第 59 卷 第 12 期/2022 年 6 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于低秩降维和稀疏重构的图像扰动防御算法

张熙凡*，于凌志

天津大学电气自动化与信息工程学院，天津 300072

摘要 在图像识别等机器视觉任务中，存在一类微弱的、不可察觉的对抗扰动，该扰动能够改变深度神经网络的输出结果。针对图像分类任务中的对抗扰动，提出了一种基于低秩降维和稀疏重构的图像对抗扰动防御算法。针对自然图像的稀疏和低秩特性，所提算法采用低秩分解削弱图像中的对抗扰动，同时利用多尺度稀疏编码对低秩图像进行重构，在滤除残余扰动的同时恢复原始图像的细节信息。采用3种攻击算法分别在黑盒攻击和灰盒攻击下验证所提算法的防御效果，并与其他4种防御算法进行了对比，实验结果表明，所提算法处理后的对抗扰动图像的Top-1分类准确率优于对比算法，且所提算法具有更好的鲁棒性。

关键词 图像处理；对抗防御；低秩降维；多尺度稀疏编码

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP202259.1210004

Image Defense Algorithm Against Adversarial Attacks Based on Low-Rank Dimensionality Reduction and Sparse Reconstruction

Zhang Xifan*, Yu Lingzhi

School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, *China*

Abstract There is a type of weak and imperceptible adversarial perturbation, which can change the output of a deep neural network in computer vision tasks such as image classification. A defense algorithm against adversarial attacks based on low-rank dimensionality reduction and sparse reconstruction is proposed to target adversarial perturbation in image classification. Because digital images are low-rank and sparse, the proposed algorithm uses low-rank decomposition to reduce adversarial perturbation. The low-rank approximated image is then subjected to multiscale sparse coding to remove residual perturbation and restore the original image's rich textural details. Three attack algorithms are used to compare the proposed algorithm's defense effect against the other four defense algorithms under black-box and gray-box attacks. The results show that the proposed algorithm processes adversarial images with the highest Top-1 accuracy of image classification compared to comparative algorithms and that the proposed algorithm is more robust.

Key words image processing; defense against adversarial attack; low-rank dimensionality reduction; multi-scale sparse coding

1 引 言

最近几年，机器学习尤其是深度学习发展迅速，在计算机视觉、语音识别等方面都有广泛应用。基于深度神经网络(DNN)的方法在某些任务中的表现甚至超过人类。但近年的研究表明，DNN极易

收稿日期：2021-04-27；**修回日期**：2021-05-21；**录用日期**：2021-06-02

基金项目：天津市科技计划(14JCQNJC01500)

通信作者：*lawalimu@163.com

受到经特殊设计的对抗扰动的攻击。Szegedy 等[1]发现，给图像添加一种不可察觉的、非随机的微弱噪声，将会改变 DNN 对该图像的分类结果。除图像分类外，Cisse 等[2]利用对抗扰动攻击了语音识别、姿态估计和语义分割等一系列应用。Kurakin 等[3]发现在现实场景中，对抗扰动的影响同样存在。对抗扰动对 DNN 输出结果的影响，已经成为 DNN 在各领域应用中（例如人脸识别应用[4]、医学应用[5]、自动驾驶应用[6]等）不可忽视的安全隐患。

近年来在图像识别领域，研究者们进行了大量的关于防御图像对抗扰动的研究。例如对图像进行 JEPG 压缩[7-9]、位深度压缩[7,10-11]、滤波[8,11-12]和稀疏编码重构[13]；或者用扰动图像对分类模型进行对抗训练[14-15]；或者修改分类模型的代价函数[14,16-17]；或者为分类模型引入可训练的预处理子网络[10,18]。

