow 0$
当$C(\boldsymbol{x}'_a) \neq y_a$ **或**$i < s$**时**
  $\epsilon \leftarrow \epsilon + \epsilon_s$
  $\boldsymbol{x}'_a \leftarrow \boldsymbol{x}_a + \epsilon \cdot$ 符号函数$(\boldsymbol{g})$
  $i \leftarrow i + 1$
结束当
返回$\boldsymbol{x}'_a$

算法2使用BIM重新攻击

输入： 检测到的对抗样本$\boldsymbol{x}_a$，扰动大小ϵ，步长α，迭代次数N

输出： 重新攻击的对抗样本$\boldsymbol{x}'_a$

$\boldsymbol{x}'_{a(0)} = \boldsymbol{x}_a$
对于$n = 0, \ldots, N-1$ **执行**
  $\boldsymbol{x}'_{a(n+1)} \leftarrow \boldsymbol{x}'_{a(n)} + \alpha \cdot$符号函数$(\nabla_{\boldsymbol{x}'_a} L(\theta, \boldsymbol{x}'_{a(n)}, y_a))$
  $\boldsymbol{x}'_{a(n+1)} \leftarrow$ 剪裁$_{\boldsymbol{x}_a, \epsilon}(\boldsymbol{x}'_{a(n+1)})$
结束循环
返回$\boldsymbol{x}'_{a(N)}$

1) 使用FGSM重新攻击

快速梯度符号法通过沿梯度方向添加单步扰动，对检测到的对抗样本$\boldsymbol{x}_a$进行重新攻击。然而，与不进行迭代的原始快速梯度符号法不同，我们在快速梯度符号法中加入线性搜索以确定移动幅度，从而确保其能够检测到标签发生变化的样本。值得注意的是，即使采用线性搜索，快速梯度符号法也只计算一次梯度，这一点不同于基本迭代法和投影梯度下降等迭代优化方法。算法 1 概述了详细算法，其中扰动的计算方式如下：

$$\boldsymbol{x}'_a = \boldsymbol{x}_a + \epsilon \cdot \mathrm{sign}(\nabla_{\boldsymbol{x}_a} L(\theta, \boldsymbol{x}_a, y_a))$$

其中$\boldsymbol{x}'_a$表示重新攻击后的图像，y_a表示$\boldsymbol{x}_a$标签，ϵ是控制扰动大小的参数，θ表示模型参数，L表示损失函数，sign 是符号函数。

2) 使用BIM重新攻击

基本迭代法通过以较小的步长沿梯度方向添加多个扰动来重新攻击检测到的对抗样本$\boldsymbol{x}_a$。算法 2 概述了详细算法，其中扰动计算如下

$$\begin{aligned} \boldsymbol{x}'_{a(0)} &= \boldsymbol{x}_a, & (1)\\ \boldsymbol{x}'_{a(n+1)} &= \mathrm{Clip}_{\boldsymbol{x}_a, \epsilon}(\boldsymbol{x}'_{a(n)} & (2)\\ &\quad + \alpha \cdot \mathrm{sign}(\nabla_{\boldsymbol{x}_a} L(\theta, \boldsymbol{x}'_{a(n)}, y_a))) \end{aligned}$$

其中α是步长。更新$\boldsymbol{x}'_{adv_{(n)}}$后，应用裁剪函数对对抗样本进行裁剪，使其在原始输入的ϵ区域内优化。

算法 3使用 DeepFool 重新攻击

输入： 检测到的对抗样本$\boldsymbol{x}_a$，步数N

输出： 重新攻击的对抗样本$\boldsymbol{x}'_a$

$\boldsymbol{x}'_{a(0)} \leftarrow \boldsymbol{x}_a$
斜体 $\leftarrow 0$
当 $\hat{k}(\boldsymbol{x}'_{a(i)}) = \hat{k}(\boldsymbol{x}_a)$ **或**$i < N$ **执行**
  对于$k : k \neq \hat{k}(\boldsymbol{x}_a)$**做**
    $\boldsymbol{w}'_k \leftarrow \nabla f_k(\boldsymbol{x}'_{a(i)}) - \nabla f_{\hat{k}(\boldsymbol{x}_a)}(\boldsymbol{x}'_{a(i)})$
    $f'_k \leftarrow f_k(\boldsymbol{x}'_{a(i)}) - f_{\hat{k}(\boldsymbol{x}_a)}(\boldsymbol{x}'_{a(i)})$
  结束循环
  $\hat{l} \leftarrow \arg\min_{k:k\neq\hat{k}(\boldsymbol{x}_a)} \frac{|f'_k|}{\|\boldsymbol{w}'_k\|_2}$
  $\boldsymbol{r}_i \leftarrow \frac{|f'_{\hat{l}}|}{\|\boldsymbol{w}'_{\hat{l}}\|_2^2} \boldsymbol{w}'_{\hat{l}}$
  $\boldsymbol{x}'_{a(i+1)} \leftarrow \boldsymbol{x}'_{a(i)} + \boldsymbol{r}_i$
  $i \leftarrow i + 1$
结束 while
返回 $\boldsymbol{x}'_{a(N)}$

3) 使用 DeepFool 再次攻击

DeepFool 通过从原始输入估计所有类别的决策边界，重新攻击检测到的对抗样本$\boldsymbol{x}_a$。它会计算朝向与原始输入边界最近的类别的扰动。详细算法如算法 3 所示。为处理非线性边界，DeepFool 方法会迭代地对边界进行线性近似，并添加使样本进入最近类别所需的最小扰动。

以下介绍使用 DeepFool 进行再次攻击的算法。设$f(\cdot)$为分类器，并将分类结果定义为$\hat{k}(\boldsymbol{x})$：

$$\hat{k}(\boldsymbol{x}) = \arg\max_k f_k(\boldsymbol{x}) \quad (3)$$

其中，$f_k(\boldsymbol{x})$是$f(\boldsymbol{x})$对应类别k的输出分数。

在每次迭代中，对于当前输入$\boldsymbol{x}'_{a(i)}$，DeepFool 会计算每个类别k对应的输出分数$f_k(\boldsymbol{x}'_{a(i)})$及其梯度$\nabla f_k(\boldsymbol{x}'_{a(i)})$。随后，对于每个其他类别$k \neq \hat{k}(\boldsymbol{x}_a)$，它通过计算$|f'_k|/\|\boldsymbol{w}'_k\|$，来近似$\boldsymbol{x}'_{a(i)}$到类别$\hat{k}(\boldsymbol{x}_a)$与类别$k$之间决策边界的距离，其中$\boldsymbol{w}'_k$和$f'_k$的计算方式如下：

$$\begin{aligned} \boldsymbol{w}'_k &= \nabla f_k(\boldsymbol{x}'_{a(i)}) - \nabla f_{\hat{k}(\boldsymbol{x}_a)}(\boldsymbol{x}'_{a(i)}) & (4)\\ f'_k &= f_k(\boldsymbol{x}'_{a(i)}) - f_{\hat{k}(\boldsymbol{x}_a)}(\boldsymbol{x}'_{a(i)}) & (5) \end{aligned}$$

接下来，它会找出与$\boldsymbol{x}'_{a(i)}$决策边界最近的类别$\hat{l}$，即：

$$\hat{l} = \arg\min_{k:k\neq\hat{k}(\boldsymbol{x}_a)} \frac{|f'_k|}{\|\boldsymbol{w}'_k\|_2} \quad (6)$$

随后，它计算一个向量，将$\boldsymbol{x}'_{a(i)}$投影到近似表示类别之间决策边界的超平面上

表 1. 所提方法中用于校正的再攻击参数。

再攻击方法	参数
FGSM	秒 $= 1{,}000, \epsilon = 1.0$
BIM	$\epsilon = 0.3, \alpha = 0.05, N = 10$
DF	$s = 100$

$\hat{}l$ 和 $\hat{}k(\boldsymbol{x}'_{a(i)})$，由此得到按如下方式计算的最小扰动 $\boldsymbol{r}_i(\boldsymbol{x}'_{a(i)})$：

$$\boldsymbol{r}_i = \frac{|f'_l|}{\|\boldsymbol{w}'_l\|_2^2}\boldsymbol{w}'_l \tag{7}$$

通过将$\boldsymbol{r}_i$ 添加到$\boldsymbol{x}'_{a(i)}$，DeepFool 更新输入并获得$\boldsymbol{x}'_{a(i+1)}$。

DeepFool 不断递增i 并重复此过程，直到达到迭代次数上限，或直到$\boldsymbol{x}'_{a(i)}$ 位于$k(\boldsymbol{x}_a)$ 区域之外。

