



面向自动驾驶系统图像分类模型的 对抗攻击及防御性能研究

唐 军, 黄文静, 李 爽, 吴自力
(中车株洲电力机车研究所有限公司, 湖南 株洲 412001)

摘 要: 图像分类模型被广泛应用于城市轨道交通、汽车、智轨等众多交通运输系统的自动驾驶系统, 以实现自主感知、自主定位等功能, 然而对抗样本会使模型产生错误的输出与判断, 对图像分类模型在自动驾驶系统的应用安全性产生较大影响。文章通过针对典型的图像分类模型 ResNet, 采用基于敏感性分析的噪声叠加攻击策略开展白盒对抗攻击, 并对攻击效果和防御性能进行研究。该研究选取 FGSM、BIM、PGD 等算法生成对抗样本, 通过调整扰动系数实现微小扰动下的攻击成功率测试; 采用 LRP、Grad-CAM 和 LIME 三种对抗解释算法对不同样本不同区域的敏感程度进行攻击原理分析, 并在基于对抗攻击效果和原理分析的基础上, 采用群智化防御、对抗训练等优化算法, 验证对抗训练后模型的分类性能, 同时通过博弈算法对弈获得攻守收益矩阵, 从而确定最佳防御策略。文章通过对抗攻击和攻防策略的研究, 提出高级辅助驾驶系统图像分类模型在运用过程中保障安全性的解决方案。

关键词: 交通运输; 自动驾驶; 网络安全; 功能安全; 图像分类; 对抗样本; 防御性能

中图分类号: U268.4; TP183 **文献标志码:** A **文章编号:** 1000-128X(2025)01-0025-10

0 引言

随着自动驾驶技术在城市轨道交通、汽车、智轨等多种交通运输系统中广泛应用, 极大提高了驾驶的易用性、便捷性和舒适性。人工智能技术与自动驾驶技术的深度耦合与交叉, 已成为自动驾驶技术未来发展的重要研究方向, 同时也引起学术界与工业界的广泛关注。尤其是, 基于图像处理的人工智能模型和算法的引入, 使自动驾驶系统的灵敏度、识别性能得到极大的提升。然而, 由于人工智能模型和算法的内生脆弱性、黑盒性、不确定性、不可解释性, 因此如何保障自动驾驶系统的应用安全性成为制约两大技术融合发展的因素。人工智能模型的训练需要依赖大量的样本数据, 数据错误或者外部干扰等因素均可能导致自动驾驶系统误判, 引起自动驾驶误动作等安全性问

题。针对上述问题, 本文基于对自动驾驶系统中大量应用的图像分类模型的分析研究, 通过对抗样本生成方式对模型进行攻击, 分析模型对不同对抗样本的抗攻击性能, 并采用对抗训练、群智化防御等方式验证图像分类模型的防御性能, 从而提出自动驾驶系统图像分类模型在使用中的相关安全性措施。

1 图像分类模型在自动驾驶系统的应用

1.1 自动驾驶系统

自动驾驶技术的典型应用场景主要集中在轨道交通和道路交通。与汽车运行环境相比, 轨道交通运行环境相对封闭且线路固定, 但城市轨道交通、智轨等半开放的运行环境存在更多的环境变化因素, 因此在传统列车运行控制系统基础上加入自主感知系统, 以增强自动驾驶系统对环境探测的敏感性与驾驶安全性。

基金项目: 国家重点研发计划项目 (2022YFB4300604)

通信作者: 李 爽, 男, 主要从事轨道交通领域功能安全技术研究; E-mail: ls1031804608@163.com

引用格式: 唐军, 黄文静, 李爽, 等. 面向自动驾驶系统图像分类模型的对抗攻击及防御性能研究[J]. 机车电传动, 2025(1): 25-34.DOI:10.13890/j.issn.1000-128X.2025.01.100.

Citation: TANG Jun, HUANG Wenjing, LI Shuang, et al. Research on adversarial attacks and defense performance of image classification models for automated driving systems[J]. Electric drive for locomotives, 2025(1): 25-34.DOI:10.13890/j.issn.1000-128X.2025.01.100.

汽车运行环境则更为多变、开放，存在更多的不确定性。根据轨道交通和道路交通两大主要自动驾驶运行场景的特点，自动驾驶系统主要分为环境感知、路径规划和决策执行三大部分，如图1所示。环境感知部分主要实现环境探测、数据处理和定位功能。环境探测一般使用激光雷达、摄像头、毫米波雷达等传感器，以获取相关图像或点云数据，并通过图像转换、解析、多种传感器数据融合分析等方式完成图像数据处理，而在图像处理过程中则会大量使用图像分类、特征提

取、点云处理等人工智能算法。对于人工智能算法处理的数据，会通过惯性定位、差分定位、通信、GPS卫星定位等方式实现定位。定位后的数据和环境数据会进一步传输给路径规划部分，通过高精地图比对、运行图分析、站点数据比对等措施完成路径规划。综合环境感知和路径规划信息，对牵引/制动控制、速度控制等自动驾驶相关控制策略进行决策，最终将决策指令发送给最终的执行机构。

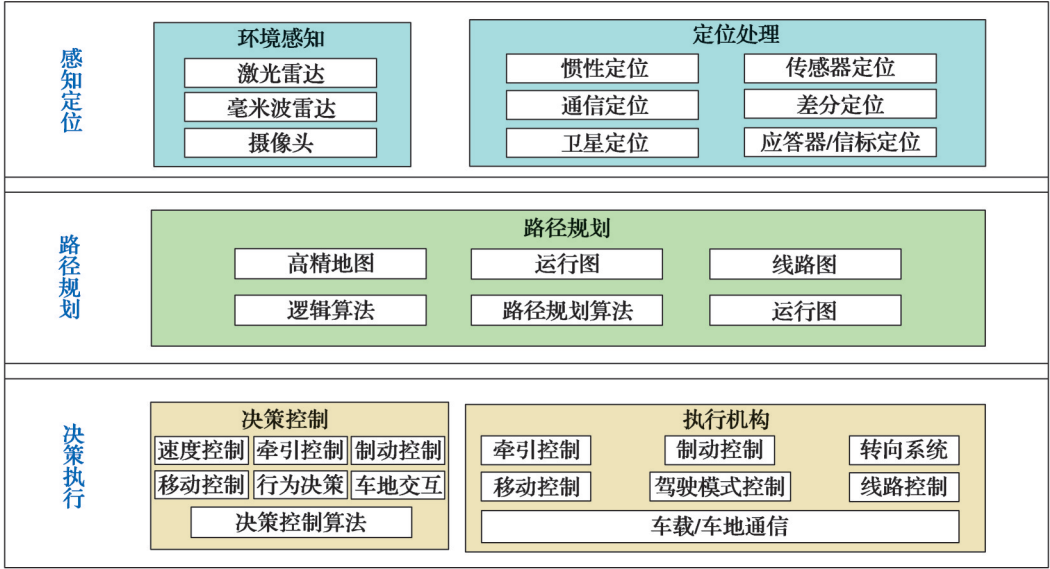


图1 自动驾驶系统架构图
Fig. 1 Architecture diagram of automated driving system

1.2 图像分类模型准确性

在自动驾驶系统中，环境感知、定位、规划、决策等关键组成部分都大量使用了图像分类模型，特别是在环境感知模块应用最为广泛。针对数据样本的识别，以图像分类模型中二分类为例，识别的结果主要分为两类，一类称为正例，另一类称为反例。图像分类模型识别根据样本真实结果和预测结果的组合可划分为4种情况，如表1所示。本文采用准确率（Accuracy）来衡量图像分类模型性能，即模型预测正确的样本数占总样本的比例，即

$$A = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FN} + N_{FP}} \tag{1}$$