本文提出了一种基于低秩降维和稀疏重构的图像对抗扰动防御算法。所提算法由离线阶段和在线防御阶段两部分组成。离线阶段，采用无扰动自然图像（下称自然图像）学习超完备字典；在线防御阶段主要对图像进行预处理，包括对图像进行低秩降维和多尺度稀疏编码两个步骤。为削弱图像中的对抗扰动，所提算法采用非负矩阵分解（NMF）[19]获取图像的低秩表示；为滤除低秩图像中的残余扰动并重建原始图像的细节信息，对图像进行多尺度稀疏编码重构，该过程使用离线阶段学习到的超完备字典进行。使用了 5 种攻击算法攻击 VGG16 模型[20]生成对抗扰动图像并进行防御测试，被攻击的图像取自 ImageNet 数据集。实验结果表明，所提算法在 12 种黑盒攻击测试中的 10 种中的表现优于对比算法，在 12 种灰盒攻击测试中的 11 种优于对比算法，且所提算法具有更好的鲁棒性。

2 对抗扰动及防御

2.1 图像的对抗扰动

对于一张包含单一物体的图像 $\boldsymbol{X}\in\mathbf{R}^{h\times w\times c}$，其正确类别标签为 l，$f(\cdot)$ 表示 DNN 分类模型，其输入为一张图像，输出为类别标签。定义对抗扰动为 $\boldsymbol{\eta}\in\boldsymbol{R}^{h\times w\times c}$，$\boldsymbol{\eta}$ 满足 $f(\boldsymbol{X}+\boldsymbol{\eta})\neq l$ 且 $\|\boldsymbol{\eta}\|_p$ 尽量小，其中 $\|\cdot\|_p$ 表示 p 范数。对抗扰动的较小范数使其难以被肉眼察觉，但却能够改变分类模型的输出结果。为了在相同条件下衡量不同攻击算法生成的对抗扰动的攻击效果，且在相同条件的对抗扰动下对比不同的防御算法的防御效果，本实验组采用对抗扰动集的平均 2 范数作为扰动大小的评估指标。给定一个包含 N 张图像的数据集 $\{\boldsymbol{X}_i|\,i=1,2,\ldots,N\}$，对应生成的对抗扰动集为 $\{\boldsymbol{\eta}_i|\,i=1,2,\ldots,N\}$，则该对抗扰动集的扰动大小 Δ 为

$$\Delta=\frac{1}{N}\sum_{i=1}^{N}\frac{\|\boldsymbol{\eta}_i\|_2}{\|\boldsymbol{X}_i\|_2}。\tag{1}$$

2.2 对抗攻击算法

Goodfellow 等[21]提出了 fast gradient sign method（FGSM），FGSM 首先计算分类网络的代价函数关于输入的梯度，然后将输入图像的每个像素值朝梯度方向偏移微小步幅形成扰动，该微小扰动使得代价函数发生较大变化从而改变输出结果。FGSM 是一种单次迭代的快速计算对抗扰动的算法，图像的每个像素仅在使代价函数增大最快的方向扰动一次。Kurakin 等[22]在 FGSM 的基础上提出了 basic iterative method（BIM），后者与前者的区别在于 BIM 会进行多次迭代，即图像的每个像素值在代价函数的梯度方向扰动多次。同时 Kurakin 等[22]还提出了与 BIM 具有相似原理的 iterative least-likely class method（ILCM），不同点在于后者是以最不可能的类别作为目标导向进行扰动。Moosavi-Dezfooli 等[23]提出了 DeepFool，与 BIM 类似，DeepFool 也是一种多步迭代算法，区别在于 DeepFool 在每次迭代时，使图像朝着距离决策边界最近的方向扰动。如果将输入图像 $\boldsymbol{X}\in\mathbf{R}^{h\times w\times c}$ 视为超空间 $\mathbf{R}^{h\times w\times c}$ 中的一个点，则 DNN 分类模型的映射函数 $f(\cdot)$ 把整个输入空间划分为 M 个区域（M 为类别总数），每个区域内的点都会输出该区域的类别标签。DeepFool 算法使得图像点每次朝距离该区域决策边界最近的方向扰动，直到其越过决策边界，导致分类错误。Carlini 等[24]提出了一种优化的 C&W 方法，该方法在攻击一些防御性蒸馏模型时，效果优于 DeepFool。