IV. 评估

为评估所提方法的有效性，我们按照下文所述进行了四项实验测试。首先，验证了所提方法在采用图像分类的深度神经网络模型时，对各种攻击方法的有效性（实验 1）。实验 1 在非定向攻击场景下针对白盒和黑盒攻击方法进行（实验 1a），并在定向场景下进行（实验 1b）。随后，我们与参考文献 [12]、[13] 中报道的最先进的纠正方法进行了比较分析（实验 2a 和 2b）。接着，结合 A^2D 方法 [11]中所述的检测器，进一步展示了所提方法的防御性能（实验 3）。最后，为展示其对其他数据模态的适用性，我们将所提方法应用于语音识别（实验 4）。

A. 实验 1a：针对包括黑盒攻击在内的各种攻击方法的校正性能（非定向攻击）

1) 设置

在本实验中，我们在非定向攻击场景下，结合多种数据集和攻击方法，评估了所提方法的校正性能。首先，我们将该方法应用于由白盒攻击生成的对抗样本，包括快速梯度符号法 [15],基本迭代法 [16], DeepFool（DF） [18], Carlini-Wagner攻击（CW） [19], 和 JSMA [20]。随后，我们使用了由黑盒攻击生成的对抗样本，例如局部搜索（LS） [21] 和 HopSkipJumpAttack（HSJA） [23]。这些攻击方法是依据防御评估指南 [46], 选取的，旨在确保方法的多样性和代表性，同时避免使用相似的方法。我们使用了 FoolBox 框架中的攻击方法实现 [47]。使用这七种方法生成对抗样本时，攻击参数均采用 FoolBox 的默认值。对于重新攻击，根据第 III-E 节所述理由，我们选择了快速梯度符号法、基本迭代法和 DeepFool。所提方法的重新攻击参数采用 FoolBox 的默认值，详见表 1。

表2。 实验 1a：校正的成功率（非定向攻击）。

数据集	再次攻击方法	攻击方法 白盒 快速梯度符号法	基本迭代法	深度欺骗	Carlini-Wagner攻击	JSMA	局部搜索	黑盒 HSJA
MNIST	FGSM	**0.999**	**0.999**	**0.978**	**1.000**	0.993	0.938	**1.000**
	BIM	0.998	**0.999**	**0.978**	**1.000**	**0.995**	0.937	**1.000**
	DF	0.993	0.998	0.944	**1.000**	0.987	**0.939**	**1.000**
CIFAR-10	FGSM	**0.992**	**1.000**	**1.000**	**1.000**	**0.994**		0.911 **1.000**
	BIM	**0.992**	**1.000**	**1.000**	**1.000**	0.993	0.911	**1.000**
	DF	0.991	0.997	0.999	0.998	**0.994**		**0.913** 0.991
ImageNet	FGSM	**0.926**	0.991	0.999	**0.994**	**0.999**	0.981	0.997
	BIM	0.919	**0.999**	**1.000**	0.992	0.998	**0.989**	**1.000**
	DF	0.923	0.997	0.997	0.993	0.998	0.987	0.997

对于实验1和实验3，我们使用了三个图像数据集：MNIST [48], CIFAR-10 [49], 和 ImageNet-1000（ILSVRC2012） [50]。MNIST 和 CIFAR-10 的分类模型基于先前研究实现 [12]，以便将所提方法与实验 2 中使用的先前方法进行公平比较；而 VGG-19 [51] 则用作 ImageNet分类器。

对于三个数据集与七种攻击方法的每种组合，我们选取了 1,000 个样本，这些样本的原始输入均被分类模型正确识别，且对抗攻击成功。我们将校正成功率的百分比定义为评估标准，据此，校正后识别对抗样本的结果与识别原始输入的结果相符。

2) 校正性能的结果

表 2 展示了所提方法结合三种再攻击方法（快速梯度符号法、基本迭代法和深度欺骗）时的校正成功率。结果表明，所提方法可有效应用于多种数据集和攻击方法，在全部三个数据集上，成功校正了由全部七种攻击方法生成的 90% 以上的对抗样本。值得注意的是，即使采用的再攻击方法与初始攻击所用的方法不同，所提方法仍能准确估计正确标签。这验证了所提方法的核心理念，表明可以利用对抗样本与决策边界之间的较近距离实现校正。

此外，我们的方法能够有效修正由 LS 和 HSJA 等黑盒攻击生成的对抗样本，这些攻击不依赖深度神经网络中的损失函数梯度。LS 的成功率低于白盒攻击方法，这是因为它缺乏在判别边界边缘发现对抗样本的机制。然而，在最差情况下，我们的方法仍能保持约 91% 的成功率；正如第 IV-D 节中与先前方法的比较所示，这一成功率已足够高。

图5通过在 ImageNet 上尝试使用再次攻击进行校正的示例，展示了我们方法的应用结果。每一行包含五张图像，从左到右依次为：输入图像、对抗样本、对抗样本的扰动、校正后的对抗样本，以及

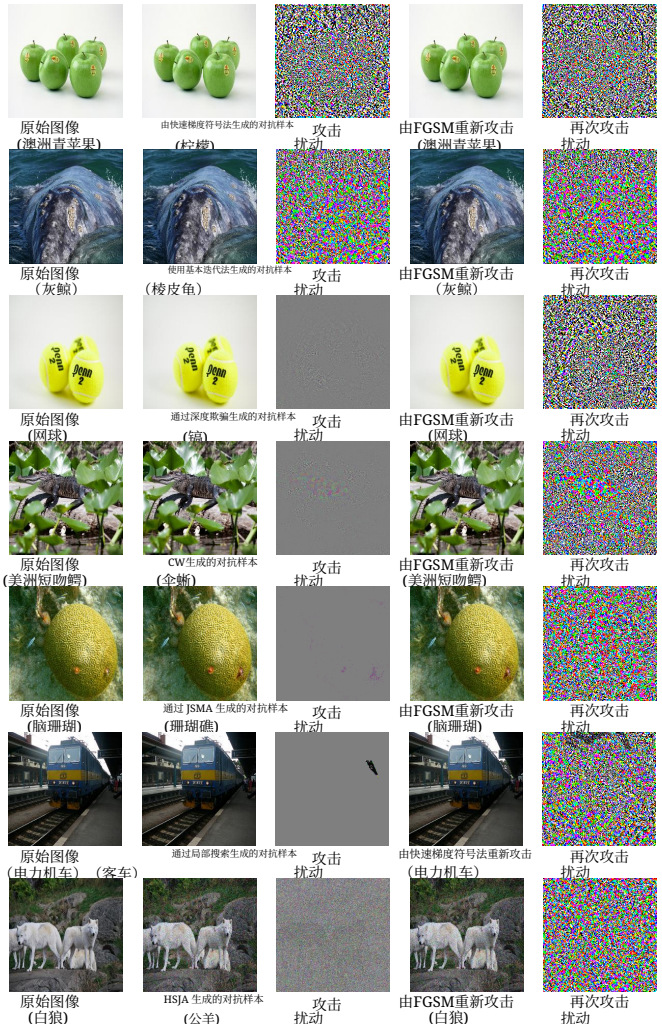


图5。 在实验 1a 中使用 FSGM 重新攻击，通过我们的方法校正的对抗样本示例。括号中的标签表示分类器的识别结果。

再次攻击扰动。例如，第一行中的第一个示例展示了针对初始快速梯度符号法攻击使用快速梯度符号法重新攻击的结果，成功校正了标签。第六个示例展示了使用快速梯度符号法针对初始 LS 攻击再次攻击的结果。这表明，尽管通过像素贪心法在车辆尾部添加了大量扰动，再次攻击仍校正了标签。

3）扰动量分析

本节探讨了所提方法为何能够校正无需依赖梯度生成的对抗样本，例如由黑盒攻击生成的样本，包括局部搜索和 HSJA。如 III-C 节所述，要有效校正对抗样本，需要以适当的方向和幅度进行再次攻击。因此，我们首先分析了再次攻击过程中引入的扰动幅度。

表3和表4分别展示了用于生成对抗样本的初始攻击以及用于校正的再次攻击的扰动量，后者——

表3。 实验 1a：对抗样本的扰动量。

数据集	攻击方法						
	白盒					黑盒	
	~~快速梯度符号法~~	~~基本迭代法~~	DF	~~Carlini-Wagner攻击~~	~~JSMA~~	LS	HSJA
MNIST	4.344	2.487	1.801	1.398	2.964	7.147	1.591
CIFAR-10	1.139	0.270	0.196	0.157	0.724	5.756	0.468
ImageNet	1.184	0.235	0.145	0.154	1.243	6.369	23.074

