基于掩膜自动编码器的对抗对比蒸馏算法

张 点[1] 董云卫[2]

[1](西北工业大学计算机学院 西安 710129)

[2](西北工业大学软件学院 西安 710129)

摘 要 随着人工智能的不断发展,神经网络对不同领域的任务都表现出了优异的性能.然而,对抗样本的存在对神经网络在安全相关领域中的应用提出了挑战.为了改善对抗训练耗时和对抗样本缺乏多样性的问题,本文提出一种使用改进掩膜自动编码器训练教师网络的对比蒸馏算法抵御对抗攻击.首先,为了减弱教师模型对图像全局特征的依赖,教师模型在改进的掩膜自动编码器中学习如何根据可见子块推理遮挡子块的特征.然后,为了减弱对抗干扰的影响,本文采用知识蒸馏和对比学习的方法提升目标模型的对抗鲁棒性,通过知识蒸馏转移教师模型的特征到学生模型减少模型对全局特征的依赖,通过对比学习提升学生模型对图像之间细节特征的识别能力.最后,本文采用标签信息对分类头进行调节确保识别准确率.在 ResNet50 和 WideResNet50 中进行的实验表明,CIFAR-10 中对抗准确率平均提升 11.50%;CIFAR-100 中对抗准确率平均提升 6.35%.实验结果证明基于掩膜自动编码器的对比蒸馏算法能够通过只生成一次对抗样本减弱对抗干扰的影响,并通过随机掩膜构建多样本视角提升样本多样性,增强神经网络对抗鲁棒性.

关键词 神经网络;对抗样本;对抗训练;掩膜自动编码器;对比蒸馏;对抗鲁棒性

中图法分类号 TP183 **DOI 号** 10.11897/SP.J.1016.2024.02274

An Adversarial Contrastive Distillation Algorithm Based on Masked Auto-Encoder

ZHANG Dian[1] DONG Yun-Wei[2]

[1](*School of Computer Science, Northwestern Polytechnical University, Xi'an* 710129)

[2](*School of Software, Northwestern Polytechnical University, Xi'an* 710129)

Abstract With the continuous development of artificial intelligence, neural networks have exhibited exceptional performance across various domains. However, the existence of adversarial samples poses a significant challenge to the application of neural networks in security-related fields. As research progresses, there is an increasing focus on the robustness of neural networks and their inherent performance. This paper aims to improve neural networks to enhance their adversarial robustness. Although adversarial training has shown great potential in improving adversarial robustness, it suffers from the drawback of long running times. This is primarily because it requires generating adversarial samples for the target model at each iteration step. To address the issues of time-consuming adversarial sample generation and lack of diversity in adversarial training, this paper proposes a contrastive distillation algorithm based on masked autoencoders (MAE) to enhance the adversarial robustness of neural networks. Due to the low information density in images, the loss of image pixels caused by masking can often be recovered using neural networks. Thus, masking-based methods are commonly employed to increase sample diversity and improve the feature learning capabilities of neural networks. Given that adversarial

收稿日期:2023-10-07;在线发布日期:2024-06-27.本课题得到国家重点研发计划(2022YFB4501801)和西北工业大学博士论文创新基金(CX2024075)资助.**张 点**,博士研究生,中国计算机学会(CCF)学生会员,主要研究方向为对比学习、神经网络可解释性、对抗攻击和对抗鲁棒性.E-mail:dianzhang@mail.nwpu.edu.cn.**董云卫**(通信作者),博士,教授,中国计算机学会(CCF)杰出会员,主要研究领域为人工智能系统测试、软件智能合成和智能代码补全等.E-mail:yunweidong@nwpu.edu.cn.

training methods often require considerable time to generate adversarial samples, this paper adopts masking methods to mitigate the time-consuming issue of continuously generating adversarial samples during adversarial training. Additionally, randomly occluding parts of the image can effectively enhance sample diversity, which helps create multi-view samples to address the problem of feature singularity in contrastive learning. Firstly, to reduce the teacher model's reliance on global image features, the teacher model learns in an improved masked autoencoder how to infer the features of obscured blocks based on visible sub-blocks. This method allows the teacher model to focus on learning how to reconstruct global features from limited visible parts, thereby enhancing its deep feature learning ability. Then, to mitigate the impact of adversarial interference, this paper employs knowledge distillation and contrastive learning methods to enhance the target model's adversarial robustness. Knowledge distillation reduces the target model's dependence on global features by transferring the knowledge from the teacher model, while contrastive learning enhances the model's ability to recognize fine-grained information among images by leveraging the diversity of the generated multi-view samples. Finally, label information is utilized to adjust the classification head to ensure recognition accuracy. By fine-tuning the classification head with label information, the model can maintain high accuracy in recognizing clean samples while improving its robustness against adversarial attacks. Experimental results conducted on ResNet50 and WideResNet50 demonstrate an average improvement of 11.50% in adversarial accuracy on CIFAR-10 and an average improvement of 6.35% on CIFAR-100. These results validate the effectiveness of the proposed contrastive distillation algorithm based on masked autoencoders. The algorithm attenuates the impact of adversarial interference by generating adversarial samples only once, enhances sample diversity through random masking, and improves the neural network's adversarial robustness.

Keywords　neural networks; adversarial examples; adversarial training; masked auto-encoder; adversarial distillation; adversarial robustness

1　引　言

神经网络已经在多个研究领域中取得了重要的突破，比如计算机视觉[1]、自动驾驶[2]和医疗领域[3]等. 尽管神经网络已经对人类社会产生了深远的影响，但神经网络脆弱性也对神经网络在安全相关领域中的应用发起了挑战. 神经网络的脆弱性是指：输入中添加人类难以察觉的微小干扰，会导致神经网络出现误判[4-5]，降低神经网络输出的可靠性. 例如，在自动驾驶中，神经网络对红绿灯或行人的误判可能引发严重的安全事故，威胁人身安全. 因此，研究如何提升神经网络的对抗鲁棒性具有重要意义.

目前，提升对抗鲁棒性的研究主要包含改进输入和神经网络两类方法. 改进输入的方法是指通过对神经网络的输入样本进行一定程度的干扰或抗干扰处理后，再对神经网络进行训练以得到更加健壮的网络模型. MagNet[6]中提出根据检测器和重整器网络处理对抗样本. 其中，检测器负责区分对抗样本，重整器负责对输入样本进行重构使其接近原始样本，确保输入样本的可靠性. APE-GAN[7]中提出使用生成器重构对抗样本，同时使用判别器确保干净样本和对抗样本的相似性. 改进神经网络的方法主要是通过对神经网络的模型结构或训练参数进行调整，提高神经网络对对抗样本的识别准确率，主要有对抗训练[8-9]、对比学习[10-11]和知识蒸馏[12-13]等算法. 为了增强神经网络在对抗样本上的识别准确率，对抗训练[14]不断将对抗样本注入到干净训练集中训练神经网络学习对抗样本的特征. 对比学习[15]关注于学习样本间特征表示，因此常常使用对抗样本构建多视角样本提升模型对抗鲁棒性. 知识蒸馏[16]采用教师网络学习样本特征，然后将特征蒸馏到学生模型中，通过梯度遮挡的方法提高神经网络鲁棒性. 为了提升模型的对抗鲁棒性，知识蒸馏也常常采用对抗样本进行训练.

随着研究的不断深入，神经网络鲁棒性越来越

关注于模型本身的性能,因此本文关注于改进神经网络的方法提升对抗鲁棒性.现有研究中的对抗训练被应用于多种类型方法中,但是对抗训练耗时的问题一直没有得到改善.对抗训练耗时的主要原因是在每一次训练过程中需要重新生成对抗样本.

由于图像中信息密度较低,掩膜导致的图像像素缺失常常可以采用神经网络进行恢复[17].因此,基于掩膜的方法[18-19]常常被用来增加样本多样性并提高神经网络的特征学习能力.现有的对抗训练方法往往需要花费大量的时间生成对抗样本,因此本文采用掩膜的方法改善对抗训练中不断生成对抗样本耗时的问题.而且通过随机遮挡图像中的部分子块能够有效提升样本的多样性,也有助于建立多视角样本克服对比学习中多视角样本特征单一的问题.