此外，Papernot 等[25]提出的 jacobian saliency map attack（JSMA）根据网络 logit 层输出计算得到的对抗性雅可比显著图实施扰动。Xiao 等[26]利用 generative adversarial network（GAN）生成对抗扰动。Su 等[27]提出了只扰动一个像素点的攻击方法 one pixel attack。Moosavi-Dezfooli 等[28]提出了一种计算通用对抗扰动（UAP）的方法。

2.3 对抗防御算法

根据防御思路的不同，现有的对抗防御算法可

分为三种类型。第1类是对神经网络进行对抗训练，通过使模型学习到对抗扰动的特征来增强鲁棒性。Zheng等[14]将原训练图像和加高斯噪声的训练图像成对输入到分类网络进行对抗训练。Metzen等[15]将原图像标记为正样本，扰动图像标记为负样本，训练对抗扰动检测器。对抗训练能提高神经网络对于对抗扰动样本的鲁棒性但效果不显著，且需耗费额外计算资源生成对抗扰动样本并进行训练。第2类是修改网络激活函数或改变、添加网络结构，从而平滑模型，增强模型稳定性。Zantedeschi等[16]将原网络Relu函数修改为定上限的Relu函数来限制扰动的逐层放大。Ross等[17]在原代价函数中增加了该函数的梯度正则项，平滑了其在输入空间的梯度。Samangouei等[18]通过训练GAN的生成器重构输入图像。此类方法耗费的额外计算资源较少，但其在面对灰盒攻击或白盒攻击时不够鲁棒。第3类方法是在图像分类前对图像进行预处理，利用预处理算法削弱对抗扰动。Jia等[10]压缩并重构子网络，先将图像位压缩至12 bit，再将图像重构。Guo等[7]用图像裁剪缩放、JPEG压缩、位压缩、总方差最小化和图像缝合5种图像处理算法去除扰动。Prakash等[12]基于鲁棒性激活图选择像素点进行邻域替换，然后对图像进行小波滤波。Xu等[11]采用位压缩和空间平滑滤波处理图像。此类方法一般不受攻击算法、分类模型和数据集的影响，在面对灰盒攻击或白盒攻击时具有较好的鲁棒性，同时由于其端到端梯度不易计算，故对于基于梯度进行的二次攻击也有较好的防御性。所提防御算法属于第3类。

3　基于低秩降维和稀疏重构的防御算法

Szegedy等[1]认为，对抗扰动在自然数据集中出现的概率极低，即对抗扰动的特征不同于自然图像的特征。基于这一区别，所提算法首先利用自然图像的低秩特性获得图像的低秩表示，过滤不同于自然图像的非低秩对抗扰动。但此时的低秩图像仍具有扰动残留，同时对图像进行低秩降维的过程也会损失原图像的部分细节信息，故所提算法对低秩图像进行多尺度稀疏编码重构。重构过程使用的字典由自然图像学习得到，保证字典原子无扰动且可组合出自然图像的细节信息。所提算法的完整流程如图1所示，图1上半部分为算法离线阶段学习字典的过程，具体包括：用训练集中的无扰动自然图像构建拉普拉斯金字塔；用金字塔相应层级提取的若干图像块学习相应层级的字典。图1下半部分为算法在线阶段的图像预处理过程，具体包括：用NMF获取扰动图像的低秩表示；将该低秩表示建立拉普拉斯金字塔并用离线阶段学得的字典对金字塔的每一层进行稀疏编码重构，最终得到重构图像。

