

基于分层结构的多任务对抗样本归因方法

孙旭, 张文琼, 龙显忠, 李云

(南京邮电大学计算机学院, 江苏 南京 210023)

摘要: 深度神经网络在计算机视觉的各种任务中表现出优异的性能。然而, 它们非常容易受到对抗攻击的威胁, 这种攻击是在推理阶段向样本中加入人眼难以察觉的扰动来完成的。为了防御对抗攻击, 目前已有部分工作研究对抗样本的欺骗逆向工程, 这也被称为对抗样本归因。通过归因生成对抗样本的攻击算法和受害者模型, 能够帮助防御者掌握攻击者拥有的知识及其攻击目标, 从而设计针对相应攻击最有效的防御算法。已有方法更多地将对抗样本归因转化为单任务学习。随着攻击算法和受害者模型可取范围的扩大, 单任务学习面临着组合爆炸问题。为了提升对抗样本归因的准确率及满足不同归因粒度的需求, 对攻击算法和受害者模型进行了分层, 并利用不同层次之间的依赖关系, 提出了一种基于分层结构的多任务对抗样本归因方法。该方法在不同层次同时执行攻击算法和受害者模型的归因任务, 并使用分层路径预测学习不同层次之间的依赖关系。实验结果表明, 相比其他归因方法, 所提方法能够获得更好的归因效果。

关键词: 深度学习; 对抗样本归因; 多任务学习; 分层结构

中图分类号: TP18

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2025009

Multi-task adversarial attribution method based on hierarchical structure

SUN Xu, ZHANG Wenqiong, LONG Xianzhong, LI Yun

School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

Abstract: Deep neural networks have demonstrated superior performance in various computer vision tasks. However, they have been found to be highly susceptible to adversarial attacks, which involve the addition of perturbations to examples during the inference phase that are imperceptible to the human eye. To defend against adversarial attacks, some works have explored the reverse engineering of adversarial examples, known as the adversarial attribution problem. By attributing the attack algorithm and victim model used to generate adversarial examples, defenders can gain insights into the attacker's knowledge and targets, thereby enabling the design of more effective defense algorithms against corresponding attacks. Existing methods have mostly approached the adversarial attribution

收稿日期: 2024-07-30; 修回日期: 2024-09-14

通信作者: 李云, liyun@njupt.edu.cn

基金项目: 国家自然科学基金 (62476137)

Foundation Item: The National Natural Science Foundation of China (62476137)

引用格式: 孙旭, 张文琼, 龙显忠, 等. 基于分层结构的多任务对抗样本归因方法[J]. 网络与信息安全学报, 2025, 11(1): 92-105.

Citation Format: SUN X, ZHANG W Q, LONG X Z, et al. Multi-task adversarial attribution method based on hierarchical structure[J]. Chinese Journal of Network and Information Security, 2025, 11(1): 92-105.

tion problem as a single-task learning problem. However, as the scope of attack algorithms and victim models has expanded, single-task learning has faced the challenge of combinatorial explosion. To improve the accuracy of adversarial attribution and meet the requirements for different attribution granularities, attack algorithms and victim models were layered, and the dependencies between different levels were utilized. A multi-task adversarial attribution method based on a hierarchical structure was proposed. This method simultaneously performed the attribution tasks of attack algorithms and victim models at different levels and employed hierarchical path prediction to learn the dependencies between these levels. Experimental results on multiple datasets demonstrate that the proposed method achieves better attribution performance compared to other attribution methods.

Keywords: deep learning, adversarial attribution, multi-task learning, hierarchical structure

0 引言

在过去二十年中, 深度神经网络 (deep neural networks, DNN) 在计算机视觉的各种任务 (如图像分类^[1]、语义分割^[2]和对象检测^[3]等) 中表现出了出色的性能。然而, 研究表明, DNN 极易受到攻击者精心设计的输入样本的攻击。攻击者在原始样本中添加一些难以察觉的对抗扰动噪声, 致使 DNN 输出错误的结果, 这种恶意生成的样本被称为对抗样本^[4]。

现有的大量研究工作致力于解决两个问题, 一个是如何生成对抗样本用来攻击 DNN, 另一个是 DNN 如何防御这些对抗样本以提高鲁棒性。针对对抗样本生成问题, 研究人员提出了多种对抗攻击方法^[5]。例如, 基于梯度的白盒攻击^[6-8], 以及基于分数和查询的黑盒攻击^[9-11]等。针对防御对抗样本问题, 现有的方案主要有对抗样本检测^[12]和对抗训练^[13]。

尽管研究人员围绕对抗攻击和对抗防御开展了大量研究工作, 但涉及对抗样本归因的研究工作较少。对抗样本归因是欺骗逆向工程 (reverse engineering of deceptions, RED) 的一部分。根据

美国国防高级研究计划局 (Defense Advanced Research Projects Agency, DARPA) 的声明, RED 旨在开发能够自动溯源攻击中工具链的技术, 如多媒体伪造、对抗机器学习攻击或其他信息欺骗攻击。对抗样本归因的目的是从对抗样本中推断出攻击者的信息 (如攻击算法和受害者模型), 从而帮助防御者采取有针对性的防御策略。对抗样本归因概览如图1所示。

目前, 研究者在对抗样本归因方面进行了初步探索^[14-15], 研究工作侧重于将对抗样本归因转化为分类问题, 利用 DNN 对对抗样本或从对抗样本中提取出来的对抗扰动进行分类, 以识别攻击算法和受害者模型。然而, 目前的研究工作大多数为一种攻击信息 (如攻击算法或受害者模型) 的归因, 随着攻击信息可取范围的扩大, 攻击信息归因组合爆炸成为亟需解决的问题。

Guo 等^[14]提出使用多任务学习框架进行对抗样本归因。该方法不仅能够有效缓解组合爆炸问题, 还有利于利用不同攻击信息之间的关系。为了提升对抗样本归因效果, Nicholson 等^[15]对具体的攻击算法进行分类, 并提出了“攻击族”的

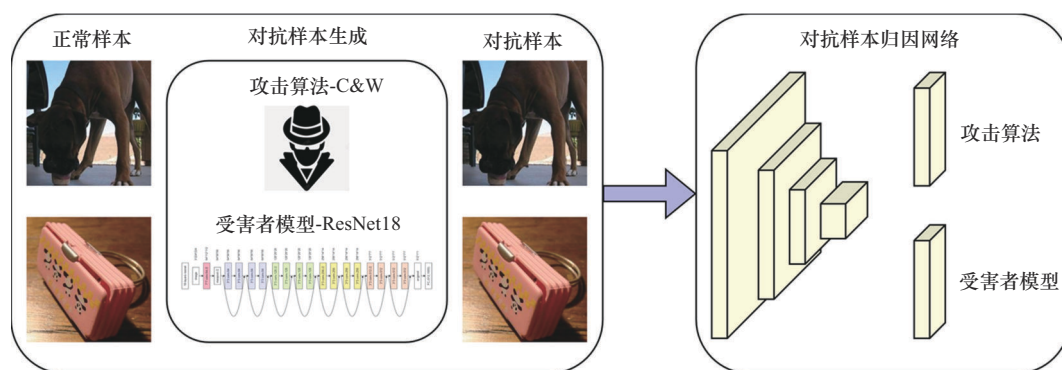


图1 对抗样本归因概览

Figure 1 Overview of adversarial sample attribution

概念, 同一攻击族的攻击算法在某些方面具有相同的模式。不过, 根据分类方法的不同, 同一攻击算法可能属于不同的攻击族。然而, Nicholson 等只是对不同攻击族内的攻击算法进行了归因, 并没有考虑攻击算法与攻击族之间的关系。