表4。 实验 1a：再次攻击的扰动量（$\times \mathbf{10^{-3}}$）。

数据集	再次攻击方法	攻击方法						
		白盒					黑盒	
		~~快速梯度符号法~~	~~基本迭代法~~	~~深度欺骗~~	~~Carlini-Wagner攻击~~	~~JSMA~~	~~局部搜索~~	~~HSJA~~
MNIST	FGSM	5.8	16.8	210.7	1.5	9.7	153.7	1.4
	BIM	5.2	15.6	184.5	0.6	8.9	141.8	0.1
	DF	110.7	13.7	317.4	1.0	113.8	169.4	7.6
CIFAR-10	FGSM	6.1	3.2	6.9	2.8	8.0	35.2	2.8
	BIM	4.6	0.8	4.7	0.5	6.4	30.5	0.1
	DF	3.2	0.5	3.0	0.3	4.1	19.4	0.6
ImageNet	FGSM	20.5	20.6	19.4	23.7	19.3	19.4	19.4
	BIM	8.4	1.0	0.5	8.0	1.4	2.6	0.1
	DF	4.4	0.8	0.4	4.6	0.9	1.4	0.1

表5。 实验 1a：攻击扰动与再次攻击逆向量之间的余弦相似度，用于表明重新攻击方向是否恰当。

数据集	再次攻击方法	攻击方法						
		白盒					黑盒	
		~~快速梯度符号法~~	~~基本迭代法~~	~~深度欺骗~~	~~Carlini-Wagner攻击~~	~~JSMA~~	~~局部搜索~~	~~HSJA~~
MNIST	FGSM	0.232	0.580	0.336	0.412	0.202	0.058	0.384
	BIM	0.232	0.583	0.359	0.408	0.204	0.061	0.384
	DF	0.266	0.503	0.438	0.887	0.490	0.117	0.827
CIFAR-10	FGSM	0.448	0.834	0.478	0.568	0.044	0.001	0.210
	BIM	0.451	0.833	0.479	0.568	0.045	0.001	0.208
	DF	0.367	0.585	0.778	0.978	0.149	0.008	0.342
ImageNet	FGSM	0.316	0.618	0.446	0.469	0.006	0.001	0.005
	BIM	0.324	0.619	0.446	0.476	0.006	0.001	0.005
	DF	0.319	0.500	0.780	0.842	0.029	0.001	0.011

前者表示对1,000个对抗样本取平均后的 $\|\boldsymbol{\delta}\|_2$，即攻击原始样本时改变标签所需的扰动量；后者表示对1,000个校正后的对抗样本取平均后的$\|\boldsymbol{\delta}'\|_2$，即校正对抗样本时纠正标签所需的扰动量。请注意，表4中的所有数值均乘以 10^{-3}。通过比较两张表中的扰动量，我们确认后者极小。

如表3所示，使用局部搜索生成的对抗样本，其范数确实大于使用其他攻击方法生成的对抗样本。在 ImageNet 上由 HSJA 生成的对抗样本范数较大，这可能与 HSJA 偶尔无法为某些分类器或输入找到合适的对抗样本有关 [23], [52]。值得注意的是，即使面对扰动幅度较大的对抗样本（即与原始图像偏差显著的对抗样本），所提方法仍能成功估计原始输入的标签。

4）重新攻击方向适当性的分析

继第IV-A3节之后，本节将探讨为什么我们的方法能够有效修正黑盒攻击生成的对抗样本

表6。 实验1b：校正成功率（定向攻击）。

（a）MNIST 中的成功率

攻击方法	再次攻击方法	目标标签 前二	前三	前四	前五
FGSM	FGSM	**0.990**	**0.299**	**0.158**	**0.172**
	BIM	**0.990**	0.298	**0.158**	**0.172**
	DF	0.979	0.297	0.157	**0.172**
BIM	FGSM	0.992	0.966	0.934	0.871
	BIM	0.992	0.966	0.932	0.871
	DF	**0.994**	**0.970**	**0.940**	**0.878**
CW	FGSM	0.993	**0.982**	0.961	**0.948**
	BIM	0.993	0.979	0.957	0.941
	DF	**0.995**	0.980	**0.962**	0.937
JSMA	FGSM	**0.988**	**0.955**	0.936	**0.870**
	BIM	**0.988**	**0.955**	**0.937**	**0.870**
	DF	0.976	0.937	0.916	0.859
HSJA	FGSM	0.994	**0.971**	0.952	**0.928**
	BIM	0.998	0.968	0.956	0.926
	DF	**0.999**	0.969	**0.958**	0.926

(b) CIFAR-10 中的成功率

攻击方法	再次攻击方法	目标标签 前二	前三	前四	前五
FGSM	FGSM	0.957	0.270	0.152	**0.122**
	BIM	0.957	**0.271**	0.152	**0.122**
	DF	**0.966**	**0.271**	**0.153**	**0.122**
BIM	FGSM	0.997	**1.000**	**0.908**	**0.856**
	BIM	**0.998**	0.937	0.898	0.848
	DF	0.994	0.936	0.901	0.847
CW	FGSM	0.997	**0.945**	**0.915**	**0.877**
	BIM	**0.998**	0.937	0.898	0.848
	DF	0.997	0.934	0.898	0.845
JSMA	FGSM	0.993	**0.896**	0.826	**0.767**
	BIM	0.993	0.895	**0.827**	0.766
	DF	**0.995**	0.891	0.823	0.758
HSJA	FGSM	0.997	**0.939**	**0.906**	**0.863**
	BIM	**0.999**	0.927	0.889	0.838
	DF	0.992	0.920	0.882	0.828

(c) ImageNet 上的成功率

攻击方法	再次攻击方法	目标标签 前二	前三	前四	前五
FGSM	FGSM	**0.917**	**0.390**	**0.251**	**0.156**
	BIM	0.915	0.388	0.242	0.142
	DF	0.915	0.385	0.241	0.144
BIM	FGSM	0.997	**0.952**	**0.915**	**0.891**
	BIM	0.997	0.924	0.895	0.857
	DF	**0.998**	0.923	0.892	0.853
CW	FGSM	0.972	0.891	**0.849**	**0.833**
	BIM	0.977	0.878	0.830	0.806
	DF	**0.987**	**0.892**	0.831	0.813
JSMA	FGSM	**0.997**	**0.914**	**0.877**	**0.852**
	BIM	0.993	0.898	0.853	0.832
	DF	0.993	0.898	0.851	0.829
HSJA	FGSM	0.728	0.602	0.565	0.517
	BIM	**0.820**	**0.692**	**0.637**	**0.568**
	DF	0.818	0.687	0.635	0.567

表7。 实验1b：定向攻击方法生成的对抗样本的扰动量。

数据集	攻击方法	目标标签 前二	前三	前四	前五
MNIST	FGSM	4.133	10.133	12.125	12.626
	BIM	2.499	2.882	3.147	3.394
	CW	1.417	1.649	1.813	1.990
	JSMA	2.707	3.007	3.324	3.497
	HSJA	1.560	1.813	1.992	2.148
CIFAR-10	FGSM	1.064	6.075	8.547	9.179
	BIM	0.266	0.348	0.399	0.444
	CW	0.161	0.211	0.242	0.268
	JSMA	0.699	0.871	0.974	1.061
	HSJA	0.432	0.546	0.607	0.655
ImageNet	FGSM	2.354	12.462	19.405	23.596
	BIM	0.226	0.286	0.309	0.334
	CW	0.160	0.186	0.190	0.201
	JSMA	1.344	1.620	1.737	1.810
	HSJA	39.999	47.810	50.421	53.445

。我们重点验证再攻击扰动方向是否恰当，具体评估其与用于生成对抗样本的原始扰动 $\boldsymbol{\delta}$ 的方向相反的程度。

表5展示了攻击扰动与再攻击扰动反向之间的余弦相似度，其计算方式如下：

$$\mathrm{sim}_{\cos} = \frac{(-\boldsymbol{\delta}, \boldsymbol{\delta}')}{||-\boldsymbol{\delta}|| \cdot ||\boldsymbol{\delta}'||} \quad s.t.\ \boldsymbol{x}_a - \boldsymbol{x} = \boldsymbol{\delta},\ \boldsymbol{x}'_a - \boldsymbol{x}_a = \boldsymbol{\delta}'$$

请注意，再攻击扰动方向已被反转；随着相似度增加，攻击与再攻击的方向将越发相反。在攻击阶段生成对抗样本是为了降低正确类别的置信度，而在再攻击阶段则对对抗样本进行校正，以降低误识别类别的置信度。即使在攻击和再攻击阶段采用相同的白盒攻击技术，所得扰动也不一定朝相反方向。

