人工智能安全理论及验证平台

功能说明文档

编写单位：浙江大学网络空间安全学院

编写时间：2022 年 6 月

目录

第 1 章 引言

近年来，随着技术发展和国家政策红利的支持，我国已经迅速涌现了一大批“AI+X”的示范应用和场景。然而，其安全隐患也逐渐暴露，一旦被攻击者加以利用，轻则损害个人利益，重则造成人身伤害，甚至危害公共与国家安全。譬如，支付宝人脸识别系统被 1153 个伪造人脸模型欺骗；截止 2022 年 6 月，Tesla 智能汽车因 AI 算法漏洞导致 271 人死亡；IBM Watson 智能医疗诊断系统疗程误诊率达到 35%；GPT-3 开发中 bug 导致训练数据污染，损失超过千万美元。

AI 系统的安全问题严重阻碍了 AI 技术的广泛应用落地。针对 AI 系统缺乏有效的安全验证、测试和增强的问题，浙江大学网络空间安全学院人工智能安全团队搭建了**人工智能安全理论及验证平台**。人工智能安全理论及验证平台凝练 AI 系统全生命周期的安全性需求和一体化架构在安全攻防方面的功能需求，研制了**全周期安全性验证系统**、**软硬件协同安全攻防测试系统**。全周期安全性验证系统能够对数据和模型中存在的安全威胁进行评估和防御；软硬件协同安全攻防测试系统集成多种协同安全功能，以满足 AI 系统的验证需求。该平台受到科技创新 2030 人工智能项目“人工智能安全理论及验证平台”、淘宝（中国）软件有限公司和湖南四方天箭信息科技有限公司资助。

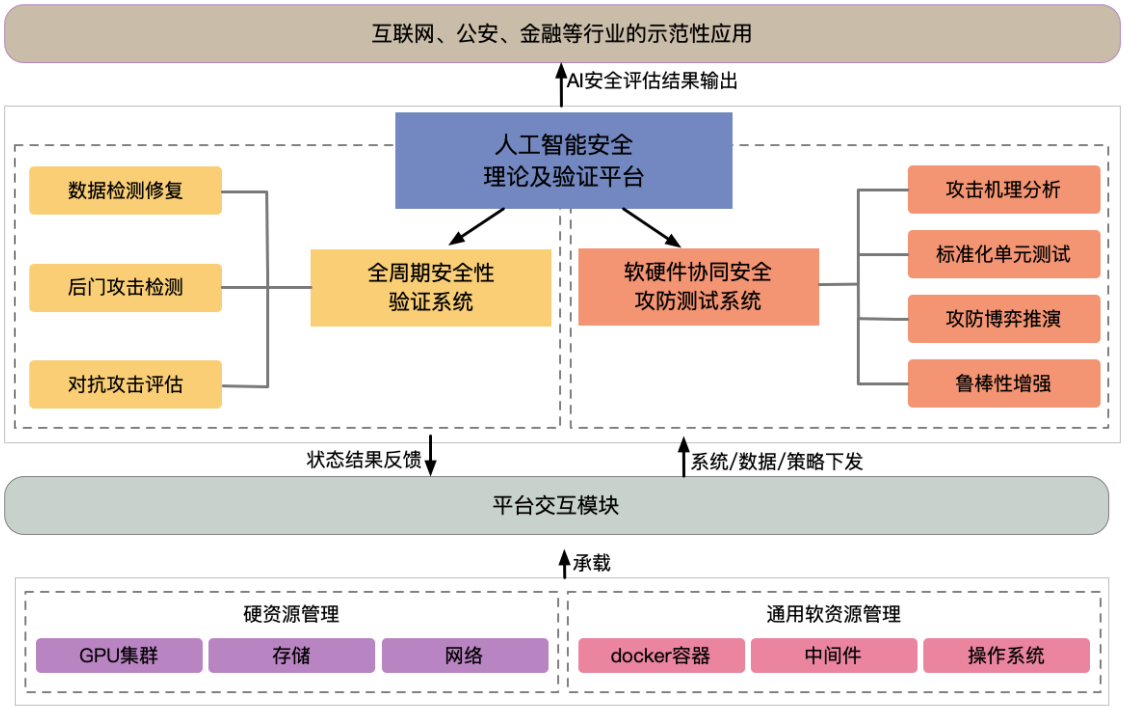

图 1.1: 人工智能安全理论及验证平台功能框架

平台架构如图 1.1所示，集成了全周期安全性验证系统和软硬件协同安全攻防测试系统，其中全周期安全性验证系统包含数据检测修复、后门攻击检测、对抗攻击评估功能；软硬件协同安全攻防测试系统包含攻击机理分析、标准化单元测试、攻防博弈推演、鲁棒性增强功

能。平台内置 MNIST、CIFAR-10 和 ImageNet 3 个数据集，支持 ResNet、VGG 和 LeNet 3 种系列 AI 模型，可调用 9 种主流对抗攻击算法和 3 种解释算法，集成 Pytorch 为主的 5 种开发框架。平台实现对数据、模型进行安全性评估的同时，提供鲁棒性增强方案、标准化单元测试准则以及攻防策略。最终输出 AI 安全评估结果，为示范性应用提供可靠的验证方案。

本文档介绍了人工智能安全理论及验证平台中使用到的基本概念、系统架构和核心功能。

第 2 章 概念和定义

2.1 数据集介绍

根据国内 T/CESA 1026-2018 《人工智能-深度学习算法评估规范》[1]及国际 ISO/IEC TR 24029-1:2021《人工智能-神经网络鲁棒性评估》[2]，ISO/IEC TR 29119-11:2020《软件和系统工程-软件测试-人工智能系统测试》[3]等标准，本平台采用 CIFAR-10、MNIST、ImageNet 作为测试数据集。

2.1.1 CIFAR-10

CIFAR-10 数据集 [8] 共有 60000 张彩色图像，这些图像尺寸为 32 × 32，分为 10 个类，每类 6000 张图。这里面有 50000 张用于训练，构成了 5 个训练批，每一批 10000 张图；另外 10000 用于测试，单独构成一批。测试批的数据里，取自 10 类中的每一类，每一类随机取 1000 张。抽剩下的就随机排列组成了训练批。注意一个训练批中的各类图像并不一定数量相同，总的来看训练批，每一类都有 5000 张图。

CIFAR-10 是国外 Kaggle 机器学习比赛中最常用数据集之一，也是 AI 研究中最常用的数据集之一，同时也被列入美国国家标准与技术研究院（National Institute of Standards and Technology，NIST）的 AI 标准化数据集。

2.1.2 MNIST

MNIST 数据集 [10] 来自美国国家标准与技术研究院。训练集（Training set）由来自 250 个不同人手写的数字构成，其中 50% 是高中学生，50% 来自人口普查局（the Census Bureau）的工作人员。测试集（Test set）也是同样比例的手写数字数据。

MNIST 是国外 Kaggle 机器学习比赛中最常用数据集之一，也是 AI 研究中最常用的数据集之一，同时也被列入美国国家标准与技术研究院的 AI 标准化数据集。

2.1.3 ImageNet

ImageNet 数据集 [18] 是一个计算机视觉数据集，是由斯坦福大学的李飞飞教授带领创建。该数据集包含 14197122 张图片。ImageNet 数据集是为了促进计算机图像识别技术的发展而设立的一个大型图像数据集。ImageNet 数据集中的图片涵盖了大部分生活中会看到的图片类别。ImageNet 最初是拥有超过 100 万张图像的数据集。

[1] http://std.samr.gov.cn/gb/search/gbDetailed?id=B06B57B38562B51DE05397BE0A0A4F3F

[2] https://www.iso.org/standard/77609.html

[3] https://www.iso.org/standard/79016.html

ImageNet（2012）为计算机视觉领域中常用数据集之一，也是每年 ILSVRC 大规模视觉识别挑战赛图像分类、目标定位、目标检测等项目的指定数据集。

2.2 模型介绍

2.2.1 ResNet

残差神经网络 ResNet [7] 由微软研究院何凯明等人提出，通过 ResBlock 添加连接跳过多个卷积层，解决了模型退化问题。目前平台支持模型有 ResNet18/ResNet34/ResNet50。图 2.1是 ResNet18 的模型示意图，包含 8 个 ResBlock，卷积层和全连接层共 18 层。

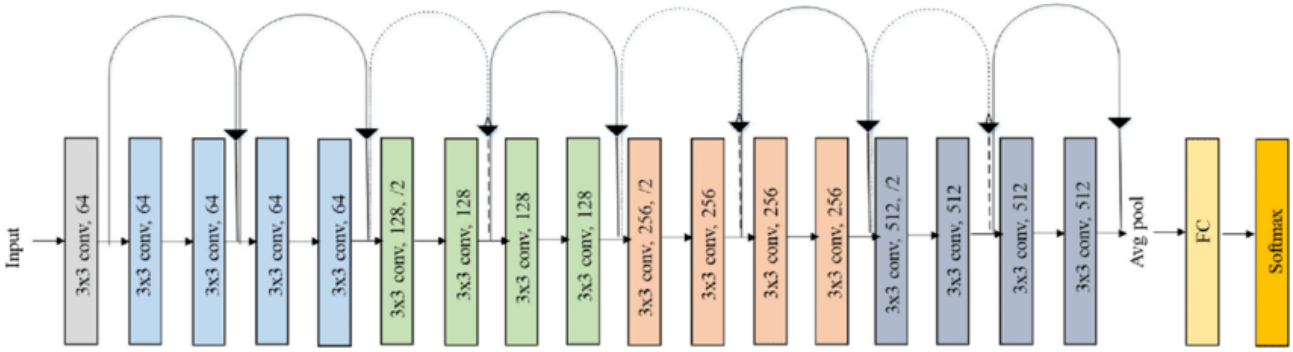

图 2.1: ResNet18 模型示意图

2.2.2 VGG

VGGNet [20] 是来自牛津大学几何组（VGG）和 DeepMind 公司的深度网络模型。VGG 采用堆积的小卷积核替代大的卷积核，因为多层非线性层可以增加网络深度来保证学习更复杂的模式，而且代价还比较小，该网络证明了增加网络的深度能够在一定程度上提升网络的性能。目前平台支持模型有 VGG13/VGG16/VGG19。图 2.2是 VGG16 的模型示意图，卷基层和全连接层共 16 层。