为了提高对抗样本归因的准确率, 以及满足不同归因粒度的需求, 本文设计了攻击算法和受害者模型归因的层次结构。以攻击算法为例, 本文将攻击算法分为 3 个层次, 不同层次对应不同的归因粒度。第 1 个层次 (以下简称层次 1) 根据攻击者对受害者模型的了解程度进行分类, 分为白盒攻击和黑盒攻击; 第 2 个层次 (以下简称层次 2) 根据攻击方式进行分类, 分为基于梯度的攻击、基于优化的攻击和基于决策的攻击等; 第 3 个层次 (以下简称层次 3) 是具体的攻击算法分类。不同层次的攻击算法之间也存在从属关系。例如, 基于梯度的攻击和基于优化的攻击隶属于白盒攻击, 而基于决策的攻击隶属于黑盒攻击。本文将这种从属关系描述为层次之间的依赖关系。在不同层次对攻击算法进行分类归因, 有助于了解更多对抗样本的相关知识。为了在不同层次对对抗样本的信息进行建模, 本文提出了一种基于分层结构的多任务对抗样本归因方法 MAA-HS (multi-task adversarial attribution method based on hierarchical structure)。

MAA-HS 的流程如图 2 所示, 其主要包含对抗特征提取模块、多分支特征提取模块以及分类模块 3 个部分。

对抗特征提取模块负责提取对抗样本的特征, 并将提取出来的特征图进行通道连接后送入多分支特征提取模块; 多分支特征提取模块的每个分支负责提取不同层次的对抗样本信息的特征表示并送入分类模块; 分类模块执行攻击算法和受害者模型分类任务。为了保证不同层次的分支提取出来的特征表示的有效性, 每个层次需添加相应的分类损失。此外, 添加分层路径预测学习不同层次之间的依赖关系。实验结果表明, 使用分层结构可以有效提高对抗样本归因的准确率。

本文主要贡献如下。

- 1) 提出了攻击算法和受害者模型的层次结构。在不同层次对攻击算法和受害者模型进行归因, 以满足不同归因粒度的需求。
- 2) 提出了一种基于分层结构的多任务对抗归因方法 (MAA-HS), 利用多分支特征提取模块提取不同层次对抗样本信息的特征表示, 通过添加分层路径预测来学习不同层次之间的依赖关系, 并结合多任务学习框架同时对攻击算法和受害者模型进行归因。
- 3) 将 MAA-HS 与已有的典型代表方法在不

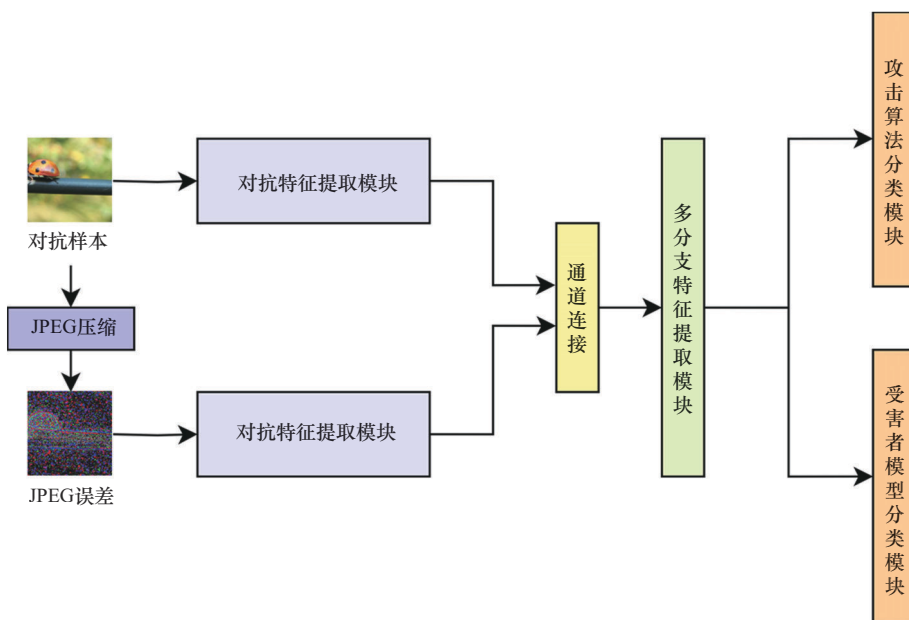


图 2 MAA-HS 的流程

Figure 2 Flow of MAA-HS

同数据集上进行了实验评估，从实验结果来看，使用分层结构有效提高了对抗样本归因的准确率，验证了MAA-HS的有效性。

1 相关工作

1.1 对抗攻击

对抗攻击由Szegedy等^[4]发现，他们揭示了深度学习模型的漏洞，即攻击者可以向图像添加视觉难以察觉的扰动来操纵其预测结果。近年来，大量的对抗攻击算法被提出。根据攻击者对受害者模型的了解程度，对抗攻击可分为白盒攻击和黑盒攻击。白盒攻击指攻击者了解有关受害者模型的一切信息（包括训练数据、输出、模型架构、权重）。根据攻击方式的不同，白盒攻击又可分为基于梯度的攻击和基于优化的攻击。基于梯度的攻击通过直接获取受害者模型损失函数的梯度生成对抗样本，而基于优化的攻击将对抗样本的生成转化为对目标函数的优化问题。黑盒攻击指攻击者无法访问受害者模型配置信息，只能依据查询和监视受害者模型的输入和输出获得的信息生成对抗样本。根据可获得信息的不同，黑盒攻击又可分为基于决策的攻击和基于分数的攻击。基于决策的攻击依据模型预测结果生成对抗样本，而基于分数的攻击依据模型预测的置信度得分生成对抗样本。通用攻击是介于白盒攻击与黑盒攻击之间的一种攻击方法。它以白盒方式产生通用对抗扰动，但攻击时不需要了解受害者模型的信息，而是直接将通用对抗扰动添加到输入样本上生成对抗样本。因此，本文将通用攻击划分为单独的攻击算法。

1.1.1 白盒攻击

快速梯度符号法^[6]（fast gradient sign method, FGSM）使用梯度生成对抗样本，攻击过程可以表述为

$$x_{\text{adv}} = x + \eta \cdot \text{sign}(\nabla_x L(x, y)) \quad (1)$$

其中， x 是带有标签 y 的干净样本， x_{adv} 是对应的对抗样本。 $L(\cdot)$ 是受害者模型的损失函数。 $\text{sign}(\cdot)$ 是符号函数。 η 代表攻击强度， $\eta \in [0, 1]$ 。

基本迭代法^[7]（basic iterative method, BIM）以迭代方式生成对抗扰动，是一种多步攻击方法。BIM可以表述为

$$x'_0 = x \quad (2)$$

$$x'_{t+1} = \text{Clip}_{x, \eta}(x'_t + \alpha \cdot \text{sign}(\nabla_x L(x'_t, y))) \quad (3)$$

其中， x 是干净样本， x'_t 是第 t 次迭代生成的对抗样本， α 为每次迭代的步长， $\text{Clip}(\cdot)$ 是一个裁剪函数，它将输入裁剪到 x 的 η 范围内，即 $\|x - x'_t\|_{\infty} \leq \eta$ 。

投影梯度下降^[8]（project gradient decent, PGD）也是一种多步攻击方法，可以表述为

$$x'_0 = x + R \quad (4)$$

$$x'_{t+1} = \text{Clip}_{x, \eta}(x'_t + \alpha \cdot \text{sign}(\nabla_x L(x'_t, y))) \quad (5)$$

其中， R 是与 x 大小相同的随机噪声，且 $\|R\|_{\infty} \leq \eta$ 。

除了初始化步骤，PGD与BIM攻击流程相同。

DeepFool^[16]将对抗攻击定义为一个优化问题，优化目标是找到跨越决策边界的最小扰动向量。DeepFool可以表述为

$$\arg \min_{\delta_i} \|\delta_i\|_2 \text{ s.t. } f(x) \neq f(x + \delta_i) \quad (6)$$