式中： N_{TP} 为真阳性（True Positive，TP）指标； N_{TN} 为真阴性（True Negative，TN）指标； N_{FP} 为假阳性（False Positive，FP）指标； N_{FN} 为假阴性（False Negative，FN）指标。

表1 混淆矩阵
Table 1 Confusion matrix

真实结果	预测结果	
	正例	反例
正例	真阳性（TP）	假阴性（FN）
反例	假阳性（FP）	真阴性（TN）

2 对抗攻击评估

2.1 对抗攻击的概念

2013年，Christian Szegedy 等人^[1]首次证明对抗样本的存在，并给出了其定义：在原始干净样本上，攻击方通过添加人为故意生成的细微扰动所形成的样本，将其输入到相应的模型后会导致模型以较高置信度输出错误结果，如大熊猫在扰动下被高置信度识别为长臂猿^[2]。图像分类模型的对抗样本通常是在原始干净图像的像素点增加噪声，生成的对抗样本会对原始图像的像素点产生一点视觉瑕疵，导致图像分类模型进行错误的分类。设定经过训练的图像分类模型为 f ，原始样本为 X ，生成的对抗样本为 X' ，其数学定义可表示为：

$$\min_{x'} \|X' - X\|_p \tag{2}$$

$$\text{s.t.} \begin{cases} f(X) = Y \\ f(X') = Y' \\ Y \neq Y' \\ x'_i \in [0, 1], i = 1, 2, \dots, n \end{cases} \tag{3}$$

式中： Y 、 Y' 分别为 X 和 X' 的输出标签； $\| \cdot \|_p$ 为 p 的范数，即2个样本之间的距离； x'_i 为图像内单个像素点的值。

设 η 为扰动, 则 $\eta = X' - X$ 表示在 X 上添加的扰动。目前对抗样本的生成算法均是对式(2)的近似求解。文献[1]提出了式(2)的求解方法“内存受限的拟牛顿法”(Limited-memory Broyden-Fletcher-Goldfarb-Shanno, L-BFGS), 但 L-BFGS 优化过程复杂, 攻击效率低。

对抗攻击的评估指标主要是攻击成功率。攻击成功率指评估算法模型面对不同对抗样本输入时的正常运行能力。对于有目标的攻击, 攻击成功率为预测目标类别样本数量与总样本数量之比; 对于无目标的攻击, 攻击成功率为预测错误样本数量与总样本数量之比。

2.2 对抗攻击方法

针对图像分类模型, 目前业内根据场景可分为数字域攻击和物理域攻击。其中, 数字域攻击是通过获取图像的数字矩阵, 并修改图像的像素点, 以此生成对抗样本; 物理域攻击是通过干扰实际物理设备(摄像头、雷达、传感器等)生成对抗样本。数字域攻击又分为白盒攻击和黑盒攻击。白盒攻击是指攻击者掌握了目标人工智能模型的结构和参数信息, 可以直接对模型内部的结构、权重、超参数、数据等元素进行解析与修改来对模型进行攻击^[2]。白盒攻击的常见方法包括基于噪声叠加和基于语义操纵的攻击策略, 其中基于噪声叠加的攻击可分为基于优化算法的攻击、基于敏感性分析的攻击和基于生成模型的攻击; 基于语义操纵的攻击可分为基于几何变换的攻击、基于色彩变换的攻击和基于参数变换的攻击^[3]。黑盒攻击是指攻击者对目标人工智能模型缺乏内部结构和参数信息, 只能通过模型的输入和输出生成对抗样本的攻击方式^[4]。对比白盒攻击方式, 黑盒的攻击者无法获得原始样本的参数, 只能通过替代模型或搜索的方式寻找会导致目标模型执行任务错误的参数, 以此实现对目标模型的攻击, 攻击成功率低于白盒攻击。黑盒攻击方式根据攻击手段分为基于梯度估计的攻击、基于迁移性的攻击和基于局部搜索的攻击^[5]。

本文主要使用数字域攻击技术白盒攻击方式中基于敏感性分析的快速梯度符号攻击(Fast Gradient Sign Method, FGSM)、基本迭代方法(Basic Iterative Method, BIM)和投影梯度下降法(Projected Gradient Descent, PGD)3种攻击方法及其衍生的RFGSM、FFGSM、MIFGSM、PGDL2、EOTPGD等5种方法对自动驾驶的图像分类模型进行对抗攻击评估。

1) FGSM

Goodfellow等人发现了深度神经网络线性假设现象, 提出了一种沿损失函数梯度方向求解扰动的攻击, 并把这种攻击方法称为快速梯度符号(Fast Gradient Sign Method, FGSM)攻击^[2]。

对于一个人工智能模型, 令 θ 为模型参数, 图像 X

为模型输入, Y_{true} 为 X 的真实类别, $L(\theta, X, Y_{\text{true}})$ 为模型的损失函数。对于扰动 $\eta = \alpha \cdot \text{sig}(\nabla_X L(\theta, X, Y_{\text{true}}))$ (α 为扰动系数, 用于控制扰动的大小), 通过FGSM方法生成的新对抗样本为

$$X^{\text{adv}} = X + \eta \quad (4)$$

公式中 $\nabla_X L(\theta, X, Y_{\text{true}})$ 为损失函数 $L(\theta, X, Y_{\text{true}})$ 关于 X 的梯度, 其中 $\text{sig}()$ 表示对梯度取符号。

文献[6]在FGSM的基础上提出了随机快速梯度符号(Random Fast Gradient Sign Method, RFGSM)攻击, 通过引入随机扰动来修改输入样本, 比FGSM具有更高的攻击成功率。文献[7]改进了FGSM, 并提出了速度更快的FFGSM算法。文献[8]提出了动量迭代快速梯度符号(Momentum Iterative Fast Gradient Sign Method, MIFGSM)攻击, 通过每一步迭代累积梯度的动量, 生成更强的对抗样本。

