

基于多维特征图知识蒸馏的对抗样本防御方法

邱宝琳, 易平

(上海交通大学网络空间安全学院, 上海 200240)

摘要: 计算机视觉领域倾向使用深度神经网络完成识别任务, 但对抗样本会导致网络决策异常。为了防御对抗样本, 主流的方法是对模型进行对抗训练。对抗训练存在算力高、训练耗时长缺点, 其应用场景受限。提出一种基于知识蒸馏的对抗样本防御方法, 将大型数据集学习到的防御经验复用到新的分类任务中。在蒸馏过程中, 教师模型和学生模型结构一致, 利用模型特征图向量作为媒介进行经验传递, 并只使用干净样本训练。使用多维度特征图强化语义信息的表达, 并且提出一种基于特征图的注意力机制, 将特征依据重要程度赋予权重, 增强蒸馏效果。所提算法在 Cifar100、Cifar10 等开源数据集上进行实验, 使用 FGSM (fast gradient sign method)、PGD (project gradient descent)、C&W (Carlini-Wagner attack) 等算法进行白盒攻击, 测试实验效果。所提方法在 Cifar10 干净样本的准确率超过对抗训练, 接近模型在干净样本正常训练的准确率。在 L2 距离的 PGD 攻击下, 所提方法效果接近对抗训练, 显著高于正常训练。而且其学习成本小, 即使添加注意力机制和多维度特征图等优化方案, 算力需求依然远小于对抗训练, 是一种轻量级的对抗防御方法。知识蒸馏作为一种神经网络学习方案, 不仅能够学习正常样本的决策经验, 而且可以提取鲁棒性特征。使用少量数据生成准确而鲁棒的模型, 可以提高泛化能力, 减少对抗训练带来的成本。

关键词: 深度学习; 对抗样本防御; 知识蒸馏; 多维度特征图

中图分类号: TP309

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2022012

Adversarial examples defense method based on multi-dimensional feature maps knowledge distillation

QIU Baolin, YI Ping

School of Cyber Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Abstract: The neural network approach has been commonly used in computer vision tasks. However, adversarial examples are able to make a neural network generate a false prediction. Adversarial training has been shown to be an effective approach to defend against the impact of adversarial examples. Nevertheless, it requires high computing power and long training time thus limiting its application scenarios. An adversarial examples defense method based

收稿日期: 2021-12-02; 修回日期: 2022-01-28

通信作者: 易平, yiping@sjtu.edu.cn

基金项目: 国家重点研发计划 (2019YFB1405000)

Foundation Item: The National Key R&D Program of China (2019YFB1405000)

引用格式: 邱宝琳, 易平. 基于多维特征图知识蒸馏的对抗样本防御方法[J]. 网络与信息安全学报, 2022, 8(2): 88-99.

Citation Format: QIU B L, YI P. Adversarial examples defense method based on multi-dimensional feature maps knowledge distillation[J]. Chinese Journal of Network and Information Security, 2022, 8(2): 88-99.

on knowledge distillation was proposed, reusing the defense experience from the large datasets to new classification tasks. During distillation, teacher model has the same structure as student model and the feature map vector was used to transfer experience, and clean samples were used for training. Multi-dimensional feature maps were utilized to enhance the semantic information. Furthermore, an attention mechanism based on feature map was proposed, which boosted the effect of distillation by assigning weights to features according to their importance. Experiments were conducted over cifar100 and cifar10 open-source dataset. And various white-box attack algorithms such as FGSM (fast gradient sign method), PGD (project gradient descent) and C&W (Carlini-Wagner attack) were applied to test the experimental results. The accuracy of the proposed method on Cifar10 clean samples exceeds that of adversarial training and is close to the accuracy of the model trained on clean samples. Under the PGD attack of L2 distance, the efficiency of the proposed method is close to that of adversarial training, which is significantly higher than that of normal training. Moreover, the proposed method is a light-weight adversarial defense method with low learning cost. The computing power requirement is far less than that of adversarial training even if optimization schemes such as attention mechanism and multi-dimensional feature map are added. Knowledge distillation can learn the decision-making experience of normal samples and extract robust features as a neural network learning scheme. It uses a small amount of data to generate accurate and robust models, improves generalization, and reduces the cost of adversarial training.

Keywords: deep learning, adversarial examples defense, knowledge distillation, multi-dimensional feature maps

0 引言

在计算机视觉方面,尤其是目标识别、无人驾驶等高精度任务下,深度学习提供了强大的算法支持。但是,此类算法的决策存有风险性。Szegedy等^[1]发现,攻击者可以根据神经网络模型参数,构造添加了微小扰动的图像。这类图像被称为对抗样本,该图像虽然不会影响到人类的视觉判断,但会改变模型分类结果。例如,在Xu等^[2]的工作中,他们仅穿上了一件添加扰动特征图像的衣服,即可蒙骗目标识别网络,顺利在摄像头下“消失”。

对抗样本在计算机视觉领域产生了深远影响,是该领域安全方向的关注重点。主流方法是通过对抗训练提高模型的防御能力。对抗训练基于模拟攻防的思想,针对目标模型生成对抗样本,再使用生成的样本训练神经网络,重复多次后获取稳定的防御模型。但是,主流的神经网络模型结构极为复杂,如ResNet^[3]和GoogleNet^[4],其具有规模庞大的参数、神经层结构。每次对抗训练都需要极大的算力支持,限制了对抗训练在低算力设备上的应用。所以研究人员期望利用已训练模型,将防御经验泛化到新的视觉任务中,减少对抗训练带来的高算力需求。

知识蒸馏^[5]是一种经验迁移方式,可以将教

师网络预训练得到的经验有效迁移到学生网络中。整个知识蒸馏过程中学生网络会调整参数,模拟教师的输出,从而达到经验转迁的效果。传统知识蒸馏通过训练,使小型网络在同任务下拥有原大型网络相近的决策能力。

本文基于上述研究工作,提出了一种优化的知识蒸馏算法,该算法将教师网络学习到的鲁棒性经验迁移到新的任务中。算法针对特征图和频域等维度进行改造,并基于类激活映射^[6]提出注意力机制,加强关注蒸馏的重要特征,提高模型的防御效果。此外,对抗训练会导致模型在干净样本的准确率有所下降,但通过本文所提算法得到的学生网络,在保持较高鲁棒性的基础上,对干净样本的检测效果显著高于对抗训练,甚至可以媲美仅使用正常样本的无扰动训练。

本文的主要贡献如下。

1) 提出了一种轻量级的对抗样本防御方法。通过知识蒸馏算法,已有鲁棒性特征可以泛化到其他数据集的任务中,无须在新的数据集上重新对抗训练即有不错的防御效果。

2) 利用多维度的特征图优化蒸馏过程,增强防御效果,并提出一种注意力机制,为特征图向量细粒度配比权重,提高泛化能力。

3) 基于方案设计了实验并生成多组对照基线,验证了本文方案的有效性。

1 相关工作

文献[7-8]对抗样本进行了一些研究,下面介绍常用对抗样本生成方法以及主流的防御方式,并阐述知识蒸馏和类激活映射等相关工作。

1.1 对抗样本生成方法

对抗样本属于图像分类任务中的逃逸攻击(evasion attack)。攻击者通过模型的可微损失函数,精心构造微小扰动,影响分类模型的准确率,主要目标为

$$\max_{\delta} L(x + \delta, y, \theta), \text{ s.t. } \|\delta\|_p \leq \varepsilon \quad (1)$$