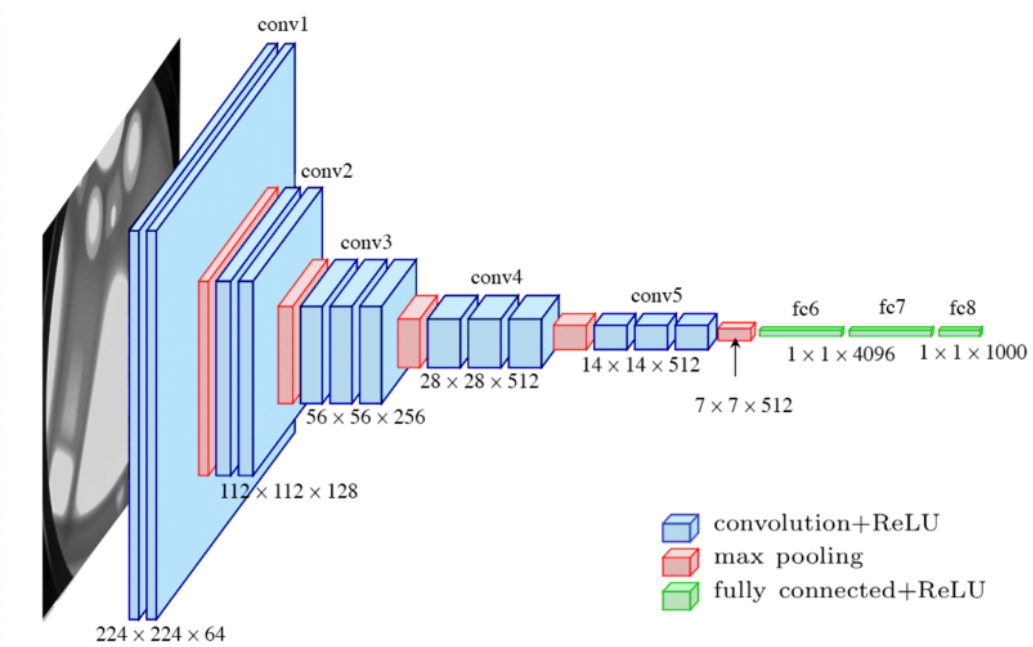

图 2.2: VGG16 模型示意图

2.2.3 LeNet

LeNet [10] 由贝尔实验室的研究员 Yan LeCun 提出，利用后向传播算法来训练多层神经网络，是最早的卷积神经网络之一。LeNet 模型如图 2.3所示，由包含两个卷积层的卷积编码器和包含三个全连接层的全连接层密集块两部分组成。

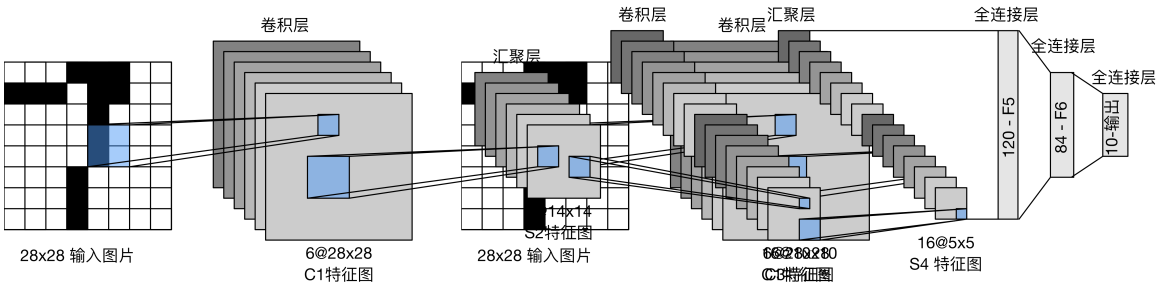

图 2.3: LeNet 模型示意图

2.3 攻击算法介绍

2.3.1 BadNets

BadNets 后门攻击算法 [6] 通过添加指定的触发器来修改某些训练样本，这些带有攻击者指定目标标签的中毒训练样本和良性训练样本将一同被输入到网络模型中进行训练，攻击过程如图 2.4所示。

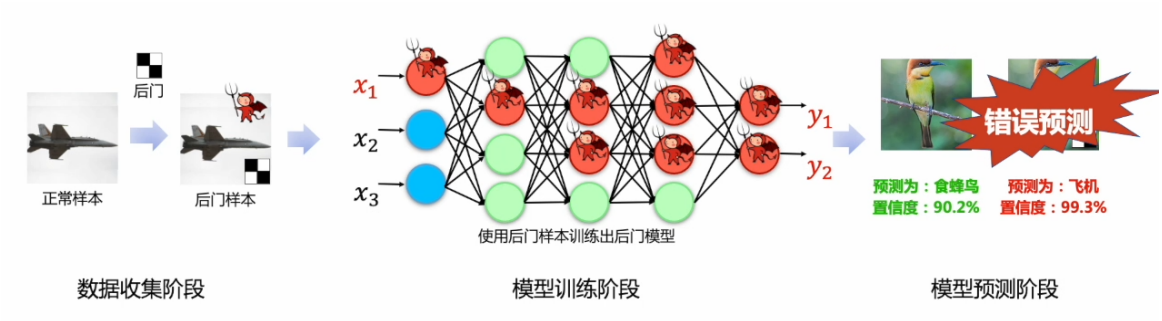

图 2.4: 后门攻击示意图

2.3.2 Trojaning Attack

Trojaning Attack 后门攻击算法 [12] 对已训练好的神经网络进行逆向，生成一个通用的后门触发器，然后利用逆向后的训练数据对模型进行再训练，将后门攻击注入到神经网络模型中。

2.3.3 FGSM

快速梯度符号法 [5]。在给定输入数据后，利用已训练的模型输出预测并计算损失函数的梯度，然后使用梯度的符号来创建使损失最大化的新数据。

2.3.4 RFGSM

随机 FGSM [21]。是 FGSM 的一种变体，在应用 FGSM 产生的对抗扰动之前，给在输入样本中增加一个小的随机扰动，这有助于避免梯度 Mask 的防御策略。

2.3.5 FFGSM

快速 FGSM [5]。是 FGSM 的一种变体，在应用 FGSM 产生的对抗扰动之前，给在输入样本中增加一个小的随机扰动。与 R-FGSM 不同的是，扰动以均匀分布代替高斯分布。

2.3.6 MIFGSM

基于动量的迭代 FGSM [3]。在 I-FGSM 的基础上，通过将动量项整合到攻击的迭代过程中，使计算结果能摆脱局部最优，并且增加了更新方向的稳定性。

2.3.7 PGD

投影梯度下降法 [13]。是 FGSM 的迭代版本，与 BIM 不同的是，它对每次迭代的结果进行裁剪，保证新样本的各个像素都在 x 的 ϵ 邻域（$L\infty$）内。

2.3.8 BIM

迭代式 FGSM [9]。是 FGSM 方法的变体，每轮迭代在上一步算得的对抗样本基础上，各像素增加（或减少）一个常数。

2.3.9 DI2FGSM

输入多样的迭代 FGSM [23] 是在 I-FGSM 的基础上，在样本输入模型之前，以一定的概率对其进行 resize、padding 等操作，增强了攻击的鲁棒性，使得黑盒攻击成功率显著增加。

2.3.10 PGDL2

L2 范数投影梯度下降法 [13] 是 PGD 算法另一个版本，其对每次的迭代结果采用 L2 范数裁剪。

2.3.11 EOTPGD

变换期望 PGD [11] 是将 EOT 的思想加入到 PGD 算法中，即迭代中用损失函数梯度的期望代替符号梯度本身。

2.4 防御算法介绍

2.4.1 STRIP

STRIP 算法 [4] 将多张不含触发器的干净图像分别叠加到测试图像上，并将叠加后的图像输入到分类模型中，对模型输出的标签概率分布结果进行熵值计算，根据熵值计算结果判断测试图像是否含有触发器。

2.4.2 Neural Cleanse

基于后门会改变模型决策边界的思想，Neural Cleanse 算法 [22] 针对分类器中的标签构造对应的触发器，并根据触发器的异常值大小判断其是否为后门触发器。

2.5 解释算法介绍

2.5.1 LRP

LRP 算法 [1] 也被称为层间相关性传播算法。LRP 计算输入样本和分类预测结果的相关性系数，想关系系数从输出层反向传播回输入层，最终以权重的形式映射到原图片每个像素中，即层间置信度相关性值，通过颜色映射算法可视化层间置信度相关性值，获得输入像素的热力图。

2.5.2 Grad-CAM

Grad-CAM 算法 [19] 获取模型网络最后一层卷积神经元输出作为提取特征，对特征分类输出值求导，导数作为权重与特征进行相乘，获得模型关注区域的权重矩阵，通过颜色映射生成热力图。

2.5.3 LIME

LIME 算法 [17] 将原始样本随机采样得到训练样本集，依据多个训练样本集训练的逻辑回归模型模拟原始模型的局部边界，通过局部模型来对复杂模型的分类进行分析，依据局部模型中的权重获得模型预测时参考特征的重要度，根据重要度可视化处理原图得到热力图。

2.6 环境介绍

平台运行环境如表格 2.1所示：

操作系统	编程语言	深度学习框架	显卡
Ubuntu 20.04 LTS	Python 3.8.5	Pytorch 1.7.1 Tensorflow 2.1 PaddlePaddle 2.1.1 Keras 2.1.1 Theano 1.0.4	GeForce RTX 3090 NVIDIA 470.82.00