2) BIM

2016年, 文献[9]提出对同一个人工智能模型多轮迭代使用FGSM的方法, 每次迭代均修改图像的像素值, 并通过梯度剪裁限制修改的像素值在原始图片的给定范围。

对于一个人工智能模型, 令 θ 为模型参数, 图像 X 为模型输入, Y_{true} 为 X 的真实类别, $L(\theta, X, Y_{\text{true}})$ 为模型的损失函数。对于扰动 $\eta = \alpha \cdot \text{sig}(\nabla_X L(\theta, X, Y_{\text{true}}))$, 通过BIM方法生成的新对抗样本为

$$\begin{aligned} X^{\text{adv}} &= X \\ X_{N+1}^{\text{adv}} &= \text{Clip}_{X, \epsilon} \{ X_N^{\text{adv}} + \alpha \cdot \text{sig}(\nabla_X L(\theta, X_N^{\text{adv}}, Y_{\text{true}})) \} \end{aligned} \quad (5)$$

式中: X^{adv} 为第 N 轮迭代所产生的对抗样本; $X_{N+1}^{\text{adv}} = \text{Clip}_{X, \epsilon} \{ \}$, 表示裁剪操作。

其中

$$\text{Clip}_{X, \epsilon} \{ X^{\eta} \} (x, y, z) = \min \{ 255, X(x, y, z) + \epsilon, \epsilon, \max \{ 0, X(x, y, z) - \epsilon, X'(x, y, z) \} \} \quad (6)$$

式中: $X(x, y, z)$ 为图像 X 在坐标 (x, y, z) 上的值; ϵ 为 X 的邻域。

添加扰动后的像素点和原图像对应像素点不会相差太大, 可以让两者在视觉上产生相像的效果。BIM方法经过多轮迭代生成对抗样本, 相比FGSM方法, 对抗样本的扰动效果更加不易被察觉, 攻击效果更好。

3) PGD

文献[10]提出一种能有效抵御一阶对抗攻击的防御方法。该方法扩展了BIM攻击方式, 攻击原理与BIM方法一致, 也是一种迭代攻击, 可以看作是FGSM方法的迭代版本, 通过迭代应用投影梯度下降(Projected Gradient Descent, PGD)优化生成对抗样本的目标函数。

对于一个人工智能模型, 令 θ 为模型参数, 图像 X 为模型输入, Y_{true} 为 X 的真实类别, $L(\theta, X, Y_{\text{true}})$ 为模型的损失函数。对于扰动 $\eta = \alpha \cdot \text{sig}(\nabla_X L(\theta, X, Y_{\text{true}}))$,

通过 PGD 方法生成的新对抗样本为

$$\left. \begin{aligned} X_0^{\text{adv}} &= X \\ X_{N+1}^{\text{adv}} &= P_{X+S} \left\{ X_N^{\text{adv}} + \alpha \cdot \text{sig}(\nabla_X L(\theta, X_N^{\text{adv}}, Y_{\text{true}})) \right\} \end{aligned} \right\} \quad (7)$$

式中： X_0^{adv} 为初始的对抗样本； X_N^{adv} 为第 N 轮迭代所产生的对抗样本； $P_{X+S}\{\}$ 为投影算子。

其中

$$P\{y\} = \arg \min_{x \in (X+S)} \left(\frac{1}{2} \|x - y\|_2^2 \right) \quad (8)$$

将得到的对抗样本输入到模型中进行训练，可以实现对攻击的抵御。由于 PGD 算法得到的对抗样本是满足损失最大的样本，因此模型如果能成功抵御 PGD 攻击，那么意味着模型能够实现对集合中所有扰动的攻击。与 BIM 方法相比，PGD 的迭代次数更多，而且加了随机化操作，使得 PGD 的攻击效果比 BIM 要好，能够生成更加不易察觉的对抗扰动。

PGDL2 算法在 PGD 的基础上使用 l_2 范数距离作为约束。文献[11]将期望转换（Expectation over Transformation, EOT）算法与 PGD 算法结合，提出了 EOTPGD 算法，使得对抗样本在选定的转换分布上保持对抗

性。

2.3 图像分类模型对抗攻击评估

针对图像分类典型模型 ResNet，使用的数据集为多种场景下轨道交通设备与自然图像。通过使用业内公开的 8 种对抗攻击方法对应算法，等间隔调整攻击扰动系数 k_{eps} 对 ResNet 模型进行微小扰动下的对抗攻击，得到 8 种攻击方法的攻击成功率，如表 2 所示，并绘制攻击成功率曲线如图 2 所示。

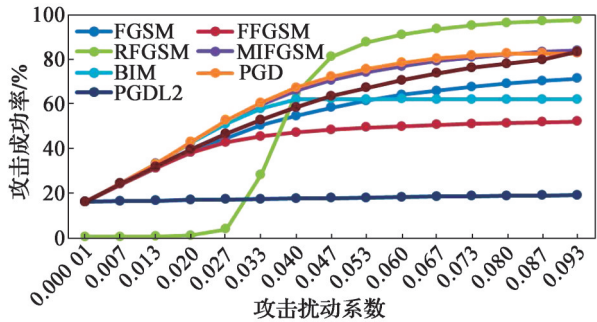


图2 攻击成功率曲线
Fig. 2 Attack success rate curves

表2 模型对抗攻击成功率

Table 2 Adversarial attack success rate of model %

攻击方法	k_{eps} 起始值 (0.000 01)	$k_{\text{eps}} = 0.007$	$k_{\text{eps}} = 0.013$	$k_{\text{eps}} = 0.020$	$k_{\text{eps}} = 0.027$	$k_{\text{eps}} = 0.033$	$k_{\text{eps}} = 0.040$	$k_{\text{eps}} = 0.047$	$k_{\text{eps}} = 0.053$	$k_{\text{eps}} = 0.060$	$k_{\text{eps}} = 0.067$	$k_{\text{eps}} = 0.073$	$k_{\text{eps}} = 0.080$	$k_{\text{eps}} = 0.087$	$k_{\text{eps}} = 0.093$
FGSM	15.90	23.72	30.99	38.05	44.31	50.18	54.34	58.12	61.27	63.80	65.73	67.44	68.92	70.01	71.12
FFGSM	15.89	23.70	30.92	37.88	42.53	45.17	47.05	48.24	49.17	49.66	50.43	50.84	51.16	51.60	51.88
RFGSM	0.10	0.15	0.33	0.73	3.55	27.92	64.63	80.90	87.45	90.88	93.33	95.00	96.17	96.87	97.56
MIFGSM	15.90	24.16	32.85	42.38	51.84	59.37	65.79	70.35	73.88	76.65	78.93	80.58	82.07	83.14	83.77
BIM	15.90	24.15	32.81	42.19	50.82	57.60	61.87	61.89	61.88	61.89	61.88	61.88	61.88	61.88	61.88
PGD	15.90	24.10	32.77	42.62	52.43	60.36	67.08	72.04	75.48	78.19	80.21	81.53	82.44	82.44	82.44
PGDL2	15.88	16.11	16.35	16.66	16.84	17.06	17.34	17.49	17.71	17.92	18.15	18.32	18.50	18.71	18.89
EOTPGD	15.90	23.97	31.64	39.20	46.30	52.52	58.22	63.27	66.93	70.41	73.47	76.09	77.74	79.65	83.06