本文提出一种基于掩膜自动编码器(Masked Auto-Encoder,MAE)的对抗蒸馏算法用于提升神经网络鲁棒性.由于对抗干扰直接被添加到图像中,因此随机遮挡会减弱对抗干扰对神经网络识别性能的影响.但严重的外观变换也会影响神经网络的任务表现,而且遮挡后的图像也不足以使用标签进行描述,因此如何提升网络对遮挡后信息的推理能力也是本文解决的一个重要的问题.因此,本文的方法描述如下:首先,在 MAE 的编码器后添加特征融合模块,使用相关的损失函数确保所学习的特征与原始图像中的特征一致性,训练教师模型学习如何根据可见子块推理图像的全局特征.然后,本文分别使用不同的遮挡尺寸和遮挡率的图像作为对比蒸馏的输入,学生模型(目标模型的主干)采用对比蒸馏学习样本间特征,并从教师模型学习特征推理能力.最后,本文使用标签信息调整分类头确保分类准确率.实验结果表明,对于CIFAR-10,本文算法在 ResNet50 的对抗准确率平均提升 12.73%,在 WideResNet50 的对抗准确率平均提升6.67%;对于CIFAR-100,本文算法在 ResNet50 的对抗准确率平均提升 3.66%,在 WideResNet50 的对抗准确率平均提升 7.13%.综上,本文的主要贡献描述如下:

本文提出了一种改进的掩膜自动编码器,能够训练模型在编码阶段根据可见子块推理全局特征表示.此方法的主要创新之处在于使用整幅图像作为输入辅助 MAE 根据可见子块推理图像的全局特征信息.

本文提出了一种基于对比蒸馏的对抗鲁棒性提升算法.知识蒸馏用于将掩膜自动编码器训练的教师模型的特征推理能力转移到目标模型的主干.与此同时,将教师模型作为对比学习模型,本文采用与教师模型不同遮挡尺寸和遮挡率的图像作为学生模型的输入使学生模型通过对比学习获得图像中的细节特征.此方法的主要创新之处在于使用结合了对比学习和知识蒸馏的方法的同时减少了对抗样本生成消耗的时间.因此,本文的方法能够将教师模型同时用于对比学习和知识蒸馏中训练学生模型提高模型对抗鲁棒性.

2 相关工作

这一部分主要介绍了神经网络中对抗攻击、对抗鲁棒性、知识蒸馏、对比学习和掩膜图像建模的相关研究.

2.1 对抗攻击

神经网络模型可以用一个函数来表示:$y=F(x;\theta)$,对于一个结构为 F、参数为 θ 的神经网络模型,给定输入 $x\in\mathbb{R}^{w\times h\times c}$,产生对应的输出 $y\in\mathbb{R}^m$.神经网络模型的训练过程可以描述为:在给定的神经网络结构中,输入训练样本 $x\in\mathbb{R}^{w\times h\times c}$,神经网络产生对应的输出 $y\in\mathbb{R}^m$,使用损失函数对预测的输出和正确的输出计算损失值,并根据损失值更新神经网络参数.神经网络模型的测试过程可以描述为:对于神经网络 $y=F(x;\theta)$ 输入 $x\in\mathbb{R}^{w\times h\times c}$,产生对应的输出 $y\in\mathbb{R}^m$.本文采用图像样本作为神经网络输入,因此其输入空间为$[0,255]$,图像的宽、高和通道数分别为 w,h,c.样本对应的标签采用正整数 N_+ 表示,N 代表标签的种类数.

对抗攻击[5]是指根据目标模型对数据集产生干扰使目标模型产生误分类的算法,常常用于产生对抗干扰误导神经网络分类.对抗攻击 Adv 可以表示为式(1)和(2).

$$I_{adv}=Adv(I,F(\cdot;\theta)) \tag{1}$$

$$C(I_{adv})=F(I_{adv};\theta)\neq C^*(I_{adv}) \tag{2}$$

式(1)中 I_{adv} 代表对抗样本;$F(\cdot;\theta)$ 代表在原始训练集下的神经网络参数为 θ 的神经网络模型.式(2)中 $C(I_{adv})$ 代表对抗样本 I_{adv} 在神经网络中预测的分类;$F(I_{adv};\theta)$ 代表神经网络 F 对输入 I_{adv} 产生的输出;$C^*(I_{adv})$ 代表 I_{adv} 对应的正确分类标签.对抗干扰 η 定义为干扰样本和原始测试样本之间的差异,可以表示为式(3).

$$\eta=abs(I_{adv}-I) \tag{3}$$

对抗攻击一般分为黑盒攻击和白盒攻击,黑盒

攻击是指在不了解目标模型的结构和参数信息的情况下的对抗攻击；白盒攻击是指针对已知目标模型的结构和参数信息下的对抗攻击. 本文中只考虑在人眼的鲁棒性范围内的对抗样本(因此本文中最大干扰强度设置为 8/255,16/255). 本文为了测试鲁棒性提升的有效性分别在干净样本和对抗样本(白盒和黑盒攻击)中进行了对比实验.

2.2　对抗鲁棒性

对抗鲁棒性是指神经网络抵抗对抗干扰的能力. 尽管对抗干扰对人眼来说往往是不可感知的,但却能够严重影响神经网络的任务表现. 迄今为止,已经有了大量的研究关注于提升神经网络的对抗鲁棒性. 文献[5]中假设神经网络的脆弱性是由其线性特性引起,并提出了快速符号梯度法(Fast Gradient Step Method,FGSM). FGSM 是一种使用模型的梯度信息生成对抗干扰的白盒攻击算法,通过最大化模型的损失函数影响模型表现. 文献[20]中通过对对抗攻击的研究证明了迭代攻击比 FGSM 算法更强,因此提出了基于迭代方式的对抗攻击方法. 为了提升同一对抗样本在多个模型上的攻击效率,文献[21]中提出了一种结合动量(momentum)的对抗攻击算法用于提升对抗样本的可迁移性.

随着对抗攻击对目标模型的攻击成功率越来越高,对抗训练作为一种防御对抗攻击的方法被提出. 对抗训练方法假设导致神经网络脆弱性的根本原因是神经网络将此类样本认为是一种新类型的样本从而无法正确识别,因此对抗训练将对抗样本和干净样本同时用于训练神经网络提高对抗鲁棒性,防御对抗样本. 随着对样本多样性的需求,对抗训练的方法也被用于一些其它类型的任务[19]中用于扩展训练集多样性. 从理论上来说,对抗训练能够促使模型学习如何正确预测样本点附近以ϵ为半径的所有样本点的标签,因此在神经网络鲁棒性研究取得了一定的成功并展现了较大的应用潜力. 文献[20]在训练过程中将对抗样本注入训练集中,使用对抗样本增大神经网络损失值,再通过不断地训练减小网络损失,使网络适应对抗样本的同时增强神经网络在对抗样本上的识别表现. TRADES[22]中分析了对抗鲁棒性和标准准确率的关系,并提出 KL 散度结合对抗训练的方法使神经网络学习更加鲁棒的隐含特征空间抵抗对抗干扰的误导. 为了解决在训练阶段中固定攻击策略的对抗样本限制了模型鲁棒性的问题,LAS-AT[23]中提出了一种基于可学习参数的对抗训练通过增加对抗样本多样性提升模型鲁棒性.

尽管对抗训练在对抗鲁棒性的研究中表现出了巨大的应用潜力,但是由于对抗训练需要在每一个迭代步中不断地对目标模型生成对抗样本,因此对抗训练存在运行时间长的缺点. 由于在对抗训练中采用多种对抗攻击生成的对抗样本会影响神经网络收敛,但单一对抗攻击产生的对抗样本缺乏多样性会限制模型提高对抗鲁棒性,因此如何使用对抗样本也是一个需要解决的重要问题.

2.3　知识蒸馏

文献[24]中首次提出了知识蒸馏,即使用教师-学生的训练方法在保留网络表现的同时压缩模型尺寸. 因此,相比较于教师模型,学生模型的尺寸往往更小. 此后,大量方法[25-26]被提出用于增强目标模型(学生模型)的鲁棒性. 文献[26]中证明了教师网络在转移信息时忽略了重要的结构信息,并提出使用基于对比目标的网络框架转移神经网络之间的结构信息. 为了解决传统知识蒸馏中点对点的单一知识转移方式,文献[27]中提出了关系知识蒸馏从距离和角度两个方面转移数据样本之间的相互关系. 为了克服传统知识蒸馏中知识转移效率低和教师模型设计困难的问题,文献[28]中提出了一种自蒸馏算法,即采用自蒸馏学习框架的方法直接将模型自身的信息作为教师模型的特征进行知识迁移,而并不是使用传统的两步式的知识蒸馏方式,即先训练教师网络,然后转移信息到学生网络中的方法.

文献[29]中首次将知识蒸馏用于神经网络对抗鲁棒性研究,称为防御蒸馏,与知识蒸馏不同之处在于防御蒸馏中使用具有相同结构的学生模型和教师模型. 由于教师模型对学生模型进行了梯度遮挡的原因,防御蒸馏能够提升学生模型对于特定攻击的防御能力. 文献[30]中研究了如何将教师模型的对抗鲁棒性通过知识蒸馏转移到学生模型,并提出使用对抗鲁棒性蒸馏用于转移教师模型的鲁棒性到学生模型中. 文献[31]提出现有的方法很难将教师模型中的对抗鲁棒性转移到学生模型,因此提出了导向对抗对比学习采用隐含特征表示的方法将教师网络所学习特征的对抗鲁棒性转移到学生模型. 由于知识蒸馏的方法能够将教师模型中的信息转移到学生模型中,因此,本文采用知识蒸馏的方法将通过 MAE 学习的特征表示转移到学生模型中,降低学生模型对图像中全局像素信息的依赖减弱对抗干扰的影响.

