

基于自适应特征优化的对抗鲁棒性增强研究

宋小雪¹, 刘万平¹, 黄东²

(1. 重庆理工大学计算机科学与工程学院, 重庆 400054;
2. 贵州大学现代制造技术教育部重点实验室, 贵州 贵阳 550025)

摘要: 现有对抗鲁棒性增强方法主要通过抑制导致模型预测错误的非鲁棒特征来提升性能, 往往忽略了鲁棒特征与非鲁棒特征的协同利用潜力。为此, 提出一种基于自适应特征优化的对抗样本防御方法, 通过联合处理机制实现鲁棒特征与非鲁棒特征的自适应协同优化。该方法包含特征分离、自适应特征优化和特征融合3个核心阶段。在特征分离阶段, 利用结合随机注意力机制的分离网络将输入特征图精准分离为鲁棒特征和非鲁棒特征; 在自适应特征优化阶段, 从鲁棒特征中筛选高级特征, 并对非鲁棒特征进行重校准, 以挖掘其潜在价值; 在特征融合阶段, 将重新校准后的非鲁棒特征与选择后的鲁棒特征进行融合, 最终生成高效的特征图。实验结果表明, 在SVHN数据集上, 所提方法针对FGSM、PGD-20、PGD-100和C&W 4种对抗攻击的鲁棒性分别提升了11.39%、13.24%、11.79%和2.6%, 集成鲁棒性提高了4.14%, 有效提升了现有对抗训练方法的鲁棒性。

关键词: 特征分离; 自适应特征优化; 特征融合; 对抗鲁棒性

DOI: 10.11907/rjtk.251105

扫描二维码阅读全文:



中图分类号: TP393

文献标识码: A

文章编号: 1672-7800(2026)005-0126-08

Research on Adversarial Robustness Enhancement Based on Adaptive Feature Refinement

SONG Xiaoxue¹, LIU Wanping¹, HUANG Dong²

(1. College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China;
2. Key Laboratory of Advanced Manufacturing Technology of the Ministry of Education, Guizhou University, Guiyang 550025, China)

Abstract: The existing methods to enhance the robustness of the confrontation mainly improve the performance by suppressing the non robust features that lead to model prediction errors, and often ignore the potential of cooperative utilization of robust features and non robust features. For this reason, a defense method against samples based on adaptive feature optimization is proposed, and the adaptive collaborative optimization of robust and non robust features is realized through joint processing mechanism. This method includes three core stages: feature separation, adaptive feature optimization and feature fusion. In the feature separation stage, the separation network combined with random attention mechanism is used to accurately separate the input feature map into robust and non robust features; In the adaptive feature optimization phase, the advanced features are screened from the robust features, and the non robust features are recalibrated to tap their potential value; In the feature fusion stage, the recalibrated non robust features are fused with the selected robust features to finally generate an efficient feature map. The experimental results show that on the SVHN dataset, the robustness of the proposed method against the FGSM, PGD-20, PGD-100 and C&W four counter attacks is improved by 11.39%, 13.24%, 11.79% and 2.6% respectively, and the integrated robustness is improved by 4.14%, effectively improving the robustness of the existing counter training methods.

Key Words: feature separation; adaptive feature optimization; feature fusion; adversarial robustness

收稿日期: 2025-04-02

基金项目: 重庆市自然科学基金项目(cstc2021jcyj-msxmX0594); 重庆理工大学研究生创新项目(gzlcx20243206)

作者简介: 宋小雪(1999-), 女, CCF会员, 重庆理工大学计算机科学与工程学院硕士研究生, 研究方向为对抗样本攻击与防御; 刘万平(1986-), 男, 博士, 重庆理工大学计算机科学与工程学院副教授, 研究方向为网络与信息安全; 黄东(1981-), 男, 博士, 贵州大学现代制造技术教育部重点实验室正高级工程师, 研究方向为网络技术与智能控制。本文通讯作者: 刘万平。

0 引言

随着深度学习理论的持续演进,深度神经网络(Deep Neural Networks, DNNs)在计算机视觉领域取得了令人瞩目的进展,在自动驾驶汽车、医疗诊断等众多应用场景中展现出强大能力^[1-2]。然而,这类模型极易受到精心构造的扰动攻击,这些扰动会致使模型作出错误决策^[3]。在深度学习应用日益广泛的现实背景下,其安全性问题愈发突出,而恶意超链接的肆意传播更是进一步加剧了安全威胁^[4]。鉴于此,提出行之有效的防御策略以应对这些安全挑战,已然成为当前研究领域亟待解决的关键问题。

针对对抗样本攻击所带来的安全挑战,目前已有许多防御技术被提出,例如特征去噪技术、对抗训练以及对抗样本检测等方法^[3,5-7]。对抗训练是目前最成功且应用最广泛的防御策略之一,其通过在训练过程中引入对抗样本来增强模型的鲁棒性。然而,即便采用了对抗训练,像素级的微小对抗扰动仍然会在模型的中间特征空间逐步累积,最终破坏模型的输出。传统对抗防御方法往往侧重于抑制非鲁棒特征,却忽视了其中可能包含的有用判别信息,同时也未能充分利用鲁棒特征中蕴含的有效判别信息。为解决这一问题,近年来的研究开始尝试对非鲁棒特征激活进行分解和去除,以降低对抗样本对模型决策的影响。

1 相关工作

1.1 对抗攻击

对抗攻击一般分为白盒攻击和黑盒攻击两类。白盒攻击指攻击者可以获取目标模型的结构、参数及梯度等信息,而黑盒攻击则指攻击者无法直接获取目标模型的具体信息,仅能通过输入输出结果进行攻击。在白盒攻击设置下,许多方法利用目标模型的梯度信息生成对抗样本,例如快速梯度下降算法(Fast Gradient Sign Method, FGSM)^[3]、投影梯度下降算法(Projected Gradient Descent, PGD)^[6]和C&W(Carlini & Wagner)攻击算法^[8]。