其中, x 为图像输入, y 为图像 x 的对应类别标签, θ 为模型参数变量, δ 为攻击者添加的扰动, $L(\cdot)$ 为分类器的损失函数,多标签分类任务通常使用交叉熵损失(cross-entropy loss)。 $\|\cdot\|_p$ 为 l_p 范数。优化目标为扰动幅度限制在 ε 的情况下,寻找扰动 δ ,使图片分类损失最大。接下来将介绍常用对抗攻击方法。

(1) FGSM 攻击

快速梯度符号法^[9](FGSM, fast gradient sign method)是一种梯度攻击方法,也是对抗样本生成算法的基础之一。在线性模型中,FGSM 期望添加的扰动变化量与模型损失函数的梯度变化方向一致,如式(2)所示。

$$x = x + \alpha \text{sign}(\nabla_x L(x, y, \theta)) \quad (2)$$

其中, α 为超参数,用以调整扰动幅度。 $\text{sign}(\cdot)$ 为符号函数, ∇_x 为损失函数的回传梯度。FGSM 具有生成对抗样本快,消耗资源少的优点。但是因为只攻击一次,所以攻击效果有限。

(2) I-FGSM 攻击

迭代快速梯度符号法^[10](I-FGSM, iterative fast gradient sign method)属于多次迭代的FGSM 攻击,初始输入为干净样本,如式(3)所示。多次迭代解决了单次攻击效果不佳的问题。

$$x_0 = x_{\text{clean}} \quad (3)$$

$\text{Clip}_x\{\cdot\}$ 为截断函数,用以将生成的扰动限制在 ε 的范围内,如式(4)所示。相较于FGSM, I-FGSM 拥有多次迭代的优势,提升了攻击成功率。

$$x_{t+1} = \text{Clip}_{x, \varepsilon} \{x_t + \alpha \text{sign}(\nabla_x L(x_t, y, \theta))\} \quad (4)$$

(3) MI-FGSM 攻击

动量迭代快速梯度符号法^[11](MI-FGSM, momentum iterative fast gradient sign method)引入动量,以累计的速度矢量指引梯度下降。

$$g_{t+1} = \mu g_t + \frac{\nabla_x L(x_t, y, \theta)}{\|\nabla_x L(x_t, y, \theta)\|} \quad (5)$$

μ 为衰减因子,当 $\mu = 0$ 时,MI-FGSM 变为 I-FGSM。式(5)的第一部分为第 t 次迭代的动量,以衰减形式累加,记忆之前梯度方向;第二部分为当前的梯度,以 l_1 距离进行归一化处理。两者相加得到最终的动量表示。每次对图像添加的扰动均使用此迭代动量,如式(6)所示。MI-FGSM 有效避免了出现过拟合的情况,在黑盒攻击下有很好的鲁棒性泛化能力。

$$x_{t+1} = x_t + \alpha \text{sign}(g_{t+1}) \quad (6)$$

(4) PGD 攻击

投影梯度下降^[12](PGD, project gradient descent)攻击类似 I-FGSM 攻击,通过多次迭代获取对抗样本。初始输入为干净样本,如式(7)所示。

$$x_0 = x_{\text{clean}} \quad (7)$$

PGD 攻击的扰动生成函数基于FGSM 攻击。迭代中对样本扰动进行约束,以 l_2 或 l_∞ 范数为限制生成扰动域,生成的对抗样本被投影到域中,即式(8)中的 P_x 。PGD 攻击针对白盒模型的效果极佳,可用作测试模型防御能力的基线算法。

$$x_{t+1} = P_x(x_t + \alpha \text{sign}(\nabla_x L(x_t, y, \theta))) \quad (8)$$

(5) DeepFool 攻击

深度欺骗攻击^[13](DeepFool attack)是基于超平面分类思想的攻击,可用于非线性模型,相较于FGSM 攻击,其具有更强的适应性。深度欺骗攻击算法基于原始图像,将分类器的决策边界线性化。添加扰动,使图像逐步逼近决策边界,并重复迭代增强效果,使样本被错误分类。作为一种早期方法,DeepFool 攻击相比FGSM 攻击,其适用范围更广,计算速度有所提高。但是需要根据决策边界添加扰动,难以对扰动幅度进行限制。

(6) OPA

单像素攻击^[14](OPA, one pixel attack)只改变少数像素点就可以获得较好的攻击效果,其目标函数为

$$\max [f(x + e(x))], \text{ s.t. } \|e(x)\|_0 \leq d \quad (9)$$

函数 $e(\cdot)$ 为根据原始图像 x 生成的扰动, 作用于原始图像, 使目标分类器决策错误标签置信度增大。 d 为扰动阈值, 当 $d=1$ 时, OPA 为严格单像素攻击。在生成扰动中, OPA 将当前扰动集合作为父集, 随机生成扰动子集合, 筛选父子集合中攻击性强的扰动, 多次迭代获得最强扰动。相较于通用扰动方法, OPA 修改的像素数量更少, 仅需要分类标签概率即可拥有较强的攻击性能, 适用于黑盒攻击下的小幅度攻击。

(7) C&W 攻击

Carlini-Wagner 攻击^[15] (C&W attack) 是针对 Hinton^[5] 的防御蒸馏网络提出的攻击方法, 具体目标函数为

$$\min [\|\delta\|_p + c \cdot f(x + \delta)], \text{ s.t. } x + \delta \in [0, 1]^n \quad (10)$$

其中, $\|\cdot\|_p$ 为扰动范数, 用以衡量扰动幅度, 使用 L_0 、 L_2 、 L_∞ 这 3 种范数进行衡量。目标函数确保扰动幅度小的同时, 模型错误分类标签的置信度足够高。C&W 攻击迁移能力强, 是用于对比实验的基线算法之一。相比于 PGD 攻击, C&W 攻击可以构造出更小的扰动, 但复杂度更高。Madry^[12] 基于 C&W 攻击损失函数提出 PGD 形式的 C&W 攻击, 该攻击方便多次迭代, 并保留了 C&W 良好的攻击性能。

1.2 对抗样本防御方法

对抗样本防御的思想是: 假设模型输入图片 x , 添加的对抗扰动为 δ , 模型输出为 $f(\cdot)$, 期望

$$f(x + \delta) = f(x) \quad (11)$$

即添加对抗扰动前后, 模型输出期望不会受到影响, 这种性质被称为模型的鲁棒性。接下来介绍目前主流的对抗防御方法。

1.2.1 基于对抗训练的防御方法

数据扩充是一种经典的神经网络训练方法, 其思想是添加数据集, 增强模型的决策经验。在对抗防御方面, 使用对抗样本扩充数据。主要思路是针对目标模型进行白盒攻击, 将生成的对抗样本添加到模型的训练集中, 对模型进行重复训练。