表5显示，通过沿梯度方向搜索扰动的方法（快速梯度符号法、基本迭代法、深度欺骗、以及Carlini-Wagner攻击）生成的对抗样本具有较高的余弦相似度，这与扰动的高维特性相符。这表明，对抗样本的修正方向与原始攻击方向相反，从而证实了我们在第III-C节中的见解。

逐像素生成贪心扰动的方法（JSMA、局部搜索）的相似度低于基于梯度的方法。这是因为局部扰动虽然抑制了可见性，但并未生成尽可能短距离的扰动，例如沿再攻击的反向方向生成的扰动。

有趣的是，不利用梯度的 HSJA 在 MNIST 和 CIFAR-10 上表现出较高的相似度值。这可能归因于其能够在以连接输入点和起始点的线段为中心的小区域内搜索最优扰动。HSJA 在 ImageNet 上的余弦相似度更接近于零，这归因于像素数量增加导致搜索空间扩大以及决策边界变得更加复杂。即使相似度接近于零，意味着其与原始攻击扰动方向正交，只要对抗样本存在于向非对抗区域凸出的凸对抗区域中，所提方法仍可通过自动调整扰动量来校正该对抗样本。

以这种方式使用所提方法，或许表明它有潜力成为一种表征对抗样本的指标，尤其是因为它不依赖于特定任务或数据模态；不过，这仍需进一步研究。

B. 实验1b：定向攻击

1) 设置

在实验1b中，我们验证了所提方法对通过定向攻击生成的对抗样本的有效性。如第 III-B 节所述，校正由定向攻击生成的对抗样本，预计比校正由非定向攻击生成的对抗样本更具挑战性。这是因为定向攻击生成的对抗样本通常与原始输入相距更远。

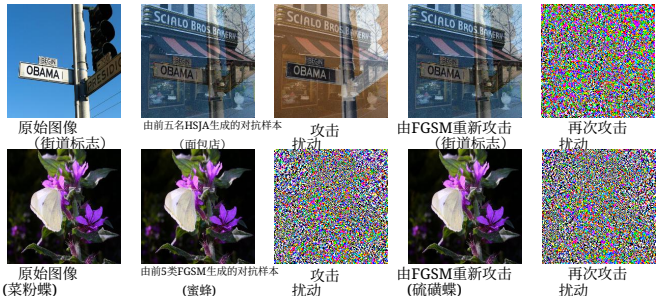
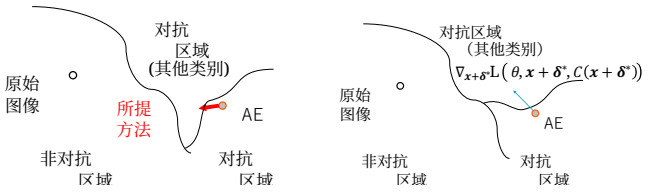


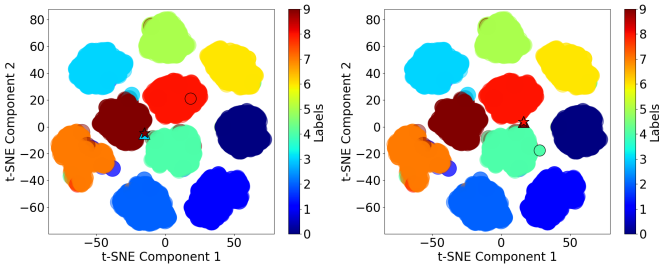
图 6。 实验1b中针对前五名标签的攻击生成的对抗样本示例。



（a）误分类区域与原始类别区域并不相邻，这是由于存在其他类别区域。

（b）重新攻击的扰动方向并不相邻。等同。

图 7。 针对定向攻击生成的对抗样本进行校正失败的情况，这些攻击会诱导样本误分类为低排名类别。



（a）误分类区域与原始类别区域并不相邻，这是由于存在其他类别区域。

(b) 扰动方向不适合重新攻击。

图 8. 针对低排名类别进行定向攻击所生成的对抗样本校正失败时的可视化结果。

在这项实验测试中，我们将分类模型输出中的 Top-2 至 Top-5 标签用作目标标签[1]；随着排名下降，这种方法预计会增加校正难度。数据集、分类模型和评估标准均与实验 1a 保持一致。请注意，与实验 1a 相同，由于正确类别未知，我们的方法在实验 1b 中进行了非定向再攻击。

采用了多种攻击类型来生成对抗样本，包括快速梯度符号法、基本迭代法、Carlini-Wagner攻击、JSMA 和 HSJA。由于局部搜索的算法行为以及 FoolBox 框架的实现限制，本实验未采用局部搜索。尽管 HSJA 攻击不依赖置信度分数，但仍可通过从属于目标类别的样本开始，并在保持分类结果 [2] 的同时最小化扰动来执行定向攻击。攻击和重新攻击参数按照实验 1a 进行配置。

[1]请注意，Top-1 对应于输入的正确标签。

2）校正性能的结果

表6展示了校正成功率。与实验 1a中的非定向攻击相比，实验1b得到的校正成功率较低，但在许多条件下仍然较高。所提出的方法成功估计了由基本迭代法、Carlini-Wagner攻击和JSMA在三个数据集上生成的对抗样本的正确标签，以及由HSJA在MNIST和CIFAR-10上生成的对抗样本的正确标签。随着目标标签从前二变为前五，原始输入与对抗样本之间的扰动增大，导致成功率降低。

表7展示了原始输入与对抗样本之间的平均扰动量。我们观察到，在所有数据集上，快速梯度符号法针对前三或更低排名目标标签生成的对抗样本，以及HSJA在ImageNet上生成的对抗样本，相较于其他方法生成的样本都包含较大的扰动。因此，它们对应的成功率低于其他条件下生成的对抗样本。快速梯度符号法定向攻击尤其对所提出的方法构成威胁。与其他迭代优化方法不同，快速梯度符号法沿直线朝目标类别添加扰动，导致对抗样本移动到特征空间中无法校正的区域。

图 6 展示了两个示例：我们将纠正方法应用于以 Top-5 标签为目标生成的对抗样本。第一行的示例是在先由 HSJA 发动初始攻击后，再使用快速梯度符号法进行攻击。该示例表明，即使对原始输入添加了较大的扰动，对抗样本仍能得到正确校正。与此同时，第二行的示例是在先使用快速梯度符号法发动初始攻击后，再使用快速梯度符号法进行攻击。在这种情况下，尽管对抗样本的扰动并不明显，我们的方法仍未能校正标签。

3）使用 t-SNE 可视化特征空间

将被误分类到排名较低类别（如 Top-4 或 Top-5）的对抗样本进行校正之所以困难，可归因于图7(a)和图7(b)所示的因素。在图7(a)中，对抗样本位于预测类别$C(\boldsymbol{x}+\boldsymbol{\delta})$与原始类别$C(\boldsymbol{x})$不相邻的区域，中间隔着其他类别区域。此外，图7(b)表明，即使$C(\boldsymbol{x}+\boldsymbol{\delta})$与$C(\boldsymbol{x})$相邻，不恰当的再次攻击扰动 $-\nabla_{\boldsymbol{x}+\boldsymbol{\delta}^*}L(\theta, \boldsymbol{x}+\boldsymbol{\delta}^*, C(\boldsymbol{x}+\boldsymbol{\delta}^*))$ 也可能使扰动远离原始类别，从而妨碍校正。

在实验1b中使用 t-SNE [54]，我们将对抗样本及其校正结果在特征空间中的位置关系可视化，以调查对抗样本校正失败的情况

[2]在本实验中，HSJA 的定向攻击从一个被分类为目标类别的随机选取的样本开始。

表 8. 实验 2a：使用可解释人工智能与最先进的纠正方法进行比较 **[12]**和图像变换方法。

数据集	方法	方法	攻击方法 FGSM BIM CW (L_∞) (L_2) (L_∞) (L_2)
MNIST	输入变换	去噪自编码器 [53]	0.500 0.760 0.581 0.621
		JPEG 压缩 [10]	0.037 0.111 0.095 0.000
		全图高斯模糊 [12]	0.389 0.616 0.459 0.389
		随机像素替换 [12]	0.280 0.320 0.280 0.260
	基于XAI的校正	先前方法（Kao 等人） [12]	0.889 0.949 0.905 0.972
	基于再攻击的纠正	所提方法（快速梯度符号法）	**0.999 0.996 0.999 1.000**
		所提方法（基本迭代法）	0.998 **0.996 0.999 1.000**
		所提方法（深度欺骗）	0.993 0.992 0.998**1.000**
CIFAR-10	输入变换	去噪自编码器 [53]	0.455 0.446 0.617 0.731
		JPEG 压缩 [10]	0.093 0.000 0.051 0.404
		全图高斯模糊 [12]	0.279 0.271 0.322 0.277
		随机像素替换 [12]	0.190 0.120 0.180 0.250
	基于XAI的校正	先前方法（Kao 等人） [12]	0.581 0.616 0.729 0.936
	基于再攻击的纠正	所提方法（快速梯度符号法）	0.990 **0.997 1.000 1.000**
		所提方法（基本迭代法）	**0.992 0.997 1.000 1.000**
		所提方法（深度欺骗）	0.991 0.995 0.997 0.998

