浙江科技学院学报，第 34 卷第 1 期，2022 年 2 月
Journal of Zhejiang University of Science and Technology
Vol. 34 No. 1, Feb. 2022

doi：10.3969/j.issn.1671-8798.2022.01.007

自动驾驶场景下对交通路标对抗攻击的防御

孙安临[1]，钱亚冠[1]，顾钊铨[2]，楼　琼[1]，李俊峰[1]
（1.浙江科技学院 理学院，杭州 310023；2.广州大学 网络空间先进技术研究院，广州 510006）

摘　要：针对现有单一对抗防御方法不能使自动驾驶视觉系统有效抵御交通路标对抗攻击的问题，提出一种多阶段对抗防御方法。首先，应用焦点损失消除正负样本数量不平衡的影响，提高对抗训练过程中模型分类的准确率；同时为使模型拥有更强的泛化能力，对数据集做混合数据增强，并在训练开始前预热学习率。然后，使用知识蒸馏算法将教师模型群的“知识信息”迁移到学生模型群。最后，以加权的方式平均学生模型群体的预测结果。经本防御方法训练后，学生模型对交通路标对抗样本的分类准确率由 8%～19%提升到了 69%～83%；同时与单一对抗防御方法相比，学生模型群体的预测准确率高达 85%，优于现有防御模型。在轻量级条件下，利用本防御方法训练的深度学习模型能有效抵御交通路标的对抗攻击，可为自动驾驶视觉系统防御对抗攻击提供参考。

关键词：自动驾驶；对抗攻击；对抗防御；知识蒸馏
中图分类号：TP393.081　　**文献标志码**：A　　**文章编号**：1671-8798(2022)01-0052-09

Defense against adversarial attack of traffic signs under autonomous driving

SUN Anlin[1], QIAN Yaguan[1], GU Zhaoquan[2], LOU Qiong[1], LI Junfeng[1]
(1. School of Sciences, Zhejiang University of Science and Technology, Hangzhou 310023, Zhejiang, China;
2. Cyberspace Institute of Advanced Technology, Guangzhou University,
Guangzhou 510006, Guangdong, China)

Abstract: In response to the problem that the existing single adversarial defense method fails to effectively prevent the autonomous driving vision system from the adversarial attack of traffic signs, a multi-stage adversarial defense method was proposed. Firstly, focal loss was applied to eliminate the influence of the quantity imbalance of positive and negative samples and improve the accuracy of model classification in the process of adversarial training. At the same time, enhancement of mixed data was performed on the data set, combined with warm up of the learning rate prior to training, for the sake of strengthening the generalization ability of the model. Then, the knowledge distillation algorithm

收稿日期：2021-03-16
基金项目：国家自然科学基金项目(61902082)
通信作者：钱亚冠(1976—　)，男，浙江省绍兴人，教授，博士，主要从事机器学习与人工智能安全研究。E-mail：qianyaguan@zust.edu.cn。

was used to transfer the knowledge information of the teacher model group to the student model group. Finally, the prediction results of the student model group were averaged in a weighted way. After being trained by this defense method, the classification accuracy of student models for adversarial samples of traffic signs has been improved from 8%～19% to 69%～83%. Compared with the single adversarial defense method, the prediction accuracy of student model group is up to 85%, superior to the existing single-stage defense model. Under the condition of light weight, the deep learning model trained by this defense method can effectively resist the adversarial attack of traffic signs, and can provide reference for enhancing the defense of autonomous driving vision system against adversarial samples.

Keywords: autonomous driving; adversarial examples; adversarial defense; knowledge distillation

图像识别是计算机视觉的研究核心，也是目标检测、图像分割、姿态估计等诸多视觉任务的基础[1]。深度学习技术的发展使该方向的研究取得了丰硕成果，被广泛应用在各个领域，例如自动驾驶、智能机器人等[2-3]。然而，对抗样本的存在严重干扰了图像识别技术的运用[4]。对抗样本像一枚隐形的“炸弹”，能轻易使绝大部分的深度学习模型在图像分类时“失准”，发生误判。