2.4　对比学习

近年来,由于人为标注数据的高成本,自监督学

习也得到了广泛关注，尤其是对比学习方法在对抗防御中也得到了广泛的应用. 由于对比学习能够与对抗训练结合的特性，因此对抗训练也广泛应用于对比学习中. 文献[32]中提出采用 Jigsaw、Rotation 和 Selfie 自监督学习算法使神经网络主干学习数据样本中含有的深层特征，然后对神经网络使用对抗训练算法提升模型的对抗鲁棒性. 文献[33]中根据自监督对比学习算法 SimCLR 提出了神经网络鲁棒性提升框架 AdvCL. 首先，AdvCL 使用对抗样本和原始样本的高频部分创建无标签数据的鲁棒性和泛化性视角；然后，AdvCL 采用 KNN 算法生成样本对应的伪标签，作为 AdvCL 的伪监督刺激去训练神经网络模型. 实验结果证明 AdvCL 能够尽可能保证模型干净数据集的准确率并提升模型的对抗鲁棒性. 由于对抗攻击算法往往是根据数据对应的标签生成对抗样本，从而通过误导神经网络判断去影响模型对抗鲁棒性，因此文献[34]中提出了基于无标签数据的对抗攻击算法，并提出了一种自监督对比学习框架，该方法能够通过使用无标签数据提升神经网络的对抗鲁棒性. 文献[34]提出的算法可概述为：首先，生成样本级的对抗样本；然后，对比学习算法被用于学习干净样本和对应对抗样本之间的相似性，学习干净样本与其它样本之间的差异性；最后，将训练后的模型使用对应的标签信息对分类头进行微调得到提升鲁棒性后的神经网络. 文献[10]认为由于训练样本附近存在不光滑的特征空间，因此训练后的神经网络中会存在对抗脆弱性干扰模型任务，因此文中结合对比学习和对抗训练的方法提出了对抗对比学习，此方法能够在不同的扩展视角下对数据进行一致性训练. 文献[35]中提出基于改进的 BYOL 的对比学习算法用于训练鲁棒性的神经网络模型，通过在 BYOL 算法中的 online 网络模型中使用对抗样本对此视角下的样本进行干扰，训练神经网络对于对抗干扰的抵抗能力，从而达到增强神经网络对抗鲁棒性的目的.

尽管在对比学习中对抗训练的方法也有着巨大的作用，但是也将对抗训练运行时间长的不足也引入其中，而且对比学习也需要建立多种类型的样本来增加样本多样性提升神经网络鲁棒性.

2.5 掩膜图像建模

掩膜图像建模(Masking Image Modelling，MIM)逐渐成为视觉表示学习中的一个有趣的话题，而且能够有助于获得更加强有力的视觉任务表现. MIM 能够随机掩盖输入图像中的一部分然后根据对应的目标图像训练模型推演遮挡部分的图像. MAE[17] 中提出使用非对称编码器-解码器结构用于重建图像中缺失部分的像素，编码器中仅处理可见的图像子块集合，解码器中根据掩膜位置重建缺失部分的像素. CAE(Context Auto-Encoder)[18] 中引入特征对齐限制，促使编码器部分学习被遮挡后的像素，使解码器对重建的像素进行恢复. 由于图像中包含语义信息密度较低，因此可以使用 MIM 的方法对图像中缺少的部分进行学习，使神经网络具备推理图像缺少部分语义信息的能力. 在本文中，MAE 的使用主要考虑如下：对抗干扰尽管对人眼几乎不可见，但在图像中可以通过遮挡的方法进行规避，而且使用遮挡后的图像训练的目标模型能够减少对于图像全局信息的依赖提升鲁棒性.

3 提出的方法

本节主要介绍如何通过基于特征蒸馏的掩膜对抗对比学习得到鲁棒性模型. 首先，特征融合模块被用于重新设计 MAE 的编码器，训练教师模型根据可见子块推理图像全局特征的能力. 然后，MAE 的编码器作为教师模型，并使用对比蒸馏转移其推理能力到学生模型(目标模型的主干). 最后，标签信息用来对分类头进行调节确保模型对样本的识别表现.

3.1 掩膜自动编码器

改进后的 MAE 如图 1 所示，由编码器、特征学习和解码器三部分组成. 首先，教师网络能够根据可

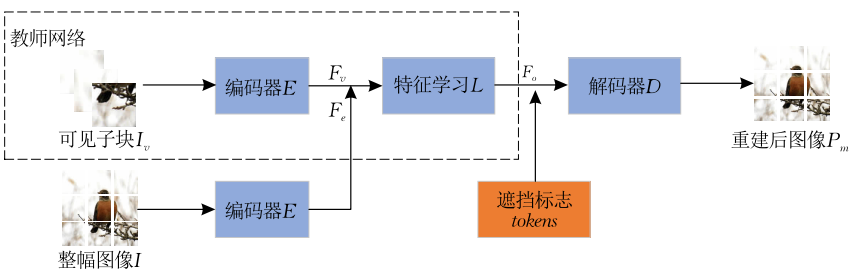

图 1 本文中的 MAE 结构

见子块、整幅图像和编码器学习如何根据可见子块推理图像的全局特征. 然后解码器将预测的特征表示映射为遮挡部分的像素确保全局特征的有效学习. 为了使 MAE 中的教师模型能够根据部分可见子块推理出全局特征，本文在 MAE 中添加了特征学习模块. 在图像输入到 MAE 之前，本文随机地将一幅图像 I 分为 2 个子块集合：可见子块集 I_v 和遮挡子块集 I_m.

编码器. 编码器 E 能够将 I_v 映射为隐含表示 L_v. 本文使用 ViT(Vision Transformer)作为 MAE 的编码器. 编码器使用线性投影和位置嵌入处理 I_v 得到 P_v，然后使用一系列 Transformer 块处理 P_v 得到特征表示 F_v. 在处理过程中，将所有遮挡的子块移除，编码器中仅处理所有可见子块集合(可见子块：遮挡子块=3∶17)和对应的标记，并不使用任何的遮挡标志(mask tokens) 和 I_m 相关信息.

特征学习. 在特征学习部分中，本文根据可见子块的特征表示 F_v 和整幅图像的特征表示 F_e 推理图像中包含遮挡子块的全局特征表示 F_o. F_e 的计算方式与 F_v 相似，即将整幅图像子块化(不进行遮挡)然后输入编码器，且处理可见子块和整幅图像采用相同的编码器. 本文使用 F_e 作为模板向量计算对应的权重，先试用池化操作将 F_e 与 F_v 转换为同一尺寸的特征表示，再使用归一化计算 F_v 对应的权重，可表示为 $\mathrm{Softmax}(Avgpool(F_e))\cdot F_v$. 最后，将权重与 F_v 进行融合作为特征学习模块的输出 F_o. 值得一提的是，特征学习中没有使用任何遮挡标志有关的信息.

解码器. 解码器 D 的输入包含特征表示 F_o 和遮挡标志. 遮挡标志 *tokens* 代表 I_m 对应图像子块的位置信息，MAE 将其转换为共享的、可学习的向量，代表将要预测的图像子块对应的位置信息. 通过对遮挡标志使用位置编码，解码器能够得到图像中预测位置的位置信息. 然后解码器 D 根据 F_o 和遮挡标志的映射重建遮挡部分的像素 P_m，即将遮挡标志与 F_o 进行整合作为 D 的输入. 在本文中，MAE 的 D 与 E 结构相似，D 由 8 层的基于自注意机制的 *transformer* 块组成，并使用线性层预测输出位置的像素信息. 解码器只接受 F_o 和遮挡标志的位置编码作为输入，并不直接使用可见的遮挡部分作为直接的输入信息.

损失函数. MAE 的损失函数由特征损失和遮挡部分像素的损失组成. 特征损失代表 F_o 和 F_e 之间的均方误差(Mean Squared Error，MSE)，这是为了确保编码器对于图像可见部分和整幅图像的特征表示尽可能的相似，保证教师网络学习到如何从局部可见子块推理全局特征. 本文使用特征损失使特征学习后的嵌入特征 F_o 包含遮挡部分的特征，减少学生模型对图像中全局特征的依赖.