Goodfellow^[9] 提出基于 FGSM 的对抗训练。训练过程中兼顾原图片和对抗样本的损失。损失

函数如下:

$$x_{\text{adv}} = x + \varepsilon \text{sign}(\nabla_x L(\theta, x, y)) \quad (12)$$

$$\text{Loss} = \alpha L(x, y) + (1 - \alpha) L(x_{\text{adv}}, y) \quad (13)$$

Madry 等^[12] 对此进行优化, 使用 PGD 攻击生成对抗样本, 并只关注对抗损失, 在白盒和黑盒攻击下均有不错的防御效果。Kannan 等^[16] 在之前工作的基础上进行优化, 其不仅约束原始图片和对抗样本的分类损失, 而且添加扰动前后图片的相似正则项, 保证分类器对原始图片的 logits 和对抗样本的 logits 一致。Wang 等^[17] 使用错分类的输入样本进行对抗训练, 提高了模型的鲁棒性, 该方法针对攻击样本的分类准确率有所提升, 但降低了干净样本的分类准确率。

对抗训练是一种非常优秀的防御方法, 但需要多次迭代, 以提高模型的鲁棒性。训练过程中需要极高的算力需求, 在小型计算设备上难以支持。

1.2.2 其他防御方法

Xie 等^[18] 对输入样本随机变换尺寸以及填充, 再输入分类模型中, 提高了模型防御能力。文献[18]提出的算法使用随机操作预处理, 使攻击者无法固定真实输入, 降低了攻击风险。随机操作额外计算量小, 可与对抗训练结合增强效果。但该算法需要足够大的随机调整空间和填充空间, 来提高攻击者的计算成本。

Xie 等^[19] 发现输入对抗样本后, 模型的特征图与原始输入差距较大, 而且对抗样本的特征图存在噪点, 这将影响神经网络模型的决策。该方法在神经网络的特征层添加去噪模块, 将特征图预处理再向后传递, 并在 ImageNet 数据集测试其白盒防御能力。同时, 文献[19]指出对抗训练是实验取得良好防御效果的必要条件, 存在高算力需求。

Xu 等^[20] 通过特征压缩检测输入图像是否为对抗样本, 从源头降低对抗样本攻击的风险。该方法减少像素的色位深度, 以及平滑像素差距来压缩图像, 缩小攻击空间。其复杂度低, 不需要对分类模型改造即有一定的防御能力。但阈值设定过于敏感会影响正常样本的输入, Ma 等^[21] 证明特征压缩对于 Cifar、ImageNet 等彩色数据集效果不佳。

非对抗训练方法通过微调、降噪、压缩等方式预处理,进行对抗防御。但预处理方案难以学习到神经网络的深层特征,需要对抗训练增强防御效果,本质上没有降低学习代价。本文对已有模型进行经验复用,减少对抗训练带来的算力需求。

1.3 知识蒸馏

大型神经网络模型算力需求高,响应速度慢,影响线上部署效率。Hinton 等^[5]提出一种迁移学习方案,小型网络通过训练,拥有原网络相似的泛化决策能力。Hinton 等对 softmax 函数进行改造,如式(14)所示。

$$q_i = \frac{\exp\left(\frac{z_i}{T}\right)}{\exp\sum\left(\frac{z_j}{T}\right)} \quad (14)$$

当 $T=1$ 时,式(14)变为原始的 softmax 变换。随着 T 的增大,教师网络和学生网络的输出更容易拟合。并以此为基础约束蒸馏中的软标签损失,如式(15)所示。

$$L_{\text{soft}} = -\sum P_i \log q_i \quad (15)$$

其中, P_i 为教师模型的 softmax 输出, q_i 为学生模型的 softmax 输出,将其添加至蒸馏的总损失中,即

$$L = L_{\text{hard}} + L_{\text{soft}} \quad (16)$$

知识蒸馏的总损失分为两类, L_{hard} 约束图片的正常分类, L_{soft} 约束蒸馏中的相似拟合。通过知识蒸馏,参数简单的学生模型模拟教师模型的输出分布,两者拥有近似的决策能力。模型压缩会使学生网络针对性学习特征,具有一定的防御效果。

Goldblum 等^[22]和 Shafahi 等^[23]根据 Hinton 等^[5]的工作,从两种角度进行对抗样本防御研究。Goldblum 等提出教师网络和学生网络的结构不同,蒸馏后结构简单的学生网络依然有较强的鲁棒性能,方便线上部署。Goldblum 等的贡献有两方面:第一根据实验证明使用干净样本进行蒸馏,学生网络可以继承教师的鲁棒性;第二提出一种对抗鲁棒特征蒸馏(ARD, adversarially robust distillation)方案,教师网络的输入为干净样本,学生网络利用自身参数生成对抗样本作为输入,

以 KL 散度约束,借鉴了 Shafahi 等^[24]的方法加速训练,得到的学生网络拥有良好的鲁棒性能。Shafahi 等^[23]提出一种鲁棒性特征迁移方案,期望模型在先前的数据集对抗训练得到的鲁棒经验,能够泛化到新的数据集任务中,减少新任务的训练消耗。蒸馏过程中教师模型和学生模型结构保持一致,但对应数据集不同,所以不能以 softmax 的输出 logits 作为蒸馏约束。Shafahi 等认为决策一致的条件中间特征层的输出相似,利用模型中间输出作为媒介,约束蒸馏过程,其在少量资源消耗的前提下获得不错的防御效果。

1.4 类激活映射

类激活映射(CAM, class activation mapping)属于语义分割(semantic segmentation)领域,可以帮助研究者理解神经网络的决策依据。本文借鉴类激活映射的思想,提出一种针对特征向量的注意力机制,加强对重要特征的关注度。

Zhou 等^[6]发现,神经网络的最后一层特征层保留着丰富的空间和语义信息。他们将最后的卷积层特征直接通过全局平均池化层(GAP, global average pooling),形成每个特征图的均值,利用全连接层获取每个特征图均值的权重。在需要生成可视化热力图时,提取对应类别的全连接层权重,与特征图加权求和,上采样获取最终图像,如图 1 所示。

类激活映射在模型可视化方面效果较好,有助于研究者深入理解神经网络模型的决策。Selvaraju 等^[25]将类激活映射扩展,使用 CNN 的最后一层卷积层的梯度信息进行判断,扩大映射的使用场景。Chattopadhyay 等^[26]优化了权重计算,使映射定位热力图更加准确。Wang 等^[27]提出一种基于等变注意力机制(SEAM, self-supervised equivariant attention mechanism)的弱监督语义分割的方法,构建孪生网络,为空间仿射变换后的类激活映射输出提供自监督,并使用注意力机制整合图像信息,在预测图像语义方面有很好的效果。

