

基于多模态特征融合的对抗样本防御方法研究

魏宣宣, 刘万平, 卢玲

(重庆理工大学计算机科学与工程学院 重庆 400054)

摘要: 神经网络模型容易受到对抗样本攻击, 尤其是在图像分类任务中, 对抗攻击会导致其产生错误的分类。为了抵御这种攻击, 已经提出多种防御方法。这些防御方法侧重于改进模型结构, 或模型仅使用对抗训练方法导致防御类型单一且损害模型分类能力。受人类视觉系统通过多模态感官感知信息的启发, 提出将文本描述信息融合到图像信息中的多模态金字塔特征融合 (MPFF) 防御方法。首先, 使用 ViT-GPT2 模型根据图像信息生成对应的文本描述信息, 通过特征金字塔网络获取多尺度信息; 然后, 利用预训练的 TF-IDF 模型提取文本描述信息的特征矩阵, 利用 ResNet50 模型提取图像特征; 接着, 将图像特征和文本特征进行加权融合, 得到最终的多模态特征; 最后, 利用分类器对融合特征进行分类检测。在 CIFAR-10 和 ImageNet 数据集上进行了对比实验, 实验结果表明, 在不同扰动强度的黑盒攻击下, 与其他方法相比, 所提方法在两个数据集上的准确率分别平均提高 21.8% 和 22.5%。

关键词: 对抗防御; 多模态特征融合; 特征金字塔网络; 特征提取; 交叉熵损失

中图分类号: TP391

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2025023

Research on adversarial examples defense method based on multi-modal feature fusion

WEI Xuanxuan, LIU Wanping, LU Ling

College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China

Abstract: In recent years, the vulnerability of neural network models to adversarial attacks has been a significant concern, particularly in image classification tasks, where such attacks can lead to incorrect classifications. To counteract these attacks, numerous defence methods have been proposed. Existing defence methods have predominantly concentrated on enhancing model structures or adopting adversarial training methods individually, resulting in a single type of defence and potentially compromising the model's classification capability. Drawing on the human visual system's ability to perceive information through multimodal sensory inputs, a multimodal pyramid feature fusion (MPFF) defence method was proposed, integrating textual descriptions of images into the image information.

收稿日期: 2024-08-25; **修回日期:** 2025-02-10

通信作者: 刘万平, wpliu@cqut.edu.cn

基金项目: 重庆市自然科学基金 (cstc2021jcyj-msxmX0594); 重庆理工大学研究生教育高质量发展行动计划 (gzlxc20243171)

Foundation Items: The Chongqing Natural Science Foundation (cstc2021jcyj-msxmX0594), Action Plan for High-Quality Development of Graduate Education of Chongqing University of Technology (gzlxc20243171)

引用格式: 魏宣宣, 刘万平, 卢玲. 基于多模态特征融合的对抗样本防御方法研究[J]. 网络与信息安全学报, 2025, 11(2): 175-188.

Citation Format: WEI X X, LIU W P, LU L. Research on adversarial examples defense method based on multi-modal feature fusion[J]. Chinese Journal of Network and Information Security, 2025, 11(2): 175-188.

Initially, ViT-GPT2 was utilized to generate corresponding textual descriptions based on image information, while a feature pyramid network captured multi-scale information. Subsequently, a pre-trained TF-IDF model was employed to extract feature matrices from the textual descriptions, and a ResNet50 model was used to extract image features. These image and text features were then weighted and fused to obtain the final multimodal features. Finally, a classifier was applied to perform classification detection using the fused features. Comparative experiments were conducted on the CIFAR-10 and ImageNet datasets. The experimental results demonstrate that the accuracy of the proposed method is improved by 21.8% and 22.5% on average compared to other methods under black-box attacks with varying disturbance intensities on the two datasets respectively.

Keywords: adversarial defense, multimodal feature fusion, feature pyramid network, feature extraction, cross entropy loss

0 引言

随着深度神经网络在多个领域的广泛应用,其安全性问题日益凸显。攻击者通过在原始样本中添加精细扰动生成对抗样本,从而欺骗深度神经网络。这些对抗样本经过精心设计,与原始样本在视觉上几乎无法区分。Goodfellow 等^[1]首次提出通过在图像中添加微小的扰动来误导神经网络。从此,对抗样本成为研究热点。研究者陆续提出各种对抗攻击方法,并发现在自动驾驶、语音识别、流量分类和恶意软件检测等众多领域^[2]均存在对抗样本攻击的潜在威胁。随后, Szegedy 等^[3]提出了更加系统化的对抗样本生成方法 FGSM (fast gradient sign method),即利用神经网络的梯度信息来生成对抗性扰动以实施攻击。如果攻击者在高隐私和高安全需求的场景中使用对抗样本进行攻击,可能会带来严重的后果。例如,在人脸识别系统中,攻击者可以通过对抗样本欺骗身份认证系统,冒充他人身份进入安全区域;在自动驾驶中,对交通标志添加扰动噪声可能导致“停车”标志被错误地识别为“直行”,从而引发交通事故;在医疗诊断中,对抗样本可能导致诊断结果错误,影响患者的治疗计划。因此,研究和发展的有效的对抗样本防御方法具有重要的现实意义。

为了有效应对对抗样本攻击,研究者提出了多种防御方法。Papernot 等^[4]通过在训练过程中引入温度参数对模型进行训练,以提高模型鲁棒性,从而减小对抗样本的影响。随后, Xu 等^[5]通过在输入样本上应用多种压缩函数来降低样本的维度,从而达到防御目的。Wang 等^[6]将高斯增强和迭代攻击策略相结合,以提高防御对抗样

本攻击的能力。然而,上述防御方法均存在一定的局限性。它们以单一图像特征信息作为输入,系统的鲁棒性提升效果会受到限制。在实际应用中,依赖单一的图像特征信息不足以保证图像被准确分类。与图像特征信息相比,图像的文本描述^[7]可以提供更加丰富的细节信息。攻击者针对图像样本发起的攻击通常对其文本描述信息的影响较小,因此,图像和文本描述这种多模态信息相比单一的图像特征信息更具鲁棒性。本文研究表明,多模态的信息可以提高模型对多源信息的理解和泛化能力,从而提升模型抵御对抗样本攻击的鲁棒性和整体性能。基于这些跨领域研究成果,本文提出利用多模态信息增加模型的复杂性,使其难以受到各种对抗攻击的影响。

本文提出一种基于图像和文本描述的多模态金字塔特征融合 (multi-modal pyramid feature fusion, MPFF) 防御方法,旨在抵御高扰动强度下的对抗攻击。首先,利用 ViT-GPT2 根据图像信息生成对应的文本描述信息,与此同时使用特征金字塔网络 (feature pyramid network, FPN) 处理图像以获取多尺度信息。随后,利用 TF-IDF 和 ResNet50 分别提取文本特征和图像特征,并进行加权融合。接下来,使用注意力机制动态调整参数,将得到的多模态特征信息用于分类。为进一步优化特征融合效果,还引入了最小化交叉熵损失函数来提升预测结果的准确性。实验结果表明,与其他防御方法相比,MPFF 防御方法的性能更优异。

本文的主要贡献如下。

1) 将图像特征与文本特征结合,通过多模态特征融合来提升模型的鲁棒性,减少传统对抗样

本防御方法对单一图像信息的过度依赖所带来的脆弱性。

2) 利用特征金字塔网络获取多尺度信息。通过加权融合文本特征与图像特征，为模型提供丰富的多模态信息，从而提高模型的检测精度。此外，引入了注意力机制和交叉熵损失函数，用来优化特征表示，提升分类器防御对抗攻击的性能。

3) 在 CIFAR-10 和 ImageNet 数据集上开展了实验，实验结果表明，在不同扰动强度的黑盒攻击（无论是 FGSM、PGD 还是 MI-FGSM 攻击）下，MPFF 方法相较于其他方法，在两个数据集上的分类准确率分别平均提高 21.8% 和 22.5%。