遮挡部分像素的损失是指 I_m 和 P_m 之间的均方误差，这主要是通过重建的像素损失确保教师网络能够准确地学习到遮挡部分像素的相关特征. 损失函数的计算方法如式(4)所示. MAE 中使用的遮挡图像、重建后图像和原始图像如图 2 所示.

$$\mathcal{L}_{mae}=MSE(F_o,F_e)+MSE(I_m,P_m) \tag{4}$$

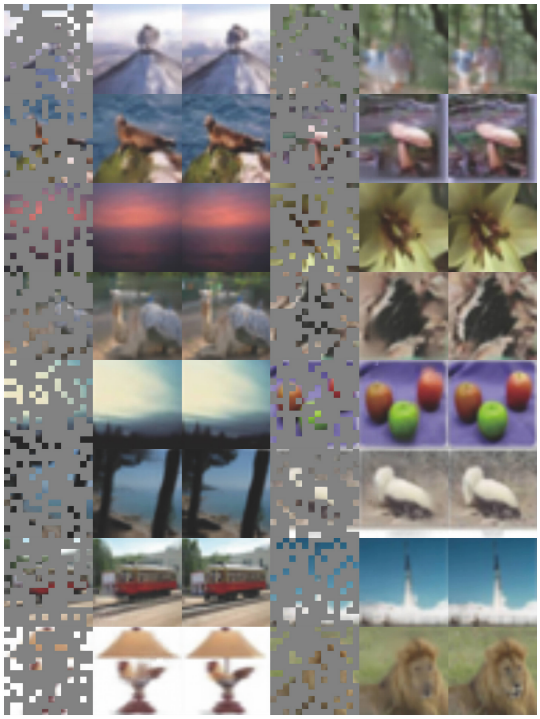

图 2　改进的 MAE 转换的图像(图中第 1 和 4 列代表随机遮挡图像，第 2 和 5 列代表重建后的图像，第 3 和 6 列代表原始样本)

MAE 的训练过程如算法 1 所示. 算法 1 中，$\mathcal{X}$ 只包含干净数据集中的样本 I_v，I_m，$tokens=SplitImage(I,patchsize,maskratio)$ 代表根据分块大小 *patchsize* 和遮挡率 *maskratio* 将图像 I 随机分割为可见子块 I_v、遮挡子块 I_m 和遮挡标签 *tokens*，训练过程中不直接使用 I_m，而是在解码器使用遮挡标签 *tokens*；其次，*MAE.encoder* 分别计算 I_v 和 I 对应的特征 F_v，F_e；然后，*MAE.featurelearning* 根据 F_v，F_e 计算全局特征 F_o；接下来，*MAE.decoder* 根据 F_o 和 *tokens* 计算遮挡部分的像素信息；最后根据损失函数更新 MAE 参数.

算法 1. 改进后 MAE 的训练.

输入：MAE：MIM 模型；$\mathcal{X}$：数据集；$epochs$：循环次数；$patchsize$：图像分块大小；$maskratio$：遮挡子块的比率

输出：MAE

FOR $i\in[0,epochs)$：

  FOR $I\in\mathcal{X}$：

    $I_v, I_m, tokens \leftarrow SplitImage(I, patchsize, maskratio)$

    $F_v, F_e \leftarrow MAE.encoder(I_v), MAE.encoder(I)$

    $F_o \leftarrow MAE.featurelearning(F_v, F_e)$

    $P_m \leftarrow MAE.decoder(F_o, tokens)$

    计算 $\mathcal{L}_{mae}$，并更新 MAE 参数

  END

END

3.2 对比知识蒸馏

本文中采用知识蒸馏的主要目的是将教师模型的信息压缩和蒸馏到学生模型，采用对比学习的目的是增强模型对于样本间特征的学习能力增强对抗鲁棒性.

本文中使用遮挡图像作为学生模型（标准模型的主干，即不包含分类头）的输入，通过知识蒸馏使学生模型从教师模型中学习从部分图像信息学习全局特征的能力，减少模型对输入图像中所有特征的依赖性.

本文采用 KL 损失将教师模型的特征表示蒸馏到学生模型中. 在对比蒸馏中，本文将对抗样本 I_{adv} 作为教师模型的输入，随机遮挡后干净样本 I_{clean} 作为学生模型的输入. 由于本文采用 MAE 的主干作为教师模型，因此 I_{adv} 输入教师网络后会随机选择一部分可见子块输入网络，而 I_{clean} 输入学生网络前将随机遮挡部分的像素全部设为 0，然后将整幅图像输入.

值得注意是，由于对于目标模型多次生成对抗样本太耗费时间而且由于针对标准模型生成的对抗样本也会干扰鲁棒性模型的表现，因此本文只是根据标准模型生成了一次对抗样本. 由于遮挡行为是随机的，因此每一次输入时都相当于创建了其它种类的对抗样本，极大的增强了输入样本的多样性. 在每次迭代过程中迭代使用同一组对抗样本，能够极大地减少算法的运行时间和耗费资源. 由于教师模型中包含特征融合模块，因此也可以将特征维度调整为与学生模型相同. 同时，本文使用对比学习的损失函数对于教师模型和学生模型的输出特征进行对比学习，确保能够学习到样本间的细节特征. 由于教师模型和学生模型都不包含分类头，因此其损失函数可以表示如式(5)，流程如图 3 所示.

$$\mathcal{L}_{robust}=KL(S_{out},T_{out})+\left\|\frac{S_{out}}{\|S_{out}\|}-\frac{T_{out}}{\|T_{out}\|}\right\|^2 \tag{5}$$

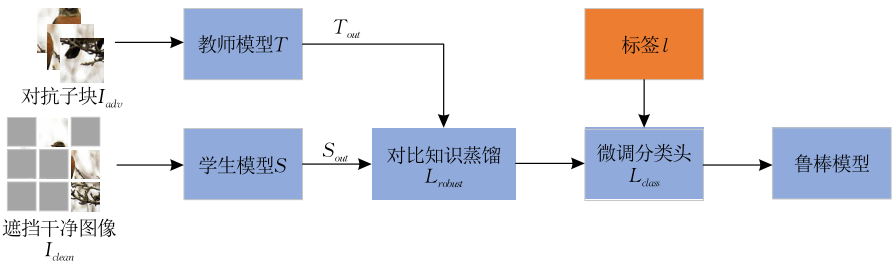

图 3 对比蒸馏算法流程

式(5)中 S_{out} 代表学生模型的输出特征，T_{out} 代表教师模型的输出特征. $KL(S_{out},T_{out})$ 表示教师模型和学生模型输出之间的 KL 损失，$\left\|\frac{S_{out}}{\|S_{out}\|}-\frac{T_{out}}{\|T_{out}\|}\right\|^2$ 表示教师模型和学生模型输出之间的对比损失. $T_{out}=S(mask\times I_{adv})$，$S_{out}=S(mask\times I_{clean})$，$mask$ 代表遮挡函数 $masked$ 生成的随机掩膜. 在迭代过程中对于教师模型输入和学生模型输入的遮挡率是不同的. 由于遮挡位置是随机的，因此每一次迭代中遮挡图像的不同位置也会增加训练样本的多样性. 在遮挡函数 $masked(imagesize, patchsize, masking\_ratio)$ 中，$imagesize$ 代表输入图像的尺寸（本文将所有图像设置为正方形），$patchsize$ 代表随机遮挡时将图像分割为不重叠的正方形子块的边长，$masking\_ratio$ 代表输入图像的遮挡比率. 在每一次对比知识蒸馏后，本文使用标签信息对模型的学习的特征进行验证的同时调节分类头确保模型的分类表现，即使用交叉熵损失计算分类损失 $\mathcal{L}_{class}$ 如式(6)所示.

$$\mathcal{L}_{class}=CE(l_{pre},l) \tag{6}$$

式(6)中 CE 代表交叉熵损失函数；l_{pre} 代表根据学生模型输出特征预测的样本标签；l 代表输入对应的正确标签；

本文中根据 MAE 教师模型训练鲁棒性模型的算法如算法 2 所示. 在算法 2 中，MAE 代表在算法 1 中已经训练完成的 MIM 模型，model 代表需要提升

鲁棒性的目标模型，$\mathcal{X}$与算法 1 中的数据集不同，$\mathcal{X}$包含对抗样本 I_{adv}、干净样本 I_{clean}和标签 l，这也是本文算法在算法 2 过程中大大减少迭代时间的主要原因. 本文只需要生成一次对抗干扰，因此可以将针对未提升鲁棒性的模型 model 生成的对抗样本迭代使用，极大程度的减少了对抗训练中需要不断生成对抗样本导致耗时的问题.

算法 2.　对比蒸馏算法.

输入：MAE：MIM 模型；model：目标模型；$\mathcal{X}$：数据集；*epochs*：循环次数；*patchsize*1：MAE 中图像分块大小；*maskratio*1：MAE 中遮挡子块的比率；*patchsize*2：学生模型中图像遮挡子块大小；*maskratio*2：学生模型中遮挡子块的比率

输出：model

FOR $i\in[0,epochs)$：

　FOR $I_{adv}, I_{clean}, l\in\mathcal{X}$：

　　$I_v, I_m, tokens \leftarrow SplitImage(I, patchsize1, maskratio1)$

　　$F_o \leftarrow MAE.featureleaning(MAE.encoder(I_v))$

　　$I_{mask} \leftarrow masked(I_{clean}, patchsize2, maskratio2)$

　　$F_s \leftarrow model.backbone(I_{mask})$

　　$l_{pre} \leftarrow model.classifier(F_s)$

　　计算$\mathcal{L}_{robust}$，$\mathcal{L}_{class}$，并更新 model 的参数

　END

END

算法 2 运行过程如下：首先，将对抗样本 I_{adv}分割为 I_v、I_m，采用与算法 1 相同的函数；其次，将 I_v输入 MAE 进行计算得到对应的全局图像特征；然后，对 I_{clean}使用与 I_{adv}不同的遮挡尺寸和遮挡率进行随机遮挡得到 I_{mask}（并没有分割图像，只是将图像中遮挡后的部分像素设置为 0）；接下来，将遮挡后 I_{clean}输入学生模型 *model.backbone* 中，得到学生模型输出特征 F_s，并根据 F_s预测 I_{mask}的标签 l_{pre}；最后，计算$\mathcal{L}_{robust}$，$\mathcal{L}_{class}$，并更新 model 的参数.