表 2.1: 环境配置描述

2.7 本文组织

人工智能安全理论及验证平台工功能说明文档包括以下章节：

- 第一章 引言：介绍了 AI 系统当前存在的安全威胁、人工智能安全理论及验证平台的整体架构和文档说明。
- 第二章 概念和定义：介绍了系统中数据集、模型与算法，配置环境和文档组织结构。
- 第三章 全周期安全验证系统：介绍了全周期安全验证系统 3 个功能的支持配置和功能实现。
- 第四章 软硬件协同安全攻防测试系统：介绍了软硬件协同安全攻防测试系统 4 个功能的支持配置和功能实现。

第 3 章 全周期安全性验证系统

全周期安全性验证系统针对 AI 系统数据收集阶段和模型训练阶段存在的安全威胁，如数据收集阶段的异常数据、模型训练阶段的数据投毒攻击和模型后门攻击，进行数据异常检测修复、后门攻击防御。系统还对训练后的模型进行不同强度和方法的对抗攻击评估，最终得到全局的安全评估结果。

数据检测修复支持 MNIST 和 CIFAR-10 2 个数据集；后门攻击检测支持检测 BadNets 和 Trojaning Attack 2 种后门攻击，支持 STRIP 和 Neural Cleanse2 种后门攻击检测算法；对抗攻击评估支持 FGSM、FFGSM、RFGSM、MIFGSM、BIM 和 PGD 6 种对抗攻击算法。

3.1 数据检测修复

3.1.1 功能描述

模型训练前需将数据集进行预处理，避免异常数据在训练时影响模型结构，导致应用时模型准确率下降。数据检测修复功能首先检测出异常数据并分类，然后对不同类型异常数据进行相应处理，最后对清洁数据集评估，为后续步骤提供数据支撑。

3.1.2 支持数据集

- CIFAR-10，详见章节 2.1.1；
- MNIST，详见章节 2.1.2。

3.1.3 实现步骤

图 3.1为数据检测修复功能示意图。功能实现步骤如下：

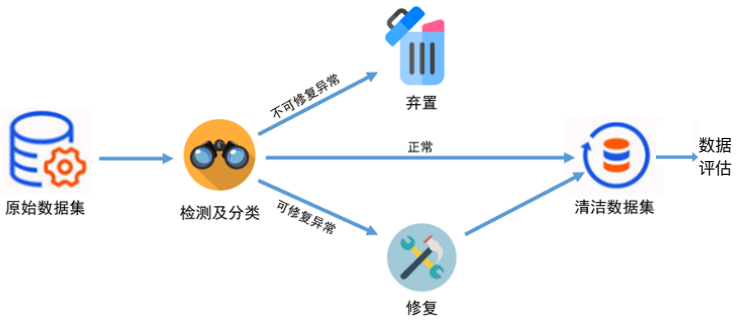

图 3.1: 数据检测修复示意图

- 异常数据检测：原始数据集会存在分布异常数据和格式异常数据，需要进行异常数据检测，对检测出的异常数据进行分类，方便后续进行异常数据处理；

- 异常数据修复：对于格式异常数据和离群数据进行修复，对于不可修复的其他类型异常数据直接弃置。格式异常数据使用均值插补、极大似然估计和多重差补处理；离群数据使用基于置信学习的 Cleanlab 方法 [14] 对错误标签样本进行修复；
- 对数据清洗后的清洁数据集进行样本标签统计和数据完整性评估。其中，数据完整性评估包括相关性、一致性、准确性和均衡性 4 个指标。

3.1.4 输出

- 数据检测修复后的清洁数据集；
- 数据集的标签统计数据，图 3.2为 CIFAR-10 数据集的标签统计结果；

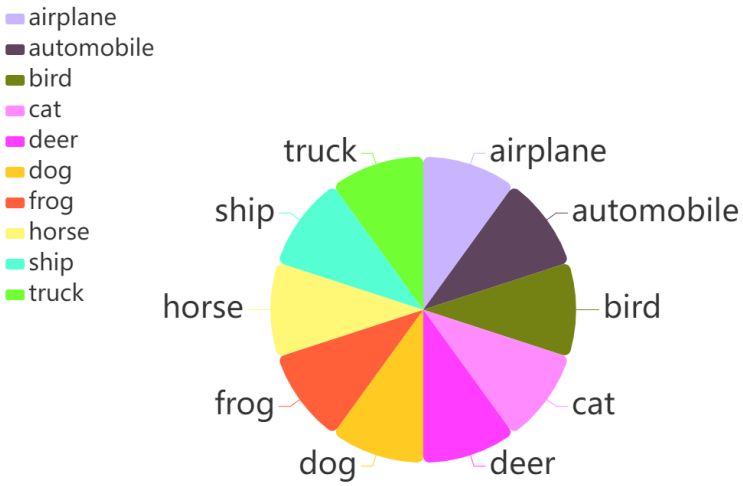

图 3.2: CIFAR-10 数据集标签统计饼状图

- 数据完整性评估结果，图 3.3为 CIFAR-10 数据集数据完整性评估结果，展示了数据相关性、一致性、准确性和均衡性。

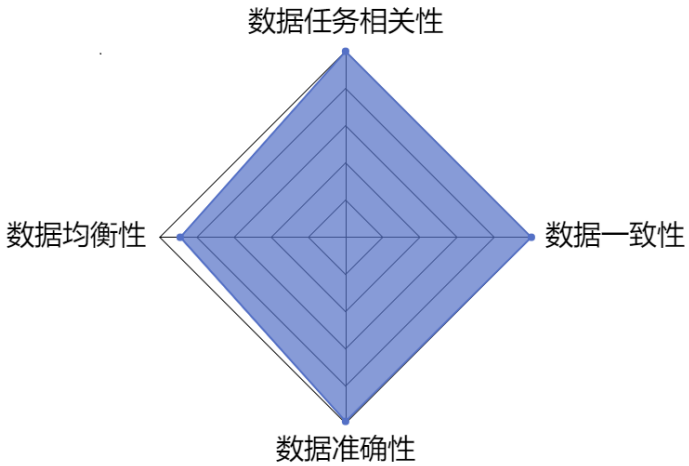

图 3.3: CIFAR-10 数据集完整性评估雷达图

3.2 后门攻击检测

3.2.1 功能描述

后门攻击是一种新兴的针对深度学习模型的攻击方式。攻击者在模型中植入后门，当后门未被激发时，模型表现正常；而埋藏的后门被攻击者激活时，模型输出变为攻击者预先指定的标签以达到恶意的目的。后门攻击检测功能能够检测模型是否存在后门攻击，进而对后门攻击进行防御。

3.2.2 支持数据集

- CIFAR-10，详见章节 2.1.1；
- MNIST，详见章节 2.1.2。

3.2.3 支持模型

- VGG16，详见章节 2.2.2。

3.2.4 支持检测的攻击算法

- BadNets 后门攻击算法，详见章节 2.3.1；
- Trojaning Attack 后门攻击算法，详见章节 2.3.2。

3.2.5 支持的防御算法

- STRIP 算法，详见章节 2.4.1；
- Neural Cleanse 算法，详见章节 2.4.2。

3.2.6 实现步骤

- STRIP 算法，详见章节 2.4.1
  - 将多张不含触发器的干净图像分别叠加到测试图像上；
  - 熵值计算：将叠加后的图像输入到分类模型中，对模型输出的标签概率分布结果进行熵值计算，计算如公式 3.1，N 为样本总数，y 为模型输出的标签概率分布；

$$entropy = \frac{-\sum_{i=1}^{N} y_i \cdot \log_2 y_i}{N} \tag{3.1}$$

  - 设置阈值：根据验证集计算出来的熵值结果，设置合理阈值，对阈值和熵值进行比较，判断测试图像是否含有触发器。

- Neural Cleanse 算法，详见章节 2.4.2
  - 获取标签：将数据集输入到待进行后门防御测试的神经网络模型中，获取数据集所有的预测标签；
  - 构造触发器：根据预测标签，反向构造出输入样本的 pattern 图像结果和 mask 图像结果，叠加出输入样本的触发器；
  - 判断：对触发器进行异常检测，根据触发器的异常值大小判断其是否为该模型的后门触发器。