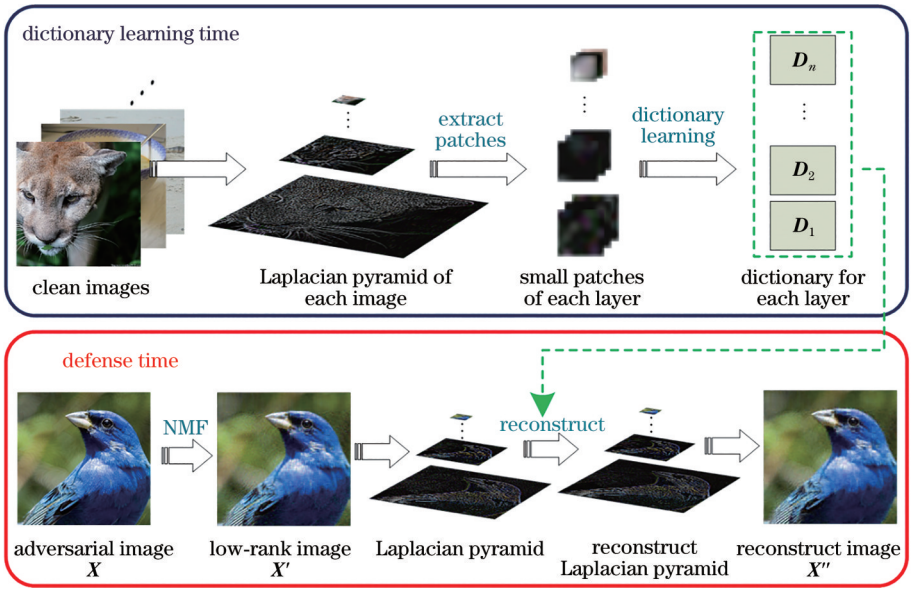

图1　所提算法流程

Fig. 1　Flow chart of proposed algorithm

3.1　低秩降维

自然图像的内容具有低秩特性，当对图像施加恶意扰动后，图像的低秩特性会遭到破坏。对扰动图像进行低秩降维能够获取该图像的低秩表示，将

图像的主要信息保留，同时在一定程度上削弱对抗扰动，从而提高分类准确率。所提算法采用NMF进行低秩降维。对于图像$\boldsymbol{X}\in\mathbf{R}^{h\times w\times c}$的某一通道$\boldsymbol{X}_c\in\mathbf{R}^{h\times w}$，其低秩表示为

$$\boldsymbol{X}_c'=\boldsymbol{W}\boldsymbol{H},\tag{2}$$

式中：基矩阵$\boldsymbol{W}\in\mathbf{R}^{h\times r}$；系数矩阵$\boldsymbol{H}\in\mathbf{R}^{r\times w}$；$r$为秩。分解过程产生的误差用分解前后的平方欧氏距离来衡量，即

$$e=\left\|\boldsymbol{X}_c-\boldsymbol{X}_c'\right\|^2=\sum_{i=0}^{h}\sum_{j=0}^{w}\left[\left(\boldsymbol{X}_c\right)_{ij}-\left(\boldsymbol{W}\boldsymbol{H}\right)_{ij}\right]^2。\tag{3}$$

在本实验中，需要调节r将e控制在合适范围，使该分解既削弱图像的对抗扰动，又尽可能保留原始图像的信息。在分解过程中，按照Lee等[19]提出的更新规则对$\boldsymbol{W}$和$\boldsymbol{H}$进行迭代更新，更新规则为

$$\boldsymbol{H}_{ai}\leftarrow\boldsymbol{H}_{ai}\frac{\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}_c\right)_{ai}}{\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{H}\right)_{ai}},\tag{4}$$

$$\boldsymbol{W}_{ja}\leftarrow\boldsymbol{W}_{ja}\frac{\left(\boldsymbol{X}_c\boldsymbol{H}^{\mathrm{T}}\right)_{ja}}{\left(\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}\right)_{ja}},\tag{5}$$

式中：$a=1,2,\ldots,r$；$i=1,2,\ldots,w$；$j=1,2,\ldots,h$。对图像每个通道分别进行NMF分解，最终得到低秩图像$\boldsymbol{X}'$。