1 相关工作

对抗样本是指在原始输入数据中添加微小扰动，使深度神经网络模型产生错误的预测或者分类。尽管对抗样本与原始样本在视觉上难以区分，但是对抗样本能显著影响模型输出结果。为抵御对抗攻击，各种防御方法应运而生。随着对抗攻击与防御策略的不断演进，模型的性能也逐步得到提升。

1.1 对抗样本攻击方法

现有的对抗攻击主要分为白盒攻击和黑盒攻击两大类。在白盒攻击中，攻击者可以访问目标模型的内部结构和参数信息。相反，黑盒攻击无法访问目标模型的内部结构和参数信息。在现实生活中，以黑盒攻击居多。

1.1.1 白盒攻击

目前，研究者已经提出多种白盒攻击方法。Szegedy 等^[3]首次证明了对抗性示例的存在性。Kurakin 等^[8]提出一种快速直接生成对抗样本的方法 FGSM。它是一种经典的基于梯度的攻击，通过增加图像分类器损失来增加对图像的不可感知扰动，生成的对抗样本如式(1)所示。

$$X^{\text{adv}} = x + \varepsilon \cdot \text{sign}(\nabla_x J(\theta, x, y_{\text{true}})) \quad (1)$$

其中， $J(\theta, x, y_{\text{true}})$ 表示交叉熵代价函数， x 是输入图像， y 是输入图像的真实标签， ε 是决定扰动大小的超参数。

Carlini 和 Wagner^[9]没有利用训练损失，而是设计了一个损失函数，并对其进行优化，以生成对抗样本。C&W 攻击被认为是最强大的攻击之

一，具体实现如式(2)、式(3)所示。

$$\min \left\| \frac{1}{2} (\tanh \cdot (x_n + 1) - x_n) \right\|_2^2 + c \cdot f \left(\frac{1}{2} \tanh \cdot (x_n) + 1 \right) \quad (2)$$

$$f(x') = \max \left(\max \{ Z(x')_{i:i \neq t} \} - Z(x'_t, -k) \right) \quad (3)$$

其中， Z 是 softmax 函数， k 是用于控制置信度的常数， t 是错误分类的目标标签， c 是通过二分搜索选择的常数， $\tanh(x)$ 是指对抗样本到 $\tanh$ 空间的映射。经过变换后， x 属于 $(-\text{inf}, +\text{inf})$ ，更有利于优化。

投影梯度下降（projected gradient descent, PGD）^[10] 是另一种迭代攻击方法。在每次迭代中，它基于对手选择的输出目标，对损失函数中的图像像素值执行梯度下降的步骤。PGD 以小步长多次迭代应用 FGSM。PGD 攻击通常比 FGSM 攻击强得多。PGD 攻击的具体实现如式(4)所示。

$$x_{i+1} = \text{clip}_\varepsilon \left(x_i + \alpha \cdot \text{sign}(\nabla L(x_i + \rho; \theta)) \right) \quad (4)$$

其中， $i \in [0, I]$ ， $X_0 = X$ ， ε 为扰动强度， clip 表示裁剪操作，经过 I 次迭代后，实现 $X_{\text{adv}} = X$ 。

1.1.2 黑盒攻击

在对抗性攻击范围内，黑盒攻击假定攻击者无法获取模型的内部结构和参数信息，只能根据输入和输出的关系进行攻击。黑盒攻击主要分为基于查询的攻击、基于迁移的攻击、基于梯度估算的攻击和基于生成对抗网络（generative adversarial network, GAN）的攻击 4 种类型。

基于查询的攻击方法通过不断查询目标模型来获取其输出信息，并生成对抗样本。在这种情况下，攻击者对模型的内部结构和参数一无所知，只能根据输入和输出的关系进行攻击。Ilyas 等^[11]提出了基于查询的黑盒对抗攻击方法，该方法通过估算输入和输出的关系来生成对抗样本。基于迁移的攻击方法利用对抗样本在不同模型间的迁移性，攻击者通常在已知模型上生成对抗样本，并利用该样本攻击目标模型。基于梯度估算的攻击方法通过估算目标模型的梯度信息生成对抗样本。在这种攻击方法中，攻击者虽然无法直接访问模型的梯度，但可以通过有限的查询和数值方法来近似估算。Yuan 等^[12]提出将自适应缩放因子与梯度攻击相结合实现黑盒攻击。基于 GAN 的对抗攻击方法通过生成器与判别器的动态对抗博弈，

实现对抗样本的协同优化生成。Xiao 等^[13]提出一种利用 GAN 进行对抗攻击的方法 AdvGAN, 通过生成器、判别器和目标神经网络生成对抗扰动, 并添加到干净样本中生成对抗样本。在实际应用中, 黑盒攻击难度更大, 且更符合现实情况, 因此研究黑盒模型更具实用价值。

1.2 对抗样本防御方法

随着对抗攻击的不断发展, 防御策略也成为深度学习领域的研究热点。主流的防御方法大致分为基于训练的防御、基于梯度的防御、基于输入输出的防御和基于知识的防御 4 种类型。

1) 在基于训练的防御方法中, Goodfellow 等^[1]通过向训练集注入对抗样本, 进行对抗训练, 以增强卷积神经网络 (convolutional neural network, CNN) 的鲁棒性。此外, 一些研究利用基于学习的方法来生成对抗样本并设计防御机制, 如使用条件 GAN 进行防御^[14]。Chen 等^[15]将浅层对抗性训练与基于注意力权重的模型相结合, 提出一种重新聚焦关键区域并强化目标轮廓的对抗性防御方法。Zhu 等^[16]提出一种基于条件变分自编码器和贝叶斯网络的双域对抗防御方法。

2) 在基于梯度的防御方法中, 考虑到许多对抗攻击依赖于梯度信息, Dabouei 等^[17]提出一种梯度相位和震级正则化相结合来防御基于梯度的攻击方法。Athalye 等^[18]将基于模糊梯度的防御方法分为破碎梯度、随机梯度和爆炸消失梯度 3 种类型。基于模糊梯度的防御方法主要通过掩盖梯度信息, 使攻击者难以获取有用的梯度, 从而实现对抗攻击的防御。

3) 基于输入输出的防御方法, 通过对输入图像进行转换来实现对抗防御。这类方法的代表有特征压缩方法^[19]、像素级防御方法^[20]、总方差最小化 (total variation minimization, TVM) 方法^[21]和稀疏变换层 (sparse transform layer, STL) 方法^[22]。一些研究通过比较模型对原始样本和压缩样本的预测结果来检测对抗样本, 从而实现基于输出的对抗防御。像素级防御方法将输入图像进行净化处理后再输入 CNN 分类器。TVM 结合像素丢失和变化最小化来进行压缩感知处理。而 STL 首先将输入图像投影到自然图像空间中, 然后将其输入网络。

4) 基于知识的防御方法已经被广泛研究并应用。Hinton 等^[23]通过提取模型的类概率知识来降低对抗性样本生成的成功率, 从而提升模型的防御能力。

现有的防御方法能够在一定程度上缓解对抗攻击对检测结果的影响^[24]。然而, 完全依赖单一图像特征信息进行检测存在局限性。一旦图像受到攻击, 其最终检测结果会受到显著影响。受人类多感官系统的启发, 本文提出一种融合图像特征与文本特征的对抗样本攻击防御方法。该方法的优势在于, 即使攻击者对图像进行攻击, 文本描述信息仍能保持较好的完整性, 从而能够显著提升分类的准确性和效率。

2 多模态金字塔特征融合防御方法

MPFF 防御方法的整体架构如图 1 所示。首先, 使用 ViT-GPT2 根据图像信息生成对应的文本描述信息, 并使用预训练 TF-IDF 模型构建文本特征矩阵。与此同时, 通过 FPN 获取多尺度信息, 结合基础特征提取器 ResNet50 提取图像特征。然后, 将图像特征和文本特征进行加权融合, 得到多模态特征。最后, 利用分类器对多模态特征进行分类检测。在 MPFF 架构中, 本文还设计了交叉熵损失函数, 用于度量多模态特征的相似性, 优化模型参数; 同时, 为了更好地实现图像特征和文本特征的多模态融合, 引入了注意力机制用来生成模态权重系数^[25]。注意力机制能够增强模型对关键信息的聚焦能力, 进一步提升特征表示的质量。