2 方案设计

本文的目标是利用知识蒸馏将模型在大数据集对抗训练得到的防御经验进行提取,并泛化到其他数据集的任务中,从而减少在其他数据集上的鲁棒性模型训练成本。方案设计主要分为 4 部

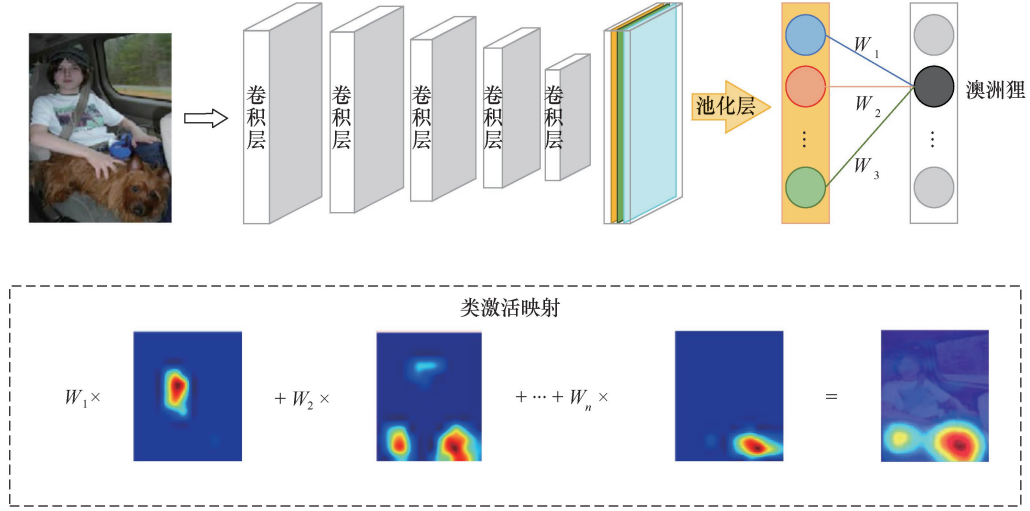


图 1 类激活映射机制
Figure 1 Mechanism of class activation mapping

分：第 1 部分利用对抗训练生成教师鲁棒模型；第 2 部分设计知识蒸馏的整体架构；第 3 部分使用多维度特征图的蒸馏优化方案；第 4 部分利用基于注意力机制的特征图完成权重优化。

2.1 对抗训练生成鲁棒模型

在操作知识蒸馏前，教师网络需要学习大数据集的鲁棒性特征，用以在其他数据集任务中进行经验传递。本文使用 Madry 等^[12]提出的方法，进行有监督的对抗训练。训练目标为

$$\min_{\theta} E_{(x,y) \sim D} \left[\max_{\delta \in S} L(\theta, x + \delta, y) \right] \quad (17)$$

式(17)是基于 min-max 优化的通用对抗训练方法。其中， θ 表示网络结构， δ 为生成的对抗扰动， (x, y) 为图像及对应标签， $L(\cdot)$ 为损失函数。具体方案如图 2 所示， $\max_{\delta \in S} L(\theta, x + \delta, y)$ 为对抗样本生成部分，期望在限定域中寻求扰动 δ ，使模型分类损失最大，从而达到分类错误的目的，在此阶段，模型结构 θ 保持不变。此后，将生成的对抗样本作为输入，训练模型结构 θ 。多次迭代后，得到最终的鲁棒模型。

Madry 等在对抗训练中使用 PGD 攻击产生对抗样本，生成的防御模型具有良好的鲁棒性。因此 Shafahi 等^[23]的知识蒸馏实验同样采取 PGD 攻击生成鲁棒模型，为方便与 Shafahi 的工作做出对比，本文采用 PGD 攻击方法生成对抗样本，如式(18)所示。

$$x_{\text{output}} = P_x \left(x_{\text{input}} + \alpha \text{sign} \left(\nabla_x L(x_{\text{input}}, y, \theta) \right) \right) \quad (18)$$

其中， x_{input} 表示上次 PGD 攻击输出的对抗样本，初始情况下 $x_{\text{input}} = x$ ， α 为超参数，用以调整扰动幅度， $\text{sign}(\cdot)$ 为符号参数， ∇_x 为损失函数的回传梯度。生成对抗样本期间神经网络参数 θ 保持不变。由于实验使用的图像识别任务是多分类任务，神经网络的损失函数使用交叉熵损失。

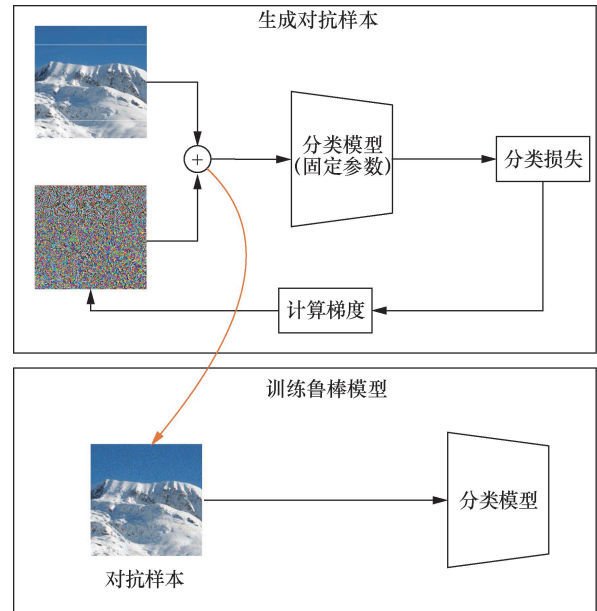


图 2 对抗训练流程
Figure 2 Process of adversarial training

2.2 蒸馏模型基础结构

本节介绍迁移的整体框架，蒸馏模型由教师网络和学生网络两部分组成。教师网络在较大数

据集上进行对抗训练, 获得鲁棒模型。之后使用其他数据集的干净样本进行蒸馏操作, 期望学生模型能够学习到教师网络的鲁棒性经验。

基于 Shafahi 等^[23]的工作, 本文使用一种优化的知识蒸馏方法, 如图 3 所示。教师网络和学生网络结构一致, 但教师参数在蒸馏过程中固定。蒸馏任务的目的是在目标数据集进行迁移学习, 训练学生网络参数。本文沿用 Hinton 等^[5]的蒸馏损失分类, 将损失分为 **hard target** 和 **soft target**。教师预训练使用的源数据集和学生目标数据集类别存在差异, 本文的 **soft target** 没有使用 logits 进行约束, 而是使用神经网络的特征图作为蒸馏项。