1.2 对抗防御

对抗训练使用对抗攻击生成的对抗样本进行训练,指导模型抵御对抗攻击,被认为是最有效的防御策略之一。可将其表述为极小极大优化问题,用公式表示为:

$$\min_{\theta} \mathbb{E}_{(x,y) \sim D} \left[\max_{\delta} L_{cls}(F_{\theta}(x + \delta), y) \right] \quad (1)$$

式中: θ 为模型参数; x 为数据集 D 的自然图像; y 为 x 的真实标签; δ 表示叠加在样本 x 上的扰动; F_{θ} 为神经网络函数; L_{cls} 为损失函数。

许多模型会从数据中学习非鲁棒特征,而对抗样本的输入扰动往往会在模型的中间层逐步累积,最终误导模型的预测^[9]。因此,有研究尝试通过操作特征激活来减少特

征映射中的异常情况,从而提升模型的鲁棒性。例如,Xie等^[5]提出特征去噪方法,通过传统去噪技术抑制异常激活单元。Xiao等^[10]提出 k -Winner-Takes-All激活机制,仅保留幅度最大的 k 个特征单元并停用其余单元,有效抑制了异常激活。Zoran等^[11]采用注意力机制强化特征图中的关键区域,从而提升模型性能。Bai等^[12]和Yan等^[13]从通道的角度研究扰动对特征激活的影响,并提出通道激活抑制(Channel Activation Suppression, CAS)方法和基于通道重要性的特征选择(Channel-wise Importance-based Feature Selection, CIFS)方法。其中,CAS抑制非鲁棒通道的激活,CIFS选择重要通道进行激活,二者均有效提升了模型的鲁棒性。Kim等^[14]提出特征分离与重新校准方法(Feature Separation and Recalibration, FSR),将特征图分离为鲁棒特征和非鲁棒特征,并对非鲁棒特征进行校准,以重新捕获潜在的判别信息。Liu等^[15]提出解耦视觉特征屏蔽(Decoupled Visual Feature Masking, DFM)方法,利用不同的屏蔽策略自适应地分离视觉判别特征与非视觉特征,并通过适当丢弃信息来破坏对抗性噪声,从而提高模型鲁棒性。

尽管上述方法在一定程度上提升了模型的鲁棒性,但仍然存在一些局限性。例如,特征去激活方法可能会丢失重要判别信息;注意力机制依赖注意力权重的学习效果;特征去噪和通道级优化方法往往未能充分考虑鲁棒特征与非鲁棒特征的协同优化。为此,本文提出一种新的联合特征处理策略,即特征分离、选择与校准(Feature Separation, Shape Structure Selection and Recalibration, FSSR)方法,通过同时对鲁棒特征和非鲁棒特征进行优化处理,充分挖掘两类特征中的判别信息。具体而言,本文方法一方面对非鲁棒特征进行重新校准,以恢复其潜在的判别信息;另一方面,从鲁棒特征中选择出具有更强判别能力的形状结构特征,为模型提供可靠的决策依据。这种双管齐下的联合特征处理策略不仅能有效利用鲁棒特征的协同增强效应,而且可充分挖掘非鲁棒特征中的有效信息,从而实现自适应特征优化,提升模型面对对抗攻击时的鲁棒性。实验结果表明,FSSR方法在多个对抗攻击场景下表现出优越的防御效果,优于现有许多特征空间对抗防御方法。

2 FSSR方法

对抗攻击通过恶意激活模型中的特定脆弱特征,诱导模型作出错误预测。然而,研究表明,部分特征具有稳健性,在对抗扰动下仍可保持稳定,从而助力模型维持正确的预测结果^[9]。此外,相关研究证实,形状结构信息在增强模型鲁棒性方面具有显著效果^[16]。因此,本文从鲁棒特征中选取形状结构特征,这些特征不仅对对抗扰动具有较强的抵抗能力,而且能为模型提供更具判别力的决策依

据,从而提升其整体性能。

如图 1 所示,FSSR 方法包括特征分离、自适应特征优化和特征融合 3 个关键步骤。具体而言,在特征分离阶段引入注意力机制,精准地将输入特征图分离为鲁棒特征和非鲁棒特征;在自适应特征优化阶段对非鲁棒特征进行重

新校准,恢复其潜在的有用判别信息,同时从鲁棒特征中选择对预测具有较强判别力的形状结构信息;在特征融合阶段将选择的鲁棒性形状结构特征与重校准后的非鲁棒特征进行融合,生成高效的特征图,从而提升模型的鲁棒性。

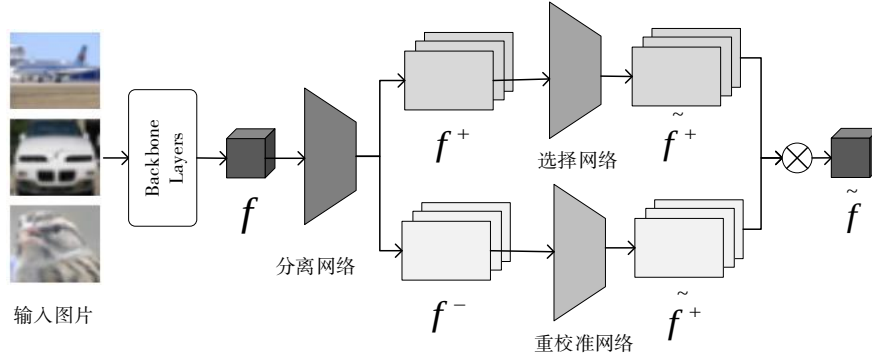


Fig. 1 Architecture of FSSR

图 1 FSSR 方法架构

基于自适应特征优化的对抗样本防御方法能有效捕获对决策至关重要的特征信息,使模型在应对对抗攻击时具备更强的鲁棒性,从而提升其整体防御能力。其伪代码为:

输入:输入图像的特征图 $f \in \mathbb{R}^{C \times H \times W}$

输出:融合特征图 $\tilde{f}$

1. 使用分离网络 S 生成鲁棒性得分图: $r \leftarrow S(f)$
2. 通过 Gumbel Softmax 计算软掩码 m :
3. $m = \text{GumbelSoftmax}(r)$
4. 正掩码: $m^+ \leftarrow m$ 负掩码: $m^- \leftarrow 1 - m$
5. 提取鲁棒特征: $f^+ \leftarrow m^+ \otimes f$
6. 提取非鲁棒特征: $f^- \leftarrow m^- \otimes f$
7. 自适应特征优化:
8. if f 是 f^+
9. $\tilde{f}^+ \leftarrow f^+ + R(f^+)$
10. else
11. $\tilde{f}^- \leftarrow f^- + m^- \otimes R(f^-)$
12. $\tilde{f} \leftarrow \text{PagFM}(\tilde{f}^+, \tilde{f}^-)$
13. return $\tilde{f}$

2.1 特征分离

与通过停用非鲁棒激活或选择重要特征来增强特征图的方法不同,本文同时充分利用鲁棒特征与非鲁棒特征。因此,在分离阶段,本文需要将输入的中间特征图分离为鲁棒激活和非鲁棒激活。为了确定哪些激活是鲁棒或非鲁棒的,本文引入分离网络,用于学习各个特征单元的鲁棒性。给定中间特征图 $f \in \mathbb{R}^{C \times H \times W}$ 作为输入,其中 C, H, W 分别表示输入特征图的通道数、高度和宽度。分离网络输出一个鲁棒性图 $r \in \mathbb{R}^{C \times H \times W}$,其表示对应特征单元的鲁棒性分数,较高的分数表示更鲁棒的特征激活。基于鲁棒性分数逐元素地将特征图分解为鲁棒特征和非鲁棒特征,即生成一个二值掩码 $b \in \{0, 1\}^{C \times H \times W}$,并将其应用于

特征图。然而,这种离散采样方式不可微且不连续,可能导致梯度屏蔽问题,使得模型的鲁棒性存在虚假提升的风险^[17-19]。为避免这一问题,本文采用 Gumbel Softmax 技术生成一个可微的软掩码 $m \in [0, 1]^{C \times H \times W}$,近似实现二值掩码^[17]。用公式表示为:

$$m = \frac{e^{\left(\frac{\log(\sigma(r)) + g_1}{\tau}\right)}}{e^{\left(\frac{\log(\sigma(r)) + g_1}{\tau}\right)} + e^{\left(\frac{\log(1 - \sigma(r)) + g_2}{\tau}\right)}} \quad (2)$$

式中: r 为鲁棒性图; σ 为 Sigmoid 函数; g_1 和 g_2 为从 Gumbel 分布中抽样的随机变量,定义为 $g = -\log(-\log(u))$, $u \sim \text{uniform}(0, 1)$ 。

通过该方法可生成一个软掩码 m ,其在高鲁棒性分数的位置取值接近 1,而在低鲁棒性分数的位置取值接近 0。使用正掩码 $m^+ = m$ 对输入特征图 f 进行元素乘积,得到鲁棒特征 $f^+ = m^+ \otimes f$ 。同理,使用负掩码 $m^- = 1 - m$ 对输入特征图 f 进行元素乘积,得到非鲁棒特征 $f^- = m^- \otimes f$ 。

在特征分离阶段往往会忽略图像中丰富的背景先验信息,因此本文引入随机注意力机制 (Shuffle Attention, SA),以进一步提升特征分离效果,其结构如图 2 所示^[20]。对于给定的输入特征映射 $f \in \mathbb{R}^{C \times H \times W}$,SA 机制首先将输入特征图在通道维度上分组为多个子特征,即 $f = [f_1, \dots, f_k, \dots, f_G]$, $f_k \in \mathbb{R}^{C/G \times H \times W}$;其次,针对每个子特征,SA 将其分为两个分支,一个分支利用通道之间的相互关系生成通道注意力图,另一个分支利用特征的空间关系生成空间注意力图;最后,SA 利用 Shuffle Unit 联合空间与通道维度上的特征依赖性将所有子特征聚合,并利用 channel shuffle 实现不同特征之间的信息交换,使模型在不同通道间保持均衡的特征表达^[21]。