2.1 文本描述信息生成

为了生成更加准确的文本描述信息, 本文选用基于 vision Transformer 架构的图像和文本描述生成模型。该模型结合了视觉和语言模型的优势, 通过 Transformer 架构将图像转换为文本。视觉转换器 (vision Transformer, ViT) 具有强大的图像理解能力, 能够有效捕捉图像中的全局信息。ViT-GPT2 生成图像的文本描述如图 2 所示, ViT 首先把原始图像分割成大小固定的图像块, 然后将这些图像块展平并映射到一维空间的特征向量中, 接着将每个图像特征向量与位置编码相加, 并通过 Transformer 架构生成图像的全局特征表示。

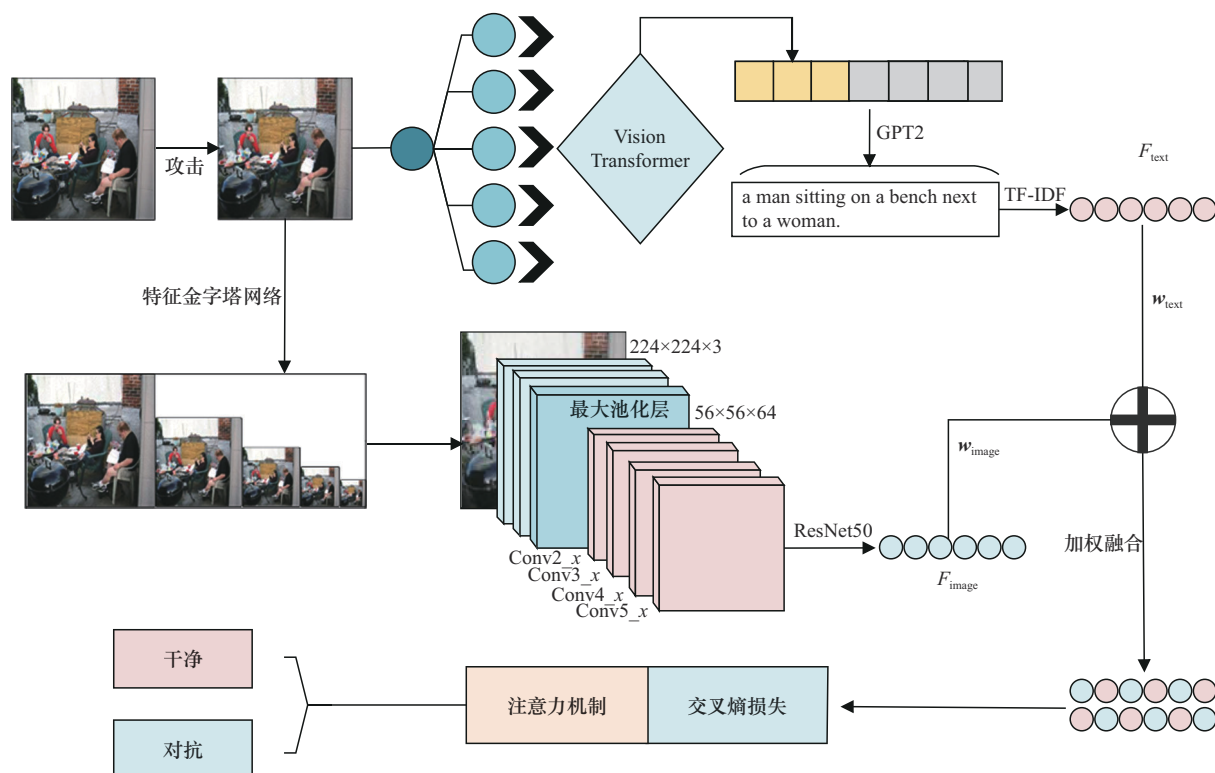


图1 MPFF防御方法的整体架构

Figure 1 Overall structure of MPFF defense method

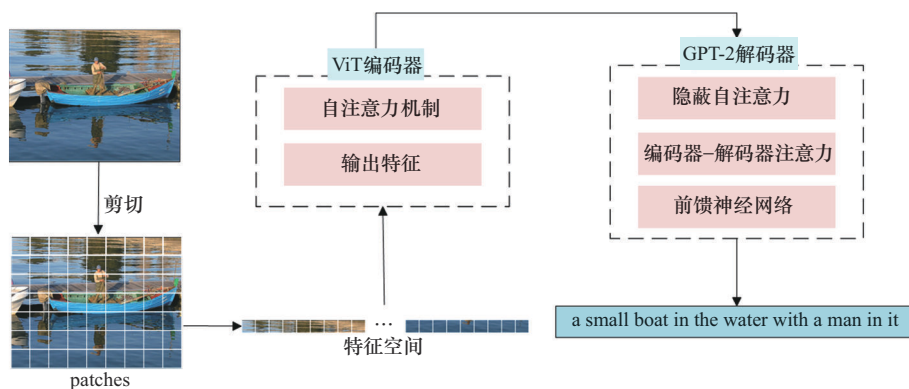


图2 ViT-GPT2生成图像的文本描述

Figure 2 ViT-GPT2 generate text description for the image

考虑到需要对图像进行处理, 本文采用计算开销较低的GPT-2^[26]作为生成文本描述的语言模型。GPT-2模型通过最大化下一个词的条件概率, 生成与图像内容相关的文本描述信息。GPT-2模型具备卓越的语言生成能力, 能够生成连贯且富有意义的文本描述信息。本文将ViT输出的图像特征向量作为GPT-2模型的初始输入, 用来引导GPT-2模型生成连贯且与图像内容相关的文本描述。ViT-GPT2模型在处理长距离依赖和并行计算

方面表现出显著优势。与传统的CNN-RNN组合相比, ViT-GPT2模型的泛化能力和性能更出色。

2.2 多模态金字塔特征提取

在图像特征提取方面, MPFF方法采用FPN与ResNet50协同架构。其中, FPN用于实现多尺度特征的提取和融合。FPN的核心思想是在多个尺度上构建图像特征金字塔, 以获取丰富的多尺度信息。具体来说, 不同层的特征图经过逐级上采样和融合, 形成完整的特征金字塔。每个尺

度的特征图经过 1×1 卷积调整通道数后进行上采样, 并与更低层特征图进行融合。同时, 采用多模态特征加权方式, 将图像特征和文本特征结合, 在多个卷积层上进行融合, 充分利用图像和文本的互补信息, 提高模型理解和处理多模态任务的能力。相比其他方法, FPN 能够更好地捕捉不同尺度信息, 提升特征表达能力。

本文将去除顶层分类器的 ResNet50 作为图像特征提取主干网络, 设置全局平均池化层以获取固定长度的特征向量, 并将其展平为二维矩阵形式。随后, 将特征向量填充到指定目标维度, 以方便后续的特征融合操作。ResNet50 的结构与

实现流程如图 3 所示, ResNet50 层级化结构包含 5 个阶段, 每个阶段为一个金字塔等级, 在每个阶段输出对应大小的特征图, 其内部卷积层之间依次连接形成特征图的多层结构。特征图层次由多个残差块中的卷积层组成。

本文采用简洁而高效的词频-逆文件频率 (term frequency-inverse document frequency, TF-IDF) 方法提取文本描述信息的特征。TF-IDF 能够有效减小常见词对文本特征的影响, 突出具有辨识度的词汇。在对抗攻击导致图像信息受到扰动的情况下, 尽管生成的文本描述信息可能存在偏差, 但文本中的核心词汇通常不会受到显著影