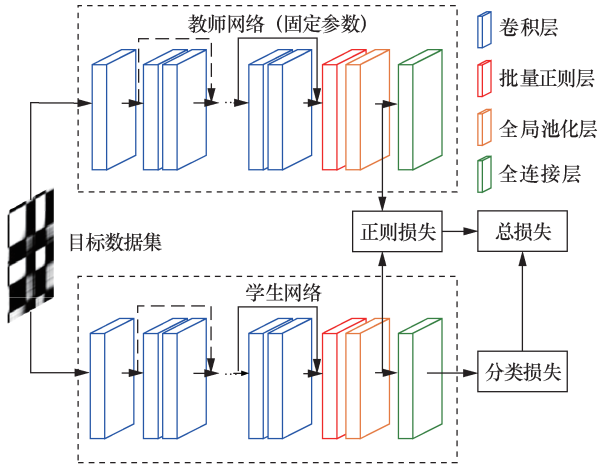


图 3 知识蒸馏结构
Figure 3 Model structure of knowledge distillation

分类神经网络中, 模型输出前通常使用全局平均池化层和全连接层进行处理。本文设定池化层前的教师网络特征图输出为 z_{teacher} , 学生网络的特征图输出为 z_{student} , 特征图 z 对应的池化层输出为 z^{gap} 。

Shafahi 等^[23]使用全局平均池化层后的特征联合训练, 损失函数如式(19)所示。

$$L(x, y) = L_{\text{ce}}(f(x; \theta), y) + \alpha \cdot d(z_{\text{teacher}}^{\text{gap}}, z_{\text{student}}^{\text{gap}}) \quad (19)$$

其中, $L_{\text{ce}}(\cdot)$ 与 Hinton 蒸馏的 **hard target** 相同, 使用交叉熵函数记为 L_{ce} 。第二部分为正则损失, 用以拟合蒸馏过程中的特征经验分布, 为蒸馏的 **soft target**, 记为 L_{kd} 。其中 α 为调整蒸馏损失的超参数, $d(\cdot)$ 为衡量教师网络和学生网络差距的距离度量, 本文使用 L_2 正则衡量。

全局平均池化层将特征压缩、高度提炼, 但这样会丢失大量空间和语义信息, 不利于鲁棒性经验的迁移。本文强化特征层的语义表达, 增强蒸馏效果。

2.3 基于多维度特征图的蒸馏优化

本节从特征图和频域两个维度强化特征层的语义表达。

首先考虑神经网络结构, Shafahi 等^[23]使用池化层之后的特征作为约束项。全局平均池化层由 Lin 等^[28]提出, 使用全局池化层代替全连接层, 将每个特征层抽象为单点输出, 这样不仅能简化模型参数, 加快训练速度, 而且减少了全连接层带来的过拟合现象。但高度抽象特征会损失大量的空间和语义信息, 也会影响蒸馏效果。因此本文使用池化操作之前的特征图作为蒸馏媒介, **soft target** 即为

$$L_{\text{kd}} = \alpha \cdot d(z_{\text{teacher}}, z_{\text{student}}) \quad (20)$$

此外, 本文利用频域特性增强特征语义。考虑到神经网络的复杂性, 本文使用全局池化层之后的特征进行实验, 具体是通过 N 维快速傅里叶变化 (FFTN, fast Fourier transform) 将该特征转化为频域 N 维向量。之后利用傅里叶变换的平移操作 (FFTShift, fast Fourier transform shift) 把零频点移到频谱的中间得到频域特征。频域特征分为实部和虚部, 二者拼接得到频域最终表示。本文以特征向量 $Z \in R^{1 \times 4}$ 进行推演, 如图 4 所示。

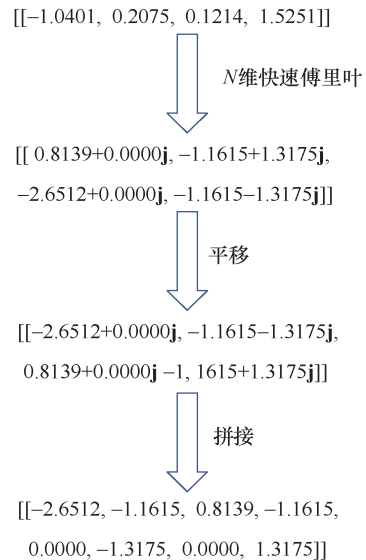


图 4 特征向量转化为频域特征样例
Figure 4 Example of feature vectors transformed into frequency domain feature

z^{gap} 为全局池化层后的输出, z^{fourier} 为蒸馏使用的频域特征表示。两者关系为

$$z^{\text{fourier}} = \text{cat}\left(\text{FFTShift}\left(\text{FFTN}\left(z^{\text{gap}}\right)\right)\right) \quad (21)$$

$\text{cat}(\cdot)$ 为拼接操作, 将两个向量按照维度 $\text{dim}=1$ 进行拼接。频域方面的 soft target 为

$$L_{\text{kd}} = \alpha \cdot d\left(z_{\text{teacher}}^{\text{fourier}}, z_{\text{student}}^{\text{fourier}}\right) \quad (22)$$

两个维度的特征强化在蒸馏过程中增加了额外计算损耗, 所以本文实验不仅验证多维度特征图方案的有效性, 而且会记录时间消耗。期望增强特征语义表达的同时, 训练时间不会显著提高。接下来, 将对特征进行细粒度划分, 关注对应类别的敏感特征, 从而提高经验利用率。

2.4 特征图权重优化

特征图向量可以理解滤波器对于输入图像的特征提取。之前的损失函数直接将教师和学生的特征图正则化比对, 每个特征向量权重比值一致。本文考虑将教师网络的经验细粒度优化, 关注相应类别的敏感特征, 以达到最佳的迁移效果。

根据 Zhou 等^[6]的工作, 可知在分类任务中, 输入特征的显著程度决定了其分类类别。模型对输入生成的类激活映射, 有助于了解特征对决策的贡献。在本文的蒸馏过程中, 不同类别的样本, 对特征层的敏感通道不一致。训练等权重特征向量时, 无法最大化利用教师经验。借鉴类激活映射的工作原理, 本文提出一种特征向量的注意力机制, 加强敏感特征的关注度。

经过 2.1 节的对抗训练后, 教师网络拥有良好的鲁棒性能。固定教师网络的参数, 使用目标数据集训练全连接层, 具体流程如图 5 所示。假设池化操作后的特征图 $\mathbf{Z} \in R^C$, 目标分类共有 B 类, 全连接层的参数 $\mathbf{M} \in R^{B \times C}$ 满足

$$\mathbf{M} = \begin{bmatrix} m_{11} & \cdots & m_{1C} \\ \vdots & \ddots & \vdots \\ m_{B1} & \cdots & m_{BC} \end{bmatrix} \quad (23)$$

对应样本标签类别为 y 的权重系数满足

$$\mathbf{M}(y) = [m_{y1} \cdots m_{yc}] \quad (24)$$

权重系数用于组织特征层向量, 使目标数据集的分类为 y , 即预训练得到的权重表征不同分类对于特征通道的关注度。本文针对特征图向量

施加注意力机制, soft target 满足

$$L_{\text{kd}} = \alpha \cdot d\left(\mathbf{z}_{\text{teacher}}^{\text{gap}} \mathbf{M}(y), \mathbf{z}_{\text{student}}^{\text{gap}}\right) \quad (25)$$