其中， δ_i 代表第 i 次迭代得到的对抗扰动， $f(\cdot)$ 代表受害者模型，即分类器。DeepFool搜索能够欺骗分类器的最小扰动，并使用几何方法来指导搜索。对于线性分类器（其决策边界是线性平面），描述分类器输出的空间区域可以用多面体表示（其平面是分类器定义的边界平面）。DeepFool在这个多面体中搜索可以改变分类器决策的最小扰动。对于一般的非线性分类器，该算法使用迭代线性化过程来获得近似多面体。

C&W^[17]是一种基于优化的多步攻击方法。C&W的目标是实现最小成本的攻击，即在 l_p 范数度量下以最小的扰动欺骗目标模型。具体来说，具有 l_2 约束的C&W攻击可以表述为

$$\min_w \left\| \frac{1}{2} (\tanh(w) + 1) - x \right\|_2^2 + c \cdot k \left(\frac{1}{2} (\tanh(w) + 1) \right) \quad (7)$$

其中 $k(x) = \max(\max\{z(x)_i; i \neq t\} - z(x), -k)$ 。 k 是置信度，用于控制攻击强度。 $z(\cdot)$ 是目标模型处于softmax层之外的子网络。 $x_{\text{adv}} = \frac{1}{2} (\tanh(w) + 1)$ 是生成的对抗样本。 c 是一个超参数，用于平衡方程式中两项的重要性。

1.1.2 黑盒攻击

简单黑盒对抗攻击^[9]（simple black-box adversarial attack, SimBA）属于基于分数的对抗样

本。该方法在预先指定的一组正交搜索方向中反复选取一个随机方向,利用置信度来检查它是指向还是远离决策边界,并通过对图像进行加或减向量来扰动图像。每次更新都会使图像远离原始图像,并朝决策边界移动。

Square Attack^[10]属于基于分数的攻击方法。它不依赖于局部梯度信息,因此不受梯度掩蔽的影响。Square Attack 基于随机搜索,但不是完全随机的噪声,而是选择方形的区域。选定区域后各个颜色通道独立在范围内选择增量。选择的区域大小可根据不同策略调整。

Boundary Attack^[11]属于基于决策的攻击方法。该方法首先通过最大熵分布采样得到一个初始的样本,该样本服从均匀分布且受害者模型对其分类错误。在迭代过程中,样本将由初始化的样本逐渐向原始样本靠近,直到找到决策边界,并在决策边界上找到与原始样本最近的对抗样本。

1.1.3 通用攻击

通用对抗扰动 (universal adversarial perturbations, UAP)^[18]生成与图像无关的通用对抗扰动 u , 可以将 u 添加到任何输入样本中,并以选定的愚弄率 fr 欺骗模型。其目标是找到满足以下两个约束的 u 。

$$\|u\|_p \leq \varepsilon \quad (8)$$

$$p(f(x+u) \neq f(x)) \geq 1 - fr \quad (9)$$

其中, ε 是最大扰动强度, $f(\cdot)$ 表示受害者模型。本文使用 DeepFool^[16] 攻击来计算 u 。

SGD-UAP^[19]使用随机梯度下降 (stochastic gradient descent, SGD) 方法来生成 UAP。SGD 是 PGD 攻击的一种变体。它是在整个批量上进行优化,而不是在单个图像上进行优化。SGD 已经被证明可以创建比最初提出的使用迭代 DeepFool 攻击更有效的 UAP。

1.2 对抗样本归因

Moayeri 等^[20]提出了一种自监督方法来检测对抗样本并进行攻击算法分类,该方法基于对预训练的自监督编码器的嵌入进行操作的线性模型。具体来说,他们提取预训练的 SimCLR 编码器的嵌入,研究使用少量样本进行对抗样本检测和攻击算法分类。

Dotter 等^[21]使用自己搭建的 11 层神经网络初探了攻击算法、受害者模型、攻击强度和范数的可归因性。然而,他们分开考虑了这 4 个攻击信息的归因,即未考虑这些攻击信息之间的关系。其次,他们仅在 MNIST 和 Cifar10 这样的小数据集上进行了实验。

Li 等^[22]提出攻击工具链库 (attack toolchain library, ATL) 以更方便地生成以机器为中心的攻击 (对抗攻击) 和以人为中心的攻击 (deepfake 攻击),并采用 ResNet50 特征提取器加上一个多层感知机 (multilayer perceptron, MLP) 分类器的结构,探索了攻击算法和受害者模型的归因。

Nicholson 等^[15]提出了攻击族的概念,并将对抗样本的逆向工程表述为监督学习问题,即根据攻击算法和超参数对对抗样本进行分类。他们将 5 种攻击算法分为白盒攻击、黑盒攻击以及补丁攻击 3 个攻击族,每种攻击算法设置 4 种攻击强度,并使用 JPEG 压缩算法提取对抗扰动,最后利用 ResNet50 分别进行攻击算法和攻击强度的归因。

Guo 等^[14]将对抗样本归因定义为归因对抗样本中隐藏的 3 种攻击信息,即攻击算法、受害者模型和攻击强度。目前的研究工作大多数为一种攻击信息 (如攻击算法或受害者模型) 的归因,随着攻击信息可取范围的扩大,攻击信息归因的组合爆炸成为不可避免的问题,即攻击算法+受害者模型+攻击强度的组合类别数的急剧增长将使得单标签分类不适用于对抗样本归因。因此,他们将对归因问题定义为多任务学习问题,并提出一种多任务对抗归因 (multi-task adversarial attribution, MTAA) 方法以同时识别 3 种攻击信息。

2 攻击算法与受害者模型层次结构

2.1 攻击算法层次结构

图 3 展示了本文提出的攻击算法的层次结构,它从粗粒度的攻击算法 (白盒攻击、黑盒攻击与通用攻击) 演变为细粒度的攻击算法 (具体的攻击算法)。考虑到通用攻击是介于白盒攻击与黑盒攻击之间的一种攻击方法,它以白盒方式产生通用对抗扰动,但攻击时不需要了解受害者模型的信息,而是直接将通用对抗扰动添加到输入样本上生成对抗样本,因此本文将通用攻击划分为单独的攻击算法。给定输入图像,MAA-HS

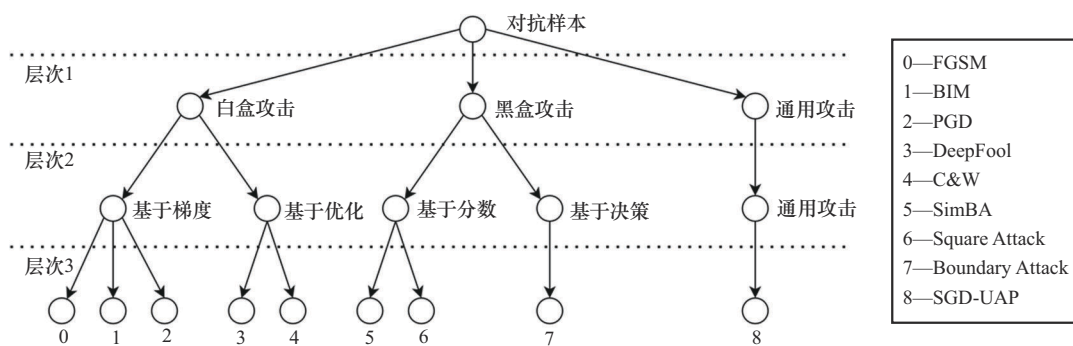


图3 攻击算法的层次结构

Figure 3 Hierarchical structure of attack algorithm

在不同层次执行不同粒度的攻击算法分类，并利用不同攻击算法之间的层次依赖关系，每个节点的分类概率取决于从根到自身的路径。基于这种层次结构，对抗样本归因模型可以学习更全面的特征表示来区分各种攻击算法。例如，C&W节点的分类概率取决于对抗样本→白盒攻击→基于优化的攻击→C&W路径中所有节点的分类概率。预测整个分层路径有助于从粗到细地理解攻击算法的信息，从而捕获不同层次之间依赖关系。