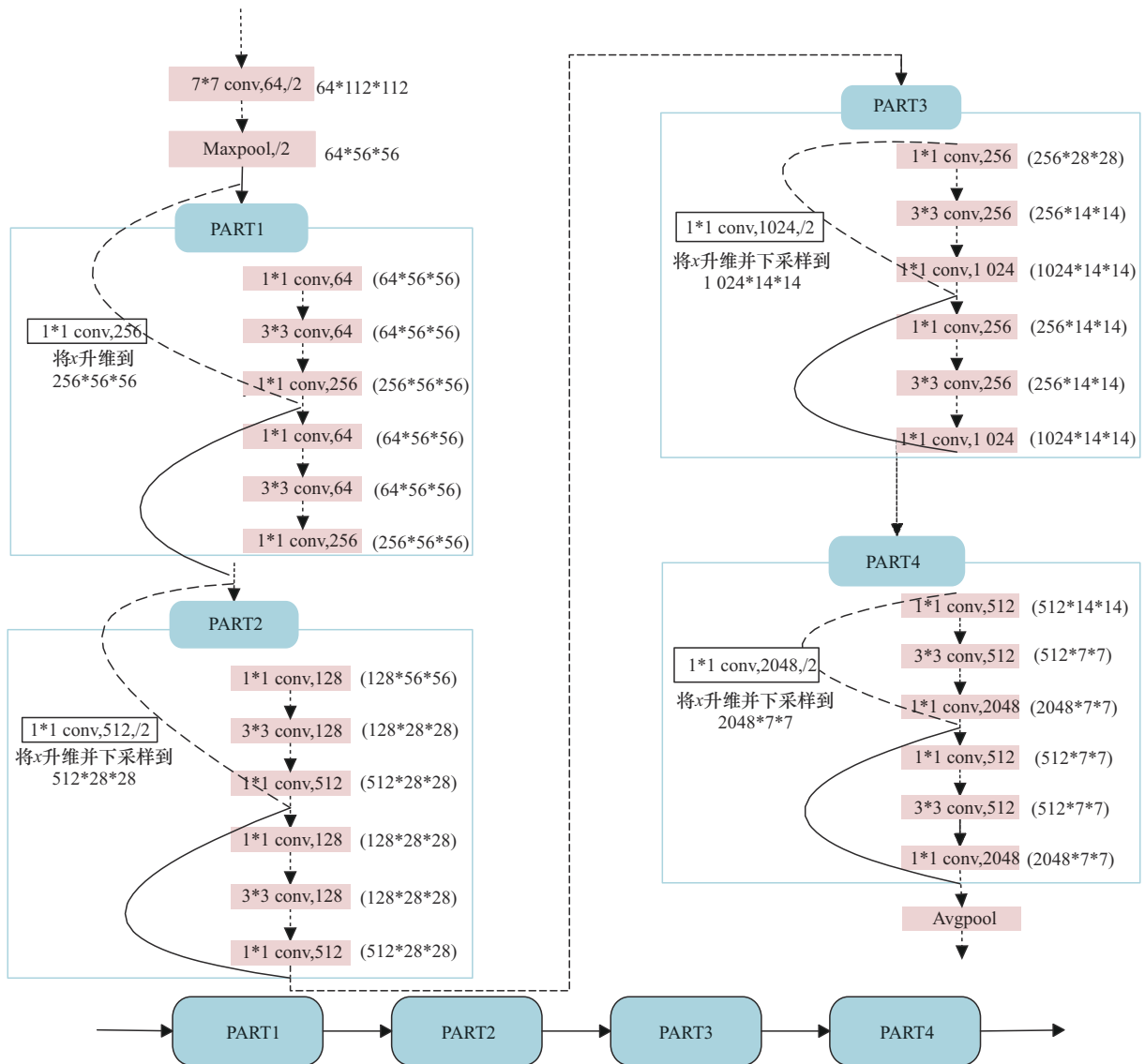


图 3 ResNet50 的结构与实现流程

Figure 3 The structure and implementation process of ResNet50

响。因此，基于TF-IDF的特征提取方式能够在一定程度上减少对抗样本攻击对图像分类结果的干扰。TF-IDF方法将每个词的值组合起来，形成特征矩阵，如式(5)所示。

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \cdot \log\left(\frac{N}{\text{DF}(t)}\right) \quad (5)$$

其中， t 是切分文本描述的单个词语， d 是文本描述语句， N 是总文本描述语句， $\text{DF}(t)$ 是包含单个词语 t 的文档数。

2.3 异构交叉熵损失与注意力机制加权融合

早期的特征融合方法通常是将图像特征和文本特征直接拼接，并利用全连接层处理拼接后的特征。然而，这些方法无法有效区分和调整不同模态的关键信息，如图像在不同分辨率下呈现的颜色与形状等视觉特征信息，以及文本中词汇的含义和上下文关系等语义信息。针对此问题，本文设计了多尺度金字塔特征提取方法，以获取图像和文本两种模态的特征信息。在加权融合之前，需要将图像特征和文本特征映射到相同维度，以便后续能够在同一维度空间内进行融合操作。这一过程通过线性变换实现，如式(6)和式(7)所示。

$$\mathbf{F}'_{\text{image}} = \mathbf{w}_{\text{image}} \mathbf{F}_{\text{image}} + b_{\text{image}} \quad (6)$$

$$\mathbf{F}'_{\text{text}} = \mathbf{w}_{\text{text}} \mathbf{F}_{\text{text}} + b_{\text{text}} \quad (7)$$

其中， $\mathbf{F}'_{\text{image}}$ 和 $\mathbf{F}'_{\text{text}}$ 分别为映射到相同维度的图像特征和文本特征， $\mathbf{w}_{\text{image}}$ 和 $\mathbf{w}_{\text{text}}$ 分别为图像特征和文本特征的映射矩阵， b_{image} 和 b_{text} 分别为图像特征和文本特征的偏置项。

将映射后的图像特征 $\mathbf{F}'_{\text{image}}$ 和文本特征 $\mathbf{F}'_{\text{text}}$ 进行拼接操作，形成联合特征 $\mathbf{F}_{\text{con}}$ 。为了实现根据不同模态特征在具体任务中的重要性进行权重分配，需要计算出注意力权重矩阵 $\mathbf{A}$ 。通过随机正态分布对权重矩阵进行初始化，通过计算 $\mathbf{QK}^T$ 反映向量之间的相关性。为避免点积过大导致经过softmax函数后出现梯度消失或梯度爆炸问题，需将结果除以 $\sqrt{d_k}$ ，得到注意力权重矩阵，如式(8)~式(12)所示。

$$\mathbf{W}_Q \sim N(0, \sigma^2 \mathbf{I}_{m \times d_q}) \quad (8)$$

$$\mathbf{W}_K \sim N(0, \sigma^2 \mathbf{I}_{m \times d_k}) \quad (9)$$

$$\mathbf{Q} = \mathbf{F}_{\text{con}} \mathbf{W}_Q \quad (10)$$

$$\mathbf{K} = \mathbf{F}_{\text{con}} \mathbf{W}_K \quad (11)$$

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d_k}}\right) \quad (12)$$

其中， $\mathbf{W}_Q$ 和 $\mathbf{W}_K$ 为权重矩阵， $\mathbf{Q}$ 、 $\mathbf{K}$ 分别为从 $\mathbf{F}_{\text{con}}$ 中通过线性变换计算出的查询矩阵和键矩阵。 $\mathbf{F}_{\text{con}}$ 表示联合特征， $\mathbf{A}$ 为注意力权重矩阵， $\mathbf{I}$ 表示单位矩阵， d_k 为缩放因子。

通过注意力机制（如图4所示）计算得到注意力权重矩阵 $\mathbf{A}$ ，其每一个元素 a_{ij} 表示在联合特征中不同位置的图像和文本特征相对于最终融合结果的重要性权重。注意力机制能够根据输入数据的实际情况动态调整权重，使模型更加聚焦于关键信息。当某一区域的图像特征向量与对应的文本特征向量点积值较大时，表明该区域与文本的相关性较高，即关键区域。在文本生成过程中，注意力机制使模型能够动态关注图像的不同区域，从而更精准地提取视觉信息。通过为不同模态信息动态分配权重，模型能够更加灵活地捕捉各模态的核心特征。此外，注意力机制通过计算特征向量之间的相关性，减少冗余信息的干扰，提高整体性能，并更好地适应不同模态特征之间的差异。然而，注意力机制的训练过程涉及大量矩阵运算和参数更新，计算复杂度较高，训练成本较高。同时，权重分配和线性变换操作对计算资源的需求较高，因此在实际应用中需要在性能提升与计算开销之间寻找平衡。