由图 2 可知，8 种攻击算法的攻击成功率随着扰动系数的增大而增加，但在不同扰动系数下不同攻击算法的攻击成功率差距较大；当 $k_{\text{eps}} \in [0.000\ 01, 0.002\ 7]$ 时，RFGSM 算法的攻击成功率最低，攻击效果最差，PGDL2 算法攻击效果次之，其他 6 种方法的攻击成功率近似，攻击效果比较好；在 $k_{\text{eps}} \in (0.002\ 7, 0.004\ 0]$ 时，PGDL2 算法的攻击成功率最低，攻击效果最差，其他 7 种方法的攻击成功率近似，攻击效果比较好，但 RFGSM 算法的攻击成功率变化率远远大于其他方法；当 $k_{\text{eps}} > 0.004\ 0$ 时，PGDL2 算法的攻击成功率最低，攻击效果最差，RFGSM 算法的攻击成功率最高，攻击效果最好，其他 6 种方法的攻击成功率近似，攻击效果比较好。

在全扰动系数范围内，PGD 和 MIFGSM 的攻击成功率均比较高，变化率比较稳定。对于图像分类模型，使用 PGD 和 MIFGSM 的攻击效果最好，图像分类模型

对通过 PGD 和 MIFGSM 方法生成的对抗样本的鲁棒性较差，这就表明通过迭代实现的对抗攻击算法对模型扰动更大，因此模型针对迭代对抗攻击进行防御可有效提高模型的鲁棒性。

3 攻击原理分析

3.1 模型可解释性

在人工智能和机器学习场景中，“模型可解释性”定义为被人类理解或解释的能力^[12]，指模型能用通俗易懂的语言进行表达，将模型的预测过程转化为具备逻辑关系的能力。“可解释性”是人类与决策模型之间的接口，可以辅助开发人员和相关业务方理解模型的内部机制和结果。

人工智能的模型可解释性可分为事前可解释性和事后可解释性两大类。其中，事前可解释性是指通过

训练结构简单、可解释性好的模型, 或将可解释性结合到具体模型结构中的自解释模型, 从而使模型本身具备可解释的能力; 事后可解释性是指通过开发可解释性技术解释已训练好的机器学习模型。本文主要采用LRP、Grad-CAM和LIME这3种事后可解释性方法对图像分类模型ResNet进行可解释性分析。

3.2 可解释性方法

1) LRP

文献[13]则提出了一种分层相关性传播方法 (Layer-Wise Relevance Propagation, LRP), 用于计算单个像素对图像分类器预测结果的贡献, 得到每个像素点与模型预测结果的相关性。LRP方法通过自定义模型反向传输规则, 将模型决策的结果归因到样本像素中, 进而可视化其中对分类有利和对分类有害的显著区域。LRP算法主要适用于需要解释神经网络模型预测结果的情况, 可以帮助理解模型对输入数据做出预测的方式, 以及确定每个输入特征对最终预测的贡献程度。

2) Grad-CAM

文献[14]提出了类激活映射 (Class Activation Mapping, CAM) 解释方法。该方法利用全局平均池化 (Global Average Pooling) 层替代传统CNN模型中除softmax层以外的所有全连接层, 并通过将输出层的权重投影到卷积特征图来识别图像中的重要区域。文献[15]对CAM方法进行了改进, 提出了一种将梯度信息与特征映射相结合的梯度加权类激活映射方法 (Gradient-weighted Class Activation Mapping, Grad-CAM)。Grad-CAM通过突出显示输入图像中有助于网络预测的重要区域, 无须修改模型架构或重新训练模型, 解决了模型的可解释性与准确性之间的权重关系, 适用于多种任务和任何基于CNN结构的人工智能模型。

3) LIME

文献[16]基于神经网络的局部线性假设, 提出了一种模型无关局部可解释方法 (Local Interpretable Model-agnostic Explanation, LIME)。LIME方法使用训练的局部代理模型对单个样本进行解释, 通过选取关注的实例样本, 在其附近进行扰动生成新的样本, 利用该实例样本和新样本训练一个可解释的人工智能模型 (如线性回归、决策树)。该解释模型与原模型的局部近似, 可基于该解释模型解释原模型对实例样本的决策依据。LIME方法可通过解释人工智能模型预测结果背后的原理, 帮助用户决定是否信任模型的结果, 为决策制定、模型改进等提供指导。

3.3 图像分类模型攻击机理分析

针对图像分类典型模型ResNet, 使用的数据集为多种场景下的轨道交通设备与自然图像, 选取FGSM、BIM、PGD等8种对抗攻击方法生成对抗样本, 采用

LRP、Grad-CAM和LIME等3种可解释性方法计算图像分类模型对于不同样本不同区域的敏感程度, 进而得出模型在决策过程中对于原始样本和对抗样本的重点关注区域, 并通过图像标注的形式加以展示, 如表3所示。

表3 图像分类模型攻击原理分析
Table 3 Analysis of the attack mechanism of image classification models

分析算法	原始图片	LRP	Grad-CAM	LIME
原始				
FGSM				
FFGSM				
RFGSM				
MIFGSM				
BIM				
PGD				
PGDL2				
EOTPGD				

由表3可知, 对于同一图像样本, 使用3种解释性方法分析图像分类模型, 对样本的重点关注区域会有所不同。相比其他2种方法, LRP方法分析的重点关注区域更多, 能够更有效地标注出图像分类模型的分类目标。同一种方法对原始样本和对抗样本的解释、图像分类模型对对抗样本分类的重点关注区域要少于原始样本, 而且关注区域与目标差异性较大。图像分类模型在对样本进行训练过程中可以根据分类目标、模型结构、样本特性等元素, 采取合适的可解释性方法对模型决策过程进行解释, 分析模型内部工作机制并优化, 以提高模型分类准确率。