2.2 受害者模型层次结构

图4展示了受害者模型的层次结构，它从粗粒度的DNN结构，即卷积神经网络（convolutional neural networks, CNN）与视觉Transformer（vision transformer, ViT）演变为特定的神经网络模型。本文提出的攻击算法层次结构与受害者模型层次结构均为3层，不同的任务可以设置不同的层次结构。

3 MAA-HS

本文提出的MAA-HS可以在不同层次执行对抗样本归因任务。MAA-HS的整体架构如图5

所示，主要包括3个部分：①对抗特征提取模块，用于提取对抗样本的特征；②多分支特征提取模块，用于在不同层次提取对抗样本的特征；③分类模块，在多分支特征提取模块的每个分支后设置两个分类层，分别用于不同层次的攻击算法分类和受害者模型分类。

3.1 对抗特征提取模块

现有的研究工作大多直接使用对抗样本进行对抗样本的归因。考虑到对抗样本本身不仅包含对抗扰动，还包含原始样本的信息和其他噪声，直接使用对抗样本进行归因效果不好，因此部分工作通过提取对抗样本中的对抗扰动进行归因。研究表明，JPEG压缩误差可以用来估计对抗扰动。现有的许多研究工作使用JPEG压缩算法去除对抗扰动^[23]。Nicholson等^[15]也使用JPEG压缩算法提取对抗扰动进行对抗样本归因。考虑到JPEG压缩算法并不能提取完整的对抗扰动，单一地使用JPEG压缩误差可能会忽略对抗样本中的部分信息，所以本文使用神经网络分别提取对抗样本和JPEG压缩误差的特征，

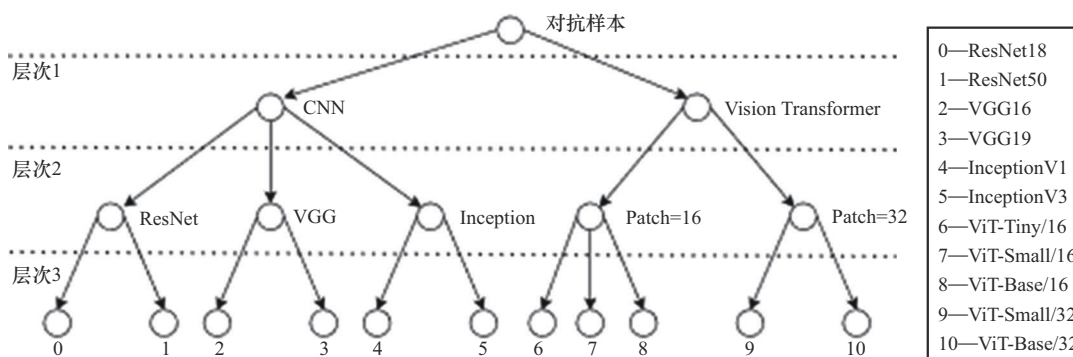


图4 受害者模型的层次结构

Figure 4 Hierarchical structure of victim models

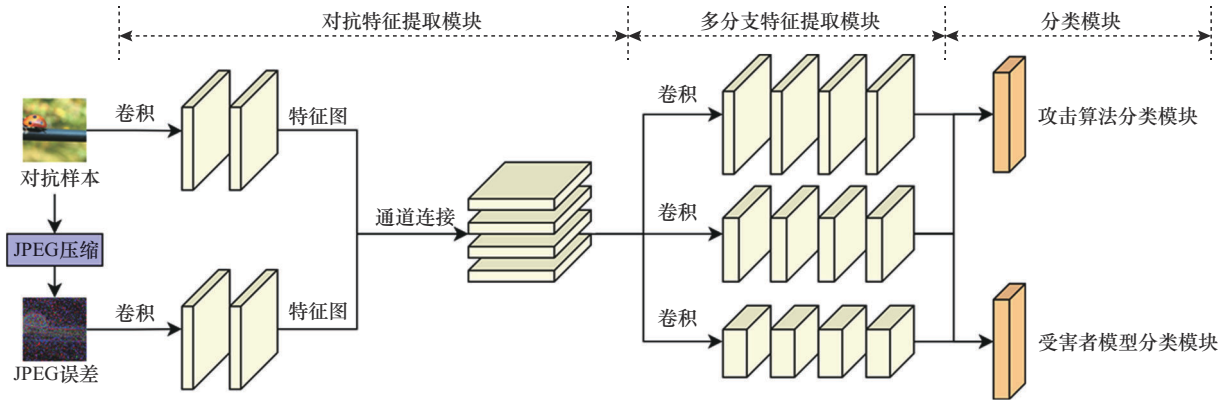


图5 MAA-HS的整体架构

Figure 5 Overall architecture of MAA-HS

并将这两部分特征直接拼接起来送入多分支特征提取模块。

3.2 多分支特征提取模块

多分支特征提取模块如图6所示。 θ_b 表示分支, 对应于攻击算法与受害者模型层次结构中的层次 b 。本文设计的攻击算法和受害者模型层次结构均为3层, 因此 $b \in \{1, 2, 3\}$ 。不同归因任务也可以使用不同的层次结构^[24]。对于多分支特征提取模块, 每个 θ_b 生成特定分辨率的特征图, 这些特征图帮助 θ_b 进行相应层次的细粒度分类。例如, 最精细的层次进行分类(即识别具体的攻击算法)时, 需要对所有空间位置的内容进行建模, 因此需要高分辨率的特征图。相反, 最粗层次进行分类(即识别白盒攻击、黑盒攻击和通用攻击)时, 使用低分辨率特征图是合理的。同时, 多分支特征提取模块并行连接从高到低的卷积流, 通过重复融合多分辨率流的特征来保证不同层次特征之间的连通性。

3.3 分类模块

分类模块如图6所示。本文将攻击算法和受害者模型归因定义为两个分类任务。分类模块的每个分支分别用于执行不同层次的分类任务。对于分支 θ_b , 最小化两个交叉熵损失来优化层次 b 的两个分类任务。

$$L_{\text{cls}_1}^b = \frac{1}{n} \sum_{x \in D} l_{\text{CE}}(\theta_b(x), y_{b_1}) \quad (10)$$

$$L_{\text{cls}_2}^b = \frac{1}{n} \sum_{x \in D} l_{\text{CE}}(\theta_b(x), y_{b_2}) \quad (11)$$

其中, D 代表对抗样本数据集, $\theta_b(x)$ 代表分支 θ_b 的输出 logits, y_{b_1} 和 y_{b_2} 分别代表层次 b 对应的攻击算法和受害者模型的真实标签, l_{CE} 表示交叉熵损失函数。

3.3.1 分层路径预测

鉴于式(10)和式(11)只能分别优化不同层次的攻击算法和受害者模型分类任务, 并没有考虑层次间的依赖关系, 因此本文在式(7)的基础

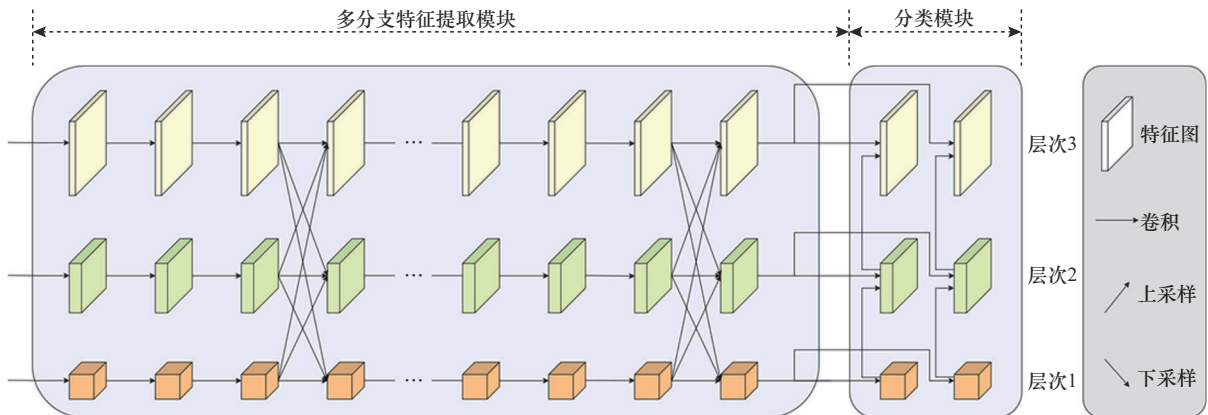


图6 多分支特征提取模块与分类模块

Figure 6 Multi-branch feature extraction module and classification module

上通过添加分层路径预测来学习不同层次之间的依赖关系。

$$p(y_b|x) = \text{softmax}(\theta_b(x) \odot (\lambda + p(y_{b-1}|x))) \quad (12)$$