因此，本文的鲁棒性提升方法可以总结为如下两部分：第一部分，使神经网络学习根据局部信息推理全局信息的能力，在这一步中本文采用了 MAE. 第二部分，增加神经网络的特征表示能力和对抗鲁棒性，这主要通过对比蒸馏学习实现. 因此，本文的方法从理论上来说能够提升神经网络的对抗鲁棒性.

4　实验结果

本节对文中提出的提升神经网络鲁棒性模型的方法的表现进行了评估. 4.1 节中描述了实验设置；4.2 节中描述了本文方法与其它方法的对比实验结果；4.3 节中描述了消融实验结果，证明了本文算法的每一部分的有效性；4.4 节中描述了运行时间的实验结果.

4.1　实验设置

数据集. 为了验证本文算法对多种数据集的有效性，本文分别在 CIFAR-10[36] 和 CIFAR-100 数据集中进行了相关实验. CIFAR-10 由 60 000 幅 32×32 彩色图像组成，其中包含 10 个类别，每个类别包含 5000 张训练图像和 1000 张测试图像. CIFAR-100 由 60 000 幅 32×32 彩色图像组成，其中包含 100 个类别，每个类别包含 500 幅训练图像、100 个测试样本. 在本文中所有的图像的默认大小设置为 224×224（本文中图像单位均为像素点）.

目标模型. 为了验证本文算法对不同模型的有效性，本文分别在 ResNet-50[37] 和 WideResNet50[38] 模型中验证了本文的算法，ResNet-50 和 WideResNet50 都是卷积神经网络分类模型，本文根据预训练的参数训练目标模型得到标准模型，并采用本文提出的方法对标准模型进行鲁棒性提升，得到了鲁棒模型. 在模型的训练过程中，本文采用随机梯度下降（Stochastic Gradient Descent，SGD）作为优化器，设置学习率为 0.003，动量因子（momentum factors）设置为 0.9，权重缩减（weight decay）设置为 0.0005. 标准模型的训练迭代了 20 次，鲁棒性模型的训练迭代了 10 次.

模型训练的细节. 本文采用 ViT（*Vision Transformer*）[39] 设计 MAE 的编码器和解码器结构. ViT 能够从一幅分辨率为 $N\times N$ 的图像中提取一个不重叠图像子块序列. 然后，ViT 应用线性转换层提取子块的标记（patch tokens），并将子块标记放入到可学习的位置嵌入特征中. 一个可学习的[CLS]标记被添加到特征序列中，此标记关注于从子块的完整序列中聚集相关特征信息. 标记序列也被输入到 *Transformer* 层中. *Transformer* 层是 ViT 的基础结构，由自注意力机制（*Self-Attention Mechanism*）和带跳跃结构的全连接层组成. 自注意力结构将注意力机制应用到输入中去学习输入之间相互的特征关系. ViT 将与[CLS]标记相关的特征表示作为输出. MAE 将 ViT 提取的子块序列分为可见子块和遮挡子块，并产生对应的子块标记（算法 1 中为了方便表示并没有描述可见子块标记），然后在不直接使用遮挡子块像素的条件下学习遮挡子块的像素.

因此,本文中基于 ViT 结构,编码器采用 12 层的 *Transformer* 结构,解码器采用 8 层的 *Transformer* 结构.特征学习模块将整幅图像的特征使用平均池化调整特征维度与可见子块特征相同,并保留其中的特征,然后将 Softmax 后特征与可见子块特征相乘,最后将此特征与可见子块特征融合使得模型能够在解码器之前根据部分可见子块学习整幅图像的特征.

在训练 MAE 过程中,本文在 CIFAR-10 和 CIFAR-100 中先对只包含编码器和解码器的 MAE 模型采用重构像素损失训练 100 次.其次,将 MAE 的参数加载到包含特征学习模块的模型中,使用 $\mathcal{L}_{mae}$ 训练 20 次后提取教师模型.然后,本文采用不同的遮盖率和子块大小使用 KL 散度损失和对比损失作为损失函数训练学生模型迭代 10 次得到鲁棒性模型.最后,使用标签信息训练模型的分类头.对于教师模型设置遮挡率为 0.75,子块大小设置为 16;对于学生模型设置遮挡率为 0.7,子块大小设置为 8.

攻击算法.为了验证本文算法对不同对抗攻击的防御效果,FGSM[5]、PGD[20]、AutoAttack[40] 和 Jitter[41] 算法被用来作为对抗攻击算法对不同方法得到的目标模型的鲁棒性进行测试.

考虑到神经网络在高维空间中的线性特性,FGSM 的主要思想是通过寻找神经网络梯度变化最大的方向,并根据此方向的反方向产生对抗干扰添加到干净样本中误导原始模型.FGSM 攻击效率较高,但是判断对抗干扰的方法比较简单.PGD 与 FGSM 相似,也使用寻找神经网络的最大梯度方向,并根据最大梯度方向的反方向产生对抗干扰欺骗神经网络.PGD 与 FGSM 之间的差异是,PGD 限制了对抗干扰到一个特定的范围中,即将对抗样本投影到干净训练集附近.AutoAttack 是一个集成参数的对抗攻击算法,集合了多种攻击方式包括 $APGD_{CE}$、$APGD^{T}_{DLR}$、FAB 和平方攻击(Square Attack).$APGD$ 是一种较快的白盒对抗攻击,在每次迭代中只要求一次正向和反向计算.FAB 能够最小化对于误导分类产生的对抗干扰.平方攻击是一种基于分数的黑盒攻击方式,能够在不计算梯度信息的情况下完成特定范数要求的对抗干扰.AutoAttack 是一种结合了多种方案的对抗攻击方法.Jitter 中设计了新的损失函数用于提升对抗攻击准确率,然后通过将模型输出的 logits 正则化到一个特定的值范围中将尺度不变性引入到损失函数中.Jitter 能够避免由具有低可信度预测值或不规则的大范围输出 logits 的神经网络产生的梯度模糊问题.

在训练过程中,PGD 算法被用来生成标准模型的对抗样本时本文设置迭代步数为 40,最大干扰设置为 8/255.在评估过程中,对抗攻击算法攻击标准模型生成对抗样本测试模型的鲁棒性.对于每一种对抗攻击算法,本文分别采用 8/255 和 16/255 作为最大干扰.对于 AutoAttack 算法,文中采用 l_2 攻击.对抗攻击中未特别指明的参数均采用默认参数.torchattack(python 包)用来产生本文中所需要的对抗样本.

比较方法.文中使用不同种类的对抗防御算法与本文提出的对比对抗蒸馏算法进行比较,比较的算法包括对抗训练、蒸馏算法和对比学习等.(1) AT(Adversarial Training)[20],AT 通过使用自然鞍点(min-max)公式将攻击和防御方法纳入一个共同的理论框架训练神经网络抵抗对抗攻击.在训练的每一次迭代过程中,AT 在内函数中生成对抗样本最大化神经网络损失,同时在外函数中训练神经网络最小化对干净和对抗样本的损失提升神经网络的鲁棒性;(2) TRADES[22],TRADES 中提出分解神经网络在对抗样本中的损失为分类损失和边界损失,并使用分类校准损失的理论提供了分类器的差分上界.TRADES 通过采用 KL 散度和对抗训练的方法提升神经网络在对抗样本上的识别准确率,也取得了对抗鲁棒性和标准准确率之间的平衡;(3) IAD(Introspective Adversarial Distillation)[42],IAD 中认为知识蒸馏过程中教师网络不可靠和对抗蒸馏不生效的主要原因是:采用教师网络中生成的对抗样本训练神经网络,并对于教师网络也能够查询学生网络对抗样本的要求太高了.因此,IAD 提出了内省式对抗蒸馏,即学生模型根据教师模型对干净数据和对应的对抗样本的表现决定是否相信教师模型;(4) MTARD[43](Multi-Teacher Adversarial Robustness Distillation)使用多个教师模型,包含对抗训练和标准训练得到的模型,在对抗训练中对较小的学生模型进行知识蒸馏;(5) AdaAD[44](Adaptive Adversarial Distillation),其中将教师模型纳入知识优化过程与学生模型交互,使教师模型能够自适应的搜索内部结果;(6) RoCL(Robust Contrastive Learning)[34].RoCL 中提出了一种使用无标签数据的对抗防御方法.RoCL 中提出了一种

自监督对比学习框架在不依赖数据标签的情况下采用对抗训练的方式抵抗对抗干扰，关注于通过最大化在随机扩增的数据样本和其对应的样本级对抗干扰之间的相似性提升神经网络的对抗鲁棒性.