3.2　多尺度稀疏编码重构

图像的低秩降维过程会损失图像细节，对抗扰动和原始图像信息会同时受到影响。使用NMF分解图像时无法保证既大量削弱对抗扰动，又大量保留原始图像信息。作为该过程中的一对矛盾，在削弱对抗扰动时，必定伴随部分原始图像信息的丢失，这会影响图像的分类准确率。为解决这一矛盾，所提算法使用稀疏编码对低秩图像进行重构，为其补充原始图像的细节信息，同时滤除残余的对抗扰动。

所提算法在重构前将低秩图像$\boldsymbol{X}'$构建为n层拉普拉斯金字塔，然后在其图像金字塔的多尺度状态下逐层重构。已有研究表明，对抗扰动具有较高的尺度敏感性。例如，Guo等[7]的实验结果表明，对扰动图像进行裁剪、缩放等操作可以有效削弱对抗扰动的攻击性；Xie等[29]先对扰动图像进行随机缩小，再随机填充像素至原来大小，经此预处理后的扰动图像的分类准确率也获得了有效的提升。针对这一特点，重构前的多尺度处理使得重构过程具有对残余扰动的自适应性，从而更好地滤除残余扰动。图2展示了$n=3$时将低秩图像$\boldsymbol{X}'$进行多尺度重构的流程，其中$\boldsymbol{X}''$为重构图像，$\uparrow_2$表示对图像进行上采样，$\downarrow_2$为下采样。

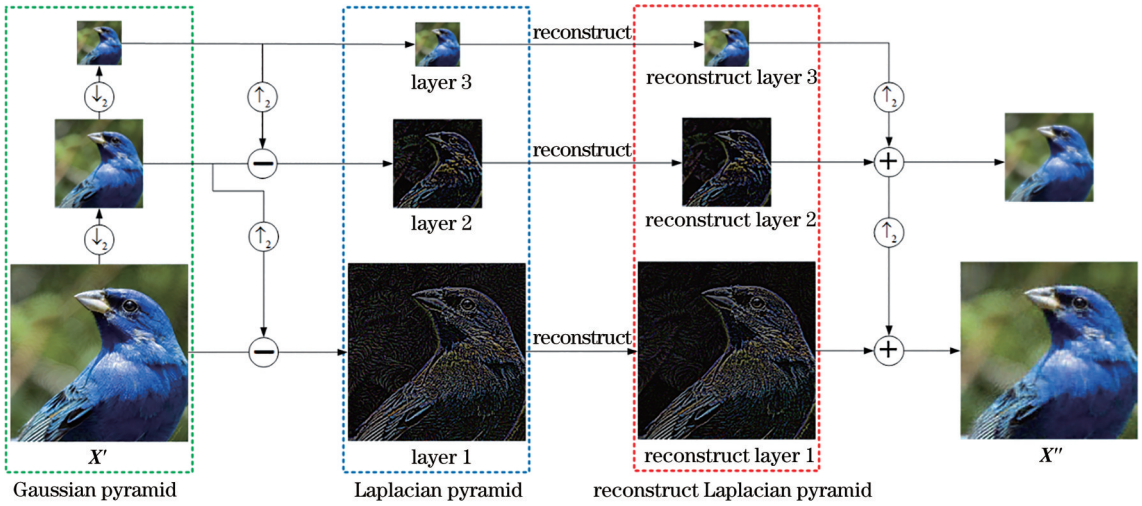

图2　多尺度重构

Fig. 2　Multi-scale reconstruction

稀疏编码重构时使用的字典是用自然图像集（下称训练集）学习得到的。该字典为超完备字典，其字典原子集包含训练图像的特征集。以该字典原子的线性组合能重构出自然图像的细节信息。字典$\boldsymbol{D}$通过K-singular value decomposition（KSVD）算法[30]学习得到，该算法解决如下优化问题：

$$\min_{\boldsymbol{D},\boldsymbol{A}}\left\|\boldsymbol{Y}-\boldsymbol{D}\boldsymbol{A}\right\|_{\mathrm{F}}^2\ \mathrm{s.t.}\left\|\boldsymbol{a}\right\|_0=t,\tag{6}$$