本文将 SA 模块引入到分离模块中,使神经网络在学习特征单元鲁棒性时能够获取更丰富的信息,从而增强模

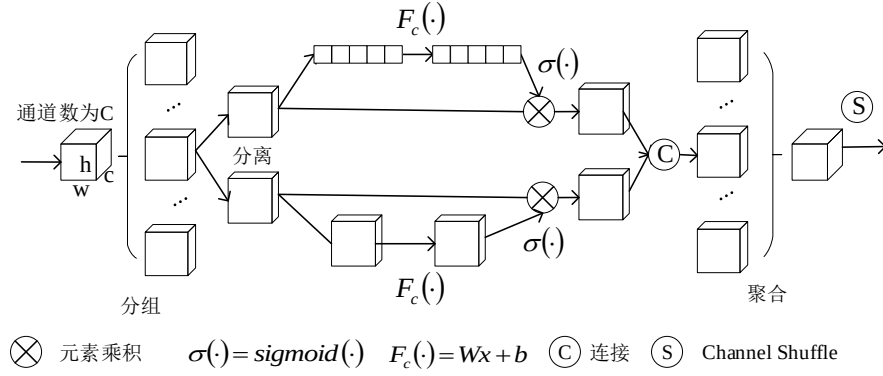


Fig. 2 Shuffle Attention structure

图2 SA结构

型专注于图像像素特征的能力。特征分离阶段流程如图3所示。

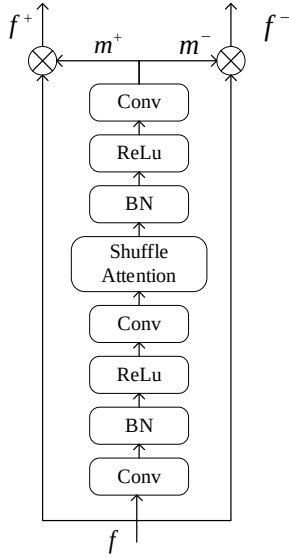


Fig. 3 Process of feature separation stage

图3 特征分离阶段流程

2.2 自适应特征优化

自适应特征优化的核心思想为通过自适应机制灵活地提取和利用特征中的有用信息,从而生成更具判别力的特征表示。本文的自适应优化主要体现在两个方面:一是从鲁棒特征中选择出形状结构信息;二是对非鲁棒特征进行重新校准。这种优化策略可以实现对鲁棒特征与非鲁棒特征的协同利用,提升模型的整体性能。

2.2.1 特征选择

本文在特征选择阶段设计了特征选择网络,其采用轻量化卷积结构,通过非线性变换和自适应筛选来重点提取鲁棒特征中的形状结构信息。该网络以鲁棒特征图作为输入,输出为经过特征选择的特征图。整体网络由3层卷积构成,第一层卷积用于初步提取特征,中间层卷积进一步提取高层次形状结构特征。层次化的卷积操作能逐步捕获与形状结构相关的重要信息,为特征选择提供有效表示。最后一层通过 1×1 卷积生成特征选择单元,并通过

Sigmoid激活函数进行归一化处理,使网络能够自适应地对输入特征图的每个通道进行加权,强调形状结构特征,同时抑制冗余特征。进一步通过全局平均池化操作增强特征选择单元对全局形状特征的敏感性,并将其叠加到原始鲁棒特征图中,最终生成经过选择的鲁棒特征图 $\tilde{f}^+$ 。表示为:

$$\tilde{f}^+ = f^+ + R(f^+) \quad (3)$$

特征选择的损失函数 L_{sel} 表示为:

$$L_{sel} = - \sum_{i=1}^N y_i \cdot \log(\tilde{p}_i^+) \quad (4)$$

式中: $\tilde{p}^+$ 为辅助层以 $\tilde{f}^+$ 为输入的预测得分。通过训练该辅助层可指导特征选择网络调整鲁棒特征的激活,使其能够提供与真实类别相关的线索。

2.2.2 特征重校准

特征重校准指对非鲁棒特征进行调整,以恢复其潜在的判别信息。本文设计了重校准网络,其以非鲁棒特征作为输入,输出为用于调整非鲁棒特征的重校准单元。对重校准单元应用负掩码 m^- ,并将结果添加到原始非鲁棒特征图中,计算得到重校准后的非鲁棒特征图 $\tilde{f}^-$ 。计算公式为:

$$\tilde{f}^- = f^- + m^- \otimes R(f^-) \quad (5)$$

为重校准后的特征图 $\tilde{f}^-$ 添加一个辅助层,并计算重校准损失 L_{rec} 。公式为:

$$L_{rec} = - \sum_{i=1}^N y_i \cdot \log(\tilde{p}_i^-) \quad (6)$$

式中: $\tilde{p}^-$ 为辅助层以 $\tilde{f}^-$ 为输入的预测得分。通过训练该辅助层可指导重校准网络调整非鲁棒特征的激活,使其能够提供与真实类别相关的线索。

2.3 特征融合

为融合鲁棒形状结构特征与重校准后的非鲁棒特征,本文引入Xu等^[22]提出的像素注意力引导融合模块(Pixel-attention-guided Fusion Module, PagFM),使P分支从I分支中选择性地学习高层语义信息,确保融合过程中更侧重于保留和利用关键语义信息。融合后的特征图不仅更加稳定,而且能够保留关键判别信息,提升模型的预测准确性

和鲁棒性。特征融合模块结构如图4所示。

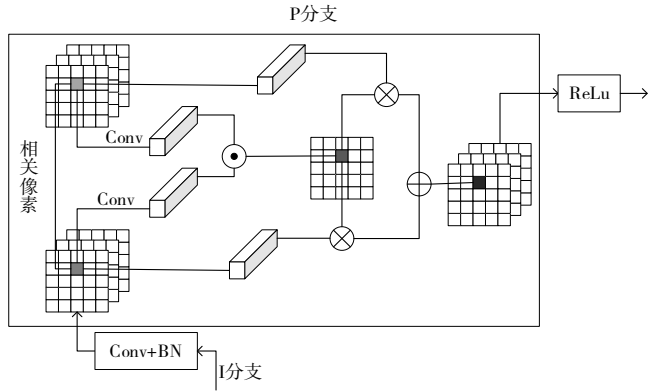


Fig. 4 Feature fusion module structure

图4 特征融合模块结构

分别定义鲁棒形状结构特征图与重校准后非鲁棒特征图中相关像素的向量为 $\vec{v}_p$ 和 $\vec{v}_i$, 通过 sigmoid 函数计算的输出为:

$$\sigma = \text{Sigmoid}\left(f_p(\vec{v}_p) \cdot f_i(\vec{v}_i)\right) \quad (7)$$