4.2　对比实验

首先，本文在表 1 和表 2 中展示了不同的对抗攻击和对抗防御算法在 CIFAR10 数据集上对于 ResNet50 和 WideResNet50 的实验结果. 8/255 和 16/255 分别代表对抗攻击算法的最大干扰强度. 标准准确率代表模型在干净数据集上的准确率. Baseline 代表在标准模型上的干净样本和对抗样本的实验结果. 设置 Baseline 的原因有以下两点：(1) 验证对抗攻击算法的有效性；(2) 确保鲁棒性提升方法的有效性.

表 1　CIFAR-10 中 ResNet50 的实验结果　　(单位：%)

对比方法	标准准确率	PGD		FGSM		AutoAttack		Jitter	
		8/255	16/255	8/255	16/255	8/255	16/255	8/255	16/255
Baseline	96.95	59.58	0.00	60.83	20.62	57.31	1.35	61.25	1.25
AT	76.31	47.50	38.34	43.25	31.91	73.31	57.64	67.72	51.98
TRADES	72.29	67.53	31.35	67.31	33.32	74.85	**65.93**	63.34	57.34
IAD	76.61	61.21	**62.54**	59.93	53.04	68.47	61.58	61.94	54.51
MTARD	84.97	58.78	58.20	64.07	48.99	64.54	55.41	58.17	57.21
AdaAD	83.24	58.63	58.34	61.13	50.39	64.34	58.32	59.34	58.34
RoCL	86.34	51.41	29.83	**72.61**	**63.31**	76.67	62.31	69.72	62.24
Ours	**94.57**	**73.62**	55.44	71.81	46.44	**78.96**	64.53	**73.33**	**73.96**

表 2　CIFAR-10 中 WideResNet50 的实验结果　　(单位：%)

对比方法	标准准确率	PGD		FGSM		AutoAttack		Jitter	
		8/255	16/255	8/255	16/255	8/255	16/255	8/255	16/255
Baseline	96.75	43.96	0.00	41.25	16.25	41.63	0.35	44.17	0.21
AT	80.35	45.50	30.34	47.63	35.31	72.98	58.78	65.34	53.21
TRADES	79.67	63.94	51.81	65.31	30.64	74.31	**62.94**	66.72	56.39
IAD	82.38	53.09	53.17	63.14	48.31	66.20	57.86	57.86	57.86
MTARD	79.92	61.26	62.09	59.45	55.50	69.26	59.55	70.76	61.98
AdaAD	80.83	62.23	**63.61**	56.17	52.50	68.14	58.66	69.73	60.93
RoCL	91.34	49.66	26.39	73.01	**60.31**	74.06	59.60	67.34	59.42
Ours	**93.74**	**71.38**	**71.62**	**73.56**	54.62	**76.88**	62.50	**72.38**	**76.00**

在表 1 中，可以观察到，本文方法对于 PGD，Jitter 在干扰强度为 8/255 和 16/255 都表现出了最好的识别准确率. 对于 FGSM 和 AutoAttack 在干扰强度为 8/255 时表现最好. 当 FGSM 和 AutoAttack 设置为 16/255 时本文方法的表现分别低于 RoCL 和 TRADES 方法. 可以观察到，不同的鲁棒性方法对于标准数据集中的识别率均有不同程度的下降，但是能够增加模型在不同对抗干扰的识别准确率.

在表 2 中，可以观察到，本文方法对于 Jitter 在干扰强度为 8/255 和 16/255 时表现出了最好的识别准确率. 对于 PGD 和 AutoAttack，本文方法在干扰强度为 8/255 时表现最好，当干扰强度为 16/255 时 IAD 和 TRADES 具有更高的准确率. 对于 FGSM，RoCL 算法在干扰强度为 8/255 和 16/255 时表现出了最高的表现. 这也证明了通过减小模型对图像中全局信息的依赖是一种有效的抵御对抗干扰的方法.

表 1 和表 2 中的实验结果表明，标准模型在干净数据集上的准确率更高，但是标准模型易被攻击算法攻击，从而导致对对抗样本的识别率急剧下降. CIFAR-10 的实验结果表明，本文的方法不仅在对抗样本中整体表现更好，而且能够极大程度地保持目标模型对标准数据集的识别准确率.

在表 3 中可以观察到，本文算法对于 FGSM、AutoAttack 和 Jitter 算法在攻击强度为 8/255 和 16/255 时都有最好的识别标准率. 对于 PGD 攻击算法，本文算法在攻击强度为 8/255 时具有更好的表现，当攻击强度为 16/255 时 IAD 算法具有更好的表现.

在表 4 中可以观察到，当攻击强度设置为 8/255 和 16/255 时，本文算法对于 PGD，AutoAttack 和 Jitter 算法具有更好的表现. 对于 FGSM 算法，本文算法在干扰强度设置为 8/255 时具有更好的表现，当攻击强度设置为 16/255 时，AdaAD 算法具有更好的表现.

表 3 CIFAR-100 中 ResNet50 的实验结果 (单位:%)

对比方法	标准准确率	PGD		FGSM		AutoAttack		Jitter	
		8/255	16/255	8/255	16/255	8/255	16/255	8/255	16/255
Baseline	82.44	20.21	0.00	19.38	5.00	18.31	5.13	17.50	0.00
AT	56.21	35.04	27.18	33.94	31.14	33.55	31.55	34.16	32.21
TRADES	56.61	30.00	26.21	35.31	34.13	33.46	31.26	31.27	30.17
IAD	57.08	32.31	29.91	36.94	37.47	35.84	34.43	32.31	33.32
MTARD	59.39	33.97	30.19	36.73	36.69	35.69	32.23	33.16	31.12
AdaAD	56.92	37.13	29.92	38.99	38.26	35.78	35.86	36.22	31.44
RoCL	64.21	36.31	**30.31**	38.17	36.17	36.61	35.50	24.36	22.13
Ours	**77.38**	**38.96**	20.00	**41.88**	**41.67**	**38.23**	**36.21**	**36.88**	**41.46**

表 4 CIFAR-100 中 WideResNet50 的实验结果 (单位:%)

对比方法	标准准确率	PGD		FGSM		AutoAttack		Jitter	
		8/255	16/255	8/255	16/255	8/255	16/255	8/255	16/255
Baseline	85.72	25.83	0.00	24.79	4.38	21.13	3.74	25.42	0.00
AT	58.31	37.98	29.58	36.74	33.05	35.61	33.36	36.18	33.74
TRADES	59.42	32.05	27.57	37.91	37.05	35.49	33.08	33.12	31.37
IAD	60.13	33.81	32.56	39.20	40.34	38.69	35.82	34.61	35.82
MTARD	69.76	36.41	34.05	40.39	40.97	42.43	40.67	38.43	36.19
AdaAD	66.79	34.87	34.55	41.87	**41.70**	44.21	41.26	36.72	35.51
RoCL	65.31	39.06	32.67	40.26	37.96	39.57	37.29	25.75	24.15
Ours	**77.33**	**43.80**	**45.42**	**44.58**	27.29	**46.88**	**45.42**	**43.32**	**48.12**

根据表 3 和表 4 的结果,可以得出与表 1 和表 2 相似的实验结论,即不同的鲁棒性的提升算法均会降低标准准确率,但同时又能够提升模型面对对抗样本时的识别准确率. 在 CIFAR-100 中,本文的方法在 ResNet50 和 WideResNet50 中具有更好的平均表现.

根据表 1 到表 4 的实验结果,可以观察到,本文的方法能够通过训练目标模型使其具备根据部分可见信息对图像中全局信息的推理能力,减弱目标模型对图像中全局信息的依赖. 本文的方法能够结合使用掩膜自动编码器结合对抗训练、对比蒸馏的方法,这是本文方法能够比对抗训练、防御蒸馏和对比学习的方法表现良好的主要原因. 因此,本文的方法能够在极大程度的保持标准准确率的情况下,保持对对抗样本的识别性能,即提高了目标模型的对抗鲁棒性.

4.3 消融实验

为了说明实验中参数的选取和本文对抗防御方法中各组件的必要性,本文进行了如下消融实验. 首先,本文在CIFAR-10中使用ResNet50进行了实验:

(1)为了说明在 MAE 训练中可见子块:遮挡子块=3:17 的合理性,即遮挡率为 3/20=0.15,本文设置图像子块大小为 8,遮挡率分别设置为 0.75、0.80、0.85、0.90、0.95 使用标准准确率作为参考标准进行选取,实验结果如表 5 所示. 表 5 中的实验结果表明,当遮挡率为 0.85 时,ResNet50 能够在 CIFAR-10 数据集中表现出最好的识别准确率,这也是本文选取可见子块:遮挡子块=3:17 的原因.