4 模型防御训练

4.1 模型防御概念

模型防御指通过对人工智能模型进行改进, 使其能更好地抵御对抗样本攻击。其目标是提高模型的鲁

棒性,使模型在面对对抗样本时仍能做出正确的预测。模型防御主要有主动防御和被动防御两种方式。其中,主动防御指在模型训练过程中查找模型的漏洞或后门,并予以解决,而被动防御指在不改变模型的任何元素基础上增加一层防御。本文主要采用对抗训练、群智能化防御和PACA自动化监测等3种方法对图像分类模型ResNet进行防御。

1) 对抗训练

文献[5]提出通过对抗训练的方法优化模型参数,增强模型鲁棒性。该方法是指在每次迭代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,从而使训练后的模型对对抗样本具有鲁棒性。通常对抗训练是基于对抗样本类型的,如模型是针对FGSM攻击进行训练的,即称之为FGSM对抗训练。

2) 群智能化防御

群智能化防御方法是基于不同攻击方法生成不同的对抗样本,然后分别将不同的对抗样本应用于模型,进行鲁棒性训练,通过多种对抗攻击放大集成训练的方式对模型鲁棒性进行加固防御,以此提高模型的防御成功率。群智能化防御方法为对抗训练方法的集成,可很好地提高模型鲁棒性。

3) PACA自动化监测

PACA自动化监测方法是基于原始图片计算图片的预测置信度信息,然后分别将原始图片与预测置信度特征分别输入结构相同的2个检测网络中,2个检测网络的输出叠加后再经过激活函数得到对最终的预测标签信息,从而识别对抗攻击^[17]。

PACA自动化监测方法的思想为:构建1个对抗样本的检测模型,该模型具有2个结构相同的子模块,分别是图像流卷积网络 F_I 和梯度流卷积网络 F_G ,子模块的输入分别为图像 X 和置信度梯度 g 。置信度梯度 g 可以通过置信度损失函数 $L(Y)$ 得到,置信度损失函数 $L(Y)$ 为

$$L(Y) = - \sum_{i=1}^n [t_i \log(y_i)], \quad y_i \in Y \quad (9)$$

$$y_i = \frac{e^{Z_i}}{\sum_{i=1}^n e^{Z_i}} \quad (10)$$

式中: n 为类别数量; Z 为logits值; t 为one-hot向量。

假设2个子模块的输出分别为:

$$\begin{cases} Z_I = F_I(x) \\ Z_G = F_G(g) \end{cases} \quad (11)$$

那么,最终输入为

$$Z = Z_I + Z_G \quad (12)$$

4.2 图像分类模型防御训练

针对图像分类典型模型ResNet,使用的数据集为多种场景下的轨道交通设备与自然图像,采用对抗训

练、群智能化防御和PACA自动化监测3种方法进行模型防御,选取8种对抗攻击方法对防御后模型进行对抗攻击评估,以此评价3种防御方法的效果。其中,对抗训练方法使用的训练样本为相应对抗攻击方法生成的对抗样本,即使用FGSM方法对防御模型进行对抗攻击,则模型对抗训练样本为FGSM方法生成的对抗样本;群智能化防御方法则是使用8种对抗攻击方法生成的对抗样本对模型进行训练。

设置扰动系数 $k_{\text{eps}} = 0.093$,对图像分类模型进行防御训练,得到原始模型和采用不同防御训练方法训练后的模型面对8种攻击方法的攻击成功率(如表4所示),并绘制攻击成功率曲线(如图3所示)和防御成功率曲线(如图4所示),以此分析3种防御方法的防御效果。

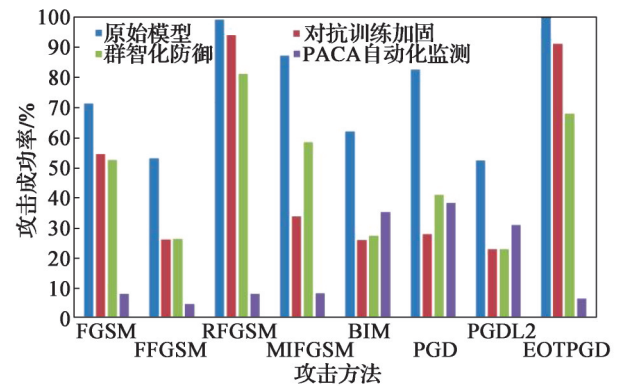


图3 训练后模型攻击成功率

Fig. 3 Post-training attack success rate of model

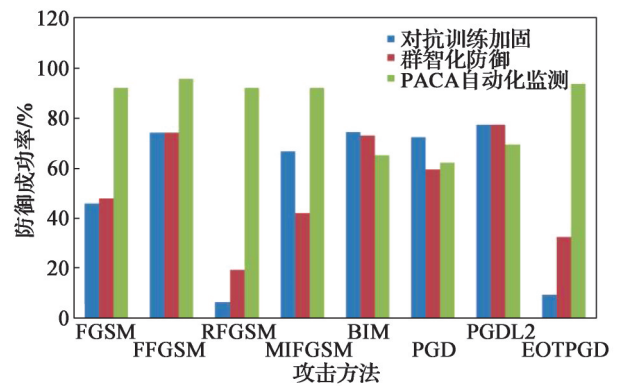


图4 训练后模型防御成功率

Fig. 4 Post-training defense success rate of model

防御成功率为模型的分类型准确率,训练后的模型防御成功率越高,分类准确率越高,则其鲁棒性越好。由图3和图4可知,经过训练后的模型攻击成功率明显下降,不同防御训练方法的防御成功率差异较大。对于FGSM、FFGSM、RFGSM、MIFGSM和EOTPGD这5种攻击方式,PACA自动化监测方法训练后的模型防御成功率远远高于其他2种方法,效果最好;对于BIM、PGD、PGDL2这3种攻击方式,3种防御训练方法的防御成功率相差不大。

表4 图像分类模型防御训练

Table 4 Defense training of image classification model								%
攻击方法	FGSM	FFGSM	RFGSM	MIFGSM	BIM	PGD	PGDL2	EOTPGD
原始模型	71.12	52.97	98.98	87.17	61.88	82.44	52.29	99.71
对抗训练加固	54.46	26.10	93.87	33.64	25.90	27.81	22.87	91.08
群智化防御	52.42	26.14	80.94	58.41	27.26	40.81	22.91	67.86
PACA 自动化监测	8.00	4.55	8.03	8.14	35.15	38.14	30.86	6.47

PACA 自动化监测方法训练后的图像分类模型对 8 种对抗攻击方法的防御成功率比较高, 对对抗样本有很好的抵御能力。在实际应用场景中, 图像分类模型采用 PACA 自动化监测方法训练加固, 可以极大地提高模型的鲁棒性, 但该方法需要构建新的对抗样本检测模型, 相比其他 2 种方法模型训练工作量有所增多。