即被误分类为低秩类别。图 8 展示了 CIFAR-10 数据集中的一些案例：使用快速梯度符号法生成的排名前五的对抗样本无法通过快速梯度符号法成功校正。每个类别以不同颜色标示，所有分类器能够正确分类的良性样本均以大圆圈表示，以指示伪分类区域。原始输入以圆圈表示，对抗样本以三角形表示，再次攻击的对抗样本以星形表示；填充颜色表示类别。例如，如果圆圈和星形颜色相同，则表示校正成功。

图 8 中的示例表明，生成的对抗样本（浅蓝色类别）无法通过再次攻击转移到原始类别区域（红色），而是迁移到了误分类类别与原始类别之间的另一个类别区域（棕色），这对应于图 7(a) 所示的情况。同样，图 8(b) 展示了一个示例：尽管对抗样本靠近原始类别区域（浅蓝绿色），再次攻击后却被错误地移入另一个类别区域（红色），这属于图 7(b) 所示的情况。

C. 实验 2A：使用可解释人工智能与最先进的纠正方法进行比较

除了四种图像变换方法——去噪自编码器 [53]，JPEG 压缩 [10]，全图高斯模糊 [12]，和随机像素替换 [12]——我们还将所提方法与利用可解释人工智能的方法进行了比较。该方法由 Kao 等人 [12],此前提出，是一种用于纠正检测到的对抗样本的最先进方法。去噪自编码器采用了 Meng 等人提出的一种名为 MagNet 的防御方法中的 reformer 模块。Dziugaite 等人提出的 JPEG 压缩是一种通过去除图像中的高频分量来实现防御的方法。全图高斯模糊是一种简单的图像模糊滤镜，它使用正态分布计算图像中每个像素的变换。随机像素替换会将对抗样本中随机选取的一些像素替换为黑色。

在本次实验评估中，我们采用快速梯度符号法、基本迭代法（L_2,L_∞）和 Carlini-Wagner攻击作为生成对抗样本的攻击类型，在 MNIST 和 CIFAR-10 数据集上进行非定向攻击。对抗样本生成的攻击参数根据先前的工作 [12]进行配置。MNIST 和 CIFAR-10 的分类模型及评估标准与实验 1a 保持一致。

表 8 列出了实验 2 的结果，其中我们参考了符合参考文献 [12] 所述条件的可解释人工智能方法的最佳结果。四种输入变换方法的结果同样取自参考文献 [12][3]。

与四种输入变换方法相比，我们提出的方法展现出更优的校正性能，足以掩盖实验设置中的任何细微差异。为了将我们的方法与 Kao 等人的方法进行比较，我们重点关注不同攻击方法下的相对成功率，因为测试样本存在差异，难以进行严格比较。由于错误解读，Kao 等人提出的方法往往会降低由快速梯度符号法和基本迭代法生成的对抗样本的校正成功率；与之相比，在 CIFAR-10 上，由 Carlini-Wagner攻击生成的对抗样本校正成功率更高，这一现象尤为明显。这可能是因为，如表3所示，由基本迭代法和快速梯度符号法生成的对抗样本包含更大的扰动，且可能比由 Carlini-Wagner攻击生成的对抗样本距离决策边界更远。相反，对于所有攻击，所提方法均表现出较高且稳定的校正性能，其成功率不会因数据集或攻击方法与再攻击方法的组合而发生显著变化。

[3]我们尽可能尝试使实验条件与其保持一致，但由于样本和分类模型权重等细节未予公开，因此无法复现完全相同的条件。即使使用相同的数据集和深度神经网络模型，样本及模型训练细节方面的差异仍应纳入考虑。

表 9. 实验 2B：与 RS&V 的比较 **[13],** 当前最先进的自然语言处理深度神经网络对抗防御方法。

数据集	防御方法	攻击方法		
		白盒		黑盒
		快速梯度符号法 基本迭代法 投影版本 Carlini-Wagner攻击 JSMA		局部搜索 HSJA
MNIST	所提方法（快速梯度符号法）	**0.999**	**0.999 0.978 1.000 0.993**	**0.938 1.000**
	RS&V($p = 0.001$)	0.002	0.078 0.530 0.169 0.000	0.000 0.551
	RS&V($p = 0.01$)	0.060	0.201 0.532 0.809 0.005	0.001 0.963
	RS&V($p = 0.1$)	0.579	0.649 0.542 0.998 0.083	0.006 0.996
	RS&V($p = 1.0$)	0.918	0.990 0.576 0.999 0.298	0.038 0.999
	RS&V($p = 10.0$)	0.548	0.770 0.531 0.775 0.407	0.206 0.650
CIFAR-10	所提方法（快速梯度符号法）	**0.992**	**1.000 1.000 1.000 0.994**	**0.911 1.000**
	RS&V($p = 0.001$)	0.001	0.139 0.032 0.038 0.004	0.001 0.525
	RS&V($p = 0.01$)	0.006	0.288 0.078 0.138 0.008	0.003 0.552
	RS&V($p = 0.1$)	0.079	0.787 0.457 0.788 0.072	0.016 0.718
	RS&V($p = 1.0$)	0.570	0.797 0.696 0.818 0.589	0.170 0.683
	RS&V($p = 10.0$)	0.139	0.152 0.147 0.149 0.149	0.040 0.144
ImageNet	所提方法（快速梯度符号法）	**0.926**	**0.991 0.999 0.994 0.999**	**0.981 0.997**
	RS&V（$p = 0.001$）	0.003	0.113 0.076 0.008 0.013	0.007 0.633
	RS&V($p = 0.01$)	0.004	0.324 0.180 0.023 0.015	0.009 0.688
	RS&V($p = 0.1$)	0.040	0.974 0.722 0.349 0.100	0.039 0.629
	RS&V($p = 1.0$)	0.761	0.988 0.973 0.886 0.701	0.266 0.772
	RS&V($p = 10.0$)	0.909	0.919 0.921 0.921 0.891	0.421 0.823

D. 实验*2B*：与自然语言处理中最先进的防御方法的比较

接下来，我们将所提方法与 RS&V [13], 进行基准比较。RS&V 是一种旨在保护自然语言处理中的深度神经网络免受对抗样本攻击的防御方法。RS&V 是一种推理时防御方法，与我们的方法类似，可通过重新攻击输入，将对抗样本校正回其原始输入类别。它通过将文本对抗样本中的词语替换为同义词，生成k 个相似句子，并根据这些句子的分类结果一致比例进行对抗样本检测和校正。由于 RS&V 的扰动方法针对语言模态设计，我们对 RS&V 中的再攻击过程进行了修改，使其向输入图像的所有像素添加随机噪声。尽管第 II-E 节中还讨论了其他方法，但由于 RS&V 无需预训练，且适用于与所提方法相似的条件，因此我们选择 RS&V 作为比较对象。

实验 2b 中使用的改进版 RS&V 生成了$k = 25$个含有随机噪声的不同派生图像，其噪声大小固定在L_2 范数p下。选择$k = 25$的依据是实证结果，即 RS&V 中派生样本的最优数量 [13]。该方法通过确定k 个派生图像中的多数分类结果来校正对抗样本。

所提方法的实验设置与第 IV-A 节相同，并采用快速梯度符号法进行重新攻击。RS&V 参数p的取值范围为 0.0001 至 10.0。

表 9 展示了与 RS&V 的比较结果。结果表明，在某些条件下，RS&V 能够纠正白盒攻击生成的 90% 以上的对抗样本，尤其是基本迭代法和 Carlini-Wagner攻击生成的对抗样本；而 LS 对对抗样本的纠正成功率极低，MNIST 上为 21%，CIFAR-10 上为 17%，ImageNet 上为 42%。相比之下，我们所提方法在所有条件下均优于 RS&V，并在所有数据集上纠正了 LS 生成的 90% 以上的对抗样本。

请注意，如附录A所示，我们所提方法几乎无需调整控制参数。相反，RS&V 需要根据问题适当调整p。参数p取值较小时，采样区域远离$C(\boldsymbol{x})$ 所在区域，因此会妨碍 RS&V 对对抗样本的校正；而p 取值较大时，由于从$C(\boldsymbol{x})$ 以外的类别中采样了许多区域，也会妨碍校正。