式中: σ 表示两个像素属于同一个对象的概率; f 表示卷积、归一化及 ReLU 等一系列计算操作。如果 σ 取值高, 则意味着 I 分支中的语义特征更具可信度; 否则, 融合过程中将加入更多 P 分支的信息。因此, PagFM 的输出表示为:

$$\text{Out}_{\text{PagFM}} = \sigma \vec{v}_i + (1 - \sigma) \vec{v}_p \quad (8)$$

3 实验方法与结果分析

3.1 实验环境

采用 Python3.9 和 PyTorch1.11.0 深度学习框架, 硬件环境为 Inter(R) Xeon(R) CPU E5-2678 v3 @2.50GHz, 搭配 24 GB 显存的 NVIDIA GeForce RTX 3090 显卡。实验参数设置见表 1。

Table 1 Experimental parameter settings

表1 实验参数设置

参数	值
Epoch	100
优化器	SGD
初始学习率	0.1(CIFAR-10) 0.01(SVHN)
Batch Size	128

3.2 数据集

采用 CIFAR-10 和 SVHN 数据集作为测试数据。CIFAR-10 数据集包含飞机、汽车等 10 个类别, 共计 60 000 张 32×32 像素的彩色图像。SVHN 数据集的格式与经典数据集 MNIST 类似, 但由于其包含更复杂的背景信息和多种数字排列方式, 比 MNIST 具有更高的挑战性。在 CIFAR-10 和 SVHN 数据集上进行测试可全面验证 FSSR 方法在不同数据分布和任务复杂度下的性能表现。其中, 在 CIFAR-10 数据集上将初始学习速率设置为 0.1, 在 SVHN 数

据集上将初始学习速率设置为 0.01, 每个数据集均在第 75 个 epoch 和第 90 epoch 时将学习率降低 10 倍。

3.3 评价方法

以 ResNet-18、VGG16 和 WideResNet34 为基线模型, 在 CIFAR-10、SVHN 两个数据集上进行比较实验。为全面评估所提方法的鲁棒性, 设置 4 种不同的对抗攻击策略, 分别为 FGSM、PGD 分别迭代 20 次和 100 次的 PGD-20 和 PGD-100, 以及 C&W 攻击算法。同时还计算了模型在不同对抗攻击下平均每个示例的集成鲁棒性(即在 FGSM、PGD-20、PGD-100 和 C&W 4 种不同对抗攻击下的平均鲁棒性)^[23]。计算公式为:

$$\text{Ensemble} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \min_{a \in \mathcal{A}} \mathbb{I}(F_{\theta}(a(x)) = y) \quad (9)$$

式中: N_{test} 为测试集中的样本数量; $\mathbb{I}(\cdot)$ 为指示函数, 当括号内的条件为真时取值为 1, 否则取值为 0; F_{θ} 为参数为 θ 的模型; y 为输入 x 的真实标签; $\mathcal{A}$ 为一系列对抗攻击, 包括 FGSM、PGD-20、PGD-100 和 C&W; $\min_{a \in \mathcal{A}}$ 表示对所有可能的攻击方法取最小值。

3.4 实验结果与分析

3.4.1 鲁棒性能分析

表 2(粗体表示最好结果, 下同)为将 FSSR 应用于 ResNet-18 上的鲁棒性。实验结果表明, 在 CIFAR-10 和 SVHN 两个数据集上, FSSR 方法均能有效提升对抗训练在不同对抗攻击下的鲁棒性。具体来说, 在 CIFAR-10 数据集上, ResNet-18 模型对 FGSM、PGD-20、PGD-100 和 C&W 4 种对抗攻击的鲁棒性分别提升了 1.59%、3.29%、3.54% 和 2.22%, 集成鲁棒性提高了 2.28%; 在 SVHN 数据集上, ResNet-18 模型对 FGSM、PGD-20、PGD-100 和 C&W 4 种对抗攻击的鲁棒性分别提升了 11.39%、13.24%、11.79% 和 2.6%, 集成鲁棒性提高了 4.14%。然而, 自然数据的准确性有所下降, 其中 CIFAR-10 数据集下降了 0.79%, SVHN 数据集下降了 0.01%。这主要归因于 FSSR 方法在特征选择过程中仅保留鲁棒特征中的形状结构特征, 限制了特征空间的表达能力, 可能导致部分关键判别信息丢失。此外, 在重校准阶段, FSSR 旨在对抗样本中的恶意激活进行调整, 以恢复潜在的判别信息。然而, 在自然图像中并不存在人为干扰的恶意激活, 对其进行重校准可能会误修改原始的有用信息, 进而影响模型对自然数据的分类性能。

如表 3 和表 4 所示, 将 FSSR 应用于对抗训练, 在

Table 2 Robustness on ResNet-18

表2 在 ResNet-18 上的鲁棒性

(%)

数据集	Resnet-18	Natural	FGSM	PGD-20	PGD-100	C&W	Ensemble
CIFAR-10	AT	85.02	56.21	48.22	46.37	47.38	45.51
	AT+FSSR(OURS)	84.23	57.80	51.51	49.91	49.60	47.79
SVHN	AT	91.21	55.55	40.85	37.54	40.61	37.41
	AT+FSSR(OURS)	91.20	66.94	54.09	49.33	43.21	41.55

VGG16 和 WideResNet34 上观察到与 ResNet-18 相似的趋势。具体来说,在 SVHN 数据集上,FSSR 方法使对抗训练在 VGG16 结构下针对 FGSM 攻击的鲁棒性提升了 7.58%,在 WideResNet34 结构下针对 FGSM 攻击的鲁棒性提升了 5.03%,验证了 FSSR 方法在不同基准模型上的适用性。