表 5 不同遮挡率对标准准确率的影响

遮挡率	标准准确率
0.75	0.8735
0.80	0.8967
0.85	**0.9363**
0.90	0.9124
0.95	0.9034

(2)为了说明对比蒸馏中图像子块大小和遮挡率选取的合理性,本文分别设置图像子块大小为 2、4、8、16,遮挡率为 0.65、0.7、0.75、0.8 以标准准确率和 PGD 对抗准确率作为参考进行比较,实验结果如表 6 所示. 表 6 中的实验结果表明,在对比蒸馏学习中,遮挡率设置为 0.7 时 ResNet50 在标准准确率和对抗准确率中均有较好的表现.

当子块尺寸设置为 4 时,ResNet50 在标准准确率上有较好的表现;子块尺寸设置为 8 时,在对抗准确率上有更好的表现. 由于本文方法是以对抗防御为目标,因此本文选取子块尺寸为 8,遮挡率为 0.7 进行对比蒸馏实验. 表 6 中的实验结果证明 MACD 中采用随机遮挡能够提升样本(对抗样本和干净样本)多样性,因此能够提高神经网络的对抗鲁棒性. 但当子块尺寸等于在 MAE 中使用的子块大小 16 时,神经网络使用蒸馏从教师模型中继承的推理能力会变弱,这表明知识蒸馏会损失教师模型中的信息,因此如何选择教师模型是一个至关重要的问题. 遮挡率为 0.7 时具有较好的表现也说明了从 MAE 中继承的遮挡推理能力有限.

表 6　不同子块尺寸和遮挡率对标准准确率和对抗准确率的影响

	子块尺寸	0.65	0.7	0.75	0.8
标准准确率	2	89.39	93.58	91.63	94.20
	4	88.49	**94.60**	84.24	89.56
	8	87.75	93.77	91.74	88.25
	16	87.17	91.77	90.19	89.03
对抗准确率	2	58.34	54.62	53.03	58.00
	4	63.25	69.62	57.44	61.69
	8	60.31	**71.47**	68.00	64.88
	16	56.06	64.19	61.28	57.09

综上所述，本文采用改进后的 MAE 作为教师模型能够提升模型对遮挡后样本的识别能力，而且在 MACD 中使用随机遮挡能够提升样本多样性，从而实现神经网络对抗鲁棒性的提升.

然后，本文在 CIFAR-10 中使用 ResNet50 和 WideResNet50 中进行了实验，为了说明训练集中引入对抗样本能够提升神经网络鲁棒性，实验结果如表 7 所示.

表 7　对抗样本对神经网络鲁棒性的影响

是否添加对抗样本	模型	标准准确率	PGD		Jitter	
			8/255	16/255	8/255	16/255
否	ResNet50	94.35	0.12	0.00	0.31	0.00
是	ResNet50	93.74	71.38	71.62	72.38	76.00
否	WideResNet50	95.61	1.25	0.00	1.47	0.00
是	WideResNet50	94.57	73.62	55.44	73.33	73.96

表 7 中的实验结果表明，在训练集中不引入对抗样本进行能够提升标准准确率，但是对抗准确率并无较大影响，因此本文在训练集中引入对抗样本能够提升神经网络鲁棒性.

最后，本文对 ResNet50 和 WideResNet50 在 CIFAR-10 和 CIFAR100 中进行了实验. 为了验证采用 MAE 训练教师网络进行知识蒸馏的方法的有效性，本文只采用不同的遮挡率遮挡对抗样本和干净样本进行对抗对比学习，实验结果如表 8 所示.

表 8　不遮挡样本的消融实验结果　　（单位：%）

数据集	模型	标准准确率	PGD		FGSM		Jitter	
			8/255	16/255	8/255	16/255	8/255	16/255
CIFAR-10	ResNet50	95.22	49.38	0.00	43.33	18.12	48.33	0.42
CIFAR-10	WideResNet50	95.88	57.71	0.00	56.25	19.79	58.33	0.42
CIFAR-100	ResNet50	82.43	23.33	0.00	20.00	6.88	20.83	0.00
CIFAR-100	WideResNet50	85.71	24.69	0.00	25.62	5.31	26.15	0.00

表 8 中实验结果表明，不遮挡样本的模型能够得到更高的标准准确率，但是仍然略低于干净准确率. 当干扰强度设置为 8/255 时，不遮挡样本的模型取得的鲁棒性表现则略高于标准模型下的表现. 当干扰强度设置为 16/255 时，则不遮挡样本的模型基本不具备正确识别对抗样本的能力. 这说明本文提出的遮挡样本的方法尽管略有降低了模型对于标准样本的识别率(CIFAR-10：2%～3%，CIFAR-100：5%～8%)，但是能够有效提升模型的对抗鲁棒性，对于对抗样本取得更好的识别准确率.

表 9 中实验结果表明，对比学习后的模型得到的准确率基本与标准模型的识别率相同，但是仍然不能有效抵抗对抗攻击. 当干扰强度设置为 8/255 时，CIFAR-10 的模型识别率基本能维持在 50%左右，CIFAR-100 的模型识别率基本维持在 20%左右，这说明遮挡图像只采用对比学习的方法不能够有效地学习样本中深层特征抵抗对抗干扰. 当干扰强度设置为 16/255 时，所有模型基本上都不能正确识别样本. 这说明采用 MAE 蒸馏的方法能够使学生模型具备从遮挡后图像中提取相关特征的能力，减少对模型全局特征的依赖，从而提升对抗鲁棒性.

表 9 对比学习实验结果 (单位:%)

数据集	模型	标准准确率	PGD		FGSM		Jitter	
			8/255	16/255	8/255	16/255	8/255	16/255
CIFAR-10	ResNet50	96.75	44.90	0.00	45.00	18.96	46.88	0.10
CIFAR-10	WideResNet50	96.95	58.33	0.00	57.60	21.35	58.23	0.21
CIFAR-100	ResNet50	82.43	20.42	0.00	17.08	4.79	23.75	0.00
CIFAR-100	WideResNet50	85.71	25.83	0.00	26.25	5.10	24.58	0.00

因此,本文算法中的 MAE 蒸馏和遮挡图像对比学习都是提升目标模型对抗鲁棒性不可缺少的一部分,而且对于提升目标模型鲁棒性具有重要的作用.

4.4 运行时间

为了验证算法的运行效率,本文对不同的鲁棒性提升算法的运行时间进行比较. 本文的算法由 MAE 和对比蒸馏两部分组成,本文将 MAE 和对比蒸馏算法进行了解耦,因此在如下的运行时间比较中本文算法的运行时间并没有添加 MAE 的运行时间,MAE 运行时间为 566 min. 对于 IAD 算法由于要预训练教师模型,本文也没有计算预训练教师模型的时间. 根据在源代码中提供的算法,AT 算法循环了 200 次,平均每次迭代 25.63 min; TRADES 算法循环了 76 次, 平均每次迭代 36.41 min; IAD 算法循环了 200 次,平均每次迭代 3.1 min; RoCL 算法循环了 150 次,平均每次迭代 38.16 min. 本文算法在对比蒸馏阶段训练 10 次,平均每次 2.72 min. 根据上述结果可以看出,AT、TRADES 和 RoCL 中在对抗训练的过程中生成对抗样本是其在每一次迭代中运行时间长的主要原因.

本文训练的 MAE 教师模型可以多次复用于具有相同学生模型输出维度的模型. 本文在每次迭代中运行时间较小的原因是本文只生成了一次对抗样本,通过随机掩膜的方法增加对抗样本多样性,减少了在训练过程中生成对抗样本所需要耗费的时间.

综上所述,本文提出的方法能够将 MAE 模型训练与对比蒸馏算法进行解耦,用于提升神经网络的对抗鲁棒性. 根据上述实验结果可以得出,对于 CIFAR-10,本文的算法在 ResNet50 的对抗准确率平均提升 13.35%,在 WideResNet50 的对抗准确率平均提升 9.64%;对于 CIFAR-100,本文算法在 ResNet50 的对抗准确率平均提升 4.33%,在 WideResNet50 的对抗准确率平均提升 8.37%. 因此,本文算法能够有效提升神经网络的对抗鲁棒性.

5 结 论

本文提出了一种基于 MAE 的对抗蒸馏对比学习算法用于提升神经网络对抗鲁棒性. 首先,在传统的 MAE 模型中添加了特征提取模块用于促使教师模型根据部分可见图像学习图像中的全局特征信息. 其次,提取 MAE 的编码器和特征提取模块作为教师模型. 然后,对学生模型的输入进行遮挡,并采用对比蒸馏的方法学习样本特征,提升目标模型鲁棒性. 最后,采用样本标签对模型的分类头进行调整,确保样本识别准确率. 在实验部分中,本文方法与其它鲁棒性提升方法进行对比实验,证明本文算法具有较好的对抗样本防御性能;消融实验证明了本文方法中每一部分都能够对神经网络对抗鲁棒性提升具有积极意义;运行时间的比较证明了本文方法能够在每一次迭代中占用较少的时间对模型进行训练,而且相比较于其它模型的训练次数,本文算法只需要训练 10 次. 因此,本文提出了一种有效的对抗鲁棒性提升算法.