由于 RS&V 添加了随机噪声，当对抗样本位于可近似为超平面的决策边界附近时，校正成功的概率预计为 50%。因此，RS&V 的校正成功率预计较低；然而，根据参数 p、数据集和攻击方法的不同，它有时在应对白盒攻击时表现出色，这与我们的直觉相反。这意味着，以对抗样本为中心、半径为 p 的超球面上有许多点位于原始类别$C(\boldsymbol{x})$ 的区域内，这表明对抗样本被一个凸决策边界所包围，该边界向$C(\boldsymbol{x})$ 的区域凸出。

E. 实验 *3*：与检测器的协同作用

在实验 3 中，我们评估了所提方法作为后处理步骤集成到传统对抗样本检测器（具体而言是 A^{2}D [11]）时的性能。由于我们的方法旨在与检测器结合使用，如果检测器未能识别出对抗样本，误将其分类为良性样本，我们的方法就会无意中对该良性样本再次发起攻击，从而生成新的对抗样本。这是所提方法设计本身固有的局限性。因此，实验 3 旨在展示我们的方法与

表 10。 实验 3：A 的检测准确率^2D 使用 Z 分数。

数据集	检测器	检测准确率 快速梯度符号法	基本迭代法	JSMA	Carlini-Wagner攻击	平均值$_a$	bng
MNIST	BIM	1.000	1.000	0.999	1.000	0.999	0.862
	基本迭代法(L_2)	1.000	1.000	0.998	1.000	0.999	0.827
	JSMA	1.000	0.997	0.999	1.000	0.999	0.881
	DBA	0.940	0.969	0.926	0.988	0.955	0.919
CIFAR-10	基本迭代法	0.812	0.997	0.956	0.996	0.940	0.866
	基本迭代法(L_2)	0.834	0.997	0.963	0.996	0.947	0.856
	JSMA	0.843	0.999	0.971	1.000	0.953	0.857
	DBA	0.524	0.961	0.651	0.967	0.775	0.964
ImageNet	BIM	1.000	0.999	1.000	0.975	0.993	0.876
	基本迭代法(L_2)	0.999	0.998	1.000	0.968	0.991	0.882
	JSMA	0.986	0.996	1.000	0.979	0.990	0.865
	DBA	0.940	0.997	0.970	0.965	0.968	0.886

表 11。 实验 3：所提方法与 A^2D使用 Z 分数时的分类器准确率。

数据集	检测器	再次攻击 方法	攻击方法 快速梯度符号法	基本迭代法	JSMA	Carlini-Wagner攻击	Avg$_a$	bng
MNIST	BIM	FGSM	**0.975**	0.983	**0.954**	**1.000**	**0.978**	0.862
		BIM	**0.975**	**0.984**	0.951	**1.000**	0.976	0.862
		DF	0.899	0.958	0.915	0.999	0.943	0.862
	BIM (L_2)	FGSM	**0.975**	0.983	0.953	**1.000**	**0.978**	0.827
		BIM	**0.975**	**0.984**	0.951	**1.000**	**0.978**	0.827
		DF	0.887	0.958	0.902	0.997	0.936	0.827
	JSMA	FGSM	**0.975**	0.981	**0.954**	**1.000**	**0.978**	0.881
		BIM	**0.975**	0.982	0.951	**1.000**	0.977	0.881
		DF	0.910	0.953	0.910	0.998	0.943	0.881
	DBA	FGSM	0.917	0.953	0.882	0.988	0.935	**0.919**
		BIM	0.917	0.954	0.879	0.988	0.935	**0.919**
		DF	0.845	0.930	0.847	0.986	0.902	**0.919**
CIFAR-10	BIM	FGSM	0.632	0.986	0.936	**0.995**	0.887	0.866
		BIM	0.631	0.990	0.935	**0.995**	0.888	0.866
		DF	0.639	0.989	0.939	0.991	0.890	0.866
	BIM (L_2)	FGSM	0.649	0.986	0.942	**0.995**	0.893	0.856
		BIM	0.649	0.990	0.941	**0.995**	0.894	0.856
		DF	0.655	0.989	0.945	0.991	0.895	0.856
	JSMA	FGSM	0.666	0.986	0.948	**0.995**	0.899	0.857
		BIM	0.663	0.990	0.948	**0.995**	0.899	0.857
		DF	**0.676**	**0.991**	**0.953**	**0.995**	**0.904**	0.857
	DBA	FGSM	0.416	0.948	0.639	0.962	0.741	**0.964**
		BIM	0.418	0.952	0.639	0.962	0.743	**0.964**
		DF	0.427	0.953	0.639	0.962	0.745	**0.964**
图像-Net	BIM	FGSM	0.954	0.994	**1.000**	0.973	0.980	0.876
		BIM	**0.956**	**0.998**	**1.000**	0.974	**0.982**	0.876
		DF	0.954	0.993	**1.000**	0.971	0.980	0.876
	BIM (L_2)	FGSM	0.953	0.993	**1.000**	0.966	0.978	0.882
		BIM	0.955	0.997	**1.000**	0.967	0.980	0.882
		DF	0.952	0.992	0.999	0.965	0.977	0.882
	JSMA	FGSM	0.940	0.991	**1.000**	0.977	0.977	0.865
		BIM	0.942	0.995	**1.000**	**0.978**	0.979	0.865
		DF	0.940	0.991	0.999	0.974	0.976	0.865
	DBA	FGSM	0.896	0.992	0.970	0.963	0.955	**0.886**
		BIM	0.898	0.996	0.970	0.962	0.957	**0.886**
		DF	0.895	0.990	0.969	0.959	0.953	**0.886**

一个检测器。

在这种情况下，所提方法会纠正 A^2D 检测到的对抗样本。我们采用了参考文献 [11], 中所述的实验设置，其中使用快速梯度符号法、基本迭代法、JSMA 和 Carlini-Wagner攻击作为攻击方法来生成对抗样本。我们基于公开代码实现了 MNIST 和 CIFAR-10 的分类模型，并针对 ImageNet 数据集实现了一个基于 ResNet-101 [55] 的模型。这些分类器按照根据参考文献 [11], 中的实验设置进行配置，与实验 1 和实验 2 所采用的设置不同。此外，我们配置了 A^2D 检测器，使用基本迭代法、基本迭代法 (L_2)、JSMA 和 DBA 攻击输入，并采用 Z 分数和 k 近邻评估输入的鲁棒性。

表 10 和表 11 分别展示了 A^2D 结合 Z 分数时的检测准确率，以及配备 A^2D 和所提方法的分类器的分类准确率。Avg$_a$ 和 bng 分别表示由四种攻击方法生成的对抗样本的平均分类准确率，以及良性样本的准确率。

对 A^2D中各种攻击方法的比较表明，使用 DBA 时良性样本的准确率最高，而针对对抗样本的准确率则相对较低。其他三种方法（基本迭代法、基本迭代法 (L_2) 和 JSMA）在各种条件下均表现出较强的对抗样本防御能力。对于良性样本，我们的方法的分类准确率与 A^2D正确分类这些样本的比率相当。请注意，表11中的分类准确率无法超过表 10 中的检测准确率，因为所提方法仅修正 A^2D检测到的对抗样本。例如，当使用 DBA 作为检测器、DF 作为再攻击方法时，CIFAR-10 上的 Avg$_a$ 为 0.745，与其他条件相比似乎较低。然而，这可归因于表 10 所示的 A^2D 检测准确率为 0.775。也就是说，在这种条件下，96% 的输入对抗样本都得到了成功修正。这些发现表明，所提方法能够修正对抗样本，同时不会降低检测器的性能。当我们的方法与使用 k 近邻的 A^2D结合时，观察到的趋势与 A^2D结合 Z 分数和我们的方法时相同。

F. 实验 *4*：应用于语音识别

为展示所提方法在图像模态之外的适用性，我们在用于语音识别的深度神经网络上开展了实验 4，以修复对抗样本。BC-ResNet-8 [56],a 是一种卷积神经网络模型，以其在标准音频分类数据集上的高准确率而闻名，并被用作受害模型。我们采用了包含 10 个语音命令类别的 Google Speech Commands 数据集 [57]，从中选取 1,000 个样本（每个类别 100 个），这些样本的原始音频均被准确识别，且对抗攻击均成功。用于生成对抗样本的攻击方法与 FoolBox 中列出的相同，也与实验 1 至实验 3 所用方法一致。鉴于 LS 攻击在 FoolBox 框架下的成功率仅为 0.3%，本研究仅关注白盒攻击。重新攻击参数与实验 1 至实验 3 中使用的参数一致，如表 1 所示。