3.2.7 输出

- STRIP 算法：将输出的熵值分布用直方图形式表示，与阈值比较判断是否含有后门触发器。

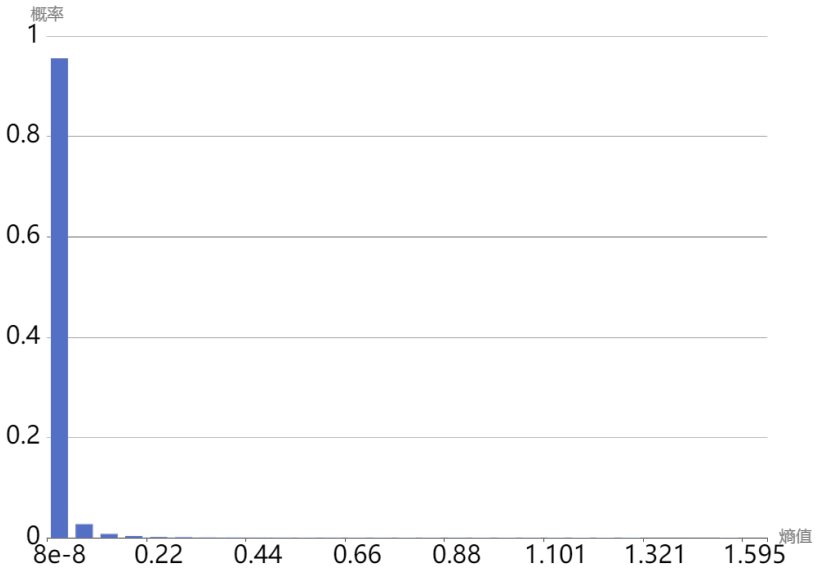

图 3.4: 输入带后门攻击样本的 STRIP 结果示意图

图 3.4所示为输入样本带后门攻击，图 3.5展示输入是干净样本，横坐标表示熵值，纵坐标表示分布概率。

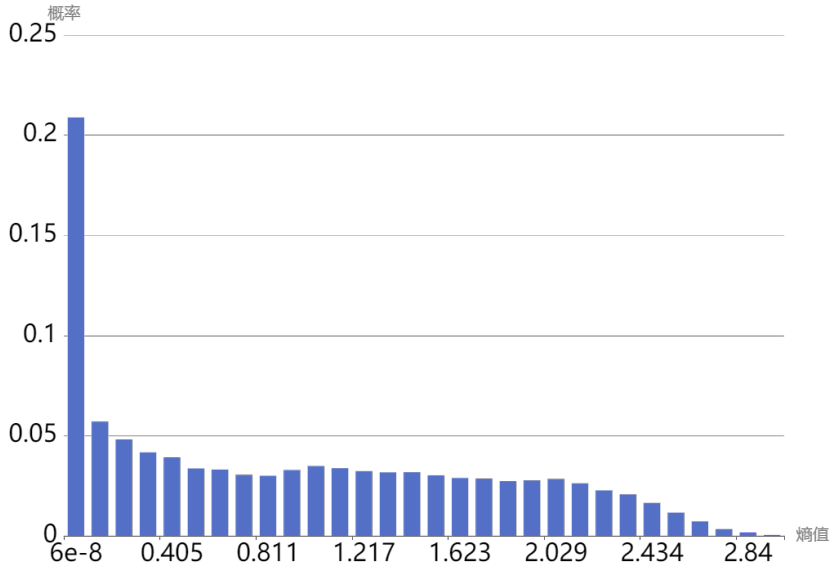

图 3.5: 输入干净样本的 STRIP 结果示意图

- Neural Cleanse 算法：输出构造的触发器，并对触发器进行异常检测判断是否存在后门攻击。

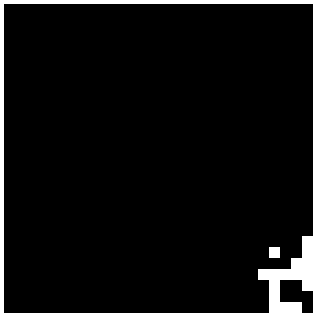

图 3.6: 检出 trigger

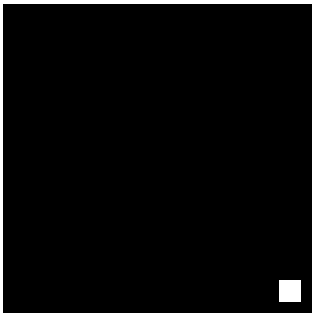

图 3.7: 真实 trigger

图 3.6和图 3.7为构造出的触发器样式，对触发器进行异常检测，根据触发器的异常值大小判断其是否为该模型的后门触发器。

3.3 对抗攻击评估

3.3.1 功能描述

对抗攻击是对输入样本故意添加一些人无法察觉的细微的干扰，导致模型以高置信度给出一个错误输出的一类攻击方法。对抗攻击评估功能模拟 AI 模型受到对抗攻击，进而实现模型的安全性评估。

3.3.2 支持数据集

- CIFAR-10，详见章节 2.1.1；
- MNIST，详见章节 2.1.2。

3.3.3 支持模型

- VGG16，详见章节 2.2.2。

3.3.4 支持的对抗攻击算法

- FGSM，详见章节 2.3.3；
- RFGSM，详见章节 2.3.4；
- FFGSM，详见章节 2.3.5；
- MIFGSM，详见章节 2.3.6；
- PGD，详见章节 2.3.7；
- BIM，详见章节 2.3.8。

3.3.5 实现步骤

- 获取攻击样本：选定数据集和对抗样本生成算法，生成不同攻击算法在不同强度噪声下的对抗攻击样本；
- 攻击 AI 模型：使用生成的对抗攻击样本攻击 AI 模型，统计分类错误数在总样本中的占比，计算攻击成功率；
- 模型安全性评估：根据对抗样本攻击 AI 模型的结果进行安全性评估，使用折线图和直方图的形式分别展示对抗攻击在不同强度噪声下的攻击成功率变化曲线和不同攻击在相同强度噪声下的攻击成功率。

3.3.6 输出

- 选定攻击算法在不同噪声干扰下的攻击效果曲线图；
  如图 3.8所示，为 FGSM，FFGSM，RFGSM，MIFGSM，BIM，PGD 攻击算法在不同噪声干扰下的攻击成功概率变化曲线。
- 相同噪声不同算法攻击效果用直方图展示。

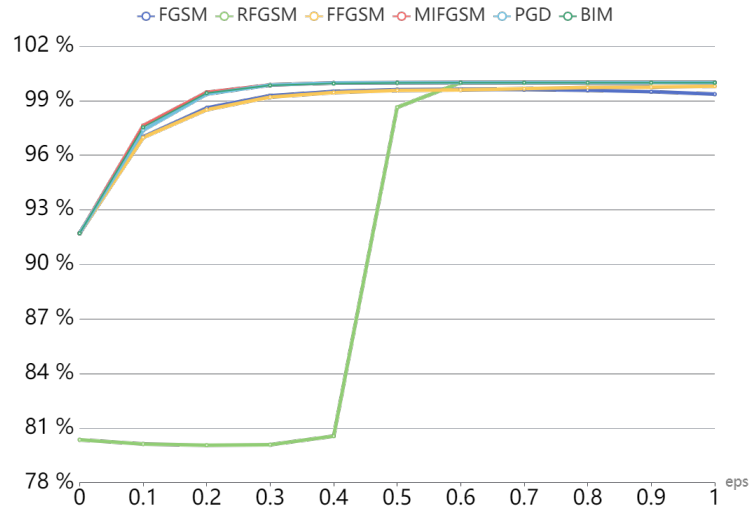

图 3.8: 不同噪声攻击效果示意图

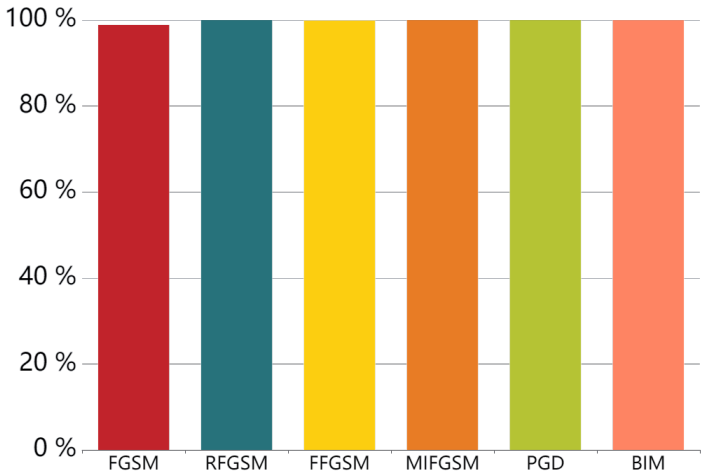

图 3.9: 不同算法攻击效果示意图

图 3.9展示了 FGSM，FFGSM，RFGSM，MIFGSM，BIM，PGD 攻击算法的攻击成功率结果直方图。

第 4 章 软硬件协同安全攻防测试系统

软硬件协同安全攻防测试系统针对一体化架构的安全性能验证给出解决方案，达到安全测试的目的。平台包含**标准化单元测试**、**攻击机理分析**、**攻防博弈推演**、**鲁棒性增强** 4 个核心模块。

其中，标准化单元测试功能模块支持重要神经元覆盖准则、多粒度神经元覆盖准则 2 种准则评估测试充分性，支持 LeNet、VGG 和 ResNet 3 种网络模型；攻击机理分析功能模块从攻击原理出发分析攻击样本特征，支持 LRP、Grad-CAM 和 LIME 3 种解释算法，支持 9 种对抗样本生成算法；攻防博弈推演功能模块采用纳什博弈算法分析攻击、防御策略的有效性，支持 6 种对抗攻击算法的推演和收益衡量；鲁棒性增强功能模块支持对抗训练算法和群智化防御 2 种模型鲁棒性增强措施，支持 6 种攻击算法的防御。

4.1 标准化单元测试

针对传统软件测试方法无法应用于 AI 模型测试的问题，平台集成两种神经元测试准则。通过所有神经元覆盖程度和重要神经元覆盖程度两种方法评估模型测试充分性，为标准化单元测试提供衡量依据。两种神经元覆盖准测分别是：

- 重要神经元覆盖准则
- 多粒度神经元覆盖准则

4.1.1 重要神经元覆盖准则

4.1.1.1 功能描述

系统使用重要神经元覆盖技术遴选出 AI 模型中对决策起关键作用的神经元，通过计算重要神经元覆盖程度，评估模型测试充分性并降低错误决策的风险。

4.1.1.2 支持数据集

- MNIST，详见章节 2.1.2；
- CIFAR-10，详见章节 2.1.1。

4.1.1.3 支持网络模型

- ResNet，详见章节 2.2.1；
- VGG，详见章节 2.2.2；
- LeNet，详见章节 2.2.3。

4.1.1.4 实现步骤

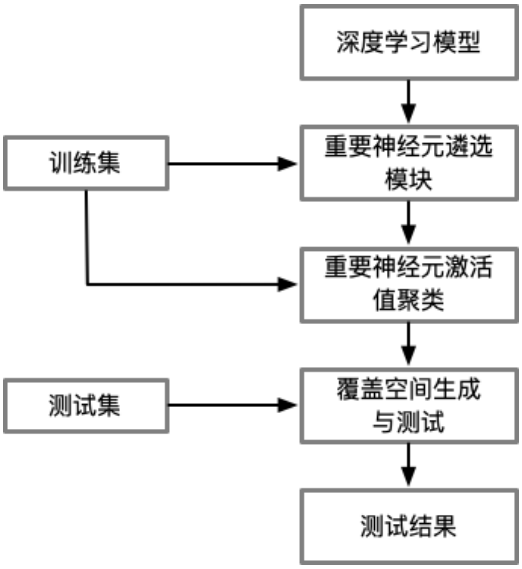

图 4.1: 重要神经元覆盖准则示意图

图 4.1所示为重要神经元覆盖准则示意图，具体实现步骤如下：

- 重要神经元遴选：运用层间相关度传递算法，计算出测试样本传递到目标层时每个神经元的激活值与相关系数，并将每个测试样本该层神经元的相关系数求和，并按照总相关系数对该层的神经元进行重要性排序；
- 激活值聚类：使用测试模型的训练集，得到上一过程中遴选出的重要神经元在训练过程中的激活值分布，并将其聚类为若干个类，聚类数由测试者自己选择，聚类数越大，代表测试粒度越细；
- 覆盖空间生成与测试：对生成的聚类结果进行处理，生成一个大小等于聚类数乘积的覆盖空间，使用测试集对其进行测试；