其中, $\theta_b(x)$ 代表分支 θ_b 的输出 logits, $p(y_b|x)$ 表示分支 θ_b 的预测概率, λ 表示权重。在计算分支 θ_b 处的概率 $p(y_b|x)$ 之前, 首先根据上一个分支的预测概率 $p(y_{b-1}|x)$ 缩放 $\theta_b(x)$, 然后强制算法学习层次依赖性。具体来说, 根据式(12), 分别按照攻击算法和受害者模型的层次结构, 将分支 θ_b 输出的所有 logits 乘以上一层 θ_{b-1} 的预测概率。

图7展示了基于决策的攻击算法和基于分数的攻击算法的分层路径预测。根据攻击算法的层次结构, 分支3(对应于层次3)的输出 logits 需要乘以分支2(对应于层次2)对应攻击算法的预测概率。SimBA 和 Square Attack 属于基于分数的攻击, 因此它们输出的 logits 需要乘以基于分数的攻击的预测概率。同理, Boundary Attack 的输出 logits 需要乘以基于决策的攻击的预测概率。

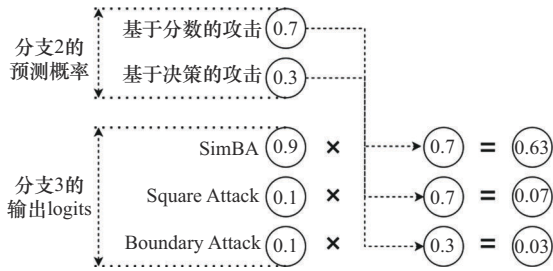


图7 分层路径预测

Figure 7 Hierarchical path prediction

3.3.2 不确定性加权损失

式(10)、式(11)和式(12)仅考虑了单独攻击算法或受害者模型的归因任务。为了缓解组合爆炸

$$\begin{aligned} L(W, \sigma_1, \sigma_2) &= \\ &= -\log(y_1 = c_1, y_2 = c_2 | f^W(x)) = \\ &= -\text{softmax}(y_1 = c_1; f^W(x), \sigma_1) \cdot \text{softmax}(y_2 = c_2; f^W(x), \sigma_2) = \\ &= -\log p(y_1 = c_1 | f^W(x), \sigma_1) - \log p(y_2 = c_2 | f^W(x), \sigma_2) = \\ &= \frac{1}{\sigma_1^2} L_{\text{cls}_1}(W) + \frac{1}{\sigma_2^2} L_{\text{cls}_2}(W) + \log \frac{\sum_{c'_1} \exp\left(\frac{1}{\sigma_1^2} f_{c'_1}^W(x)\right)}{\sum_{c'_1} \exp\left(f_{c'_1}^W(x)\right)^{\frac{1}{\sigma_1^2}}} + \log \frac{\sum_{c'_2} \exp\left(\frac{1}{\sigma_2^2} f_{c'_2}^W(x)\right)}{\sum_{c'_2} \exp\left(f_{c'_2}^W(x)\right)^{\frac{1}{\sigma_2^2}}} \approx \\ &= \frac{1}{\sigma_1^2} L_{\text{cls}_1}(W) + \frac{1}{\sigma_2^2} L_{\text{cls}_2}(W) + \log \sigma_1 + \log \sigma_2 \end{aligned} \quad (16)$$

问题, 同时利用攻击算法和受害者模型这两种攻击信息之间的关系, 本文使用多任务学习模型同时进行攻击算法和受害者模型的归因任务。

受 Kendall 等^[25]的启发, 多任务学习模型的性能在很大程度上取决于不同任务之间的权重。因此, 本文采用了标准不确定性加权损失, 它利用不依赖于输入数据而是依赖于任务不确定性的同方差不确定性。按照最大似然推理的步骤, 首先将分类任务的概率模型描述为

$$p(y|f^W(x), \sigma) = \text{softmax}\left(\frac{1}{\sigma^2} f^W(x)\right) \quad (13)$$

其中, $f^W(x)$ 是具有参数 W 和输入 x 的多任务学习模型的输出。分类似然 $p(y|f^W(x), \sigma)$ 对分类层的输出 logits 进行缩放, 缩放参数为 σ^2 , 使其满足玻尔兹曼分布。

然后对式(13)进行因式分解。

$$p(y_1, \dots, y_k | f^W(x), \sigma) = p(y_1 | f^W(x), \sigma) \cdots p(y_k | f^W(x), \sigma) \quad (14)$$

其中, $y_1, \dots, y_k$ 是多任务模型 k 个任务的输出, 如对抗样本归因中的攻击算法和受害者模型分类任务。

最后取似然函数 $p(y|f^W(x), \sigma)$ 的对数进行最大似然推理。分类任务的最大似然推理如式(15)所示。

$$\log p(y = c | f^W(x), \sigma) = \frac{1}{\sigma^2} f_c^W(x) \quad (15)$$

$$-\log \sum_{c'} \exp\left(\frac{1}{\sigma^2} f_{c'}^W(x)\right)$$

其中, c 为输出 y 的正确标签, $f_c^W(x)$ 是向量 $f^W(x)$ 的第 c 个元素。

在对抗样本归因任务中, 通过将不确定性加权损失添加到联合损失函数并遵循最大似然推理, 可以正式描述组合损失函数如式(16)所示。

其中, $L_{\text{cls}_1}(W)$ 和 $L_{\text{cls}_2}(W)$ 分别代表参数为 W 的攻击算法 y_1 和受害者模型 y_2 分类任务的交叉熵损失, c_1 和 c_2 分别代表 y_1 和 y_2 的正确标签。式(16)中第 4 行使用了式(15)的近似函数。可以通过优化参数 W 、 σ_1 和 σ_2 来最小化联合损失函数。在式(16)中使用显式简化假设 $\frac{1}{\sigma^2} \exp\left(\frac{1}{\sigma^2} f_c^W(x)\right) \approx$

$$\left(\sum_c \exp(f_c^W(x))\right)^{\frac{1}{\sigma^2}}。 \sigma_1 \text{ 和 } \sigma_2 \text{ 用来衡量攻击算法分}$$

类任务和受害者模型分类任务的不确定性, 较大的不确定性度量值会导致对应损失函数的贡献较低。这两个度量值由式(16)中的最后两个对数项调节, 度量值过高时会对其进行惩罚。

3.3.3 损失函数

在训练中, 多分支特征提取模块的每个分支都将针对相应层次的分类任务进行优化。在最后的分类模块, 对攻击算法和受害者模型层次结构的 3 个层次分别使用 3 个多任务分类损失。

$$L_1(W, \sigma_{11}, \sigma_{12}) = \frac{1}{\sigma_{11}^2} L_{\text{cls}_1}^1(W) + \quad (17)$$

$$\frac{1}{\sigma_{12}^2} L_{\text{cls}_2}^1(W) + \log \sigma_{11} + \log \sigma_{12}$$

$$L_2(W, \sigma_{21}, \sigma_{22}) = \frac{1}{\sigma_{21}^2} L_{\text{cls}_1}^2(W) + \quad (18)$$

$$\frac{1}{\sigma_{22}^2} L_{\text{cls}_2}^2(W) + \log \sigma_{21} + \log \sigma_{22}$$

$$L_3(W, \sigma_{31}, \sigma_{32}) = \frac{1}{\sigma_{31}^2} L_{\text{cls}_1}^3(W) + \quad (19)$$

$$\frac{1}{\sigma_{32}^2} L_{\text{cls}_2}^3(W) + \log \sigma_{31} + \log \sigma_{32}$$