表12详细列出了使用所提方法校正对抗样本的成功率，表13则展示了重新攻击期间的扰动量。所有攻击方法的校正成功率均超过97%，

表12. 实验 4：语音识别中的校正性能。

数据集	再次攻击方法	攻击方法 ~~快速梯度符号法 基本迭代法 深度欺骗 Carlini-Wagner攻击 JSMA~~
Google 语音命令	FGSM	0.979 0.998 0.997 1.000 1.000
	BIM	0.979 0.997 0.996 1.000 1.000
	DF	0.979 0.998 0.998 1.000 1.000

表13。 实验 4：语音识别中重新攻击的扰动量。

数据集	再次攻击方法	攻击方法 ~~快速梯度符号法 基本迭代法 深度欺骗 Carlini-Wagner攻击 JSMA~~
Google 语音命令	FGSM	0.158 0.135 0.316 0.082 0.240
	BIM	0.114 0.077 0.222 0.018 0.195
	DF	0.052 0.036 0.098 0.011 0.094

且所需扰动极小，这凸显了所提方法对于语音识别深度神经网络的鲁棒性。

V. 结论

本研究介绍了一种简单而有效的方法，通过重新攻击对抗样本，使其原始输入获得正确的分类结果，从而纠正对抗样本。所提方法利用对抗样本脆弱性来纠正对抗样本，因此无论输入信号类型如何，例如图像或音频，均可应用于深度神经网络。通过一系列实验，我们成功证明，与传统方法相比，所提方法在纠正由各种攻击方法生成的对抗样本时更加稳定。研究结果凸显了所提方法对黑盒攻击和定向攻击所生成对抗样本的有效性。

未来，预计将重点研究如何扩大所提方法的应用范围，将其拓展至包括语言在内的各种模态。此外，我们还将探究所提方法作为对抗样本特征指标的可行性。

附录A 所提方法对参数取值的鲁棒性

由于所提方法专门用于校正对抗样本，因此在重新攻击时，其运行不受所用控制参数具体设置的影响。为验证这一点，我们在该方法中使用快速梯度符号法进行重新攻击时，对参数 ϵ 进行了调整并开展实验。

表 14 展示了在 CIFAR-10 上，当 ϵ设置为 0.001、0.01、0.1、1.0 和 10.0 时的校正成功率。与表 9 所示的 RS&V 相比，即使 ϵ 偏离了 FoolBox 中 1.0 的默认值，我们的方法仍能保持较高的成功率。

表 14. 使用快速梯度符号法的所提方法在控制参数 ϵ下的鲁棒性。

ϵ 在快速梯度符号法中用于再次攻击	攻击方法 白盒 ~~快速梯度符号法 基本迭代法 深度欺骗 Carlini-Wagner攻击 JSMA~~	黑盒 局部搜索 HSJA
0.001	0.992 0.999 0.981 1.000 0.994 0.762	1.000
0.01	0.992 1.000 1.000 1.000 0.994 0.911	1.000
0.1	0.992 1.000 1.000 1.000 0.994 0.911	1.000
1.0	0.992 1.000 1.000 1.000 0.994 0.911	1.000
10.0	0.993 0.999 1.000 1.000 0.995 0.914	0.999

参考文献

[1] Christian Szegedy、Wojciech Zaremba、Ilya Sutskever、Joan Bruna、Dumitru Erhan、Ian Goodfellow 和 Rob Fergus。《神经网络的有趣特性》。*arXiv* 预印本 *arXiv:1312.6199*，2013年。[2] Ranjie Duan、Xiaofeng Mao、A Kai Qin、Yuefeng Chen、Shaokai Ye、Yuan He 和 Yun Yang。《对抗性激光束：瞬间对深度神经网络发起有效的物理世界攻击》。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 16062–16071 页，2021年。

[3] Qi Sun、Arjun Ashok Rao、Xufeng Yao、Bei Yu 和 Shiyan Hu。自动驾驶中的对抗攻击反制。载于《国际计算机辅助设计会议第*39*届论文集》，第1–7页，2020年。[4] Abhiram Gnanasambandam、Alex M Sherman 和 Stanley H Chan。光学对抗攻击。载于《*IEEE/CVF*国际计算机视觉会议论文集》，第92–101页，2021年。[5] Takami Sato、Sri Hrushikesh Varma Bhupathiraju、Michael Clifford、Takeshi Sugawara、Qi Alfred Chen 和 Sara Rampazzi。不可见的反射：利用红外激光反射攻击交通标志感知。*arXiv*预印本 *arXiv:2401.03582*，2024年。[6] Yisroel Mirsky 和 Wenke Lee。深度伪造的制作与检测：一项综述。*ACM*计算调查（*CSUR*），54(1):1–41，2021年。[7] Shahroz Tariq、Sowon Jeon 和 Simon S Woo。我是真人还是虚假的名人？评估人脸识别与验证API在深度伪造冒充攻击下的表现。载于*2022*年*ACM*网络会议论文集，第512–523页，2022年。[8] Le Qin、Fei Peng、Min Long、Raghavendra Ramachandra 和 Christoph Busch。无人值守人脸验证系统易受基于面部组件的呈现攻击影响：一项实证研究。*ACM*隐私与安全学报，25(1):1–28，2021年。[9] Cihang Xie、Jianyu Wang、Zhishuai Zhang、Zhou Ren 和 Alan Yuille。通过随机化减轻对抗效应。*arXiv*预印本 *arXiv:1711.01991*，2017年。[10] Gintare Karolina Dziugaite、Zoubin Ghahramani 和 Daniel M Roy。研究JPG压缩对对抗图像的影响。*arXiv*预印本 *arXiv:1608.00853*，2016年。[11] Zhe Zhao、Guangke Chen、Jingyi Wang、Yiwei Yang、Fu Song 和 Jun Sun。以攻为守：利用鲁棒性刻画对抗样本。载于第*30*届*ACM SIGSOFT*国际软件测试与分析研讨会论文集，第42–55页，2021年。[12] Ching-Yu Kao、Junhao Chen、Karla Markert 和 Konstantin Böttinger。使用可解释人工智能技术纠正对抗输入。载于*2022*年第*30*届欧洲信号处理会议（*EUSIPCO*），第573–577页。IEEE，2022年。[13] Xiaosen Wang、Xiong Yifeng 和 Kun He。通过随机替换与投票检测文本对抗样本。载于人工智能中的不确定性，第2056–2065页。PMLR，2022年。[14] Heng Yang 和 Ke Li。最好的防御就是攻击：修复文本对抗样本中的语义。*arXiv*预印本 *arXiv:2305.04067*，2023年。[15] Ian J Goodfellow、Jonathon Shlens 和 Christian Szegedy。解释并利用对抗样本。*arXiv*预印本 *arXiv:1412.6572*，2014年。[16] Alexey Kurakin、Ian J Goodfellow 和 Samy Bengio。物理世界中的对抗样本。载于人工智能安全与安保，第99–112页。Chapman and Hall/CRC，2018年。[17] Aleksander Madry、Aleksandar Makelov、Ludwig Schmidt、Dimitris Tsipras 和 Adrian Vladu。迈向能够抵御对抗攻击的深度学习模型。*arXiv*预印本 *arXiv:1706.06083*，2017年。[18] Seyed-Mohsen Moosavi-Dezfooli、Alhussein Fawzi 和 Pascal Frossard。DeepFool：一种欺骗深度神经网络的简单而准确的方法。载于*IEEE*计算机视觉与模式识别会议论文集，第2574–2582页，2016年。[19] Nicholas Carlini 和 David Wagner。迈向神经网络鲁棒性的评估。载于*2017*年*IEEE*安全与隐私研讨会（*SP*），第39–57页。IEEE，2017年。[20] Nicolas Papernot、Patrick McDaniel、Somesh Jha、Matt Fredrikson、Z Berkay Celik 和 Ananthram Swami。深度学习在对抗环境中的局限性。载于*2016*年*IEEE*欧洲安全与隐私研讨会（*EuroS&P*），第372–387页。IEEE，2016年。