5 模型攻防博弈

模型攻防博弈指以对抗样本的生成和防御算法作为博弈的双方, 其中选定的不同对抗样本生成算法作为攻击方的可选策略集, 以不同攻击算法生成的对抗样本进行对抗训练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。本文使用 8 种攻击方法对图像分类模型 ResNet 进行攻击, 获得准确率收益矩阵、鲁棒性收益矩阵和博弈策略收益矩阵, 以此进行攻防博弈分析。

5.1 准确率收益矩阵

准确率收益矩阵主要是使用对抗训练方法在不同扰动系数下训练加固后的模型, 测试该模型对正常样本的分类准确率, 以此分析图像分类模型的最佳防御策略。图像分类模型 ResNet 使用 8 种攻击方法在不同扰动系数下生成的对抗样本进行对抗训练, 测试训练后的模型对原始样本的分类准确率。针对每种防御训练方法, 选取 3 组最佳的分类准确率数据, 如表 5 和图 5 所示。

表5 图像分类模型准确率收益矩阵

Table 5 Accuracy gain matrix of image classification model								%
指标	原始	FGSM			FFGSM			
		$k_{\text{eps}}=0.141\ 18$	$k_{\text{eps}}=0.156\ 86$	$k_{\text{eps}}=0.172\ 55$	$k_{\text{eps}}=0.123\ 53$	$k_{\text{eps}}=0.137\ 25$	$k_{\text{eps}}=0.150\ 98$	
分类准确率	73.65	80.28	80.89	81.61	82.78	82.19	82.62	
指标	原始	RFGSM			MIFGSM			
		$k_{\text{eps}}=0.113\ 65$	$k_{\text{eps}}=0.126\ 27$	$k_{\text{eps}}=0.138\ 90$	$k_{\text{eps}}=0.118\ 24$	$k_{\text{eps}}=0.131\ 37$	$k_{\text{eps}}=0.144\ 51$	
分类准确率	73.65	74.69	78.67	80.00	79.64	78.99	79.45	
指标	原始	BIM			PGD			
		$k_{\text{eps}}=0.231\ 18$	$k_{\text{eps}}=0.256\ 86$	$k_{\text{eps}}=0.282\ 55$	$k_{\text{eps}}=0.141\ 18$	$k_{\text{eps}}=0.156\ 86$	$k_{\text{eps}}=0.172\ 55$	
分类准确率	73.65	82.67	82.67	82.67	80.57	80.57	80.57	
指标	原始	PGDL2			EOTPGD			
		$k_{\text{eps}}=0.900\ 00$	$k_{\text{eps}}=1.000\ 00$	$k_{\text{eps}}=1.100\ 00$	$k_{\text{eps}}=0.270\ 00$	$k_{\text{eps}}=0.300\ 00$	$k_{\text{eps}}=0.330\ 00$	
分类准确率	73.65	82.64	82.69	82.31	74.99	75.00	73.77	

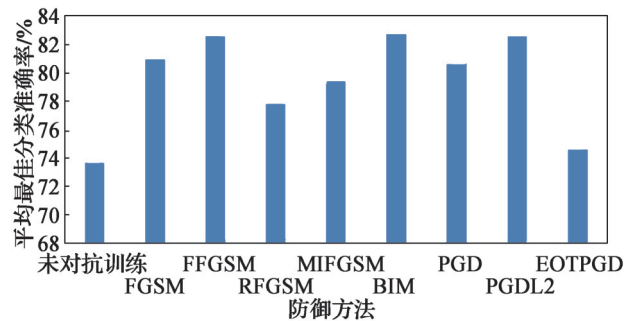


图5 图像分类模型各防御方法平均准确率收益图

Fig. 5 Bar chart of average accuracy gain of defense methods for image classification model

由表 5 和图 5 可知, 未经过对抗训练的图像分类模型对原始样本的分类准确率为 73.65%, 而经过对抗训

练的模型对原始样本的分类准确率明显提高, 并且对于同一种对抗训练方法, 使用高扰动系数下生成的对抗样本对模型进行训练, 模型对原始样本的分类准确率有所提高。对于图像分类模型, 使用 FGSM、FFGSM、BIM、PGDL2 等 4 种方法进行防御训练可以有效提高模型对原始样本的分类准确率, 因此图像分类模型对于原始样本的最佳防御策略为 FGSM、FFGSM、BIM、PGDL2 等 4 种方法。

5.2 鲁棒性收益矩阵

鲁棒性收益矩阵主要是使用对抗训练方法训练加固后的模型, 测试该模型在不同扰动系数下对对抗样本的鲁棒性, 以此分析图像分类模型的最佳防御策略。

鲁棒性是指在某些参数略微改变或控制变量稍微偏离最优值时系统仍能保持稳定性和有效性。本文所提的模型鲁棒性收益为对抗样本对模型的攻击失败的概率，对于无目标的攻击，模型鲁棒性收益为模型分类准确率。图像分类模型 ResNet 使用 8 种攻击方法对在不同

扰动系数下生成的对抗样本进行对抗训练，测试经过训练后的模型再次使用 8 种攻击方法在相同的扰动系数下进行攻击。根据准确率收益矩阵，每种攻击方法选择相同的扰动系数，确定模型的鲁棒性收益，计算结果如表 6 和图 6 所示。