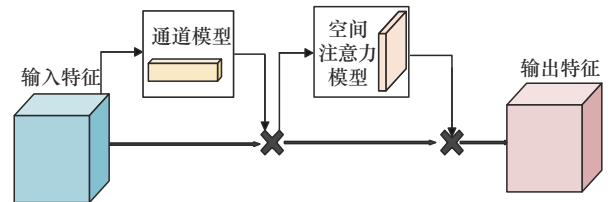


图4 注意力机制

Figure 4 Attention mechanism

通过注意力机制确定融合权重，对不同模态信息进行加权融合，如式(13)所示。

$$\mathbf{F}_{\text{fusion}} = \mathbf{A}_{\text{image}} \mathbf{F}'_{\text{image}} + \mathbf{A}_{\text{text}} \mathbf{F}'_{\text{text}} \quad (13)$$

其中， $\mathbf{F}'_{\text{image}}$ 和 $\mathbf{F}'_{\text{text}}$ 分别表示图像和文本特征， $\mathbf{A}_{\text{image}}$ 和 $\mathbf{A}_{\text{text}}$ 分别表示从注意力权重矩阵 $\mathbf{A}$ 中分解出的图像和文本注意力权重。在模型训练过程中，通过反向传播算法动态调整 $\mathbf{A}_{\text{image}}$ 和 $\mathbf{A}_{\text{text}}$ ，以达到最佳融合效果。

为了增强MPFF防御方法的鲁棒性，选择交叉

熵损失函数对模型进行训练。在模型训练过程中,将加权融合后的特征输入多层感知机分类器,得到模型对每个类别的预测概率。使用交叉熵损失函数衡量模型预测标签分布与真实标签分布之间的差异,通过不断迭代训练缩小差异值,使融合特征的预测结果更加接近真实标签,实现对模型参数的逐步优化,使模型能够更好地拟合训练数据。

MPFF 防御方法通过整合交叉熵损失函数和注意力机制,在处理多模态信息(如图像和文本)时具有更好的适应性和可靠性,进而提高分类性能和泛化能力。一方面,注意力机制根据特征间的相关性动态调整图像和文本特征的权重;另一方面,交叉熵损失函数通过衡量模型预测的准确性,在模型训练过程中引导模型动态调整注意力机制和特征映射的参数,确保特征融合的合理性。

MPFF 防御方法的实现过程如下。

首先,利用 FPN 从原始图像和对抗样本中获取多尺度图像特征信息 F_S 和 F'_S , 利用 ViT-GPT2 根据原始图像和对抗样本的图像信息生成原始图像和对抗样本的文本描述 T 和 T' , 并利用 TF-IDF 得到原始图像和对抗样本的文本特征矩阵 F_T 和 $F_{T'}$ 。

然后,利用注意力机制权重 A_S 和 $A_{T'}$ 融合图像特征和文本特征,得到融合特征矩阵 F_m 。

最后,通过多次迭代训练和总样本数 N 逐步缩小图像和文本融合的总损失值 L_{total} 。通过不断分类总样本数的预测标签获得图像的防御准确率 C 。

上述过程如算法 1 所示。

算法 1 MPFF 防御方法

输入 原始图像 S , 图像标签 i

输出 文本描述 T , 防御准确率 C , 总损失 L_{total}

- 1) $F_S \leftarrow \text{FPN}(S)$;
- 2) $F'_S \leftarrow F_S$;
- 3) $T \leftarrow \text{get_caption}(S), T' \leftarrow \text{get_caption}(S')$;
- 4) $F_T \leftarrow \text{TF}(t, d) \cdot \log\left(\frac{N}{\text{DF}(t)}\right), F_{T'} \leftarrow \text{TF}(t', d') \cdot \log\left(\frac{N}{\text{DF}(t')}\right)$;
- 5) $F_m \leftarrow A_S \cdot F_S + A_{T'} \cdot F_{T'}$;
- 6) **for** epoch=1 to epochs **do**
- 7) **for** $a=1$ to total sample N **do**

$$8) \quad L_{\text{total}} \leftarrow L_S + L_{T'};$$

9) **end for**

$$10) \quad C \leftarrow i/N;$$

11) **end for**

12) **Return:** T, C , and L_{total}

3 实验与结果分析

3.1 实验设置

3.1.1 数据集

为了评估 MPFF 防御方法的性能,使用以下 3 个公开数据集进行实验。

1) CIFAR-10 数据集: 该数据集包含 60 000 张图像,其中 50 000 张图像用于训练(共包含 10 类,每类 5 000 张)、10 000 张图像用于测试。

2) ImageNet 数据集: 该数据集包含 1 400 多万张图像,本文实验选用该数据集的 ILSVRC2012 子集进行训练和测试,随机选取 30 000 张图像作为训练集、12 000 张图像作为测试集。

3) Flickr30k 数据集: 该数据集包含 31 783 张广泛应用于计算机视觉研究的图像,可用于训练各种图像识别和处理系统。

选择 CIFAR-10 数据集和 ImageNet 数据集进行训练和测试, Flickr30k 数据集则用于消融实验,具体如表 1 所示。

表 1 3 个数据集的分布
Table 1 Distribution of three datasets

数据集	训练数据集/张	测试数据集/张
CIFAR-10	50 000	10 000
ImageNet	30 000	12 000
Flickr30k	—	31 783

3.1.2 攻击设置

实验中使用多种攻击算法处理数据集。由于白盒攻击方法具有较强的攻击性,使用经典白盒攻击算法作为替代模型生成对抗样本,并利用迁移性对模型进行防御测试。对于 CIFAR-10 数据集,使用 FGSM、20 次迭代的 PGD 和 20 次迭代的 MI-FGSM 攻击方法生成对抗样本; ImageNet 数据集采用基于梯度的 FGSM 攻击、20 次迭代的 PGD 攻击和高迁移性的 AdvGAN 黑盒攻击生成对抗样本。在 CIFAR-10 数据集和 ImageNet 数据

集实验中,对于FGSM攻击,分别设置扰动强度 $\varepsilon=2/255$ 、 $8/255$ 、 $16/255$;对于20次迭代的PGD攻击,分别设置扰动强度 $\varepsilon=2/255$ 、 $8/255$ 、 $16/255$;对于20次迭代的MI-FGSM攻击和高迁移性的AdvGAN黑盒攻击,设置扰动强度 $\varepsilon=0.01$ 。将生成的对抗样本输入替代模型,以便进行防御性测试和评估。

3.1.3 防御方法

为了评估MPFF防御方法的性能,本文选用特征压缩、像素级防御、ComDefend^[27]、防御性GAN^[28]、DIPDefend^[29]、PCLC^[30]和HFFN^[31]共7种仅利用图像信息进行防御的方法与MPFF进行对比实验。特征压缩通过减少图像的颜色深度、将相邻像素合并来减少对抗扰动的影响,在本文实验中颜色深度设置为4、使用 2×2 的像素合并。像素级防御通过随机扰乱图像部分像素位置来干扰攻击效果,在本文实验中参数设置为每张图像扰动20个像素,并在每个图像块内进行微调。ComDefend通过图像压缩与重构来消除部分对抗扰动,在本文实验中JPEG压缩率设置为75%。防御性GAN利用GAN来生成去噪后图像,在本文实验中生成器步长设置为5。DIPDefend利用未被攻击影响的图像结构信息来重建原始图像。PCLC基于概率对比学习,在本文实验中校准参数设置为0.5、学习率设为0.001。HFFN融合图像高层次和低层次特征。在实验过程中,对这些防御方法均进行了严格的参数调试,以便在不同攻击方法和扰动强度下,为评估MPFF方法的有效性提供对照依据。

3.1.4 环境设置

实验环境如下:处理器为i7-8700,显卡为NVIDIA GeForce RTX3090,内存为32 GB,操作系统为Windows10,使用Pycharm2023.1.3进行开发。考虑到黑盒攻击相比白盒攻击更加贴近现实场景(即攻击者对模型的内部结构和参数信息一无所知),所有攻击实验均在黑盒条件下进行,以便更好地测试模型在未知攻击下的鲁棒性。随机选取CIFAR-10数据集中的10 000张图像训练替代模型,训练时长约3 h,内存平均占用4 GB。将原始图像和被攻击后的图像输入ViT和GPT-2模型中生成文本描述信息作为文本数