4.1.1.5 输出

- 随着测试进行，各个神经元重要程度值的变化过程；

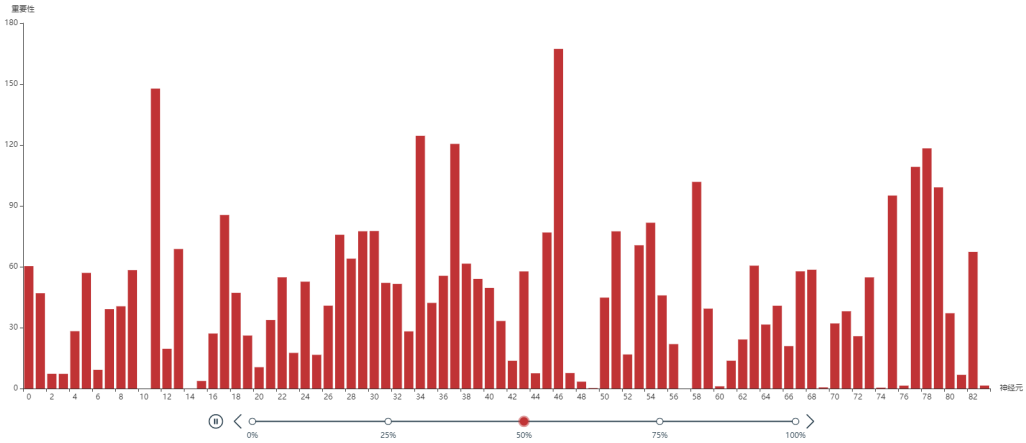

图 4.2: 神经元重要程度变化示意图

如图 4.2所示，随着测试过程进行，每个神经元的重要程度在变化，每条线高度代表其重要程度值，线越高表示该神经元越重要。

- 测试完成，重要神经元覆盖率以百分比形式展示在系统界面。

4.1.2 多粒度神经元覆盖准则

4.1.2.1 功能描述

粒度指统计的粗细程度，多粒度神经元覆盖准则包括单个神经元测试准则（细粒度）和神经网络层测试准则（粗粒度）。系统选择测试集对 AI 模型进行测试，在两个粒度上对神经元覆盖率进行计算，从而衡量模型的测试充分程度。

4.1.2.2 支持数据集

- MNIST，详见章节 2.1.2；
- CIFAR-10，详见章节 2.1.1。

4.1.2.3 支持网络模型

- ResNet，详见章节 2.2.1；
- VGG，详见章节 2.2.2；
- LeNet，详见章节 2.2.3。

4.1.2.4 实现步骤

图 4.3所示为多粒度神经元覆盖准则示意图，具体步骤如下：

- 输入预处理：输入测试数据和 AI 模型，对 AI 模型添加检测代码，检测代码负责在模型前向传播的过程中记录所有隐藏层神经元的输出结果；

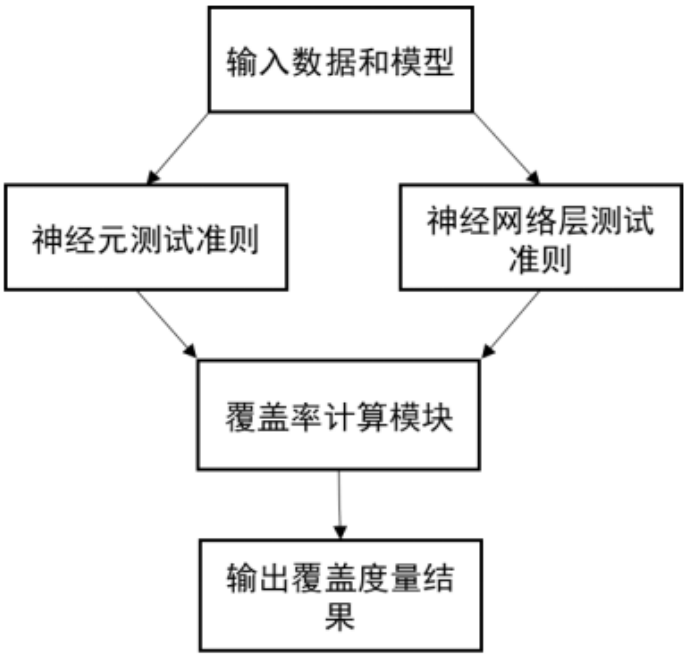

图 4.3: 多粒度神经元覆盖准则示意图

- 使用单个神经元测试准则和神经网络层测试准则作为指导；
  - 单个神经元测试准则：把神经元的输出范围分成长度相同的多个区间，每个区间用来表征神经元的某个特征行为，通过判定神经元的输出值是否在某个区间之内，来判断该逻辑行为是否被测试数据覆盖，对 AI 模型的测试充分性进行细粒度的衡量；
  - 神经网络层测试准则：使用最活跃神经元及其组合来衡量神经网络层在模型中的行为，从每层神经元数值输出之间相互关系的角度，对 AI 模型的测试充分性进行粗粒度的衡量；
- 计算模型的神经元覆盖率，对模型的测试充分程度进行衡量；

$$NeuronCoverage = \frac{|ActivatedNeurons|}{TotalNeurons} \tag{4.1}$$

如公式 4.1所示，Neuron Coverage 表示神经元覆盖率，Activated Neurons 代表测试集激活/覆盖的神经元总数，Total Neurons 代表模型种所有神经元数量。

4.1.2.5 输出

- 模型测试中神经元覆盖率变化效果图进行平台动态展示；
- 平台展示模型最终的神经元覆盖率参数。

4.2 攻击机理分析

4.2.1 功能描述

对抗样本等各类攻击误导模型的内在原因不明确，通过可视化展示对抗样本攻击效果的攻击机理分析功能，可以研究对抗噪声在模型传递过程中对网络决策造成的影响，分析噪声所导致模型关注区域的偏移情况，为安全性验证与防御策略构建等工作提供理论指导。

4.2.2 支持数据集

- ImageNet，详见章节 2.1.3；
- CIFAR-10，详见章节 2.1.1。

4.2.3 支持深度学习模型

- ResNet，详见章节 2.2.1；
- VGG，详见章节 2.2.2。

4.2.4 使用解释算法

- LRP，详见章节 2.5.1；
- Grad-CAM，详见章节 2.5.2；
- LIME，详见章节 2.5.3。

4.2.5 支持的对抗样本生成算法

- FGSM，详见章节 2.3.3；
- RFGSM，详见章节 2.3.4；
- FFGSM，详见章节 2.3.5；
- MIFGSM，详见章节 2.3.6；
- DI2FGSM，详见章节 2.3.9；
- PGD，详见章节 2.3.7；
- PGDL2，详见章节 2.3.10；
- EOTPGD，详见章节 2.3.11；
- BIM，详见章节 2.3.8。

4.2.6 肯德尔等级相关系数

肯德尔等级相关系数 [16] 是统计学中重要的相关性计算方法，主要用于测量两个随机变量相关性的统计值，其值分布在 −1 到 1 之间，当取值为 1 时，表示两个随机变量拥有一致的等级相关性；当取值为 −1 时，表示两个随机变量拥有完全相反的等级相关性；而当取值为 0 时，表示两个随机变量是相互独立的。

4.2.7 实现步骤

- 获取对抗样本：选定数据集和对抗样本生成算法，生成不同对抗样本生成数据；
- 特征可视化：使用 LRP、Grad-CAM 和 LIME 三种解释算法对对抗样本数据进行解释，获得可视化热力图数据结果；
- 攻击机理分析：使用三种解释算法解释后的特征结果，可直观观察到模型分类时所关注的图像区域，并将解释算法的肯德尔等级相关系数可视化输出，计算正常解释图和对抗解释图之间的相似度。从视觉和统计学角度分析对抗噪声对样本特征的作用机理。

4.2.8 输出

- 正常样本和对抗样本三种特征解释算法解释后的特征热力图：

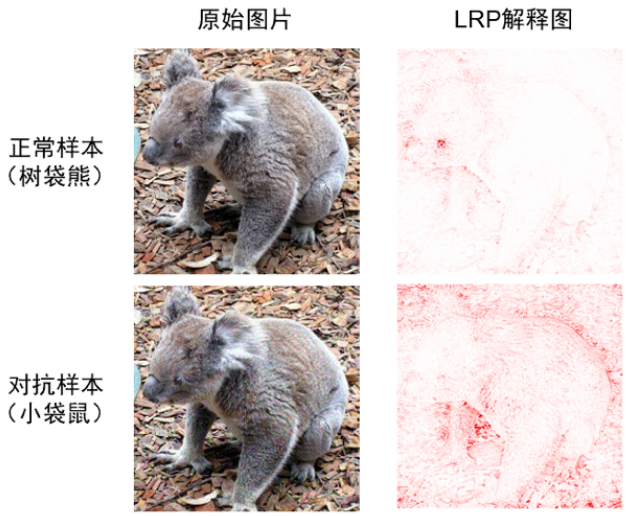

图 4.4: LRP 解释效果图

- 图 4.4所示为 LRP 解释效果图。对于正常样本来说，模型预测其为树袋熊的主要依据是其头部特征，然而由于对抗噪声的存在，模型在分类对抗样本时无法提取头部特征，而是提取到了背景中的其他特征；