表 6 图像分类模型鲁棒性收益矩阵

Table 6 Robustness gain matrix of image classification model %

防御方法		FGSM	FFGSM	RFGSM	MIFGSM	BIM	PGD	PGDL2	EOTPGD
FGSM	$k_{\text{eps}}=0.141\ 18$	47.00	66.99	37.95	30.12	63.53	45.50	68.84	46.21
	$k_{\text{eps}}=0.156\ 86$	45.54	64.18	43.46	25.79	61.15	40.90	67.88	45.45
	$k_{\text{eps}}=0.172\ 55$	44.07	63.22	48.20	25.18	60.25	39.33	67.08	50.32
FFGSM	$k_{\text{eps}}=0.123\ 53$	47.36	73.44	16.43	45.94	73.16	63.44	76.50	30.78
	$k_{\text{eps}}=0.137\ 25$	47.61	73.90	17.31	46.22	73.34	63.82	76.64	31.81
	$k_{\text{eps}}=0.150\ 98$	47.11	74.11	17.03	46.00	73.54	63.64	76.77	32.02
RFGSM	$k_{\text{eps}}=0.113\ 65$	33.60	60.63	6.88	26.34	57.45	40.44	63.25	8.64
	$k_{\text{eps}}=0.126\ 27$	33.93	61.67	6.13	27.42	60.00	42.21	65.91	8.26
	$k_{\text{eps}}=0.138\ 90$	28.04	58.45	2.85	20.46	55.30	32.96	62.86	2.35
MIFGSM	$k_{\text{eps}}=0.118\ 24$	65.97	76.92	55.72	67.90	76.72	74.09	77.58	62.34
	$k_{\text{eps}}=0.131\ 37$	64.90	75.65	54.57	66.36	75.58	72.43	76.62	61.97
	$k_{\text{eps}}=0.144\ 51$	63.87	75.82	52.49	65.23	75.66	72.09	76.80	59.83
BIM	$k_{\text{eps}}=0.231\ 18$	49.33	74.51	19.23	48.48	74.10	65.10	76.93	34.19
	$k_{\text{eps}}=0.256\ 86$	49.33	74.51	19.23	48.48	74.10	65.10	76.93	34.19
	$k_{\text{eps}}=0.282\ 55$	49.33	74.51	19.23	48.48	74.10	65.10	76.93	34.19
PGD	$k_{\text{eps}}=0.141\ 18$	61.63	76.60	46.05	63.32	76.50	72.19	77.76	56.45
	$k_{\text{eps}}=0.156\ 86$	61.63	76.60	46.05	63.32	76.50	72.19	77.76	56.45
	$k_{\text{eps}}=0.172\ 55$	61.63	76.60	46.05	63.32	76.50	72.19	77.76	56.45
PGDL2	$k_{\text{eps}}=0.900\ 00$	46.64	73.82	14.59	44.69	73.31	62.78	76.72	28.51
	$k_{\text{eps}}=1.000\ 00$	48.13	74.27	17.51	47.14	73.83	63.97	77.13	31.99
	$k_{\text{eps}}=1.100\ 00$	50.86	74.89	21.65	50.57	74.42	66.18	77.23	36.71
EOTPGD	$k_{\text{eps}}=0.270\ 00$	38.06	63.97	13.74	33.83	62.65	49.13	66.91	18.94
	$k_{\text{eps}}=0.300\ 00$	32.51	59.32	7.34	26.16	57.25	40.34	63.11	8.92
	$k_{\text{eps}}=0.330\ 00$	31.48	58.46	6.36	24.83	56.41	38.58	62.11	7.31

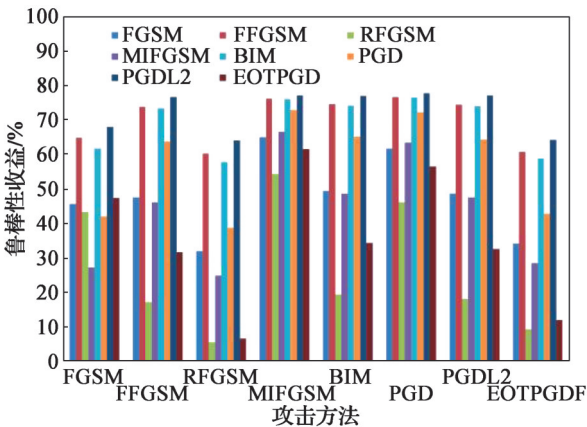


图 6 图像分类模型各防御方法的平均鲁棒性收益图

Fig. 6 Bar chart of average robustness gain of defense methods for image classification model

由图 6 可知，经过对抗训练的模型对对抗样本的分类准确率远低于对原始样本的分类准确率，但对比原始模型的鲁棒性有了较大提升。对于图像分类模型，经过 MIFGSM 和 PGD 攻击方法训练后的模型对 8 种对抗攻击

方式生成的对抗样本的分类准确率比较高，鲁棒性收益优于其他 6 种方法，因此图像分类模型对于对抗样本的最佳防御策略为 MIFGSM 和 PGD 这 2 种攻击方法。

5.3 博弈策略收益矩阵

博弈策略收益矩阵主要是选择最佳防御方法训练后的模型在不同鲁棒性权重下对对抗样本的收益，以此分析图像分类模型的最佳博弈策略。模型博弈策略是指防御方采用不同的防御方式在不同的扰动系数下对模型进行防御加固，攻击方采用不同的攻击方式在不同的扰动系数下对防御的模型进行对抗攻击，计算模型分类准确率和鲁棒性，确定防御方的得分。博弈测试收益公式为

$$G=P \cdot A+(1-P) R \quad (13)$$

式中： G 为博弈策略收益； P 为鲁棒性权重，当 $P=0$ 时，模型抗干扰能力最强，鲁棒性最强，而当 $P=1$ 时，模型抗干扰能力最弱，鲁棒性最弱； A 为分类准确率； R 为鲁棒性。

在线性变化的鲁棒性权重下, 图像分类模型ResNet对采用8种攻击方法在不同扰动系数下生成的对抗样本进行对抗训练, 训练后的模型再次使用8种攻击方法在相同的扰动系数下进行攻击, 测试图像分类模型在不同鲁棒性权重下的博弈收益, 计算结果如表7和图7所示。

表7 图像分类模型博弈策略收益矩阵

Table 7 Gain matrix of game strategy for image classification model								%
权重 P	FGSM	FFGSM	RFGSM	MIFGSM	BIM	PGD	PGDL2	EOTPGD
0	65.97	76.92	55.72	67.90	76.72	74.09	77.76	62.34
10	67.34	77.19	58.11	69.07	77.01	74.65	78.04	64.07
20	68.70	77.46	60.50	70.25	77.31	75.20	78.32	65.80
30	70.07	77.79	62.90	71.42	77.72	75.76	78.80	67.53
40	71.44	78.19	65.29	72.60	78.13	76.31	79.35	69.26
50	72.81	78.60	67.68	73.77	78.54	76.87	79.91	70.99
60	74.17	79.41	70.07	74.94	79.24	77.42	80.47	72.72
70	75.54	80.22	72.46	76.19	80.10	78.06	81.02	74.45
80	76.91	81.04	74.93	77.29	80.96	79.16	81.58	76.18
90	79.34	81.85	78.27	79.25	81.82	80.91	82.15	78.48
100	82.78	82.78	82.78	82.78	82.78	82.78	82.78	82.78

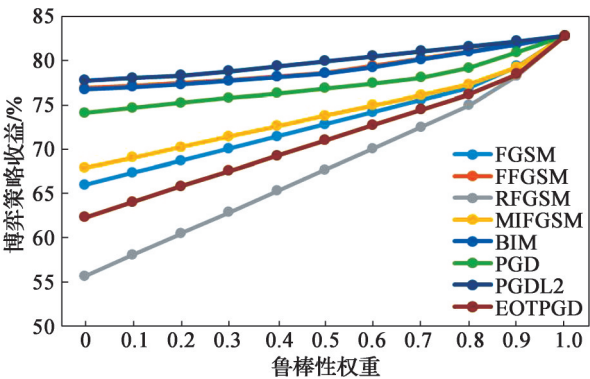


图7 图像分类模型博弈策略收益曲线

Fig. 7 Gain curve of game strategy for image classification model

由图7可知, 图像分类模型博弈策略收益随着鲁棒性权重的增大而增加, 不同攻击方法的博弈策略收益不一致, 使用PGD、BIM对抗防御训练后的模型的博弈策略收益较高。当鲁棒性权重 $P=0$ 时, 8种方法的博弈策略收益鲁棒性权重最低; 当 $P=100\%$ 时, 8种方法的博弈策略收益为82.78%, 此时博弈策略收益为分类准确率, 但模型的鲁棒性为0, 模型对抗干扰的能力很差。在实际应用中, 图像分类模型应针对不同场景需求, 选取不同的鲁棒性权重进行防御训练, 以此达到分类准确率和鲁棒性的最优博弈。