近些年来，自动驾驶方兴未艾，随着相关研究的深入，自动驾驶的安全问题越来越受关注[5]，现有的对抗攻击方法种类多且攻击性强，如快速梯度符号攻击(fast gradient sign method，FGSM)[6]、映射梯度攻击(project gradient descent，PGD)[7]、基础迭代攻击(basic interative method，BIM)[8]等。实证研究表明，当对抗样本被恶意生成去攻击自动驾驶视觉系统时，自动驾驶的视觉系统会变得脆弱，甚至失去作用[9]，进而引发严重的交通事故。因此，如何防御对抗样本攻击是自动驾驶亟须解决的难题。Goodfellow 等[6]首先提出针对性对抗防御方法，以 FGSM 攻击时产生的对抗样本去训练深度神经网络，以此达到防御对抗攻击的目的；与此思想类似的还有 Madry 等[7]提出使用 PGD 攻击产生对抗样本训练深度神经网络。但是，这类对抗防御方法是单一的，受限于对抗样本的产生方式，以及数量、种类，无法防御多类型的对抗攻击。由此，Kurakin 等[10]提出了集成对抗训练(ensemble adversarial training，EAT)，该方法利用多个预先训练的模型中转移的对抗样本进行对抗训练；Xie 等[11]采用更直接的对抗防御方法，将训练图片进行随机调整大小和填充，以此增强模型的鲁棒性，弱化对抗攻击效果；Liu 等[12]提出一种随机噪声机制(random self-ensemble，RSE)来防御对抗样本攻击，即在每个卷积层之前添加一个噪声层，并集成预测结果以确保模型的鲁棒性；Dhillon 等[13]提出了随机修剪(stochastic activation pruning，SAP)的方法进行对抗防御，通过随机修剪网络中每一层的部分不活跃的神经元，以此增强对抗样本攻击的防御力。上述方法虽能在一定程度上抵御多类型的对抗攻击，但对抗训练缺乏充分性、多样性，且模型不够轻量，防御效果一般。

综上所述，现有的对抗防御方法大都是单一的、单阶段的，具有一定的局限性，无法有效应对各种各样的对抗样本攻击[14-15]，且它们的试验对象都是拥有庞大参数计算量的深度神经网络，不适用于计算力有限的车载系统。因此，在轻量级网络框架下，我们提出多阶段对抗防御方法。防御方法分为 3 个不同的阶段，生成有效防御对抗样本攻击的轻量级网络集群。

1　防御模块

1.1　对抗训练

对抗样本问题由 Szegedy 等[16]于 2013 年首次提出，即在图像分类时，通过在测试图片上添加一些细微的噪声扰动，使得深度学习模型分类出错。后续研究表明，按照某种规则生成的扰动噪声可定向使得深度学习模型输出攻击者预期的结果。2 种不同的对抗样本生成实例如图 1 所示。

对抗训练是一种有效防御对抗样本攻击的方法[17]，具体而言是在原样本上添加各种各样的扰动信息生成对抗样本集，并用对抗样本集训练深度神经网络模型的过程。对抗样本实例对应图 1(c)。在对深度神经网络模型进行对抗训练的过程中，将对抗样本与原样本一起输入深度神经网络模型，使模型学习

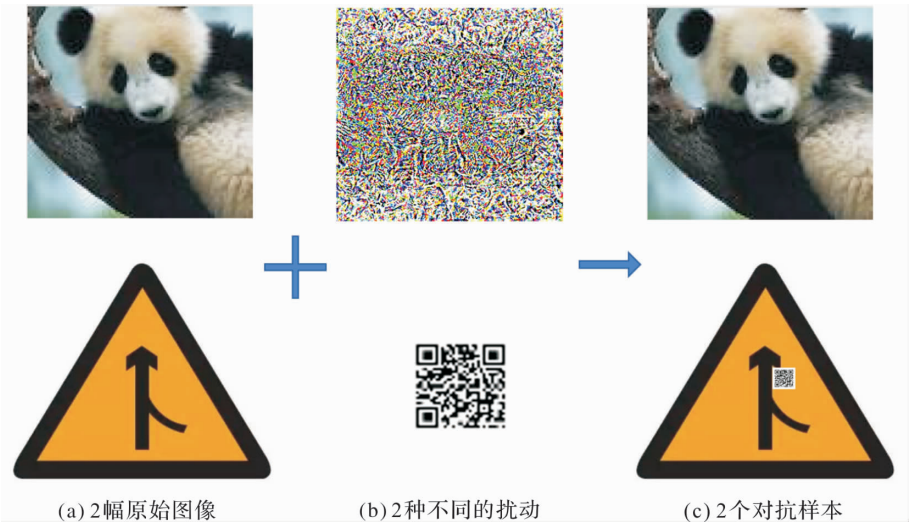


图 1 2 种不同的对抗样本生成实例

Fig. 1 Generation of two kinds of different adversarial samples

更丰富、更复杂的信息，从而对干扰信息形成抵抗力，提高模型的鲁棒性。

1.2 知识蒸馏

知识蒸馏是模型压缩的一种有效方式[18]，假设有一个复杂模型 $G(X)$ 和简单模型 $l(X)$，知识蒸馏就是将 $G(X)$ 的“知识”迁移到简单模型 $l(X)$ 上，使 $l(X)$ 在具备复杂模型的判断能力的同时又降低复杂度。这种“知识迁移”的实现原理是使 $l(X)$ 的 softmax 概率输出分布向 $G(X)$ 靠近，即 $l(X)$ 学习的类别标签是 $G(X)$ 预测输出的，不是真实的类别标签。通常情况下，我们称复杂模型 $G(X)$ 为教师模型，简单模型 $l(X)$ 为学生模型。softmax 激活函数的表达式如下：