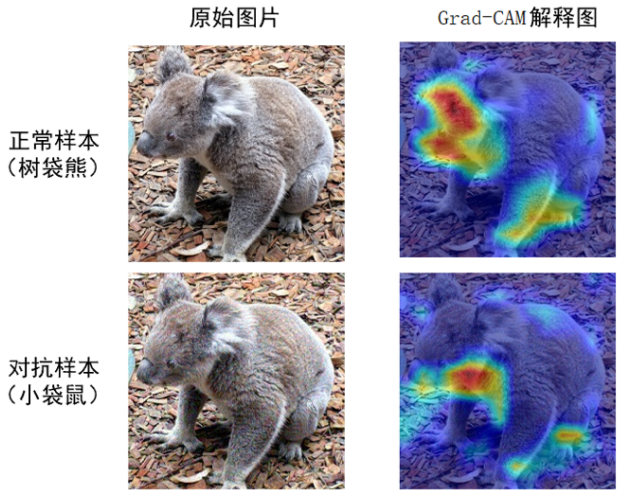

图 4.5: Grad-CAM 解释效果图

- 图 4.5所示为 Grad-CAM 解释效果图，热力图中红色代表权重高的区域，也是模型分类时主要参考的特征，而蓝色则相反。可以明显观察到，由于对抗噪声的出现，模型对样本特征的关注区域发生了偏移，因而无法识别到正确特征，进而做出错误分类;

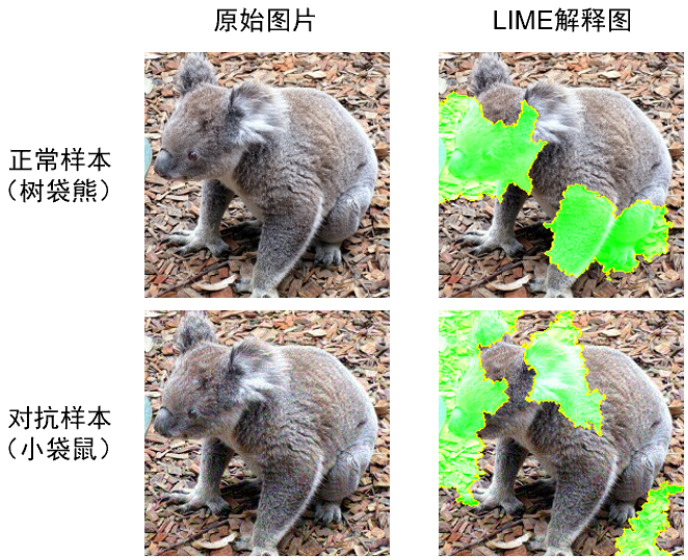

图 4.6: LIME 解释效果图

- 图 4.6所示为 LIME 解释效果图，可以明显发现，噪声的出现导致模型关注区域从目标物体中迁移到了背景中，说明模型在预测时关注的特征区域发生了偏移。

- 用肯德尔等级相关系数量化正常解释图和对抗解释图的数据分布。
  图 4.7箱型图是通过肯德尔等级指数计算了不同攻击算法生成的对抗样本与正常样本间解释图的相关性，计算值分布在 $0-1$ 之间，值越大则表示二者解释图越相似，进而反应了对抗样本欺骗模型效果越好。

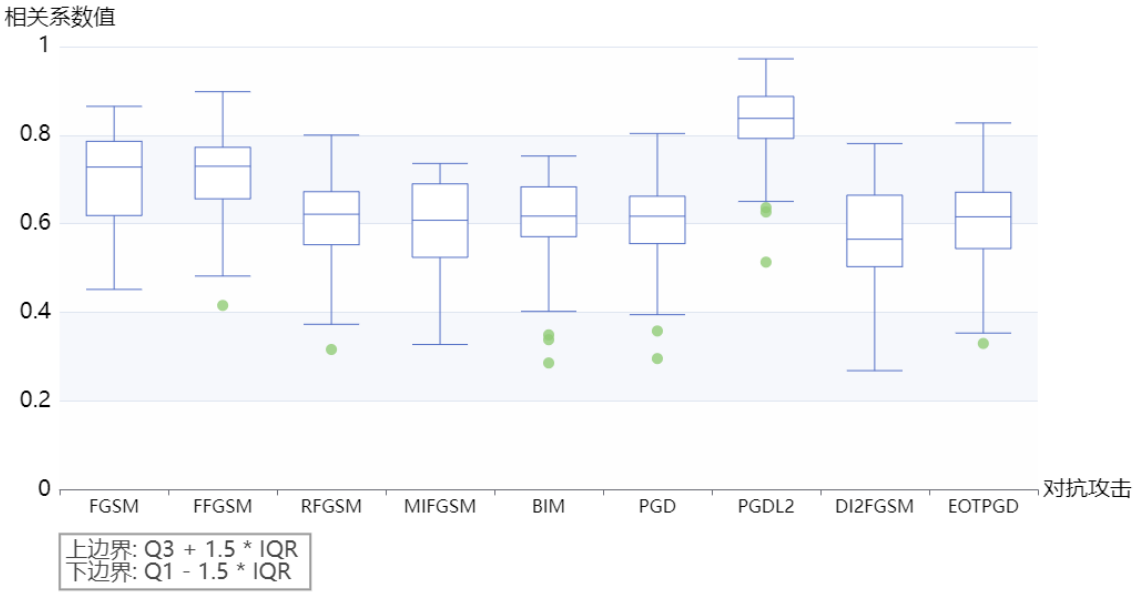

图 4.7: LRP 肯德尔等级相关系数

4.3 攻防博弈推演

4.3.1 功能描述

攻防博弈推演是以对抗样本生成算法和防御算法作为博弈的双方。不同对抗样本生成算法为攻击方的可选策略集，不同攻击算法生成的对抗样本进行训练后的模型作为防守方的可选策略集，最终经过博弈获得攻守收益矩阵，从而确定特定场景下的最佳攻/防策略。

4.3.2 支持数据集

- CIFAR-10，详见章节 2.1.1。

4.3.3 支持深度学习模型

- VGG，详见章节 2.2.2；
- ResNet，详见章节 2.2.1。

4.3.4 支持的攻防推演算法

- 纳什博弈：在一个博弈过程中，无论对方的策略选择如何，当事人一方都会选择某个确定的策略，则该策略被称作支配性策略。如果两个博弈的当事人的策略组合分别构成各自的支配性策略，那么这个组合就被定义为纳什均衡。

4.3.5 支持的对抗攻击算法

- FGSM，详见章节 2.3.3；
- RFGSM，详见章节 2.3.4；
- FFGSM，详见章节 2.3.5；
- MIFGSM，详见章节 2.3.6；
- DI2FGSM，详见章节 2.3.9；
- PGD，详见章节 2.3.7；
- PGDL2，详见章节 2.3.10；
- EOTPGD，详见章节 2.3.11；
- BIM，详见章节 2.3.8。

4.3.6 实现步骤

- 构建攻击测试模块：使用 6 种对抗样本攻击算法（FGSM、FFGSM、RFGSM、MIFGSM、BIM、PGD）生成对抗样本；
- 构建防御加固模块：使用上述 6 种对抗样本攻击算法 3 组扰动系数生成的对抗样本分别混合干净样本组成 18 个训练集，对预训练模型进行鲁棒性训练；
- 计算准确率和鲁棒性：使用标准测试集输入到不同的模型中获得模型准确率收益矩阵，使用不同的对抗样本集攻击不同的模型获得鲁棒性收益矩阵；
- 绘制攻防博弈收益曲线：模拟攻击者使用对抗样本攻击模型，防御者采用不同鲁棒训练后的模型进行防御，攻防博弈后所得到最终收益兼顾模型准确率和模型鲁棒性两个指标，计算公式：