参 考 文 献

[1] Wong A, Wu Y, Abbasi S, et al. Fast GraspNeXt: A fast self-attention neural network architecture for multi-task learning in computer vision tasks for robotic grasping on the edge//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Vancouver, Canada, 2023: 2293-2297

[2] Choi J, Kim T, Ryu D, et al. Just-in-time defect prediction for self-driving software via a deep learning model. Journal of Web Engineering, 2023, 22(2): 303-326

[3] Sun J, Li C, Wang Z, Wang Y. A memristive fully connect neural network and application of medical image encryption based on central diffusion algorithm. IEEE Transactions on Industrial Informatics, 2023, 9: 1-11

[4] Christian S, Wojciech Z, Ilya S, et al. Intriguing properties of neural networks//Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada, 2014: 1-10

[5] Ian J G, Jonathon S, Christian S. Explaining and harnessing adversarial examples//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA, 2015: 1-11

[6] Bardia E, Alireza A, Mohammad S. Maximising robustness and diversity for improving the deep neural network safety. Electronics Letters, 2021, 57(3): 116-118

[7] Jin G, Shen S, Zhang D, et al. APE-GAN: Adversarial perturbation elimination with GAN//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK, 2019: 3842-3846

[8] Bai Tao, Luo Jinqi, Zhao Jun, et al. Recent advances in adversarial training for adversarial robustness//Proceedings of the International Joint Conference on Artificial Intelligence. Montreal, Canada, 2021: 4312-4321

[9] Wong E, Rice L, Kolter J Z. Fast is better than free: Revisiting adversarial training//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020: 1-17

[10] Jiang Z, Chen T, Chen T, Wang Z. Robust pre-training by adversarial contrastive learning//Advances in Neural Information Processing Systems. Virtual, 2020: 16199-16210

[11] Zhang C, Zhang K, Zhang C, et al. Decoupled adversarial contrastive learning for self-supervised adversarial robustness//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 725-742

[12] Zi B, Zhao S, Ma X, Jiang Y G. Revisiting adversarial robustness distillation: Robust soft labels make student better //Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 16443-16452

[13] Chung I, Park S, Kim J, Kwak N. Feature-map-level online adversarial knowledge distillation//Proceedings of the International Conference on Machine Learning. Virtual, 2020: 2006-2015

[14] Shafahi A, Najibi M, Ghiasi M A, et al. Adversarial training for free!//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2019: 3353-3364

[15] Du P, Zheng X, Qi M, et al. Towards adversarial robust representation through adversarial contrastive decoupling//Proceedings of the IEEE International Conference on Multimedia and Expo. Taipei, China, 2022: 1-6

[16] Heo B, Lee M, Yun S, Choi J Y. Knowledge distillation with adversarial samples supporting decision boundary//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Hawaii, USA, 2019: 3771-3778

[17] He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 15979-15988

[18] Chen X, Ding M, Wang X, et al. Context autoencoder for self-supervised representation learning. International Journal of Computer Vision, 2024, 132(1): 208-223

[19] Shi Y, Siddharth N, Torr P H S, Kosiorek A R. Adversarial masking for self-supervised learning//Proceedings of the International Conference on Machine Learning. Baltimore, USA, 2022: 20026-20040

[20] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018: 1-23

[21] Dong Y, Liao F, Pang T, et al. Boosting adversarial attacks with momentum//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 9185-9193

[22] Zhang H, Yu Y, Jiao J, et al. Theoretically principled trade-off between robustness and accuracy//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019: 7472-7482

[23] Jia Xiaojun, Zhang Yong, Wu Baoyuan, et al. LAS-AT: Adversarial training with learnable attack strategy//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 13388-13398

[24] Hinton G E, Vinyals O, Dean J. Distilling the knowledge in a neural network. CoRR, abs/1503.02531, 2015

[25] Heo B, Kim J, Yun S, et al. A comprehensive overhaul of feature distillation//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 1921-1930

[26] Tian Y, Krishnan D, Isola P. Contrastive representation distillation//Proceedings of the International Conference on Learning Representations (ICLR 2020). Addis Ababa, Ethiopia, 2020: 1-19

[27] Park W, Kim D, Lu Y, Cho M. Relational knowledge distillation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 3967-3976

[28] Zhang L, Song J, Gao A, et al. Be your own teacher: Improve the performance of convolutional neural networks via self distillation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 3712-3721

[29] Papernot N, McDaniel P D, Wu X, et al. Distillation as a defense to adversarial perturbations against deep neural networks//Proceedings of the IEEE Symposium on Security and Privacy. San Jose, USA, 2016: 582-597

[30] Goldblum M, Fowl L, Feizi S, Goldstein T. Adversarially robust distillation//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA, 2020: 3996-4003

[31] Bai T, Zhao J, Wen B. Guided adversarial contrastive distillation for robust students. IEEE Transactions on Information Forensics and Security, 2023: 1-14

[32] Chen T, Liu S, Chang S, et al. Adversarial robustness: From self-supervised pre-training to fine-tuning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 696-705

[33] Fan L, Liu S, Chen P, et al. When does contrastive learning preserve adversarial robustness from pretraining to finetuning? //Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2021: 21480-21492

[34] Kim M, Tack J, Hwang S J. Adversarial self-supervised contrastive learning//Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2020: 1-14

[35] Grill J-B, Strub F, Altché F, et al. Bootstrap your own latent—A new approach to self-supervised learning//Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2020: 1-14

[36] Krizhevsky A, Hinton G, et al. Learning multiple layers of features from tiny images. Toronto: University of Toronto,

Technical Report: 0, 2009

[37] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[38] Zagoruyko S, Komodakis N. Wide residual networks// Proceedings of the British Machine Vision Conference. New York, UK, 2016: 1-15

[39] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. Virtual Event, Austria, 2021: 1-21

[40] Croce F, Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks// Proceedings of the 37th International Conference on Machine Learning. Virtual Event, 2020: 2206-2216

[41] Schwinn L, Raab R, Nguyen A, et al. Exploring misclassifications of robust neural networks to enhance adversarial attacks. Applied Intelligence, 2023, 53(17): 19843-19859

[42] Zhu J, Yao J, Han B, et al. Reliable adversarial distillation with unreliable teachers//Proceedings of the International Conference on Learning Representations. Virtual Event, 2022: 1-15

[43] Zhao S, Yu J, Sun Z, et al. Enhanced accuracy and robustness via multi-teacher adversarial distillation//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 585-602

[44] Huang B, Chen M, Wang Y, et al. Boosting accuracy and robustness of student models via adaptive adversarial distillation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 24668-24677

ZHANG Dian, Ph. D. candidate. His main research interests include contrastive learning, adversarial attacks, adversarial robustness and the testing of neural networks.

DONG Yun-Wei, Ph. D., professor. His main research areas include artificial intelligence system testing, software intelligent synthesis, and intelligent code completion.

Background

Owing to the inherent vulnerability of neural networks, their deployment across diverse domains has faced a certain level of impediment. Through the integration of masked autoencoders, contrastive learning, and knowledge distillation, this paper introduces an algorithm with the objective of fortifying the adversarial resilience of neural networks. This algorithm is meticulously crafted to mitigate the impact of adversarial interference within neural network inputs.

The neural network's vulnerability pertains to the phenomenon wherein the introduction of imperceptible, subtle perturbations into the input can result in misclassifications, thereby eroding the reliability of neural network outputs. As a result, the pursuit of enhancing the adversarial robustness of neural networks holds paramount significance.

The pursuit of enhancing adversarial robustness in research primarily falls into two categories: input refinement and enhancements to neural networks themselves. The 'input refinement' approach involves subjecting neural network input samples to a certain degree of perturbation or anti-interference processing. Subsequently, the network undergoes training to yield a more resilient model. The 'neural network enhancement' method primarily entails adjusting the model's architecture or training parameters to enhance the neural network's accuracy in recognizing adversarial samples. Noteworthy algorithms in this category include adversarial training, contrastive learning, and knowledge distillation.

This paper introduces an enhanced masked autoencoder capable of inferring global feature representations during the encoding phase by training the model to deduce global feature representations based on visible sub-blocks. The paper achieves this by incorporating the features obtained from encoding the entire image as auxiliary inputs and employing a cross-attention mechanism. This cross-attention mechanism facilitates the integration of the globally encoded features into the auxiliary model's inference process, thereby enhancing the extraction of global feature information from the image.

The authors of this paper have conducted a series of studies on neural network vulnerability, including adversarial attacks, robustness assessment, and adversarial defense. Some of our achievements have already been published in *Neural Networks*.

This research was supported by the National Key Research and Development Program of China (2022YFB4501801) and the Doctoral Dissertation Innovation Fund of Northwestern Polytechnical University (CX2024075).