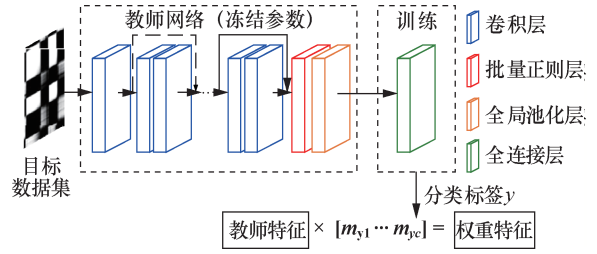


图5 教师特征权重配比流程

Figure 5 Process of obtaining weight for teacher features

对池化前的特征图 $\mathbf{Z} \in R^{C \times H \times W}$ 进行蒸馏的场景下, 需要将 $\mathbf{M}(y)$ 进行扩展, 如下。

$$\mathbf{M}_{\text{fm}}(y) = \begin{bmatrix} \begin{bmatrix} m_{y1} & \cdots & m_{y1} \\ \vdots & \ddots & \vdots \\ m_{y1} & \cdots & m_{y1} \end{bmatrix}_{H \times W} & \cdots & \begin{bmatrix} m_{yc} & \cdots & m_{yc} \\ \vdots & \ddots & \vdots \\ m_{yc} & \cdots & m_{yc} \end{bmatrix}_{H \times W} \end{bmatrix} \quad (26)$$

蒸馏的 soft target 表达式为

$$L_{\text{kd}} = \alpha \cdot d\left(\mathbf{z}_{\text{teacher}} \mathbf{M}_{\text{fm}}(y), \mathbf{z}_{\text{student}}\right) \quad (27)$$

本文将权重系数应用到频域方面。频域特征是针对池化特征的增强, 所以训练全连接层时, 可将拼接后的频域特征作为全连接层的输入, 得到最终的权重系数。训练流程与池化特征处理类似, 在此不再赘述。

2.5 小结

本节对蒸馏过程中的特征图进行优化, 丰富其空间和语义信息, 加强表达效果, 并基于类激活映射提出注意力机制, 针对不同类别强化部分特征, 达到更好的蒸馏效果。最终的损失函数为

$$L(x, y) = L_{\text{ce}} + \alpha \cdot d\left(\mathbf{z}_{\text{teacher}} \mathbf{M}(y), \mathbf{z}_{\text{student}}\right) \quad (28)$$

$\mathbf{z}_{\text{teacher}}$ 、 $\mathbf{z}_{\text{student}}$ 为广义特征向量, $\mathbf{M}(y)$ 为对应特征形式的权重, α 为控制正则损失的超参数。

3 实验设计和结果分析

3.1 实验环境设置

本文实验训练及测试环境为 Centos 7.7 操作系统, 处理器型号为 Intel Xeon Scalable Cascade Lake 8168(2.7 GHz, 24 核), 内存 1.5 TB

DDR4 ECC REG 2666, GPU 为 NVIDIA Tesla V100。

本文使用 Cifar100 和 Cifar10 数据集进行效果展示, 图片尺寸为 32×32 。教师网络在 Cifar100 数据集上进行对抗训练, 获取鲁棒模型。之后固定教师网络参数, 使用 Cifar10 训练集进行蒸馏训练, 在多种算法下测试迁移效果。

预训练阶段, 本文使用添加动量(momentum)的 SGD^[29]算法, momentum 为 0.9, weight_decay 值为 2×10^{-4} 。对抗样本使用 L_∞ 的 PGD 攻击算法生成, 扰动幅度 $\varepsilon = \frac{8}{255}$, 单步最大扰动为 $\frac{2}{255}$, 迭代次数 step=7。

知识蒸馏阶段, 教师网络和学生网络均选取 wide-resnet28^[3], 学生网络的初始参数与教师保持一致, 来加快蒸馏训练。在完成 15 个 epoch 训练后, 训练集上的损失基本稳定, 学生网络能够顺利收敛。以添加权重 mask 的频域蒸馏训练为例, 超参数 $\alpha = 10$, 模型在训练集的准确率和损失曲线如图 6 所示。

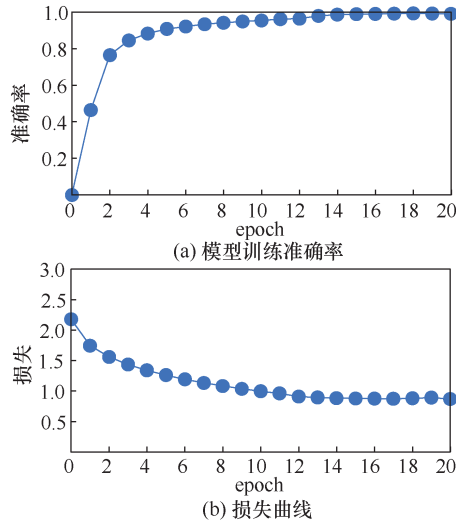


图 6 基于频域特征蒸馏的准确率与损失曲线
Figure 6 Accuracy and loss curve on knowledge distillation with frequency

为了证明本文方法的有效性, 共设置了 3 组模型用以比对。第一组使用 Cifar10 直接训练得到的正常模型; 第二组使用 Cifar10 对抗训练得到的鲁棒性模型; 第三组复现 Shafahi^[23]的工作, 用以比对蒸馏效果。3 组基线初始参数与学生网络相同, 从而确保训练的公平性。

3.2 算法效果

本文使用 Cifar10 测试集在模型上的准确率, 作为判断鲁棒性特征学习效果的依据。模型在样本的分类上准确率越高, 说明学习到的鲁棒性特征越强, 防御更有效。本实验使用正常输入、FGSM、PGD、C&W 等多种攻击测试防御效果。其中, FGSM 最大扰动幅度 $\varepsilon = \frac{8}{255}$ 。PGD 攻击选取 L_2 、 L_∞ 两种范数攻击, PGD- L_2 最大扰动幅度 $\varepsilon = 1$, PGD- L_∞ 最大扰动幅度 $\varepsilon = \frac{8}{255}$, 步长均为 $\frac{\varepsilon}{5}$ 。C&W 攻击采用 Madry^[12]提出的迭代形式, 扰动幅度与 PGD- L_∞ 相同。PGD 和 C&W 攻击测试 5、10、20 这 3 种迭代次数下的防御效果。

Shifahi^[23]蒸馏和特征图蒸馏均使用超参数 $\alpha = 500$ 。频域特征转化幅度较大, 不适用同样参数, 使用 $\alpha = 10$ 进行测试。之后进行消融实验去除超参数 α 的影响。在 Cifar10 上的准确率如表 1~表 4 所示。