Table 3 Robustness on VGG16

表 3 在 VGG16 上的鲁棒性 (%)

数据集	VGG16	Natural	FGSM	PGD-20	PGD-100	C&W	Ensemble
CIFAR-10	AT	80.56	53.47	47.17	45.58	45.82	43.71
	AT+FSSR(OURS)	80.53	54.38	48.70	47.42	45.84	44.36
SVHN	AT	89.59	54.88	40.27	36.90	39.46	36.62
	AT+FSSR(OURS)	91.74	62.46	44.13	37.62	41.35	37.14

Table 4 Robustness on WideResNet34

表 4 在 WideResNet34 上的鲁棒性 (%)

数据集	WideResNet34	Natural	FGSM	PGD-20	PGD-100	C&W	Ensemble
CIFAR-10	AT	87.49	59.47	50.72	48.75	50.42	48.52
	AT+FSSR(OURS)	87.55	61.07	52.37	50.48	51.39	49.65
SVHN	AT	92.05	61.24	46.93	43.50	45.57	43.06
	AT+FSSR(OURS)	92.23	66.27	49.29	43.37	47.23	42.63

3.4.2 与现有方法的比较

为验证 FSSR 方法在提升模型鲁棒性方面的有效性,本文在 ResNet-18 结构上使用 SVHN 数据集进行实验,对 FSSR 方法与现有特征处理类对抗样本防御方法进行比较,分别为对抗训练、CAS、CIFs 和 FSR。实验结果见表 5。

Table 5 Comparison of robustness of different methods

表 5 不同方法鲁棒性比较 (%)

Resnet-18	Natural	FGSM	PGD-20	PGD-100	C&W	Ensemble
AT	91.21	55.55	40.85	37.54	40.61	37.41
CAS	94.08	65.24	48.47	43.75	46.15	-
CIFs	93.94	66.24	52.02	47.49	50.13	-
FSR	91.28	60.46	43.94	39.01	43.22	38.81
FSSR	91.20	66.94	54.09	49.33	43.21	41.55

可以看出,FSSR 方法在大多数白盒攻击下的表现优于其他 4 种方法,能有效提升对抗训练的鲁棒性。与 FSR 方法相比,FSSR 在 FGSM、PGD-20 以及 PGD-100 攻击下的鲁棒性分别提升了 6.48%、10.15%、10.32%,在集成鲁棒性上提升了 2.74%,表明 FSSR 方法在不同攻击场景下具备更强的防御能力。尽管 FSSR 在大多数攻击下表现优异,但在防御 C&W 攻击时,CAS 和 CIFs 方法表现更出色。这可能是由于 CAS 和 CIFs 通过辅助分类器的权重或 logit 输出的梯度调整特征激活,扩大了特征空间的预测边界,从而在 C&W 攻击下取得了更好的防御效果。相比之下,FSSR 在应对 C&W 攻击时可能未能充分优化决策边界,从而导致防御能力相对较弱。

虽然 FSSR 方法提升了模型的对抗鲁棒性,但在自然数据上的分类准确率有所下降。这可能是由于特征选择策略主要侧重于鲁棒特征中的形状结构信息,限制了特征

空间的覆盖范围,因而忽略了其他对自然数据分类十分重要的特征,如颜色等。此外,在重校准阶段,FSSR 旨在调整对抗样本中的异常激活,以恢复潜在的判别信息。然而,在自然图像中,并不存在人为干扰的恶意激活,FSSR 可能会误修改原始的有效特征,影响模型的自然数据分类能力。相比之下,CAS 和 CIFs 方法的特征调整策略允许更广泛的特征空间覆盖且不会对有效特征进行修改,因此在自然数据上的分类性能更优。实验结果表明,FSSR 在自然数据上的分类准确率为 91.20%,略低于 CAS(94.08%)、CIFs(93.94%)和 FSR(91.28%),表明 FSSR 在提升鲁棒性的同时可能牺牲了部分自然数据的分类性能。图 5 可视化展示了不同防御方法在各类攻击下的比较结果。

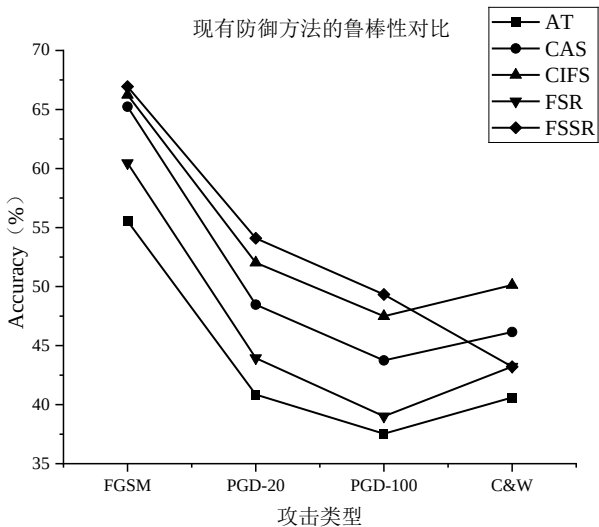


Fig. 5 Comparison of robustness of different defense methods under various attacks

图 5 不同防御方法在各类攻击下的鲁棒性比较

3.4.3 消融实验

为评估 FSSR 方法中各个组件的有效性,本文在 ResNet-18 上使用 CIFAR-10 数据集进行实验,通过测量集成鲁棒性对模型性能进行分析。同时,将 FSSR 方法各阶段生成的特征图替换为最终输出的特征图,并传递到后续层,通过比较这种操作下模型鲁棒性的变化来分析不同特征图在增强模型鲁棒性方面的作用。