$$p_i = \frac{\exp(z_i)}{\sum_j^m \exp(z_j)}, i = 1,2,\cdots,m。 \tag{1}$$

式(1)中：p_i 为相应的概率输出值；z_i 为第 i 个样本的前向传播输出值。

知识蒸馏时教师模型的分类层激活函数 softmax 的表达式如下：

$$p_i = \frac{\exp\frac{z_i}{T}}{\sum_j^m \frac{z_j}{m}}, i = 1,2,\cdots,m; T > 0。 \tag{2}$$

式(2)中：T 为知识蒸馏调节参数，取值为正整数。

教师模型分类层输出的概率向量 $\boldsymbol{y}_{\text{soft}} = \{p_1, p_2, \cdots, p_m\}$，称为软标签。在蒸馏过程中，学生模型的损失函数可由软硬标签两部分组成，其表达式如下：

$$L = \partial J(\boldsymbol{y}, \text{softmax}_{\text{student}}) + (1-\partial) J(\boldsymbol{y}_{\text{soft}}, \text{softmax}_{\text{student}}) T^2。 \tag{3}$$

式(3)中：$\partial \in [0,1]$；$J(\cdot)$ 为交叉熵损失函数。

1.3 模型投票

模型投票是一种决策机制，是将多个模型的预测结果综合平均，相比单个模型具有更高的准确率、更强的鲁棒性。设模型集合 $W = f_1(x), f_2(x), \cdots, f_N(x)$，若模型的输出类型为数值，则可按简单平均法和加权平均法 2 种方式综合模型的输出结果，其表达式分别为

$$H(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x); \tag{4}$$

$$J(x) = \sum_{i=1}^{N} w_i f(x)。 \tag{5}$$

式(4)～(5)中：$H(x)$、$J(x)$ 为模型组合 W 简单平均法和加权平均法对应的平均决策结果；N 为样本个

数；$w_i \in [0,1]$，$\sum w_i = 1$。

1.4　焦点损失

以分类任务为例，训练模型的样本集中往往包含诸多类别，而这些类别的数量往往是不一致的，有时甚至差距很大，以致正负样本类别失衡，且样本中还会有像素数较少的小目标，这些小目标使模型难以正确分类。因此，在不能继续扩大训练样本集时，可通过修改损失函数来消除正负样本类别不平衡、小目标难训练等问题。焦点损失就是一种有效的方法，以图像二分类为例，焦点损失可定义如下：

$$L_{\mathrm{F}} = \begin{cases} -\alpha(1-p)^{\gamma}\lg(p), y = 1; \\ -(1-\alpha)p^{\gamma}\lg(1-p), y = 0。 \end{cases} \tag{6}$$

式(6)中：y 为类别标签，$y=1$ 为样本正类，$y=0$ 为样本负类；$p \in [0,1]$为模型输出概率值；$\gamma \in [0,+\infty]$为调节因子，控制易分类样本对损失函数值的影响，γ 越大，易分类样本对损失函数的贡献越小；反之越大。$\alpha \in [0,1]$为平衡因子，控制正负样本总体对损失函数值的相对贡献大小。

1.5　预热学习率

在模型进行深度学习的过程中，学习率的重要性是毋庸置疑的，其衰减方式往往对模型最终的精度具有直接影响。相关研究[19]表明预热学习率在多个数据集上可以取得最佳效果，因此，我们在训练模型时使用这种学习率衰减方式，其定义如下：

$$\gamma_{\mathrm{L}} = \begin{cases} c, 0 < S \leqslant v; \\ g, v < S \leqslant f_1; \\ \Gamma(g), f_1 < S \leqslant f_2; \\ \Gamma(\Gamma(g)), f_2 < S \leqslant f_3; \\ \quad\vdots \qquad\qquad \vdots \\ o(g), f_{i-1} < S \leqslant f_i。 \end{cases} \tag{7}$$