其中, σ_{ij} 用于衡量第 i 层次的第 j 个任务的不确定性。 $L_{\text{cls}_1}^b$ 和 $L_{\text{cls}_2}^b$ 分别代表层次 b ($b \in \{1, 2, 3\}$) 攻击算法分类和受害者模型分类的交叉熵损失函数。该架构结合分层路径预测, 以每层不同的学习率进行端到端训练, 总的损失函数为

$$L_{\text{total}} = L_1(W, \sigma_1, \sigma_2) + \quad (20)$$

$$L_2(W, \sigma_1, \sigma_2) + L_3(W, \sigma_1, \sigma_2)$$

4 实验设置与结果分析

4.1 实验设置

本文实验使用 Cifar10 和 Mini-ImageNet 两个数据集。Cifar10 数据集是一个日常物品的小型

数据集, 共有 60 000 张尺寸为 32×32 的图片, 这些图片被分成 10 个类别。Mini-ImageNet 数据集是 ImageNet 数据集的一个子集, 共有 100 个类别, 每个类别有 600 张图片, 每张图片的尺寸为 224×224 。本文实验使用 9 种攻击算法和 11 种受害者模型生成对抗样本, 具体的对抗归因场景如表 1 所示。

表 1 对抗归因场景
Table 1 Adversarial attribution scenario

攻击算法	超参数	受害者模型
FGSM	攻击强度: 8/255	
BIM	攻击强度: 8/255 迭代次数: 20	ResNet18
PGD	攻击强度: 8/255 迭代次数: 50	ResNet50
DeepFool	迭代次数: 50	VGG16
C&W	置信度: 0 迭代次数: 100	VGG19
SimBA		InceptionV1
Square Attack	迭代次数: 5 000	InceptionV3
Boundary Attack		Vit-Tiny/16
UAP-SGD	攻击强度: 10/255	Vit-Tiny/32
		Vit-Small/16
		Vit-Small/32
		Vit-Base/32

对于 FGSM、BIM 和 PGD 攻击, 攻击强度设置为 8/255, BIM 攻击的迭代次数为 20 次, PGD 攻击的迭代次数为 50 次。对于 DeepFool 攻击和 C&W 攻击, 使用原始论文的参数生成对抗样本。对于 SimBA、Square Attack 和 Boundary Attack, 考虑到时间成本, 攻击迭代次数均设置为 5 000 次。由于 UAP-SGD 攻击是通用攻击, 需要使用较大的扰动强度才能保证攻击成功, 所以攻击强度设置为 10/255。

预训练的 Vision Transformer 只能处理具有较大分辨率的图像, 因此将 Cifar10 数据集中的图片放大到 224×224 。考虑到需要归因的攻击算法与受害者模型数量较多, 对于每种攻击算法+受害者模型, 从数据集中随机选择 200 个样本作为训练集, 50 个样本作为测试集。

本文实验在 PyTorch 框架下进行训练和测试。处理器为 Intel Xeon Platinum 8352V (主频为 2.10 GHz, 内存为 120 GB)、GPU 为 NVIDIA

RTX4090（显存为24 GB）。对于分层结构的多任务学习方法，使用JPEG压缩得到对抗样本的JPEG压缩误差，压缩率设置为90%，并使用两个卷积层分别提取对抗样本和与其对应的JPEG压缩误差的特征，然后将这两部分特征图连接起来送入多分支特征提取模块。对于多分支特征提取模块，分支数设置为3，每个分支的输入特征图分辨率分别为 224×224 、 112×112 、 56×56 。对于分类模块，使用两个卷积层作为任务特定层。对于分层路径预测，式(12)中的 λ 设置为1。优化器使用Adam，初始学习率设置为0.000 01，采用余弦退火学习率调节算法，共训练100次，每个批量大小为50。

4.2 评价指标与基准模型

本文归因场景主要是分类，因此使用分类准确率来评估攻击算法和受害者模型归因的性能。本文将所提MAA-HS方法与SMC（supervised meta-classifier）^[22]、fingerprinting^[15]、MTAA^[14]和HiFi-Net^[24]4种方法进行了比较。SMC方法使用ResNet50特征提取器和MLP（multilayer perceptron）分类器进行对抗样本归因；fingerprinting使用ResNet50作为主干网络，使用JPEG压缩算法提取对抗扰动进行对抗样本归因；MTAA使用ResNet101特征提取器，利用多任务学习框架进行对抗样本归因；HiFi-Net通常用于深度伪造检测和定位，也可用于对抗样本归因。考虑到不同方法使用了ResNet的不同版本作为特征提取器，本文实验统一使用ResNet101作为特征提取器。

4.3 实验结果与分析

表2和表3分别展示了本文所提MAA-HS方法在Cifar10和Mini-ImageNet数据集上对抗归因的结果。对于第3层次的攻击算法和受害者模型分类任务，MAA-HS方法在测试数据集上的表现良好。在Cifar10数据集上分别达到99.85%和72.74%的准确率，在Mini-ImageNet数据集上分别达到98.36%和63.19%的准确率。然而，Cifar10是只有10个分类的小型数据集，主干网络的性能会饱和，因此MAA-HS在Cifar10数据集上的优秀表现不如在Mini-ImageNet数据集上显著。表2为Cifar10数据集的对抗归因结果对比，

表3为Mini-ImageNet数据集的对抗归因结果对比。得益于JPEG压缩对对抗扰动的估计，与fingerprinting相比，SMC和MTAA的对抗归因准确率较低。尽管HiFi-Net通常被用于深度伪造检测，但在本文实验中的准确率依然较高。与深度伪造检测不同，HiFi-Net的特征提取方法并不完全适用于对抗归因。同时，从表2和表3中可以看出，层次1和层次2的分类准确率均高于层次3的分类准确率，即粗粒度攻击算法归因比细粒度攻击算法归因更简单，受害者模型同理。这验证了MAA-HS方法的合理性，即最精细级别（识别具体的攻击算法）分类使用高分辨率特征图，最粗级别（识别白盒攻击、黑盒攻击和通用攻击）分类使用低分辨率特征图。

表2 Cifar10数据集的对抗归因结果对比
Table 2 Comparison of adversarial attribution results on Cifar10 dataset

对抗归因方法	攻击算法			受害者模型		
	层次3	层次2	层次1	层次3	层次2	层次1
SMC	96.09%	99.57%	99.57%	68.72%	78.18%	95.03%
fingerprinting	98.03%	99.63%	99.71%	70.90%	79.93%	95.19%
MTAA	96.42%	99.69%	99.81%	69.61%	78.72%	95.29%
HiFi-Net	98.26%	99.72%	99.85%	70.86%	80.04%	95.21%
MAA-HS	99.85%	99.97%	99.97%	72.74%	80.48%	95.63%

表3 Mini-ImageNet数据集的对抗归因结果对比
Table 3 Comparison of adversarial attribution results on Mini-ImageNet dataset

对抗归因方法	攻击算法			受害者模型		
	层次3	层次2	层次1	层次3	层次2	层次1
SMC	91.07%	98.80%	98.46%	58.38%	68.14%	84.30%
fingerprinting	95.27%	99.31%	99.45%	61.52%	70.62%	86.74%
MTAA	91.59%	99.31%	99.09%	59.09%	68.70%	85.37%
HiFi-Net	95.79%	99.53%	99.57%	61.03%	69.13%	86.02%
MAA-HS	98.36%	99.89%	99.89%	63.19%	71.23%	87.49%