(1)分离阶段评估。实验测量了鲁棒特征 f^+ 与非鲁棒特征 f^- 在捕获有助于正确预测的线索方面的效果,并通过集成鲁棒性指标来评估模型的性能,结果如表 6 所示。可以看出,鲁棒特征 f^+ 在对抗攻击下的集成鲁棒性能比非鲁棒特征 f^- 高出 5.01%,而仅依赖其中任何一种特征(鲁棒或非鲁棒)均会降低模型的集成鲁棒性(3.01%或 8.02%)。这表明鲁棒特征 f^+ 能有效捕获模型正确分类所需的鲁棒性线索,而非鲁棒特征 f^- 包含了在对抗攻击下容易导致模型预测错误的脆弱线索。因此,FSSR 方法的分离阶段达到了预期目的,成功地将输入特征图分离为鲁棒特征和非鲁棒特征,为后续特征选择与重校准步骤奠定了

基础。

Table 6 Model robustness after replacing the final feature map with different feature maps

表 6 使用不同特征图替换最终特征图后模型的鲁棒性 (%)

Resnet-18	Ensemble
f^+	44.78
f^-	39.77
$\tilde{f}^+$	45.26
$\tilde{f}^-$	46.49
FSSR(OURS)	47.79

(2)特征选择阶段评估。为验证 FSSR 方法中特征选择阶段的有效性,实验评估了从鲁棒特征中选择的形状结构特征在模型预测中的作用。具体而言,采用选择出的鲁棒性形状结构特征 $\tilde{f}^+$ 替换最终特征图,并重新测量模型的集成鲁棒性。实验结果如表 6 所示。可以看出,与鲁棒特征相比,筛选出的鲁棒性形状结构特征 $\tilde{f}^+$ 在集成鲁棒性上提升了 0.48%,表明特征选择阶段能有效提取有助于正确预测的关键特征,从而提升模型的对抗鲁棒性。

(3)特征重校准阶段评估。为验证 FSSR 方法中特征重校准阶段的有效性,实验评估了对非鲁棒特征进行重校准在模型中的作用。具体而言,采用经过重校准后的非鲁棒特征 $\tilde{f}^-$ 替换最终特征图,并重新测量模型的集成鲁棒性。实验结果如表 6 所示。可以看出,与非鲁棒特征相比,重校准后的非鲁棒特征 $\tilde{f}^-$ 在集成鲁棒性上提升了 6.72%,表明对非鲁棒特征进行重校准可以恢复潜在的对模型预测有用的信息,从而提升模型的对抗鲁棒性。

3.4.4 特征分离、选择、重校准与融合的有效性

为验证 FSSR 方法中不同模块对模型鲁棒性提升的必要性,本文在 ResNet-18 上使用 CIFAR-10 数据集进行实验,比较移除特征分离、特征选择、特征重校准或特征融合 4 个不同模块后模型的鲁棒性。

如表 7 所示,移除特征分离模块可导致模型鲁棒性下降,在 FGSM、PGD-20、PGD-100 和 C&W 攻击下的鲁棒性分别下降了 0.65%、0.66%、0.51% 和 0.17%,集成鲁棒性下降了 0.69%。这是由于在没有特征分离的情况下,选择网络和重校准网络需要直接处理包含鲁棒性和非鲁棒性信息的完整特征图,导致选择网络和重校准网络的次优训练。即使没有特征分离阶段,FSSR 方法依然在一定程度上提升了对抗训练的鲁棒性,这表明特征选择和重校准能有效捕获与模型预测相关的重要线索。移除特征选择模块后,模型在 FGSM、PGD-20、PGD-100 和 C&W 攻击下的鲁棒性分别下降了 2.37%、2.93%、2.78% 和 0.72%,集成鲁棒性下降了 0.99%。这是由于没有特征选择,鲁棒特征中可能包含无关或冗余特征,干扰有效特征的学习,导致模型鲁棒性下降。移除特征重校准模块后,模型在 FGSM、PGD-20、PGD-100 和 C&W 攻击下的鲁棒性分别下降了 1.06%、1.12%、0.82% 和 0.78%,集成鲁棒性下降了 0.56%。这是由于重校准非鲁棒特征中有用的判别信息可以提升

模型的鲁棒性。移除特征融合模块可导致模型鲁棒性下降,在 FGSM、PGD-20、PGD-100 和 C&W 攻击下的鲁棒性分别下降了 1.02%、0.69%、0.20% 和 0.68%,集成鲁棒性下降了 0.69%。这是由于特征融合阶段通过自适应加权机制将鲁棒特征与重校准后的非鲁棒特征有机结合,避免了因特征丢失导致的性能下降。综上所述,FSSR 方法的每个模块均在提升模型鲁棒性方面发挥了至关重要的作用。

Table 7 Comparison of model robustness after removing separation, selection, recalibration, or fusion, respectively

表 7 分别移除分离、选择、重校准或融合后模型鲁棒性比较 (%)

模型	FGSM	PGD-20	PGD-100	C&W	Ensemble
AT	56.21	48.22	46.37	47.38	45.51
+FSSR	57.80	51.51	49.91	49.60	47.79
w/o Sep	57.15	50.85	49.40	49.43	47.10
w/o Fu	56.78	50.82	49.71	48.92	47.10
w/o Sel	55.43	48.58	47.13	48.88	46.80
w/o Rec	56.74	50.39	49.09	48.82	47.23