式(7)中：γ_{L} 为学习率；c 为学习率预热初始值；g 为学习率初始值，$c<g$，c 变为 g 的过程称为学习率预热；$\Gamma(\cdot)$为衰减函数；$o(\cdot)$为无穷小值；S 为模型迭代次数；v 为学习率预热阈值；区间$[f_{i-1}, f_i]$为学习率衰减 $i-1$ 次后所对应的区间，$i \in [0, S-v]$。

1.6　混合数据增强

混合数据增强[20]可充当模型训练过程中的正则化和对抗训练，可有效缓解深度神经网络的过拟合现象，降低模型对对抗样本的敏感性。通过图像混合方式可建立数据集中不同类别的样本之间的联系，构造出虚拟样本，不同类别的样本图片生成的虚拟样本如图 2 所示。图像混合的定义如下：

$$\begin{cases} \bar{x} = \lambda x_i + (1-\lambda)x_j; \\ \bar{y} = \lambda y_i + (1-\lambda)y_j。 \end{cases} \tag{8}$$

式(8)中：x_i 和 x_j 为取自数据集的 2 个样本，$i,j \in [0,N]$)，N 为训练集样本总数；y_i 和 y_j 分别为 x_i 和 x_j 对应的类别标签；$\lambda \in [0,1]$。

(a) 样本图片1　(b) 样本图片2　(c) 虚拟样本

图 2　不同类别的样本图片生成的虚拟样本

Fig. 2　Virtual sample generated with different categories of sample pictures

2 防御方法

2.1 框架

我们的防御方法分为 3 个阶段，依次为焦点损失对抗训练阶段（第 1 阶段）、知识蒸馏阶段（第 2 阶段）、模型投票决策阶段（第 3 阶段）。前 2 个阶段为深度学习训练模型过程，其中第 1 阶段是基于焦点损失函数消除正负样本不平衡类别的对抗训练，第 2 阶段是对第 1 阶段的教师模型群体进行知识蒸馏，将其“知识信息”迁移到学生模型群体；第 3 阶段是对第 2 阶段的学生模型群体的预测结果进行加权平均。同时，我们在第 1 阶段的对抗训练中还使用了以下技术：混合数据增强、预热学习率。防御方法的定义如下：

假定有 n 个深度神经网络模型 $G_1(X),G_2(X),G_3(X),\cdots,G_n(X)$，在图像分类任务中，对所有 $G_i(X)$ 进行焦点损失的对抗训练，生成 n 个对抗训练后的模型 $F_1(X),F_2(X),F_3(X),\cdots,F_n(X)$；选定 n 个结构简单、计算量小的轻量级网络 $l_1(X),l_2(X),l_3(X),\cdots,l_n(X)$ 作为学生模型，使用知识蒸馏算法分别将 $F_1(X),F_2(X),F_3(X),\cdots,F_n(X)$（教师模型）的“知识信息”迁移到各学生模型上，之后对各学生模型的预测结果加权平均，$\sum w_i l_i(X), w_i \in [0,1]$。多阶段防御方法过程如图 3 所示。

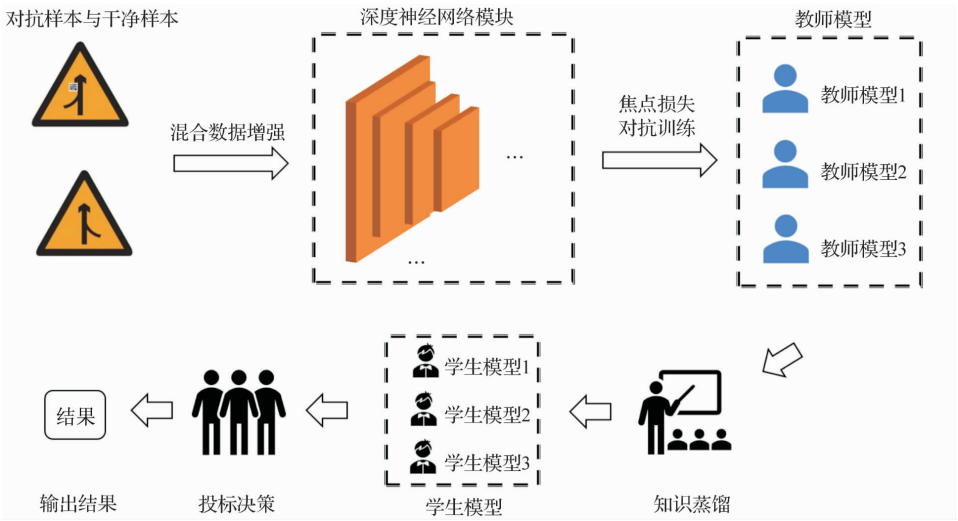


图 3 多阶段防御方法过程示意图

Fig. 3 Process diagram of multi-stage defense method

2.2 损失函数

我们的任务基于自动驾驶路标识别分类场景下，所选用的基础损失函数为交叉熵损失（式(9)）。构造好交通路标的对抗样本集后便可进行对抗训练，对抗训练损失函数如式(10)。为消除数据集正负样本类间的不平衡影响，同时为了提高模型的鲁棒性，我们在原对抗训练损失函数式(10)中加入焦点损失构造出最终的损失函数（式(11)）。