式中：$\boldsymbol{Y}$为由若干自然图像块构成的信息矩阵；$\boldsymbol{A}$为系数矩阵；$\boldsymbol{a}$表示系数矩阵的列向量；t为稀疏度约束，即为重构信息所使用的原子数量。获得字典$\boldsymbol{D}$

后，便可使用 orthogonal matching pursuit（OMP）算法[31]和字典 $\boldsymbol{D}$ 重构图像，该算法解决如下优化问题：

$$\min_{\boldsymbol{a}}\left\|\boldsymbol{Y}'-\boldsymbol{Y}''\right\|_{\mathrm{F}}^{2}=\min_{\boldsymbol{a}}\left\|\boldsymbol{Y}'-\boldsymbol{D}\boldsymbol{a}\right\|_{\mathrm{F}}^{2}\ \text{s.t.}\left\|\boldsymbol{a}\right\|_{0}=t,\tag{7}$$

式中：$\boldsymbol{Y}'\in\mathbf{R}^{3d^2}$ 为低秩图像的图像块向量，图像块的大小为 d；$\boldsymbol{Y}''$ 为其重构。在式(7)中，图像块的重构目标为低秩图像信息 $\boldsymbol{Y}'$，式(6)中字典学习的目标为自然图像信息 $\boldsymbol{Y}$，故当式(7)中的稀疏度约束 t 与式(6)保持一致时，字典原子的线性组合重构 $\boldsymbol{Y}''$ 将具有最接近自然图像的细节信息。由于低秩图像的多尺度处理，对应地，所提算法也将对金字塔的每一层分别进行字典学习和重构。在线防御阶段算法流程如图 3 所示。

Proposed defense Algorithm

Input: Adversarial image $\boldsymbol{X}$, number of pyramid levels n, patch size $\{d_l\,|\,l=1,2,\ldots,n\}$, sparsity $\{t_l\,|\,l=1,2,\ldots,n\}$, dictionary $\{\boldsymbol{D}_l\,|\,l=1,2,\ldots,n\}$

Output: Reconstruct image

for each channel of $\boldsymbol{X}$ **do** $\boldsymbol{X}'_{\mathrm{c}}\leftarrow\boldsymbol{W}\boldsymbol{H}$

Combine each channel and get $\boldsymbol{X}'$

for layer $l=1\rightarrow n$ **do**

　　if $l<n$ **do** $\boldsymbol{X}'_l\leftarrow\downarrow_2^{(l-1)}(\boldsymbol{X}')-\uparrow_2\left[\downarrow_2^{(l)}(\boldsymbol{X}')\right]$

　　else do $\boldsymbol{X}'_l\leftarrow\downarrow_2^{(l)}(\boldsymbol{X}')$

for each layer l **do**

　　Partition image $\boldsymbol{X}'_l$ into patches

　　for each vectorized patch $\boldsymbol{Y}'\in\mathbf{R}^{3d_l^2}$ **do**

　　$\boldsymbol{a}\leftarrow\min_{\boldsymbol{a}}\left\|\boldsymbol{Y}'-\boldsymbol{D}_l\boldsymbol{a}\right\|_{\mathrm{F}}^{2}\ \text{s.t.}\left\|\boldsymbol{a}\right\|_0=t_l$

　　$\boldsymbol{Y}''\leftarrow\boldsymbol{D}_l\boldsymbol{a}$

　　Combine reconstruct patch vectors into image $\boldsymbol{X}''_l$

for layer $l=n\rightarrow 1$ **do**

　　if $l=n$ **do** $\boldsymbol{X}_{\text{temp}}\leftarrow\boldsymbol{X}''_l$

　　else do $\boldsymbol{X}_{\text{temp}}\leftarrow\boldsymbol{X}''_l+\uparrow_2\left(\boldsymbol{X}_{\text{temp}}\right)$