据。ResNet50在中小型数据集上表现出色,且模型较小易于训练,符合实验要求,因此选择其作为图像特征提取的主干网络。

3.2 实验结果与分析

本文分别使用ImageNet和CIFAR-10数据集进行实验。针对每个被攻击后的图像样本,需要生成与之相应的文本描述信息,如图5所示。在黑盒攻击设置中,采用FGSM借助迁移攻击生成对抗样本。具体而言,首先在自身构建模型上生成对抗样本,而后将其用于对目标模型的攻击。由于对抗样本在不同模型间的迁移性,这些在其他模型上生成的样本在目标模型上依然能够产生攻击效果。本文将经防御处理后能够正确分类对抗样本的比例(即准确率)作为评估对抗样本防御机制有效性的关键指标。

3.2.1 针对CIFAR-10数据集的防御实验

本文在CIFAR-10数据集上开展了特征压缩、像素级防御、ComDefend、防御性GAN、DIPDefend、PCLC、HFFN和MPFF的对比实验,评估指标采用准确率,实验结果如表2所示。MPFF防御方法在3种不同扰动强度的FGSM攻击中均表现出显著的优势,尤其在扰动强度 $\varepsilon=16/255$ 时,MPFF防御方法的准确率达到62.36%,远高于其他方法。这表明MPFF防御方法在高强度攻击下具有极高的鲁棒性。随着扰动强度的增加,各防御方法的准确率均有所下降,但MPFF防御方法在高扰动强度下依然保持了较高的准确率。这证明即使在严苛的攻击环境中,MPFF防御方法仍能有效抵抗攻击。在迭代次数20的PGD攻击中,尽管MPFF防御方法在低扰动强度下的准确率低于HFFN方法,但随着扰动强度的增加,其他方法的准确率明显下降,而MPFF防御方法展现出显著的优势。这表明MPFF防御方法在高扰动强度的情况下具备更高的稳定性。同样地,在扰动强度 $\varepsilon=2/255$ 的MI-FGSM攻击下,尽管MPFF防御方法与HFFN方法相比,准确率低0.49%,但随着扰动强度的增加,MPFF防御方法逐渐展现出优于HFFN和PCLC方法的防御效果。在没有攻击的情况下,各种防御方法的准确率均较高,其中MPFF防御方法的准确率为92.14%。在无攻击场景中,MPFF防御方法虽然表现出优异性能,但与其他方法相



图 5 TF-IDF 根据图像信息生成对应的文本描述信息

Figure 5 Image generates text description information through TF-IDF

表 2 8 种防御方法在不同攻击下的准确率对比 (CIFAR-10 数据集)

Table 2 Comparison of accuracy of eight defense methods under different attacks (CIFAR-10 dataset)

防御方法	原始样本	FGSM			PGD			MI-FGSM		
		2/255	8/255	16/255	2/255	8/255	16/255	2/255	8/255	16/255
特征压缩	84.77%	31.05%	20.80%	18.24%	40.32%	3.87%	0	—	—	—
像素级防御	85.54%	73.35%	46.78%	24.12%	56.40%	13.78%	27.98%	—	—	—
ComDefend	93.30%	75.19%	52.23%	34.99%	59.25%	24.21%	13.83%	—	—	—
防御性 GAN	90.23%	81.80%	63.96%	52.08%	75.44%	45.09%	36.89%	—	—	—
DIPDefend	90.0%	74.48%	67.27%	47.69%	77.18%	55.10%	47.28%	—	—	—
PCLC	93.94%	82.82%	46.08%	30.49%	78.74%	20.29%	4.10%	80.21%	18.74%	3.88%
HFFN	91.14%	81.52%	64.36%	46.50%	81.45%	66.09%	46.61%	81.18%	65.33%	40.56%
MPFF	92.14%	87.23%	68.22%	62.36%	80.78%	69.25%	65.45%	80.69%	79.89%	54.12%

比, 防御效果并不是最好的。总体来说, 在 CIFAR-10 数据集上, MPFF 防御方法的准确率平均提高 21.8%, 可见该方法具有较好的防御效果。

如图 6 所示, 各种防御方法的防御效果随着扰动强度 ϵ 值的增大呈现下降趋势, 说明各种防御方法在面对更强攻击时的防御效果普遍降低。然而, 整体来看, MPFF 防御方法始终维持在较高的水

平, 尤其是在扰动强度 $\epsilon=16/255$ 时, 与其他防御方法相比, 展现出明显优势。MPFF 防御方法通过融合文本描述信息, 来减小图像信息对分类结果的影响, 而其他防御方法仅依赖图像信息, 一旦受到高强度扰动, 分类结果将全面受影响。相比之下, MPFF 防御方法的文本信息不会受到图像攻击的直接干扰, 进一步凸显该方法的优越性。

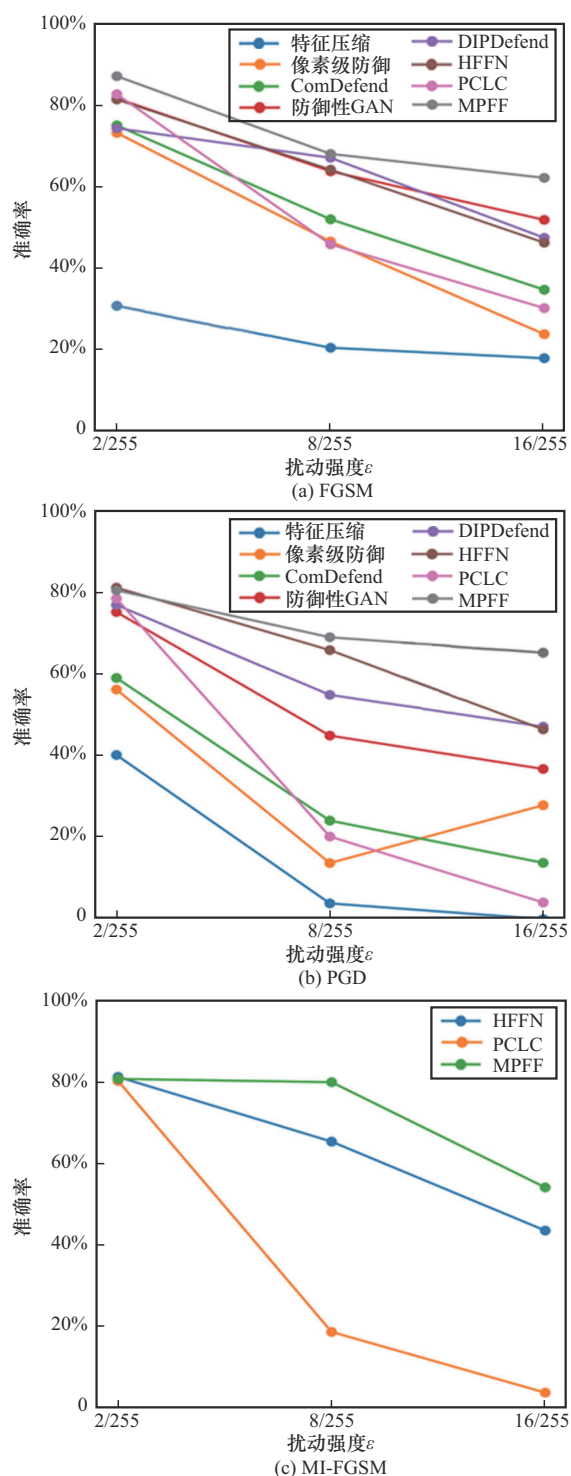


图6 各种防御方法在不同扰动强度的FGSM、PGD、MI-FGSM攻击下的准确率对比(数据集=CIFAR-10)

Figure 6 Accuracy comparison of various defense methods under FGSM, PGD and MI-FGSM attacks with different disturbance intensities (dataset = CIFAR-10)