$$L(X,y) = -\sum_{i=1}^{C} y_i^X \lg(r_i^X); \tag{9}$$

$$L = \alpha L(X,y) + (1-\alpha) L(\widetilde{X},y); \tag{10}$$

$$L = -\sum_{i=1}^{C} \alpha(1-r_i^X) y_i^X \lg(r_i^X) + (1-\alpha)(1-r_i^{\widetilde{X}}) y_i^{\widetilde{X}} \lg(r_i^{\widetilde{X}})。 \tag{11}$$

式(9)～(11)中：X 为输入的干净样本；$\widetilde{X}$ 为输入的对抗样本；C 为目标类别数目；y_i^X 为输入样本 X 所对应的第 i 类别标签；r_i^X 为输入样本 X 模型预测输出第 i 类别概率值；同理可知 $y_i^{\widetilde{X}}$、$r_i^{\widetilde{X}}$。

3 试验结果及分析

3.1 数据集

自动驾驶视觉系统对交通路标的正确识别是确保自动驾驶汽车安全上路的第一步，且有研究[21]表

明通过在交通路标上粘贴“补丁”便可实现对自动驾驶视觉系统的对抗攻击，使其发生误判，因此我们的试验将针对交通路标的对抗补丁攻击进行防御。

目前国内公开的大规划交通路标数据集较少，因此本文使用德国道路交通数据集（The German Traffic Sign Detection Benchmark，G-TSDB），此数据集包含 43 个类别，一共 50 000 张图像，我们按照 7∶2∶1 的比例将数据集划分为训练集、验证集和测试集。在 3 个子数据集各取 1/2 的图片生成补丁对抗样本如图 1 所示。

3.2　软硬件试验平台及评价标准

试验评价标准遵从 ImageNet[22] 数据集的标准，即 top-k，k 为大于 1 的整数，代表输入一张照片，模型依据预测的类别概率值大小依次输出图片类别，前 k 个输出中包含正确类别。在本试验中，取 k 值为 1，即 top-1，代表模型一次命中正确类别的概率，即分类准确率。

所有试验的平台相同，硬件设施如下：GPU 为 2080Ti x4，CPU 为 48 核 Inter(R) Xeon(R)，2.20 GHz；软件设施包括服务器操作系统型号、深度学习框架版本、编程语言类型版本，依次为 Ubuntu 16.04、Pytorch 1.5、Python 3.7。

3.3　教师模型对抗训练

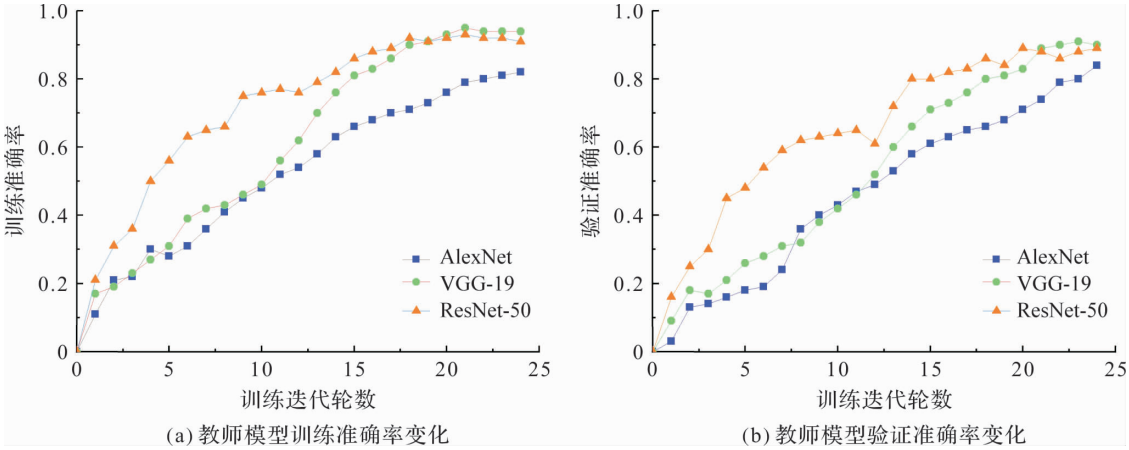
教师模型通常是体量巨大的深度神经网络模型，拥有极强的拟合数据能力，在分类问题背景下，经典的深度神经网络 AlexNet、VGG-19、ResNet-50、ResNet-101 都是合适的教师模型。因此，我们对这 4 种教师模型分别做补丁攻击的对抗训练。训练 25 个周期，使用随机梯度下降算法更新网络权重参数，初始学习率为 10^{-3}，图像输入大小[224，224]，随机旋转[−30°，30°]。这些教师模型在 G-TSDB 数据集上对抗训练前后的性能对比见表 1。