干净样本和 FGSM 攻击在 Cifar10 上的准确率结果如表 1 所示。替换特征层后, 语义信息更加丰富, 偏重拟合鲁棒性特征, 导致在干净样本和 FGSM 下准确率有所下降。使用权重系数 mask 能有效控制非关键特征, 蒸馏过程更具有针对性, 准确率均明显提升。权重系数在 Shafahi 的工作同样有效, FGSM 攻击下准确率由 46.13% 提高到 61.30%。本文的方法(采用频域+mask 训练)在干净样本上的准确率甚至高于对抗训练, 达到 94.23%。

表 1 干净样本和 FGSM 攻击在 Cifar10 上的准确率
Table 1 Accuracy of the clean and FGSM attack on Cifar10

训练方式	数据集	干净样本	FGSM
正常训练	Cifar10	94.77%	7.1%
对抗训练	Cifar10	87.42%	58.04%
Shafahi 蒸馏	Cifar100→Cifar10	92.93%	46.13%
特征图蒸馏	Cifar100→Cifar10	85.67%	40.02%
频域蒸馏	Cifar100→Cifar10	87.81%	40.56%
Shafahi+mask	Cifar100→Cifar10	94.79%	61.30%
特征图+mask	Cifar100→Cifar10	94.17%	62.44%
频域+mask	Cifar100→Cifar10	94.23%	63.57%

表2 PGD- L_2 攻击算法在 Cifar10 上的准确率
Table 2 Accuracy of the PGD- L_2 attack on Cifar10

训练方式	数据集	PGD-5	PGD-10	PGD-20
正常训练	Cifar10	0.01%	0	0
对抗训练	Cifar10	45.57%	31.06%	24.95%
Shafahi 蒸馏	Cifar100→Cifar10	13.15%	4.7%	2.35%
特征图蒸馏	Cifar100→Cifar10	33.27%	18.26%	12.91%
频域蒸馏	Cifar100→Cifar10	27.27%	13.07%	8.39%
Shafahi+mask	Cifar100→Cifar10	29.44%	17.71%	13.32%
特征图+mask	Cifar100→Cifar10	37.41%	28.63%	23.49%
频域+mask	Cifar100→Cifar10	39.94%	30.44%	25.14%

表3 PGD- L_∞ 攻击算法在 Cifar10 上的准确率
Table 3 Accuracy of the PGD- L_∞ attack on Cifar10

训练方式	数据集	PGD-5	PGD-10	PGD-20
正常训练	Cifar10	0.01%	0	0
对抗训练	Cifar10	57.76%	49.31%	47.38%
Shafahi 蒸馏	Cifar100→Cifar10	11.63%	4.78%	2.61%
特征图蒸馏	Cifar100→Cifar10	35.95%	26.15%	23.93%
频域蒸馏	Cifar100→Cifar10	30.49%	20.26%	17.46%
Shafahi+mask	Cifar100→Cifar10	25.95%	15.89%	12.25%
特征图+mask	Cifar100→Cifar10	34.56%	26.17%	21.87%
频域+mask	Cifar100→Cifar10	37.09%	27.82%	23.33%

表4 C&W 攻击算法在 Cifar10 上的准确率
Table 4 Accuracy of the C&W attack on Cifar10

训练方式	数据集	CW-5	CW-10	CW-20
正常训练	Cifar10	0.015%	0	0
对抗训练	Cifar10	53.84%	46.54%	45.27%
Shafahi 蒸馏	Cifar100→Cifar10	8.71%	3.27%	1.77%
特征图蒸馏	Cifar100→Cifar10	32.85%	24%	22.46%
频域蒸馏	Cifar100→Cifar10	27.35%	17.98%	16.15%
Shafahi+mask	Cifar100→Cifar10	20.93%	13.19%	9.83%
特征图+mask	Cifar100→Cifar10	28.44%	20.39%	16.64%
频域+mask	Cifar100→Cifar10	32.17%	23.44%	19.66%

本实验验证了 Cifar10 在 PGD 攻击和 C&W 攻击下的准确率,结果如表2~表4所示。相较于 FGSM 攻击, PGD 和 C&W 攻击强度更高,增强语义的效果更加明显。特征层权重系数也有效提高算法的准确率,在 PGD- L_2 下本文的方法略微高于对抗训练,达到了 25.14%, 对抗训练只有 24.95%。

此外,本文以 PGD- L_∞ 为例,验证高轮次迭代下的算法效果,结果如图7所示。即使是100次

的迭代攻击,频域蒸馏和特征图蒸馏依旧有很高的鲁棒性。

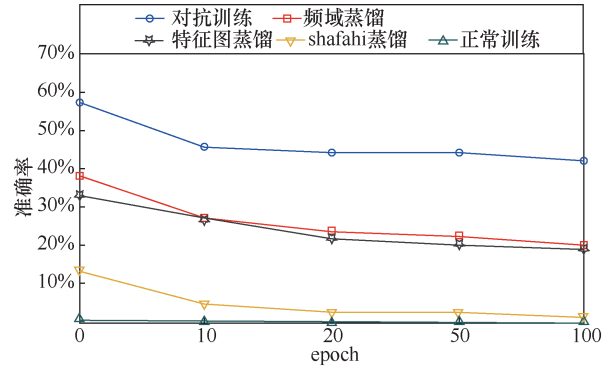


图7 PGD- L_∞ 迭代强度与准确率关系

Figure 7 The relationship between epoch of PGD- L_∞ attack and accuracy

3.3 算法耗时

本文对不同算法在 Cifar10 上的耗时进行对比,如表5所示。本文蒸馏耗时与 Shafahi 类似,接近正常训练,远小于对抗训练耗时。这种轻量级的蒸馏算法相比于 Shafahi 工作,防御效果显著提高,算力要求较低。

表5 不同算法在 Cifar10 的耗时情况
Table 5 Time-consuming for Different Algorithms on Cifar10

训练方式	数据集	耗时
正常训练	Cifar10	19min
对抗训练	Cifar10	5.5h
Shafahi 蒸馏	Cifar100→Cifar10	24min
特征图蒸馏	Cifar100→Cifar10	29min
频域蒸馏	Cifar100→Cifar10	27min
Shafahi+mask	Cifar100→Cifar10	29min
特征图+mask	Cifar100→Cifar10	35min
频域+mask	Cifar100→Cifar10	33min

总体而言,本文通过多维特征图和权重系数 mask 进行优化,增强防御效果。相比 Shafahi 的蒸馏有着不错的提升,即使在多次迭代的 PGD 攻击下,模型依旧拥有良好的防御能力。算法计算耗时较低,接近正常训练。

3.4 消融实验

为了控制超参数 α 带来的影响,本文对其进行消融实验。选取干净样本、FGSM、PGD- L_∞ 这3种攻击测试不同权重 α 的蒸馏效果,如表6~表8所示。

表 6 Shafahi+mask 基于不同 α 在 Cifar10 上的分类准确率
Table 6 Accuracy based on α with shafahi used mask on Cifar10