$$\text{收益} = p \cdot \text{准确率} + (1-p) \cdot \text{鲁棒性} \tag{4.2}$$

如公式 4.2，p 为准确率系数，准确率从准确率收益矩阵中获取，鲁棒性从鲁棒性收益矩阵中获取；

4.3.7 输出

- 准确率收益直方图，如图 4.8 展示标准测试集在不同鲁棒性训练模型上的准确率；

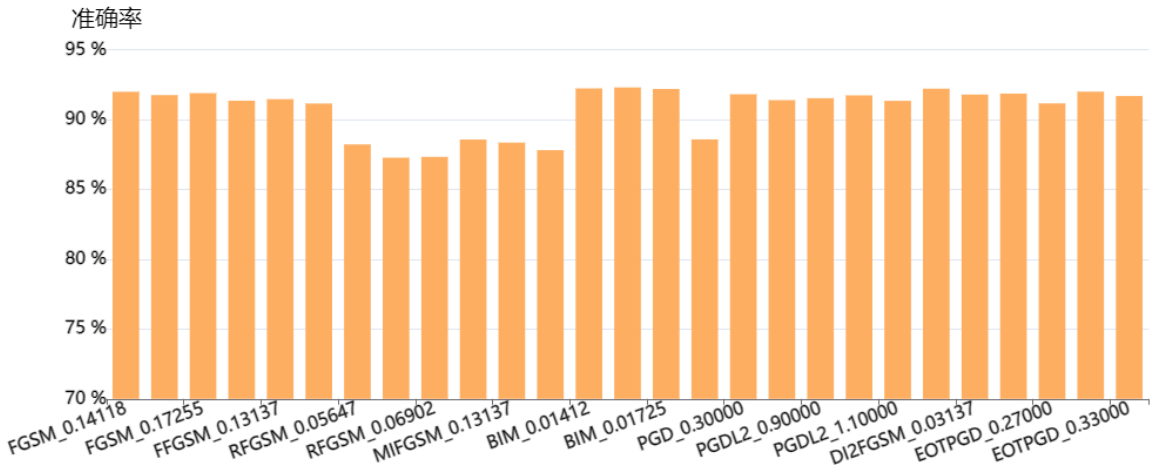

图 4.8: 准确率收益直方图

- 鲁棒性收益矩阵，如图 4.9展示采用不同鲁棒性训练算法训练的模型对不同对抗样本攻击算法的鲁棒性；

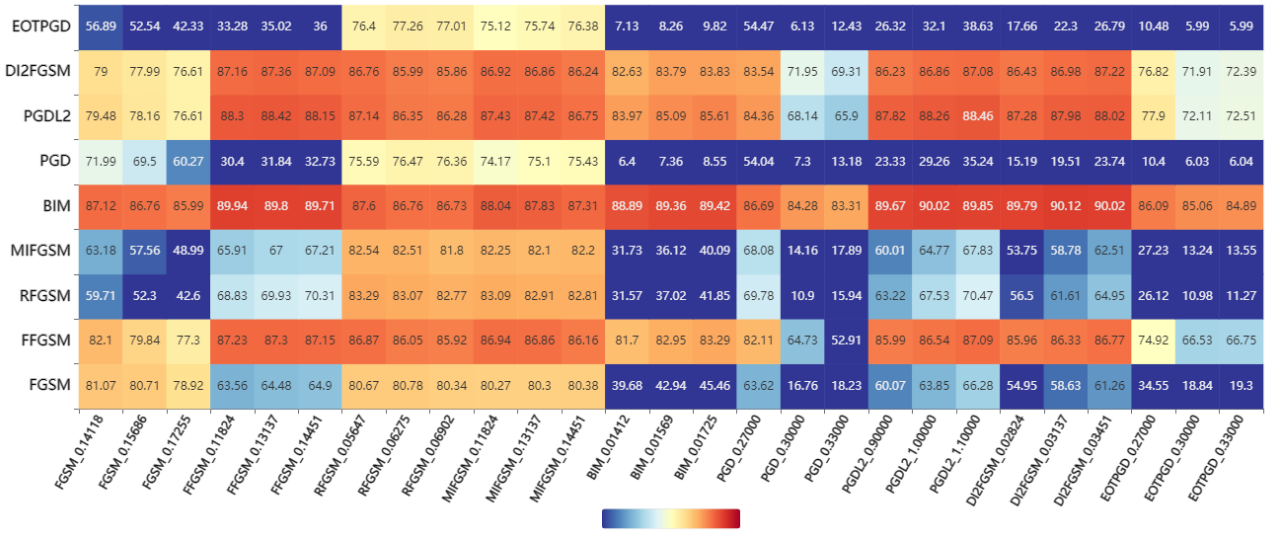

图 4.9: 鲁棒性收益矩阵

- 博弈策略收益曲线，如图 4.10所示，曲线表示衡量鲁棒性和准确率权值后的收益变化趋势，曲线图中每条线代表一种对抗攻击算法，横轴 p 为准确率系数。

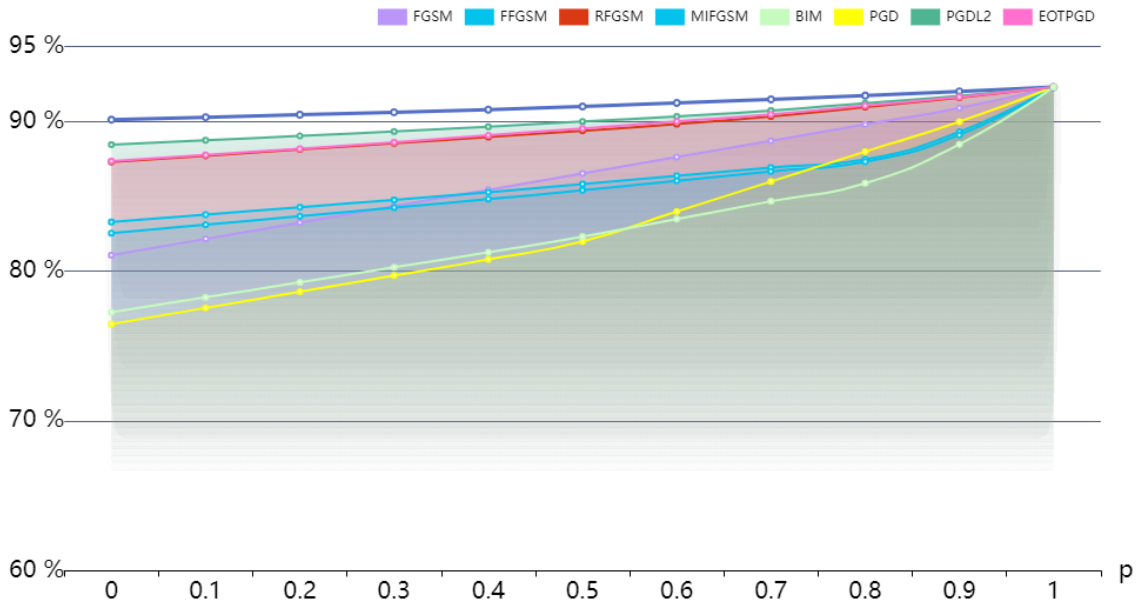

图 4.10: 博弈策略收益曲线

4.4 鲁棒性增强

4.4.1 功能描述

AI 模型通过鲁棒性训练获得一个健壮的深度神经网络，避免外来的恶意攻击导致模型失效。鲁棒性增强功能采用对抗训练或群智化防御对模型进行鲁棒性加固，并对原始模型和鲁棒性增强后 AI 模型进行评估，验证鲁棒性增强效果。

4.4.2 支持数据集

- CIFAR-10，详见章节 2.1.1。

4.4.3 支持深度学习模型

- ResNet，详见章节 2.2.1；
- VGG，详见章节 2.2.2。

4.4.4 可防御的攻击算法

- FGSM，详见章节 2.3.3；
- RFGSM，详见章节 2.3.4；
- FFGSM，详见章节 2.3.5；
- MIFGSM，详见章节 2.3.6；
- DI2FGSM，详见章节 2.3.9；

- PGD，详见章节 2.3.7；
- PGDL2，详见章节 2.3.10；
- EOTPGD，详见章节 2.3.11；
- BIM，详见章节 2.3.8。

4.4.5 支持的鲁棒性增强措施

- 对抗训练

  如图 4.11所示，在模型的训练过程中构建对抗样本，并将对抗样本和原始样本混合一起训练模型，从而使训练后的模型对对抗样本具有鲁棒性。

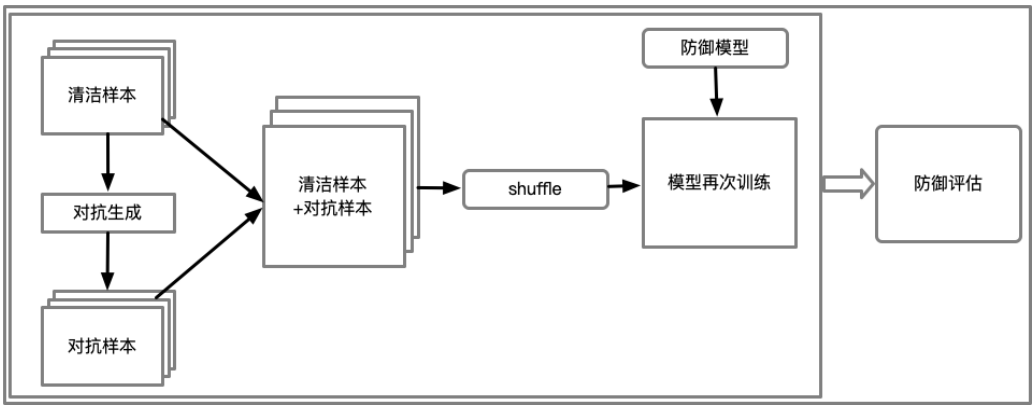

图 4.11: 对抗训练示意图

- 群智化防御

  针对特定的机器学习模型，如图 4.12所示，首先使用多种对抗样本生成算法生成各自独立的对抗样本数据集，与原始样本混合后，分别进行鲁棒性训练，生成对应每种攻击算法的鲁棒性模型，使用模型进行预测时，分别将数据输入每个鲁棒性模型，各模型的输出结果求和后经过 Softmax 层得到群智化防御模型的输出。

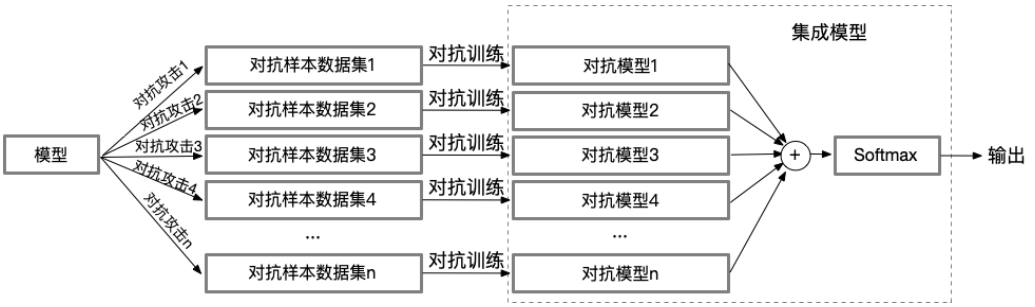

图 4.12: 群智化防御示意图

4.4.6 攻击检测方法

- PACA 自动化测试检测：基于原始图片计算图片的预测置信度信息，如图所示 4.13，分别将原始图片与预测置信度特征分别输入结构相同的两个检测网络中，两个检测网络的