表 1　教师模型在 G-TSDB 数据集上对抗训练前后的性能对比

Table 1　Performance comparison of teacher model after and before adversarial training on G-TSDB data set

模型	参数量/MB	计算量/(10 亿次·s^{-1})	对抗训练前分类准确率/%		对抗训练后分类准确率/%	
			测试准确率	训练准确率	验证准确率	测试准确率
AlexNet	61.10	0.77	12	82	84	79
VGG-19	143.67	19.77	23	94	90	89
ResNet-50	25.56	4.14	32	91	89	87
ResNet-101	44.55	7.87	28	86	84	83

由表 1 可知，这些教师模型在含有补丁对抗样本的交通路标数据集上进行对抗训练后，其测试结果总体上有较高的准确率，这说明针对交通路标对抗样本的对抗训练是有效果的。但从表 1 中也看出，ResNet-101 网络的性能不如 ResNet-50 优良，因此，我们采用 AlexNet、VGG-19、ResNet-50 这 3 种教师模型来训练对抗样本，同时，将它们在对抗训练过程中的准确率变化情况可视化，如图 4 所示。



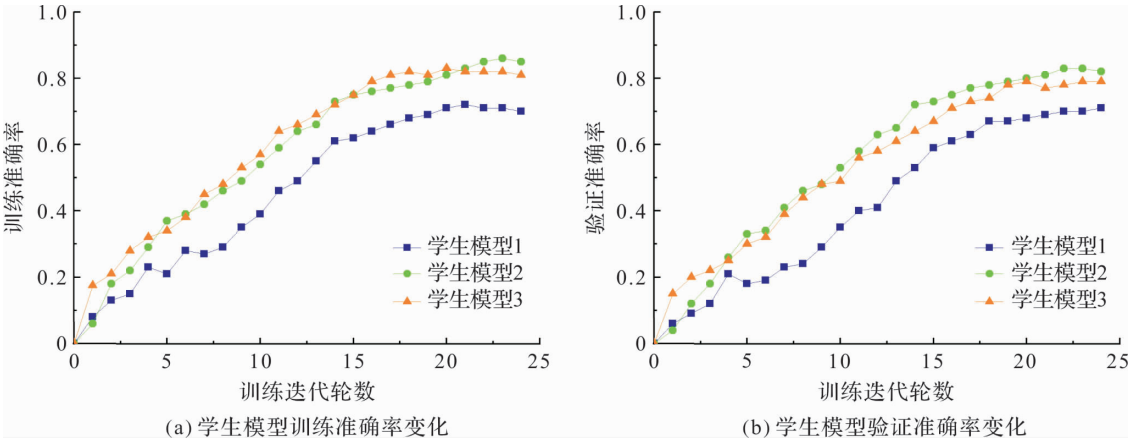
(a) 教师模型训练准确率变化　　(b) 教师模型验证准确率变化

图 4　教师模型对抗训练过程中的准确率变化情况

Fig. 4　Change of accuracy in teacher model adversarial training

3.4 知识蒸馏

有研究[23]表明，知识蒸馏时超参数 T 的变化会影响最终学生模型的性能，而根据一般经验，$T=20$ 时学生网络的性能最佳，也较为稳定。为减少试验次数，我们确定超参数 T 取值 20。AlexNet、VGG-19、ResNet-50 作为教师网络，以其预测输出作为软标签训练学生模型 1（SqueezeNet[24]）、学生模型 2（MobileNet[25]）、学生模型 3（ShuffleNet[26]），以完成知识的迁移。知识蒸馏过程中学生模型蒸馏过程中的准确率变化如图 5 所示。