权重 α	Clean	FGSM	PGD-5	PGD-10	PGD-20
5×10^{-1}	94.65%	28.31%	0.55%	0.02%	0
5	94.80%	38.39%	2.15%	0.17%	0.05%
50×10^1	94.95%	52.78%	9.21%	2.61%	1.00%
5×10^2	94.79%	61.3%	25.95%	15.89%	12.25%
5×10^3	93.99%	58.28%	27.41%	17.54%	13.22%

表 7 特征图+mask 基于不同 α 在 Cifar10 上的分类准确率
Table 7 Accuracy based on α with feature map used mask on Cifar10

权重 α	Clean	FGSM	PGD-5	PGD-10	PGD-20
5×10^{-1}	94.63%	33.29%	0.89%	0.05%	0.01%
5	95.12%	45.31%	3.38%	0.44%	0.09%
5×10^1	94.79%	58.92%	15.71%	6.21%	3.37%
5×10^2	94.17%	62.44%	34.56%	26.17%	21.87%
5×10^3	91.67%	54.38%	29.78%	19.03%	13.32%

表 8 频域+mask 基于不同 α 在 Cifar10 上的分类准确率
Table 8 Accuracy based on α with frequency used mask on Cifar10

权重 α	Clean	FGSM	PGD-5	PGD-10	PGD-20
0.01	94.75%	43.68%	1.73%	0.12%	0.02%
0.1	95.03%	51.42%	7.87%	1.78%	0.43%
1	95.05%	62.78%	27.53%	17.40%	13.36%
10	94.23%	63.57%	37.09%	27.82%	23.33%
20	93.20%	57.47%	33.34%	28.40%	25.92%
30	87.89%	42.49%	19.55%	15.11%	11.96%

α 的增大有助于蒸馏关注 soft target, 使模型更具有鲁棒性, 提高在攻击算法上的准确率, 但会降低在干净样本上的准确率。频域特征对于超参数敏感程度更高, 小量级下就有优秀的表现。最终, 本文选取频域特征下的超参数 $\alpha = 10$, 其他蒸馏下超参数 $\alpha = 5 \times 10^{-1}$ 。

4 结束语

本文提出一种基于知识蒸馏的轻量级对抗样本防御方法, 用以解决多任务下经验复用的问题, 更快获取鲁棒性模型。一方面, 使用多维度的特征图, 获取更丰富的空间和语义信息, 增强蒸馏效果; 另一方面, 使用权重系数对特征层参数细粒度配比, 有效利用教师经验。之后, 通过对照

实验证明了所提算法的有效性。本文算法在高强度的白盒攻击下, 准确率均显著高于 Shafahi 的蒸馏方案, 在干净样本上的准确率甚至高于对抗训练, 接近模型直接训练的准确率。

本文算法还可以在很多方向进行完善。第一, 可以测试在更大数据集如 ImageNet 方面的迁移学习效果。第二, 目前仅利用单层特征进行蒸馏, 之后可以拓展到多层同步提取。第三, 将对应样本类别的全连接系数作为权重 mask, 可以研究对于非该类别的 mask 能否有效。

参考文献:

- [1] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv:1312.6199, 2013.
- [2] XU K D, ZHANG G Y, LIU S J, et al. Adversarial T-shirt evading person detectors in a physical world[C]//Proceedings of Computer Vision – ECCV 2020. 2020: 665-681.
- [3] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.
- [4] SZEGEDY C, VANHOUCHE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2818-2826.
- [5] HINTON G E, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.
- [6] ZHOU B L, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2921-2929.
- [7] 易平, 王科迪, 黄程, 等. 人工智能对抗攻击研究综述[J]. 上海交通大学学报, 2018, 52(10): 1298-1306.
- [8] YI P, WANG K D, HUANG C, et al. Adversarial attacks in artificial intelligence: a survey[J]. Journal of Shanghai Jiao Tong University, 2018, 52(10): 1298-1306.
- [9] 王科迪, 易平. 人工智能对抗环境下的模型鲁棒性研究综述[J]. 信息安全学报, 2020, 5(3): 13-22.
- [10] WANG K D, YI P. A survey on model robustness under adversarial example[J]. Journal of Cyber Security, 2020, 5(3): 13-22.
- [11] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv:1412.6572, 2014.
- [12] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial machine learning at scale[J]. arXiv preprint arXiv:1611.01236, 2016.
- [13] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 9185-9193.
- [14] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint

- arXiv:1706.06083, 2017.
- [13] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. Deep-Fool: a simple and accurate method to fool deep neural networks[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2574-2582.
- [14] SU J W, VARGAS D V, SAKURAI K. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841.
- [15] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of 2017 IEEE Symposium on Security and Privacy. 2017: 39-57.
- [16] KANNAN H, KURAKIN A, GOODFELLOW I. Adversarial logit pairing[J]. arXiv preprint arXiv:1803.06373, 2018.
- [17] WANG Y, ZOU D, YI J, et al. Improving adversarial robustness requires revisiting misclassified examples[C]//Proceedings of International Conference on Learning Representations. 2019.
- [18] XIE C, WANG J, ZHANG Z, et al. Mitigating adversarial effects through randomization[J]. arXiv preprint arXiv:1711.01991, 2017.
- [19] XIE C H, WU Y X, VAN DER MAATEN L, et al. Feature denoising for improving adversarial robustness[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 501-509.
- [20] XU W L, EVANS D, QI Y J. Feature squeezing: detecting adversarial examples in deep neural networks[J]. arXiv preprint arXiv:1704.01155, 2017.
- [21] MA S Q, LIU Y Q, TAO G H, et al. Nic: detecting adversarial samples with neural network invariant checking[C]//Proceedings of the 26th Network and Distributed System Security Symposium (NDSS 2019). 2019.
- [22] GOLDBLUM M, FOWL L, FEIZI S, et al. Adversarially robust distillation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4): 3996-4003.
- [23] SHAFABI A, SAADATPANAH P, ZHU C, et al. Adversarially robust transfer learning[J]. arXiv preprint arXiv:1905.08232, 2019.
- [24] SHAFABI A, NAJIBI M, GHIASI A, et al. Adversarial training for free![J]. arXiv preprint arXiv:1904.12843, 2019.
- [25] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[J]. International Journal of Computer Vision, 2020, 128(2): 336-359.
- [26] CHATTOPADHAY A, SARKAR A, HOWLADER P, et al. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks[C]//Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision. 2018: 839-847.
- [27] WANG Y D, ZHANG J, KAN M N, et al. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 12272-12281.
- [28] LIN M, CHEN Q, YAN S. Network in network[J]. arXiv preprint arXiv:1312.4400, 2013.
- [29] BOTTOU L. Large-scale machine learning with stochastic gradient descent[C]//Proceedings of COMPSTAT'2010. 2010.

[作者简介]



邱宝琳 (1995-), 男, 山东潍坊人, 上海交通大学硕士生, 主要研究方向为人工智能安全和对抗样本。



易平 (1969-), 男, 河南洛阳人, 博士, 上海交通大学副教授, 主要研究方向为人工智能安全和网络对抗。