3.2.2 针对 ImageNet 数据集的防御实验

本文在 ImageNet 数据集上开展了 ComDe-

fend、防御性GAN、DIPDefend和MPFF的对比实验,实验结果如表3所示。在AdvGAN黑盒攻击下,MPFF防御方法的准确率达到72.16%。在扰动强度 $\epsilon=2/255$ 的FGSM攻击下,在4种方法中,MPFF防御方法的准确率最高(达到63.69%)。在无攻击下,MPFF防御方法的准确率低于其他方法。随着扰动强度的增加,在相同扰动强度下MPFF防御方法的表现更出色。尤其在扰动强度 $\epsilon=16/255$ 的FGSM攻击下,MPFF防御方法的准确率为39.5%,相比ComDefend、防御性GAN和DIPDefend方法分别高出27.89%、17.51%和23.7%。在扰动强度 $\epsilon=16/255$ 的PGD攻击下,MPFF防御方法的准确率为45.38%,而ComDefend的准确率仅为1.9%。可见,高扰动强度对模型的干扰更大,使模型难以有效区分原始样本和对抗样本,模型的分类准确率显著降低。在面对强扰动时,尽管MPFF防御方法的准确率有所下降,但相比其他方法,MPFF防御方法所受的影响较小,其依然展现出显著的优势。各种防御方法在不同扰动强度的FGSM、PGD攻击下的准确率对比如图7所示,可以看出,在FGSM和PGD攻击下,MPFF防御方法均展现出极高的鲁棒性,特别是在高强度攻击下,其防御效果尤为突出。总体来说,在ImageNet数据集上,MPFF防御方法与其他防御方法相比,准确率平均提高22.5%。综上所述,MPFF防御方法具有较强的抗扰动能力,能够在保持较高准确率的同时有效抵御攻击,展现出显著的鲁棒性和稳定性。

表3 4种防御方法在不同攻击下的准确率对比
(ImageNet数据集)

Table 3 Comparison of accuracy of four defense methods under different attacks (ImageNet dataset)

防御方法	原始样本	FGSM		
		2/255	8/255	16/255
ComDefend	69.22%	31.41%	13.0%	11.61%
防御性GAN	75.83%	38.33%	28.25%	21.99%
DIPDefend	68.17%	50.43%	29.74%	15.8%
MPFF	73.76%	63.69%	50.3%	39.5%
防御方法	AdvGAN	PGD		
		2/255	8/255	16/255
ComDefend	29.46%	42.35%	2.87%	1.9%
防御性GAN	11.36%	32.55%	28.6%	26.14%
DIPDefend	38.56%	54.4%	43.52%	33.5%
MPFF	72.16%	53.89%	47.56%	45.38%

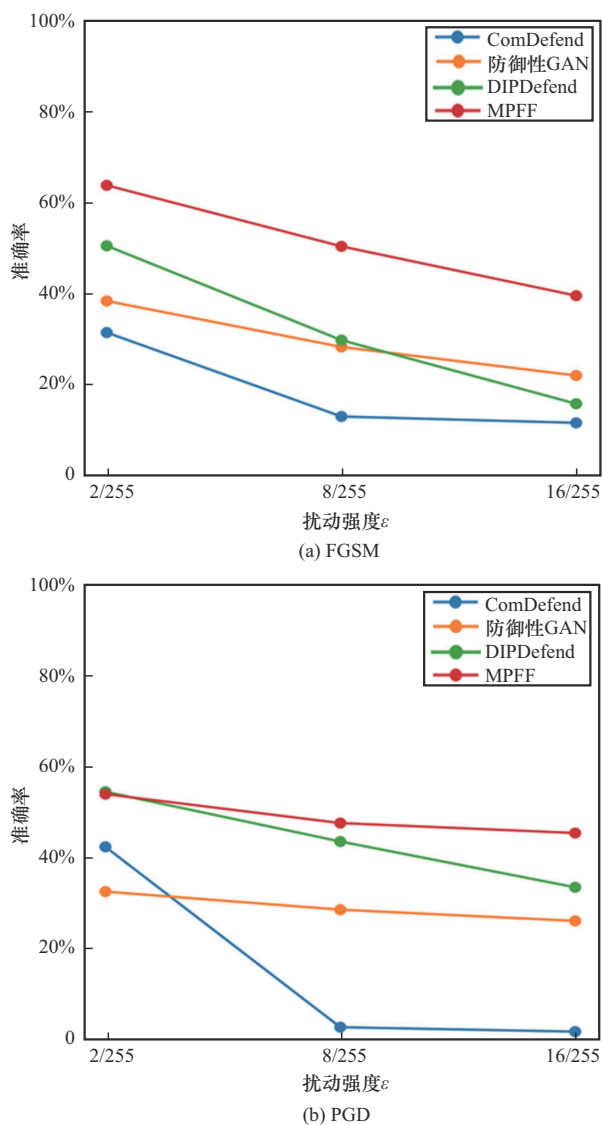


图 7 各种防御方法在不同扰动强度的FGSM、PGD攻击下的准确率对比(ImageNet数据集)

Figure 7 Accuracy comparison of various defense methods under FGSM and PGD attacks with different disturbance intensities (ImageNet dataset)

3.3 消融实验

本文进行了消融实验,以评估各个模块对MPFF模型的贡献。将Flickr30k数据集作为消融实验测试集,使用扰动强度 $\epsilon=2/255$ 的FGSM方法进行数据预处理。FGSM攻击在生成对抗样本的效率、攻击效果等方面表现出较好的平衡性。扰动强度 $\epsilon=2/255$ 是经过大量相关研究及实际应用场景的综合分析确定的。在众多对抗样本防御研究中,该扰动强度被广泛应用,因为其能够展现出对抗样本在不

同情况下的特性。为了研究FPN对MPFF模型的影响,将MPFF模型去除FPN模块,称此模型为 M_I ;将加权特征融合替换为简单特征拼接,称此模型为 M_{II} ;去除MPFF模型中的交叉熵损失函数,称此模型为 M_{III} ;模型 M_{IV} 则仅考虑图像特征信息,不进行多模态特征融合。通过对比每个模型在Flickr30k和CIFAR-10数据集上的准确率,判断各模块对模型性能的贡献。

表4为各个模型在消融实验中的结果。实验均在扰动强度 $\epsilon=2/255$ 的FGSM攻击下进行。去除FPN模块的模型 M_I 在CIFAR-10数据集上的准确率下降14.94%,在Flickr30k数据集上的准确率下降15.94%。用简单的特征拼接替代加权融合的模型 M_{II} 在CIFAR-10数据集上的准确率降低27.11%,在Flickr30k数据集上的准确率下降7.04%。去除交叉熵损失函数的模型 M_{III} 在CIFAR-10和Flickr30k数据集上的准确率分别下降9.72%、20.93%。不使用多模态特征的 M_{IV} 模型,仅依赖图像特征,其在CIFAR-10和Flickr30k数据集上的准确率分别下降7.12%和12.72%。从实验结果可以看出,MPFF模型中各个模块均对模型性能提升起到关键作用。尽管在两个数据集上的实验结果存在差异,但总体上每个模块的去除都会导致模型的准确率显著下降,证明这些模块能够有效提高MPFF的防御效果。尤其是多模态特征的引入,使MPFF在抵御对抗攻击时展现出更强的鲁棒性。但多模态特征的引入会增加模型的计算复杂度。在RTX3090显卡设置下,不同防御模型的训练时间对比如表5所示,其中MPFF的训练时间约20 h。

表 4 消融实验

Table 4 Ablation experiment

模型	Flickr30k	CIFAR-10
M_I	68.73%	72.29%
M_{II}	77.63%	60.12%
M_{III}	63.74%	77.51%
M_{IV}	71.95%	80.11%
MPFF 模型	84.67%	87.23%