(a) 学生模型训练准确率变化　　(b) 学生模型验证准确率变化

图 5 学生模型蒸馏过程中的准确率变化

Fig. 5 Change of accuracy in student model distillation process

从图 5 可以看出，随着蒸馏过程的进行，3 种学生模型的分类准确率稳步上升，最终分别达到 71%、82%、79%的验证准确率。学生模型的收敛速度比教师模型快很多，其原因是：1）学生模型采用学习软标签，使损失函数的数值有所减小；2）学生模型参数数量比教师模型有所减少。表 2 是 3 种学生模型在 G-TSDB 数据集上的性能对比，对抗训练后各学生模型测试准确率分别达到 69%、83%、78%。

表 2 3 种学生模型在 G-TSDB 数据集上的性能对比

Table 2 Performance comparison of three student models on G-TSDB data set

模型	参数量/MB	计算量/(10 亿次·s^{-1})	对抗训练前分类准确率/%		对抗训练后分类准确率/%	
			测试准确率	训练准确率	验证准确率	测试准确率
学生模型 1	1.25	0.35	19	71	71	69
学生模型 2	3.5	0.33	13	85	82	83
学生模型 3	3.5	0.31	8	82	79	78

分析表 1 和表 2 可知，经过知识蒸馏后各学生模型能达到与各教师模型相近的性能，同时参数数量和计算量比教师模型大为减少，3 种学生模型参数总量与计算总量也远小于单个 ResNet-50 教师模型，而更少的参数量和计算量意味着更快的运行速度。

3.5 模型投票决策

1.3 节中我们介绍过 2 种主要的模型投票表决方式。一种是少数服从多数的投票原则，融合各种模型的知识，在特征提取中可以保留不同模型的特征；另一种是平均各种模型的输出结果，即平均投票，是一种比较客观的方法，能综合各种模型的预测结果。基于这 2 种投票方式，我们在包含对抗样本的交通路标数据集上进行试验，结果见表 3。

表 3 基于 2 种投票方式的试验结果

Table 3 Experimental results based on two voting methods %

模型集成方式	最终分类准确率	
	验证准确率	测试准确率
第 1 种投票方式	81	82
第 2 种投票方式	91	85

通过比较表 2 和表 3，我们发现表 3 使用投票表决方式的学生模型的性能，均超过表 2 中单个学生模型的性能。由表 3 可知，第 2 种投票方式即平均投票法能给学生模型带来更高的性能提升，所以针对交通标志的补丁对抗攻击，我们选择第 2 种投票方式来综合学生模型的预测结果，以此防御交通路标的对抗攻击。

3.6　防御方法性能对比

我们的防御方法是多阶段的，而现有的防御方法大都是单阶段的，从中取几种经典对抗防御方法与我们的方法进行试验对比，结果见表 4。其中，评价指标为各模型在对抗样本攻击下的交通路标分类准确率，学生模型 1、2、3 分别为 SqueezeNet、MobileNet、ShuffleNet，试验数据集及其他条件同上。

表 4　经典对抗防御方法与我们的方法在对抗样本攻击下的性能对比

Table 4　Performance comparison between classical adversarial defense method and the proposed method under attack of adversarial samples　　%

对抗防御方法	分类准确率			
	学生模型 1	学生模型 2	学生模型 3	模型投票表决
FGSM[7]	53	59	61	
PGD[8]	67	72	70	
FreeAT[12]	76	71	82	
我们的方法	69	83	78	85

4　结　论

自动驾驶近几年发展遇到了很多挑战，安全问题始终是人们关注的焦点。对抗样本的存在对自动驾驶的安全构成了极大的威胁，本研究针对交通路标的对抗补丁攻击做相应防御，基于图像分类模型提出一种轻量级的多阶段对抗防御算法。在含有对抗样本的德国道路交通数据集 G-TSDB 上对我们的方法进行防御性能的试验，结果显示经我们的方法训练后的学生模型群体取得了 85％的分类准确率，单个学生模型也分别取得了 69％、83％、78％的分类准确率，具有较好的防御效果。可见，多阶段的防御方法能有效抵抗对抗样本的攻击，提高了模型的鲁棒性。未来，我们会进一步推进本方法的研究，将其实际应用到自动驾驶的车载系统上，以帮助自动驾驶克服障碍。

参考文献：

[1] LECUN Y, BOTTOU L. Gradient based learning applied to document recognition[J]. Proceedings of the IEEE, 1998,86(11):2279.

[2] GOODFELLOW I, BENGIO Y, COURVILLE A. Deep learning[M]. Cambridge, MA: The MIT Press Cambridge, 2016.

[3] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet classification with deep convolutional neural networks [C]//Conference and Workshop on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc, 2012:1110.