3.4.5 超参数对模型鲁棒性的影响

为深入分析超参数(学习率、批量大小)对模型鲁棒性的影响,在 ResNet-18 上使用 CIFAR-10 数据集进行实验,分别调整学习率(0.1、0.01、0.001)和批量大小(64、128),在多种白盒攻击下评估模型的对抗鲁棒性。其余超参数均保持一致,以确保公平性。实验结果如表 8 所示。可以看出,降低学习率和批量大小会导致模型鲁棒性下降。例如,在 PGD-20 攻击下,学习率为 0.1 时的鲁棒性相较于 0.01 和 0.001 时分别提高了 5.44% 和 7.02%,表明较高的学习率有助于模型更有效地学习特征。在相同学习率下,批量大小从 128 降至 64 时,模型鲁棒性略有下降(如 PGD-20 攻击下下降 0.74%),表明较大批量大小有助于提升模型的稳定性。综合来看,学习率为 0.1 且批量大小为 128 时,模型的对抗鲁棒性最优。

Table 8 Robustness under different learning rates and batch sizes

表 8 不同学习率与批量大小下的鲁棒性 (%)

学习率	批量大小	FGSM	PGD-20	PGD-100	C&W	Ensemble
0.1	128	57.80	51.51	49.91	49.60	47.79
0.01	128	53.95	46.07	44.51	45.21	43.36
0.001	128	51.09	44.49	42.77	42.81	41.12
0.1	64	57.22	50.77	49.27	49.07	47.42

4 结语

本文提出一种新的自适应特征优化方法——FSSR,旨在通过协同优化鲁棒特征与非鲁棒特征,提升对抗训练模型的鲁棒性。实验结果表明,FSSR 方法在 CIFAR-10 和 SVHN 数据集上的鲁棒性优于现有对抗防御方法,并在多种攻击类型下均表现出更强的防御能力。然而,由于该方法在特征选择过程中侧重于形状结构信息,在重校准阶段可能误修改了自然数据中有助于正确分类的特征,从而导致自然数据的分类准确率下降。为减少这一负面影响,未

来研究可以从以下几个方面入手:①引入多尺度特征融合策略,结合形状结构、颜色、纹理等关键信息增强模型对自然数据的判别能力;②借鉴课程学习(Curriculum Learning)的思想,通过逐步调整特征选择和重校准策略,使其更好地适应不同攻击强度的样本,从而在提升鲁棒性的同时减少对自然数据分类性能的影响。

参考文献:

- [1] WENG X, IVANOVIC B, WANG Y, et al. PARA-Drive: parallelized architecture for real-time autonomous driving[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 15449–15458.
- [2] MANZARI O N, AHMADABADI H, KASHIANI H, et al. MedViT: a robust vision transformer for generalized medical image classification[J]. Computers in Biology and Medicine, 2023, 157: 106791.
- [3] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//San Diego: International Conference on Learning Representations, 2015.
- [4] LIU W, ZHONG S. Web malware spread modelling and optimal control strategies[J]. Scientific Reports, 2017, 7: 42308.
- [5] XIE C, WU Y, MAATEN L, et al. Feature denoising for improving adversarial robustness[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 501–509.
- [6] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]//Vancouver: International Conference on Learning Representations, 2018.
- [7] PANG T, DU C, DONG Y, et al. Towards robust detection of adversarial examples[J]. Advances in Neural Information Processing Systems, 2018, 31: 4584–4594.
- [8] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//2017 IEEE Symposium on Security and Privacy, 2017: 39–57.
- [9] ILYAS A, SANTURKAR S, TSIPRAS D, et al. Adversarial examples are not bugs, they are features[J]. Advances in Neural Information Processing Systems, 2019, 32: 125–136.
- [10] XIAO C, ZHONG P, ZHENG C. Enhancing adversarial defense by k-winners-take-all[C]//Online: International Conference on Learning Representations, 2020.
- [11] ZORAN D, CHRZANOWSKI M, HUANG P S, et al. Towards robust image classification using sequential attention models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 9480–9489.
- [12] BAI Y, ZENG Y, JIANG Y, et al. Improving adversarial robustness via channel-wise activation suppressing[C]//Online: International Conference on Learning Representations, 2021.
- [13] YAN H, ZHANG J, NIU G, et al. Cifs: improving adversarial robustness of cnns via channel-wise importance-based feature selection[C]//International Conference on Machine Learning, 2021: 11693–11703.
- [14] KIM W J, CHO Y, JUNG J, et al. Feature separation and recalibration for adversarial robustness[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 8183–8192.
- [15] LIU D, CHEN T, PENG C, et al. Improving adversarial robustness via decoupled visual representation masking[DB/OL]. <https://arxiv.org/abs/2406.10933>.
- [16] SUN M, LI Z, XIAO C, et al. Can shape structure features improve model robustness under diverse adversarial settings?[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 7506–7515.
- [17] JANG E, GU S, POOLE B. Categorical reparameterization with gumbel-softmax[C]//Paris: International Conference on Learning Representations, 2016.
- [18] XIE Y, DAI H, CHEN M, et al. Differentiable top-k with optimal transport[J]. Advances in Neural Information Processing Systems, 2020, 33: 20520–20531.
- [19] ATHALYE A, CARLINI N, WAGNER D. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples[C]//International Conference on Machine Learning, 2018: 274–283.
- [20] ZHANG Q L, YANG Y B. Sa-net: shuffle attention for deep convolutional neural networks[C]//2021 IEEE International Conference on Acoustics, Speech and Signal Processing, 2021: 2235–2239.
- [21] ZHANG P Y, MA Q M. Brain tumor segmentation algorithm based on cross-modal dual attention[J]. Computer Systems Applications, 2024, 33(1): 119–126.
- [22] XU J, XIONG Z, BHATTACHARYYA S P. PIDNet: a real-time semantic segmentation network inspired by PID controllers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 19529–19539.
- [23] DONG Y, DENG Z, PANG T, et al. Adversarial distributional training for robust deep learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 8270–8283.

(责任编辑:尹晨茹)