表5 不同防御模型的训练时间对比
Table 5 Comparison of training time of different defense models

防御模型	总训练时间/h
ComDefend	约14
防御性GAN	约10
DIPDefend	约11
MPFF	约20

3.4 讨论

从CIFAR-10和ImageNet数据集的对比实验结果来看,相较于其他方法,MPFF防御方法在大部分情况下表现出色。然而,无论是FGSM、PGD还是MI-FGSM攻击,当扰动强度 $\epsilon=2/255$ 时,MPFF防御方法的准确率提高相对较小,甚至有所降低(在PGD和MI-FGSM攻击下,MPFF防御方法的准确率相较于其他方法有所降低)。这是因为低强度的扰动对图像的影响较小,此时多模态特征对防御结果的影响更明显,与单纯依赖图像特征相比,MPFF防御方法的准确率有所下降。因此,MPFF防御方法在面对较高强度扰动攻击时的表现更加优异,而在低强度扰动下的防御效果并未保持显著优势。通过消融实验分析,证明了MPFF防御方法的合理性。无论是金字塔特征提取还是加权融合策略,均对实验结果有着重要影响。尤其是交叉熵损失函数的引入,对CIFAR-10和ImageNet两个数据集的分类准确率提升起到关键作用。这验证了MPFF防御方法具有良好的泛化能力和鲁棒性。因此,本文提出的MPFF防御方法在一定程度上提升了防御对抗攻击的效果,尤其在应对高强度扰动攻击时,展现出更高的性能和可靠性。从实验结果来看,尽管MPFF防御方法在特征融合和防御效果方面表现优异,但增加了计算复杂度。在大规模数据集上进行训练和测试时,模型的训练时间明显延长,这主要归因于多模态输入和复杂特征融合网络需要大量计算资源。

4 结束语

为了提升对抗样本的防御效果并减少对单一图像信息的依赖,本文提出一种基于多模态特征融合的对抗样本防御方法。该方法通过在图像信息中引入文本描述信息来增强模型的鲁棒性。具

体来说,先使用FPN进行图像特征提取,并将图像特征与文本特征进行加权融合,利用注意力机制动态调整融合权重,缩小多模态特征间的差异。将多模态特征作为模型的输入,进行分类,并使用交叉熵损失函数进行训练,减小对分类结果的影响。通过比较不同攻击方法和扰动强度的防御效果,评估模型分类结果。实验结果表明,相比其他防御方法,MPFF防御方法在面对高强度扰动的对抗攻击时表现更出色,分类准确率大幅提高。然而,在低强度扰动或无攻击场景下,MPFF防御方法的准确率并未显著提升。未来研究将聚焦于以下方面:①如何缩小多模态特征融合距离,减小异构特征对分类结果的影响,引入动态调整机制和自适应策略,结合多种防御技术,在原始样本和低强度扰动下保持优异的防御效果。②如何降低模型的计算复杂度,减少计算开销。例如,考虑采用模型压缩和量化技术减少不必要的模型参数,降低模型对计算资源的需求。③如何优化现有多模态特征融合网络架构,使其在保证性能的前提下更加简单和高效。另外,未来MPFF防御方法可以用来解决其他多模态问题。

参考文献:

- [1] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//Proceeding of the 3rd International Conference on Learning Representations (ICLR 2015). 2015.
- [2] LIU W P, ZHONG S M. Web malware spread modelling and optimal control strategies[J]. Scientific Reports, 2017, 7: 42308.
- [3] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//Proceeding of International Conference on Learning Representations (ICLR). 2014.
- [4] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]//Proceedings of the IEEE Symposium on Security and Privacy. Piscataway:IEEE Press, 2016: 582-597.
- [5] XU W, EVANS D, QI Y. Feature squeezing: detecting adversarial examples in deep neural networks[C]//Proceeding of International Conference on Network and Distributed System Security Symposium. Virginia: the Internet Society, 2018:1-15.
- [6] WANG D N, CHEN W, YANG Y, et al. Defense method of adversarial training based on gaussian enhancement and iterative attack[J]. Computer Science, 2021, 48(6A):509-513,537.
- [7] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: a neural image caption generator[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 3156-3164.
- [8] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[C]//Proceedings of the 5th International

- Conference on Learning Representations. 2018: 99-112.
- [9] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of the IEEE Symposium on Security and Privacy (SP). Piscataway:IEEE Press, 2017: 39-57.
- [10] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]//Proceeding of International Conference on Learning Representations(ICLR). 2018.
- [11] ILYAS A, ENGSTROM L, ATHALYE A, et al. Black-box adversarial attacks with limited queries and information[C]//Proceedings of International Conference on Machine Learning. 2018: 2137-2146.
- [12] YUAN Z, ZHANG J, JIANG Z, et al. Adaptive perturbation for adversarial attack[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(8): 5663-5676.
- [13] XIAO C, LI B, ZHU J, et al. Generating adversarial examples with adversarial networks[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence. 2018: 3905-3911.
- [14] GAUTHIER J. Conditional generative adversarial nets for convolutional face generation[J]. Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter Semester, 2014(5): 2.
- [15] CHEN J Y, ZHENG H B, CHEN R X, et al. RCA-SOC: a novel adversarial defense by refocusing on critical areas and strengthening object contours[J]. Computers & Security, 2020, 96: 101916.
- [16] ZHU J L, PENG G H, WANG D W. Dual-domain-based adversarial defense with conditional VAE and bayesian network[J]. IEEE Transactions on Industrial Informatics, 2020, 17(1): 596-605.
- [17] DABOUEI A, SOLEYMANI S, TAHERKHANI F, et al. Exploiting joint robustness to adversarial perturbations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press, 2020: 1122-1131.
- [18] ATHALYE A, CARLINI N, WAGNER D. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples[C]//Proceedings of International Conference on Machine Learning. 2018: 274-283.
- [19] PRAKASH A, MORAN N, GARBER S, et al. Protecting JPEG images against adversarial attacks[C]//Proceedings of Data Compression Conference. Piscataway:IEEE Press, 2018: 137-146.
- [20] SONG Y, KIM T, NOWOZIN S, et al. PixelDefend: leveraging generative models to understand and defend against adversarial examples[C]//Proceedings of the 6th International Conference on Learning Representations.2018.
- [21] ZHANG C, GAO P. Countering adversarial examples: combining input transformation and noisy training[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway:IEEE Press, 2021: 102-111.
- [22] SUN B, TSAI N, LIU F C, et al. Adversarial defense by stratified convolutional sparse coding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press, 2019: 11447-11456.
- [23] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. Computer Science, 2015, 14(7): 38-39.
- [24] 肖镇杰, 黄诗瑀, 叶锋, 等. 基于 PatchTracker 的对抗补丁防御算法[J]. 网络与信息安全学报, 2024, 10(1):169-180.
- XIAO Z J, HUANG S Y, YE F, et al. Adversarial patch defense algorithm based on PatchTracker[J]. Chinese Journal of Network and Information Security, 2024, 10(1): 169-180.
- [25] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of Advances in Neural Information Processing Systems 30(NIPS 2017). Piscataway: IEEE Press, 2017: 5998-6008.
- [26] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press, 2016: 770-778.
- [27] JIA X, WEI X, CAO X, et al. ComDefend: an efficient image compression model to defend adversarial examples[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press, 2019: 6084-6092.
- [28] LAYKAVIRIYAKUL P, PHAISANGITTISAGUL E. Collaborative defense -GAN for protecting adversarial attacks on classification system[J]. Expert Systems with Applications, 2023, 214: 118957.
- [29] DAI T, FENG Y, WU D X, et al. DIPDefend: deep image prior driven defense against adversarial examples[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 1404-1412.
- [30] LI X, LI X R, PAN D, et al. Improving adversarial robustness via probabilistically compact loss with logit constraints[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 8482-8490.
- [31] ZHANG J H, MAEDA K, OGAWA T, et al. Defense against black-box adversarial attacks via heterogeneous fusion features[C]//Proceedings of ICASSP 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2023: 1-5.

[作者简介]



魏宣宣 (2000-), 女, 河南周口人, 重庆理工大学硕士生, 主要研究方向为网络与信息安全、深度学习。



刘万平 (1986-), 男, 重庆人, 博士, 重庆理工大学副教授, 主要研究方向为网络与信息安全。



卢玲 (1975-), 女, 重庆人, 重庆理工大学教授, 主要研究方向为深度学习、自然语言处理。