为了更直观地展示MAA-HS方法的攻击算法和受害者模型分类准确率，本文绘制了混淆矩阵。图8和图9分别展示了MAA-HS方法在Cifar10数据集和Mini-ImageNet数据集上攻击算法和受害者模型分类准确率。

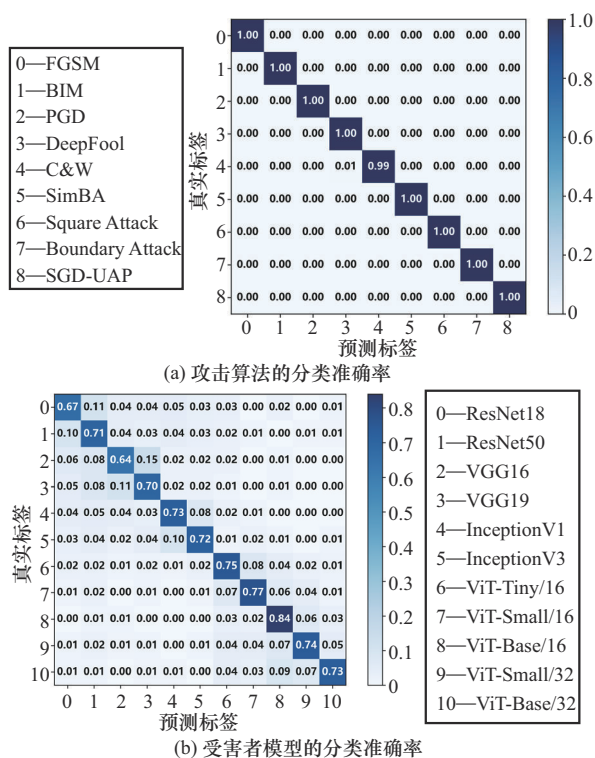


图8 Cifar10数据集上攻击算法和受害者模型的分类准确率
Figure 8 Classification accuracy of attack algorithms and victim models on Cifar10 dataset

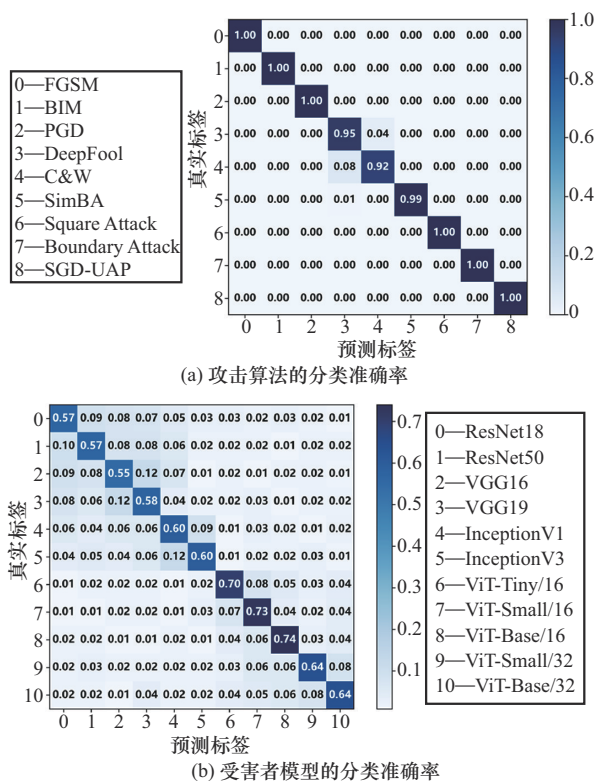


图9 Mini-ImageNet数据集上攻击算法和受害者模型的分类准确率
Figure 9 Classification accuracy of attack algorithms and victim models on Mini-ImageNet dataset

从混淆矩阵可以看出,除了DeepFool和C&W外的所有攻击生成的对抗样本均被正确分类。与基于梯度的攻击不同,DeepFool和C&W攻击尽管可以访问受害者模型,但不直接使用受害者模型损失函数的梯度,因此生成的对抗扰动难以被准确地识别和分类。黑盒攻击一般采用添加噪声方式生成对抗样本,不同的攻击具有不同的模式,很容易被对抗归因网络识别出来。同样,通用攻击产生相同的对抗扰动,因此更容易分类。

从混淆矩阵可以观察到, Vision Transformer的分类准确率均高于CNN。Vision Transformer独特的网络结构使得其相比CNN更容易归因。同时还可以观察到,图3中具有相同父节点的攻击算法之间的分类误判率较高。例如,DeepFool和C&W均属于基于优化的攻击,它们之间的分类误判率明显高于其他攻击算法。这验证了本文对不同攻击算法和模型结构的划分方法的合理性。

从上述实验结果可以看出,受害者模型归因效果一般。为了探究分类准确率较低的原因,表4和表5分别展示了Cifar10和Mini-ImageNet数据集上不同对抗样本受害者模型的归因结果。

表4 Cifar10数据集不同对抗样本受害者模型的归因结果

Table 4 Attribution results of victim models for different adversarial examples on Cifar10 dataset

攻击算法	层次3	层次2	层次1
FGSM	99.81%	100%	100%
BIM	100%	100%	100%
PGD	99.45%	100%	100%
DeepFool	72.00%	80.72%	96.54%
C&W	92.90%	98.36%	99.45%
SimBA	31.45%	47.27%	81.09%
Square Attack	29.09%	51.09%	95.45%
Boundary Attack	30.18%	46.36%	88.18%
SGD-UAP	100%	100%	100%

从表4和表5可以看出,除DeepFool和C&W外,黑盒攻击的受害者模型分类准确率远低于平均准确率。由于黑盒攻击不能直接访问受

害者模型的结构信息，只能依据受害者模型的输出结果生成对抗样本，而不同受害者模型输出结果相近，因此黑盒攻击的受害者模型分类准确率较低。至于DeepFool和C&W，它们不直接使用受害者模型损失函数的梯度，导致其受害者模型分类准确率低于基于梯度的攻击和通用攻击的受害者模型分类准确率。

表5 Mini-ImageNet数据集不同对抗样本受害者模型的归因结果对比

Table 5 Attribution results of victim models for different adversarial examples on Mini-ImageNet dataset

攻击算法	层次3	层次2	层次1
FGSM	98.90%	99.81%	100%
BIM	98.18%	99.63%	100%
PGD	98.36%	99.27%	99.81%
DeepFool	51.81%	60.36%	87.45%
C&W	52.00%	63.63%	88.90%
SimBA	21.45%	37.27%	63.45%
Square Attack	23.45%	37.63%	70.36%
Boundary Attack	23.63%	41.63%	76.18%
SGD-UAP	100%	100%	100%

4.4 消融实验

为了评估MAA-HS中不同组件（如多任务学习、JPEG压缩率、分层路径预测等）的性能，本文开展了消融实验。表6为Cifar10数据集上消融实验结果，表7为Mini-ImageNet数据集上消融实验结果。在表6和表7中，第1行是不使用多任务学习的结果（即分别进行攻击算法与受害者模型分类），第2行至第4行是分别使用压缩率为95%、85%、80%的JPEG压缩的实验结果，第5行是无JPEG压缩的实验结果（即只使用对抗样本进行归因），第6行是只使用对抗样本的JPEG误差进行归因的结果，第7行是无分层路径预测的实验结果，第8行是没有层次1、层次2分类损失的结果（即只保留层次3的分类损失），第9行是本文模型的归因结果。