[21] Nina Narodytska 和 Shiva Prasad Kasiviswanathan。用于深度网络的简单黑盒对抗扰动。arXiv预印本 arXiv:1612.06299，2016年。[22] Wieland Brendel、Jonas Rauber 和 Matthias Bethge。基于决策的对抗攻击：针对黑盒机器学习模型的可靠攻击。arXiv预印本 arXiv:1712.04248，2017年。[23] Jianbo Chen、Michael I Jordan 和 Martin J Wainwright。HopSkipJumpAttack：一种查询高效的基于决策的攻击。载于2020年IEEE安全与隐私研讨会（SP），第1277–1294页。IEEE，2020年。[24] Ali Shafahi、Mahyar Najibi、Mohammad Amin Ghiasi、Zheng Xu、John Dickerson、Christoph Studer、Larry S Davis、Gavin Taylor 和 Tom Goldstein。免费的对抗训练！《神经信息处理系统进展》，32，2019年。[25] Hongyang Zhang、Yaodong Yu、Jiantao Jiao、Eric Xing、Laurent El Ghaoui 和 Michael Jordan。鲁棒性与准确率之间具有理论依据的权衡。载于《国际机器学习会议》，第7472–7482页。PMLR，2019年。[26] Ying Meng、Jianhai Su、Jason O'Kane 和 Pooyan Jamshidi。Athena：一个基于多种弱防御构建对抗防御的框架。arXiv预印本 arXiv:2001.00308，2020年。[27] Chuan Guo、Mayank Rana、Moustapha Cisse 和 Laurens Van Der Maaten。使用输入变换反制对抗图像。arXiv预印本 arXiv:1711.00117，2017年。[28] Jacob Buckman、Aurko Roy、Colin Raffel 和 Ian Goodfellow。温度计编码：抵御对抗样本的一种独热方法。载于《国际学习表征会议》，2018年。[29] Reuben Feinman、Ryan R Curtin、Saurabh Shintre 和 Andrew B Gardner。从伪影中检测对抗样本。arXiv预印本 arXiv:1703.00410，2017年。[30] Xingjun Ma、Bo Li、Yisen Wang、Sarah M Erfani、Sudanthi Wijewickrema、Grant Schoenebeck、Dawn Song、Michael E Houle 和 James Bailey。利用局部内在维度刻画对抗子空间。arXiv预印本 arXiv:1801.02613，2018年。[31] Huiyan Wang、Jingwei Xu、Chang Xu、Xiaoxing Ma 和 Jian Lu。Dissector：通过跨层剖析验证深度学习应用的输入。载于2020年IEEE/ACM第42届国际软件工程会议（ICSE），第727–738页。IEEE，2020年。[32] Jingyi Wang、Guoliang Dong、Jun Sun、Xinyu Wang 和 Peixin Zhang。通过模型变异测试检测深度神经网络的对抗样本。载于2019年IEEE/ACM第41届国际软件工程会议（ICSE），第1245–1256页。IEEE，2019年。[33] Matthias Rottmann、Kira Maag、Mathis Peyron、Natasa Krejic 和 Hanno Gottschalk。通过反制攻击检测迭代对抗攻击。arXiv预印本 arXiv:2009.11397，2020年。[34] Simin Chen、Zihe Song、Lei Ma、Cong Liu 和 Wei Yang。Attackdist：通过反制攻击刻画零日对抗样本。2021年。[35] Dan Cireşan、Ueli Meier、Jonathan Masci 和 Jürgen Schmidhuber。用于交通标志分类的多列深度神经网络。《神经网络》，32:333–338，2012年。[36] Hadi Salman、Andrew Ilyas、Logan Engstrom、Sai Vemprala、Aleksander Madry 和 Ashish Kapoor。非对抗样本：为鲁棒视觉设计物体。《神经信息处理系统进展》，34:15270–15284，2021年。[37] Aochuan Chen、Peter Lorenz、Yuguang Yao、Pin-Yu Chen 和 Sijia Liu。用于对抗鲁棒性的视觉提示。载于ICASSP 2023—2023年IEEE国际声学、语音与信号处理会议（ICASSP），第1–5页。IEEE，2023年。[38] Weijia Wang、Chao Zhou、Da Lin 和 Yuan-Gen Wang。FeConDefense：通过特征一致性损失逆转对抗攻击。《计算机通信》，211:263–270，2023年。[39] Chengzhi Mao、Mia Chiquier、Hao Wang、Junfeng Yang 和 Carl Vondrick。自然监督能够逆转对抗攻击。载于IEEE/CVF国际计算机视觉会议论文集，第661–671页，2021年。[40] Maximilian Mozes、Pontus Stenetorp、Bennett Kleinberg 和 Lewis D Griffin。用于检测文本对抗样本的频率引导词语替换。arXiv预印本 arXiv:2004.05887，2020年。[41] Wenqi Wang、Run Wang、Jianpeng Ke 和 Lina Wang。TextFirewall：针对情感分类中的对抗文本进行全方位防御。《IEEE Access》，9:27467–27475，2021年。[42] Edoardo Mosca、Shreyash Agarwal、Javier Rando 和 Georg Groh。"这反应很可疑！"：解读logits变化以检测自然语言处理对抗攻击。arXiv预印本 arXiv:2204.04636，2022年。

[43] Fan Yin、Yao Li、Cho-Jui Hsieh 和 Kai-Wei Chang。Addmu：利用数据和模型不确定性估计检测远边界对抗样本。*arXiv*预印本 *arXiv:2210.12396*，2022年。[44] Hyun Kwon、Hyunsoo Yoon 和 Ki-Woong Park。Acoustic-decoy：通过修改语音识别系统中的音频来检测对抗样本。*Neurocomputing*，417：357–370，2020年。[45] Kun Wang、Xiangyu Xu、Li Lu、Zhongjie Ba、Feng Lin 和 Kui Ren。{FraudWhistler}：一种用于说话人识别的弹性、鲁棒、即插即用的对抗样本检测方法。载于第*33*届 *USENIX* 安全研讨会（*USENIX Security 24*），第7303–7320页，2024年。[46] Nicholas Carlini、Anish Athalye、Nicolas Papernot、Wieland Brendel、Jonas Rauber、Dimitris Tsipras、Ian Goodfellow、Aleksander Madry 和 Alexey Kurakin。关于对抗鲁棒性的评估。*arXiv*预印本 *arXiv:1902.06705*，2019年。[47] Jonas Rauber、Wieland Brendel 和 Matthias Bethge。Foolbox：用于基准测试机器学习模型鲁棒性的 Python 工具箱。*arXiv*预印本 *arXiv:1707.04131*，2017年。[48] Li Deng。用于机器学习研究的手写数字图像 MNIST 数据库。*IEEE*信号处理杂志，29 142，2012年 6 141( ): – 。[49] Alex Krizhevsky、Geoffrey Hinton 等。学习微型图像的多层特征。2009年。[50] Olga Russakovsky、Jia Deng、Hao Su、Jonathan Krause、Sanjeev Satheesh、Sean Ma、Zhiheng Huang、Andrej Karpathy、Aditya Khosla、Michael Bern-stein、Alexander C. Berg 和 Li Fei-Fei。ImageNet大规模视觉识别挑战赛。国际计算机视觉杂志（*IJCV*），115(3)：211–252，2015年。[51] Karen Simonyan 和 Andrew Zisserman。用于大规模图像识别的超深卷积网络。*arXiv*预印本 *arXiv:1409 1556*，2014年。。[52] Viet Quoc Vo、Ehsan Abbasnejad 和 Damith C Ranasinghe。Ramboattack：一种鲁棒且查询高效的深度神经网络决策攻击。*arXiv*预印本 *arXiv:2112.05282*，2021年。[53] Dongyu Meng 和 Hao Chen。Magnet：针对对抗样本的双管齐下的防御。载于*2017*年 *ACM SIGSAC*计算机与通信安全会议论文集，第135–147页，2017年。[54] Laurens Van der Maaten 和 Geoffrey Hinton。使用 t-SNE 可视化数据。机器学习研究杂志，9(11)，2008 年。[55] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，第770–778页，2016年。[56] Byeonggeun Kim、Simyung Chang、Jinkyu Lee 和 Dooyong Sung。用于高效关键词检测的广播式残差学习，2023年。[57] Pete Warden。Speech Commands：用于有限词汇语音识别的数据集，2018年。



FUMIYA MORIMOTO于2023年获得日本鹿儿岛大学工程学学士学位。目前，他是鹿儿岛大学理工学研究科工程系硕士生。其研究兴趣包括人工智能安全，尤其是对抗防御。



RYUTO MORITA于2023年获得日本国立高等专门学校机构鹿儿岛工学院工程学学士学位。目前，他是鹿儿岛大学理工学研究科工程系硕士生。其研究兴趣包括人工智能安全，尤其是面向语音识别的对抗防御。



SATOSHI ONO 于2002年在筑波大学获得工程学博士学位。2001年至2003年，他担任日本学术振兴会（JSPS）研究员。随后，他加入鹿儿岛大学理工学研究科信息与计算机科学系，担任研究副职。他目前是鹿儿岛大学信息科学与生物医学工程系教授。

他曾获2023年日本人工智能学会年度大会奖、IWAIT2020最佳论文奖、TAAI2019优秀论文奖、日本信息处理学会山下SIG研究奖（2012年）等。他是IEEE、日本信息处理学会、电子信息通信学会和日本人工智能学会的成员。目前，他的研究重点是进化计算、机器学习及其在现实世界问题中的应用。