6 结束语

本文针对自动驾驶领域典型的图像分类模型ResNet, 使用多种场景下轨道交通设备和自然图像作为数据集, 选取FGSM、BIM、PGD三类8种对抗攻击算法生成对抗样本, 并对其进行攻击评估, 并采用LRP、Grad-CAM、LIME这3种可解释性方法计算图像分类模型对于不同样本不同区域的敏感程度, 开展攻击

原理分析, 得出不同的对抗攻击方法对模型的扰动程度不同, 生成的噪声像素针对的图像区域也有所不同。基于对抗攻击效果和攻击原理分析, 采用对抗训练、群智能化防御和PACA自动化监测3种防御方法对模型进行防御训练, 分析3种方法的防御效果, 其中PACA对8种对抗攻击的防御效果最好。最后, 通过博弈算法对弈获得攻守收益矩阵, 确定最佳防御策略。本文通过规模化测评、选择合适的模型可解释性方法和模型防御训练方法, 确定模型最佳攻防博弈策略, 能够有效地提高模型抗攻击能力, 增强鲁棒性, 有效保障自动驾驶系统的可靠性与安全性, 为自动驾驶系统图像分类模型的开发和使用者提供可靠性与安全性参考。

参考文献:

[1] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//ICLR. 2nd International Conference on Learning Representations, ICLR 2014. Banff: ICLR, 2014: 1-10.

[2] GOODFELLOW I J, SHLENS J, SZEGEDY C, et al. Explaining and harnessing adversarial examples[C]//ICLR. ICLR 2015. San Diego: ICLR, 2015: 1-11.

[3] WANG Jia, WANG Chengyu, LIN Qiuzhen, et al. Adversarial attacks and defenses in deep learning for image recognition: a survey[J]. Neurocomputing, 2022, 514: 162-181.

[4] WANG Chenxu, ZHANG Ming, ZHAO Jinjing, et al. Black-box adversarial attacks on deep neural networks: a survey [C]//IEEE. 2022 4th International Conference on Data Intelligence and Security. Shenzhen: IEEE, 2022: 88-93.

[5] 王志波, 王雪, 马菁菁, 等. 面向计算机视觉系统的对抗样本攻击综述[J]. 计算机学报, 2023, 46(2): 436-468.

WANG Zhibo, WANG Xue, MA Jingjing, et al. Survey on adversarial example attack for computer vision systems[J].

- Chinese journal of computers, 2023, 46(2): 436-468.
- [6] TRAMÈR F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: attacks and defenses[EB/OL]. (2020-04-26) [2024-10-12]. <https://arxiv.org/abs/1705.07204>.
- [7] WONG E, RICE L, KOLTER J Z. Fast is better than free: revisiting adversarial training[EB/OL]. (2020-01-12) [2024-10-12]. <https://arxiv.org/abs/2001.03994>.
- [8] DONG Yinpeng, LIAO Fangzhou, PANG Tianyu, et al. Boosting adversarial attacks with momentum[C]//IEEE. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 9185-9193.
- [9] KURAKIN A, GOODFELLOW I J, BENGIO S, et al. Adversarial examples in the physical world[C]//ICLR. 5th International Conference on Learning Representations, ICLR 2017. Toulon: ICLR, 2017: 1-13.
- [10] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]//ICLR. 6th International Conference on Learning Representations, ICLR 2018. Vancouver: ICLR, 2018: 1-28.
- [11] ZIMMERMANN S R. Comment on "adv-BNN: improved adversarial defense through robust bayesian neural network"[EB/OL]. (2019-07-01) [2024-10-12]. <https://arxiv.org/abs/1907.00895>.
- [12] GUIDOTTI R, MONREALE A, RUGGIERI S, et al. A survey of methods for explaining black box models[J]. ACM computing surveys, 2018, 51(5): 1-42.
- [13] BACH S, BINDER A, MONTAVON G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation[J]. PLoS one, 2015, 10(7): e0130140.
- [14] ZHOU Bolei, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization[C]//IEEE. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 2921-2929.
- [15] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[C]//IEEE. 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 618-626.
- [16] RIBEIRO M T, SINGH S, GUESTRIN C. "Why should i trust you?": explaining the predictions of any classifier[C]//Association for Computing Machinery. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: Association for Computing Machinery, 2016: 1135-1144.
- [17] CHEN Kejiang, CHEN Yuefeng, ZHOU Hang, et al. Adversarial examples detection beyond image space[C]//IEEE. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto: IEEE, 2021: 3850-3854.

Research on adversarial attacks and defense performance of image classification models for automated driving systems

TANG Jun, HUANG Wenjing, LI Shuang, WU Zili

(CRRC Zhuzhou Institute Co., Ltd., Zhuzhou, Hunan 412001, China)

Abstract: Image classification models have been widely applied to facilitate functions such as autonomous perception and positioning for automated driving in many transportation systems, including automobiles, autonomous rail and urban rail transit systems. However, output and judgment errors generated by these models due to the presence of adversarial examples, impose a great impact on the security and safety associated with their applications in automated driving systems. A sensitivity analysis-based noise superposition attack strategy was employed to perform white-box adversarial attacks against ResNet, a typical image classification model. Subsequent studies evaluated the attack effects and defense performance. Firstly, algorithms such as FGSM, BIM, and PGD were selected to generate adversarial examples and tests were conducted through adjusting perturbation coefficients to determine attack success rates under small perturbations. Then, sensitivity analyses were carried out across different regions and examples to identify attack mechanisms using three adversarial interpretation algorithms: LRP, Grad-CAM, and LIME. Based on these analysis results, optimization algorithms such as swarm intelligence defense and adversarial training were adopted to verify the classification performance of the model following adversarial training. A benefit matrix for attack and defense was established using a game algorithm, leading to the development of an optimal defense strategy. Finally, a solution was proposed to ensure security and safety associated with the application of image classification models in advanced driver assistance systems, based on the study findings of adversarial attacks and attack-defense strategies.

Keywords: transportation; automated driving; cyber security; functional safety; image classification; adversarial examples; defense performance