[4] NGUYEN A, YOSINSKI J, CLUNE J. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images[C]//Conference on Computer Vision and Pattern Recognition. Boston: IEEE,2015:430.

[5] KELLER C G, DANG T, FRITZ H, et al. Active pedestrian safety by automatic braking and evasive steering[J]. IEEE Transactions on Intelligent Transportation Systems,2011,12(4):1298.

[6] GOODFELLOW I, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//International Conference on Learning Representations. San Diego: Computer Science Bibliography,2015:1.

[7] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]// International Conference on Learning Representations. Vancouver: Computer Science Bibliography,2018:1.

[8] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial machine learning at scale[C]//International Conference on Learning Representations. Toulon: Computer Science Bibliography,2017:1.

[9] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world[EB/OL]. (2017-02-11) [2021-03-14]. https://arxiv.org/abs/1607.02533.

[10] TRAMER F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: attacks and defenses[EB/OL]. (2017-03-19)[2021-03-14]. https://arxiv.org/abs/1705.07204.

[11] XIE C, WANG J, ZHANG Z, et al. Mitigating adversarial effects through randomization[EB/OL]. (2018-02-28)[2021-03-14]. https://arxiv.org/abs/1711.01991.

[12] LIU X, CHENG M, ZHANG H, et al. Towards robust neural networks via random self-ensemble[C]// Proceedings of the 2018 European Conference on Computer Vision. Munich: Springer,2018:381.

[13] DHILLON G S, AZIZZADENESHELI K, LIPTON Z C, et al. Stochastic activation pruning for robust adversarial defense[EB/OL]. (2018-03-05)[2021-03-14]. https://arxiv.org/abs/1803.01442.

[14] 张嘉楠,赵镇东,宣晶,等.深度学习对抗样本的防御方法综述[J].信息安全与技术,2019,10(8):93.

[15] 陈晋音,邹健飞,苏蒙蒙,等.深度学习模型的中毒攻击与防御综述[J].信息安全学报,2020(4):14.

[16] 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.

[17] SHAFAHI A, NAJIBI M, GHIASI A, et al. Adversarial training for free! [C]//Conference and Workshop on Neural Information Processing Systems. Vancouver: Curran Associates Inc,2019:1.

[18] BA L J, CARUANA R. Do deep nets really need to be deep? [J]. Advances in Neural Information Processing Systems,2014,3(1):2654.

[19] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE,2016:770.

[20] ZHANG H, CISSE M, DAUPHIN Y N, et al. Mixup: beyond empirical risk minimization[C]//International Conference on Learning Representations. Vancouver: Computer Science Bibliography,2018:2.

[21] LIU A, LIU X, FAN J, et al. Perceptual-sensitive GAN for generating adversarial patches[J]. Proceedings of the AAAI Conference on Artificial Intelligence,2019,33(1):1028.

[22] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE,2009:248.

[23] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. Computer Science,2015,14(7):38.

[24] IANDOLA F N, HAN S, MOSKEWICZ M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size[EB/OL]. (2016-11-04)[2021-03-14]. https://arxiv.org/abs/1602.07360v1.

[25] HOWARD A G, ZHU M, CHEN B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[C]//Conference on Computer Vision and Pattern Recognition. Hawaii: IEEE,2017:432.

[26] ZHANG X, ZHOU X, LIN M, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices[C]//Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE,2018:116.

(上接第24页)

[13] MAHER D, SEYED R M, NASIM H, et al. Enhancing droplet deposition through in-situ precipitation[J]. Nature Communications,2016,8(30):12560.

[14] ZAMEER S, ALI J, VOHORA D, et al. Development, optimization and evaluation of chitosan nanoparticles of alendronate against Alzheimer's disease in intracerebroventricular streptozotocin model for brain delivery[J]. Journal of Drug Targeting,2020:1.

[15] NAGPAL K, SINGH S K, MISHRA D N. Chitosan nanoparticles: a promising system in novel drug delivery[J]. Chemical & Pharmaceutical Bulletin,2010,58(11):1423.

[16] HUANG J, DENG Y, REN J, et al. Novel in situ forming hydrogel based on xanthan and chitosan re-gelifying in liquids for local drug delivery[J]. Carbohydrate Polymers,2018,186(4):54.

[17] 王未,黄从建,张满成,等.我国区域性水体农药污染现状研究分析[J].环境保护科学,2013,39(5):5.

[18] 景亮亮,柴军发,高强,等.6种喷雾助剂对3种药剂表面张力与接触角的影响[J].浙江农业学报,2020,32(10):1.