从表6和表7的第1行可以看出，多任务学习不仅可以缓解组合爆炸问题，同时有利于利用

不同攻击信息之间的关系提升归因性能。表6和表7的第2到第4行可以看出，不同的JPEG压缩率会对归因性能产生一定的影响，因此选择合适的JPEG压缩率同样重要。表6和表7的第5和第6行提供了缺少JPEG压缩误差和原始对抗样本的实验结果，相比原始对抗样本，缺少JPEG压缩误差对归因性能的影响更大。表6和表7的第7行结果表明，执行细粒度分类，而不强制各层之间的依赖关系，会损害分层对抗信息的学习，导致归因性能下降。表6和表7的第8行是在训练中仅保留第3级细粒度分类损失的实验结果，可以看出归因性能明显下降，这验证了在不同层次上执行细粒度分类的有效性。

表6 Cifar10数据集上消融实验结果

Table 6 Results of ablation experiments on Cifar10 dataset

模型结构	攻击算法			受害者模型		
	层次3	层次2	层次1	层次3	层次2	层次1
1	99.49%	99.95%	99.95%	72.26%	80.42%	95.41%
2	99.09%	99.93%	99.95%	72.32%	80.58%	95.29%
3	99.59%	99.95%	99.95%	71.83%	79.69%	95.23%
4	99.15%	99.91%	99.93%	71.31%	78.76%	95.17%
5	98.80%	99.89%	99.91%	70.74%	79.73%	95.11%
6	99.41%	99.91%	99.93%	71.97%	79.87%	95.27%
7	98.28%	99.87%	99.89%	71.65%	79.39%	95.35%
8	98.40%	—	—	70.30%	—	—
9	99.85%	99.97%	99.97%	72.74%	80.48%	95.63%

表7 Mini-ImageNet数据集上消融实验结果

Table 7 Results of ablation experiments on Mini-ImageNet dataset

模型结构	攻击算法			受害者模型		
	层次3	层次2	层次1	层次3	层次2	层次1
1	98.30%	99.85%	99.87%	63.11%	71.23%	87.49%
2	97.41%	99.81%	99.83%	62.57%	71.06%	87.13%
3	98.20%	99.83%	99.83%	62.98%	70.14%	87.31%
4	97.81%	99.79%	99.81%	61.42%	69.92%	87.14%
5	95.73%	99.53%	99.57%	61.12%	70.33%	86.12%
6	97.49%	99.77%	99.85%	62.52%	71.01%	87.11%
7	97.23%	99.77%	99.79%	61.52%	70.84%	87.23%
8	97.05%	—	—	60.62%	—	—
9	98.36%	99.89%	99.89%	63.19%	71.25%	87.55%

5 结束语

对抗样本归因用于识别对抗样本背后的信息(如攻击算法和受害者模型),是逆向工程的重要组成部分。现有方法更多地将对抗样本归因转化为单任务学习。为了提高对抗样本归因的准确率,以及满足不同归因粒度的需求,本文提出了一种基于分层结构的多任务对抗样本归因方法 MAA-HS。具体来说,首先提出了攻击算法和受害者模型的层次结构,用来对攻击算法和受害者模型进行分层;然后提出了针对对抗样本归因的分层路径预测,用于在不同层次对每个对抗样本进行不同粒度的归因;最后结合多任务学习框架同时对攻击算法和受害者模型进行归因。在 Cifar10 和 Mini-ImageNet 数据集上对所提方法进行了实验验证,实验结果证明了本文所提方法的有效性。

在测试时,若出现训练集中未包含的攻击算法和受害者模型,本文方法无法对这些未知对抗样本进行拒绝识别操作。因此,后续的工作将着重于开集场景下的对抗样本归因,在保证已知类别的分类准确率的同时,对未知类别进行拒识,从而实现未知对抗样本归因。

参考文献:

- [1] 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.
- [2] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 3431-3440.
- [3] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 580-587.
- [4] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//Proceedings of the 2nd International Conference on Learning Representations. 2014: 1-10.
- [5] 潘文雯, 王新宇, 宋明黎, 等. 对抗样本生成技术综述[J]. 软件学报, 2020, 31(1): 67-81.
PAN W W, WANG X Y, SONG M L, et al. Survey on generating adversarial examples[J]. Journal of Software, 2020, 31(1): 67-81.
- [6] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//Proceedings of the International Conference on Learning Representations. 2018: 1-11.
- [7] KURAKIN A, GOODFELLOW I J, BENGIO S. Artificial intelligence safety and security[M]. New York: Chapman and Hall/CRC, 2018: 99-112.
- [8] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]//Proceedings of the 6th International Conference on Learning Representations. 2018: 1-28.
- [9] GUO C, GARDNER J, YOU Y R, et al. Simple black-box adversarial attacks[C]//Proceedings of the 36th International Conference on Machine Learning. 2019: 2484-2493.
- [10] ANDRIUSHCHENKO M, CROCE F, FLAMMARION N, et al. Square attack: a query-efficient black-box adversarial attack via random search[C]//Proceedings of European Conference on Computer Vision. 2020: 484-501.
- [11] BRENDEN W, RAUBER J, BETHGE M. Decision-based adversarial attacks: reliable attacks against black-box machine learning models[C]//Proceedings of the 6th International Conference on Learning Representations. 2018: 1-12.
- [12] 严飞, 张铭伦, 张立强. 基于边界值不变量的对抗样本检测方法[J]. 网络与信息安全学报, 2020, 6(1): 38-45.
YAN F, ZHANG M L, ZHANG L Q. Adversarial examples detection method based on boundary values invariants[J]. Chinese Journal of Network and Information Security, 2020, 6(1): 38-45.
- [13] 刘西蒙, 谢乐辉, 王耀鹏, 等. 深度学习中的对抗攻击与防御[J]. 网络与信息安全学报, 2020, 6(5): 36-53.
LIU X M, XIE L H, WANG Y P, et al. Adversarial attacks and defenses in deep learning[J]. Chinese Journal of Network and Information Security, 2020, 6(5): 36-53.
- [14] GUO Z Y, HAN K J, GE Y, et al. Attribution of adversarial attacks via multi-task learning[C]//Proceedings of the 30th International Conference on Neural Information Processing. Singapore: Springer Nature Singapore, 2023: 81-94.
- [15] NICHOLSON D A, EMANUELE V. Reverse engineering adversarial attacks with fingerprints from adversarial examples[J]. arXiv preprint arXiv:2301.13869, 2023.
- [16] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. Deepfool: a simple and accurate method to fool deep neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 2574-2582.
- [17] 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.

- [18] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 1765-1773.
- [19] CO K T, MUÑOZ-GONZÁLEZ L, KANTHAN L, et al. Universal adversarial robustness of texture and shape-biased models[C]//Proceedings of 2021 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE Press, 2021: 799-803.
- [20] MOAYERI M, FEIZI S. Sample efficient detection and classification of adversarial attacks via self-supervised embeddings[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2021: 7677-7686.
- [21] DOTTER M, XIE S, MANVILLE K, et al. Adversarial attack attribution: discovering attributable signals in adversarial ML attacks[C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence Workshop. 2021:1-9.
- [22] LI Y. Supervised classification on deep neural network attack tool-chains[D]. Boston: Northeastern University, 2021.
- [23] GUO C, RANA M, CISSE M, et al. Countering adversarial images using input transformations[C]//Proceedings of International Conference on Learning Representations. 2018:1-12.
- [24] GUO X, LIU X H, REN Z Y, et al. Hierarchical fine-grained image forgery detection and localization[C]//Proceedings of the IEEE CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 3155-3165.
- [25] KENDALL A, GAL Y, CIPOLLA R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7482-7491.

[作者简介]



孙旭（1999-），男，江苏连云港人，南京邮电大学硕士生，主要研究方向为对抗样本归因、对抗样本检测。



张文琼（1998-），女，山东日照人，南京邮电大学博士生，主要研究方向为机器学习公平性、因果推断。



龙显忠（1985-），男，河南信阳人，南京邮电大学副教授，主要研究方向为自监督学习、对抗机器学习。



李云（1974-），男，安徽安庆人，南京邮电大学教授、博士生导师，主要研究方向为对抗机器学习、数据挖掘、自然语言处理、模式识别。