输出叠加后再经过 Softmax 激活函数得到对最终的预测标签信息，从而检测输入是否为对抗样本。

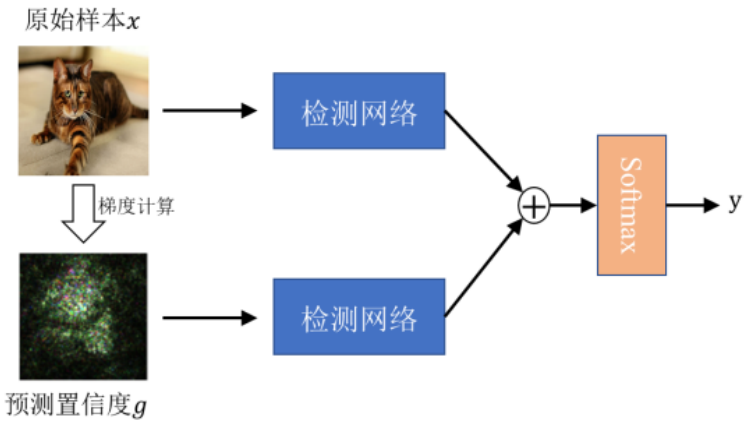

图 4.13: PACA 自动化测试检测示意图

4.4.7 实现步骤

- 获取对抗样本：选定数据集和对抗攻击方法，获取对应对抗攻击方法的攻击样本数据。
- 模型鲁棒性增强：对清洁数据集和对抗样本根据鲁棒性训练方法各自进行混合、置信度预测等处理，通过对抗训练、群智化防御两种鲁棒性增强措施提高 AI 模型鲁棒性。
- 防御效果验证：AI 模型鲁棒性训练完成后，使用直方图形式对鲁棒性训练后的防御成功率和 PACA 检测出的对抗攻击成功率结果进行展示，通过数据直观评估鲁棒性增强效果。

4.4.8 输出

- 鲁棒性加固后的模型；
- 攻击成功率直方图，如图 4.14展示对抗攻击算法对不同鲁棒性加固措施后的 AI 模型攻击成功率；

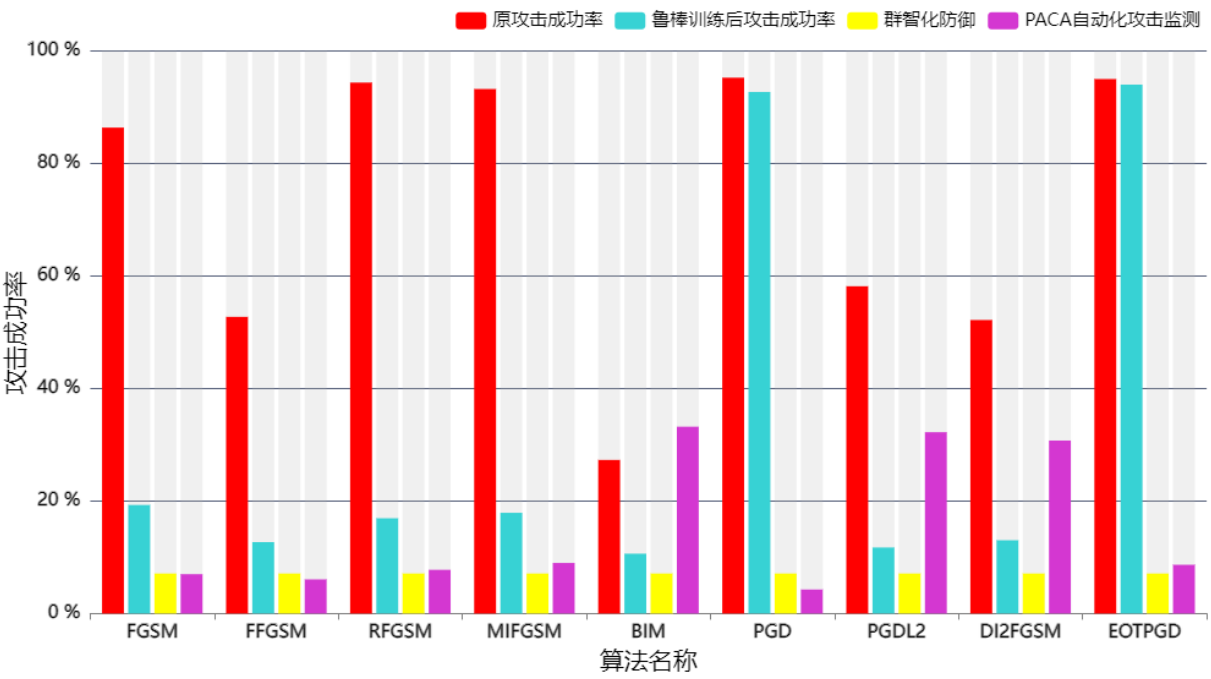

图 4.14: 攻击成功率直方图

- 模型分类精度直方图，如图 4.15展示原始模型和受不同攻击算法攻击后模型分类精度的变化。

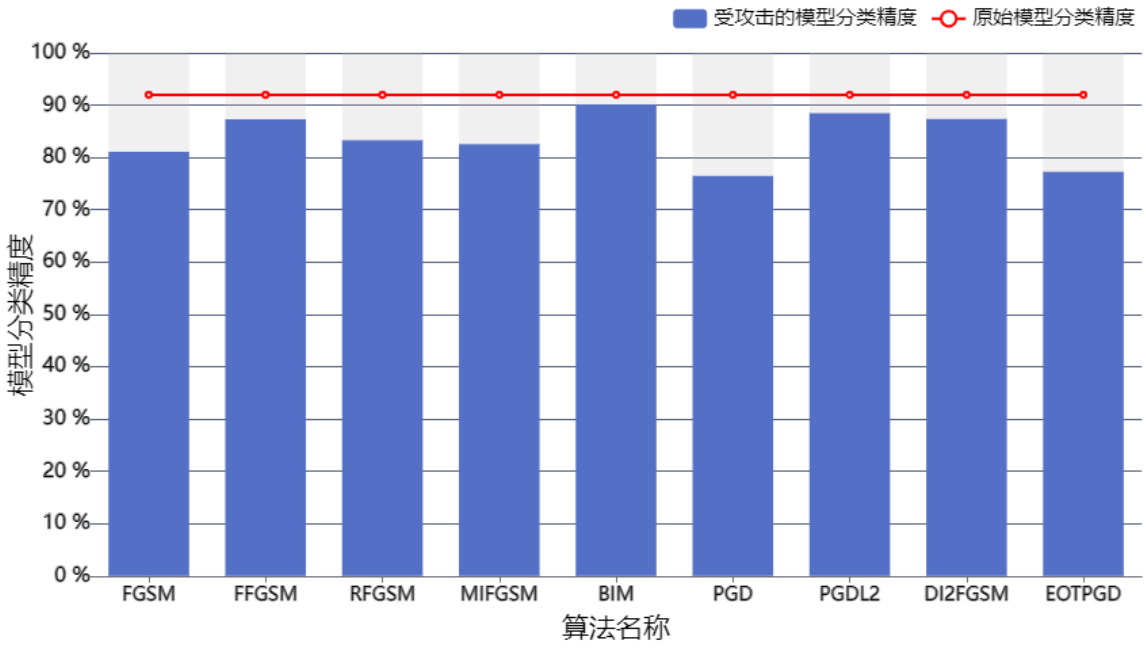

图 4.15: 模型分类精度直方图

参考文献

[1] Sebastian Bach et al. “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation”. In: *PloS one* 10.7 (2015), e0130140.

[2] Francesco Croce and Matthias Hein. “Minimally distorted adversarial examples with a fast adaptive boundary attack”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 2196–2205.

[3] Yinpeng Dong et al. “Boosting adversarial attacks with momentum”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 9185–9193.

[4] Yansong Gao et al. “Strip: A defence against trojan attacks on deep neural networks”. In: *Proceedings of the 35th Annual Computer Security Applications Conference*. 2019, pp. 113–125.

[5] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples”. In: *arXiv preprint arXiv:1412.6572* (2014).

[6] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. “Badnets: Identifying vulnerabilities in the machine learning model supply chain”. In: *arXiv preprint arXiv:1708.06733* (2017).

[7] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[8] Alex Krizhevsky, Geoffrey Hinton, et al. “Learning multiple layers of features from tiny images”. In: (2009).

[9] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. “Adversarial examples in the physical world”. In: *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[10] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[11] Xuanqing Liu et al. “Adv-bnn: Improved adversarial defense through robust bayesian neural network”. In: *arXiv preprint arXiv:1810.01279* (2018).

[12] Yingqi Liu et al. “Trojaning attack on neural networks”. In: (2017).

[13] Aleksander Madry et al. “Towards deep learning models resistant to adversarial attacks”. In: *arXiv preprint arXiv:1706.06083* (2017).

[14] Curtis Northcutt, Lu Jiang, and Isaac Chuang. “Confident learning: Estimating uncertainty in dataset labels”. In: *Journal of Artificial Intelligence Research* 70 (2021), pp. 1373–1411.

[15] Adam Paszke et al. “Automatic differentiation in pytorch”. In: (2017).

[16] Review et al. “Rank Correlation Methods. by M. G. Kendall”. In: *Journal of the Royal Statistical Society. Series D (The Statistician)* 20.3 (1971).

[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “” Why should i trust you?” Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[18] Olga Russakovsky et al. “Imagenet large scale visual recognition challenge”. In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[19] Ramprasaath R Selvaraju et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[20] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[21] Florian Tramèr et al. “Ensemble adversarial training: Attacks and defenses”. In: *arXiv preprint arXiv:1705.07204* (2017).

[22] Bolun Wang et al. “Neural cleanse: Identifying and mitigating backdoor attacks in neural networks”. In: *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2019, pp. 707–723.

[23] Cihang Xie et al. “Improving transferability of adversarial examples with input diversity”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2730–2739.

[24] Baolin Zheng et al. “Black-box adversarial attacks on commercial speech platforms with minimal information”. In: *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 2021, pp. 86–107.