$\boldsymbol{X}''\leftarrow\boldsymbol{X}_{\text{temp}}$

图 3　所提防御算法

Fig. 3　Proposed defense algorithm

4　实　　验

4.1　实验设置

实验选择预训练的 VGG16[20]和 ResNet50[32]作为分类模型，它们在 ImageNet[33]测试集上的 Top-1 分类准确率分别为 71.5% 和 75.2%。字典学习训练集使用 ImageNet 数据集中随机挑选的 1000 张图像。为了验证所提算法的鲁棒性，实验一共生成 12 类测试集，并采用 FGSM、BIM 和 DeepFool 3 种攻击算法，分别攻击 VGG16 模型，分别生成 4 个不同层次扰动大小（$\Delta=0.01, 0.02, 0.03, 0.04$）的测试集，攻击算法通过开源工具 Cleverhans[34]实现，其攻击图像均来自 ImageNet 数据集。为了能更直观地展示防御算法的有效性，每个测试集的图像均为筛选后的 1000 张能攻击成功的扰动图像，即所有测试集在 VGG16 模型上的分类准确率均为 0。为了对比所提算法的防御效果，在与上述测试环境完全相同的条件下选择了其他 4 种防御算法进行对比实验，分别是 JPEG 压缩（JPEG）[7]、总方差最小化（TVM）[7]、pixel deflection and wavelet denoise（PDWD）[12]和 Comdefend[10]，前三种算法属于传统图像处理算法，后一种属于基于卷积神经网络的图像处理算法，以上对比算法均使用其对应文献[7, 10, 12]提供的开源代码和最佳参数设置进行实验。5 种算法都在黑盒攻击和灰盒攻击两种攻击模式下分别进行防御实验。

本实验的相关参数中，拉普拉斯金字塔的层数 n 决定了图像分解中的尺度范围，为均衡计算复杂度和对细节的描述能力，将 n 设置为 3。参数 r、稀疏编码中的字典原子数 k、单位图像块大小 d 和稀疏度约束 t 影响图像重构过程对原始图像细节信息的重建程度，这些参数分别通过二分搜索法在其合理范围内选择具有最优防御效果的取值，r 设定为 40，其他参数如表 1 所示。

表 1 稀疏编码的参数设置

Table 1 Parameter settings for sparse coding

Layer 1			Layer 2			Layer 3		
k	*d*	*t*	*k*	*d*	*t*	*k*	*d*	*t*
1000	14	6	300	7	5	100	4	4

4.2 实验结果及分析

防御实验在黑盒攻击和灰盒攻击两种模式下进行。黑盒攻击为攻击方在未知防御方防御策略、分类模型及参数等任何信息的情况下进行的攻击。为了模拟黑盒攻击场景，设定防御方使用ResNet50进行分类，而攻击方在未知防御方任何信息的情况下攻击VGG16模型。灰盒攻击为攻击方在已知防御方防御策略或分类模型及参数等部分信息的情况下有所针对的攻击。为了模拟灰盒攻击场景，设定防御方使用VGG16进行分类，而攻击方知晓了这一信息，针对性地对VGG16模型进行攻击生成扰动图像。在12类测试集中，5种算法在防御黑盒攻击实验和防御灰盒攻击实验中的Top-1分类准确率结果如表2所示，其中“/”左侧为防御黑盒攻击准确率，右侧为防御灰盒攻击准确率，粗体表示表现最好的数据。

表 2 各防御算法的 Top-1 分类准确率。

Table 2 Top-1 classification accuracy of each defense algorithm unit: %

Attack algorithm	Δ	Proposed algorithm	JPEG	TVM	PDWD	ComDefend
FGSM	0.01	**36.4/34.0**	30.8/28.9	35.0/32.8	28.0/26.5	31.5/27.3
	0.02	34.2/**32.3**	27.0/18.0	**34.6**/30.0	26.1/19.0	27.0/20.1
	0.03	**31.1/26.0**	18.6/13.4	30.9/24.0	22.7/14.1	21.3/13.3
	0.04	**29.2/21.9**	15.4/12.0	28.7/21.5	17.1/10.4	16.4/11.7
	Average	**32.7/28.6**	23.0/19.6	32.3/27.1	23.5/17.5	24.1/18.1
BIM	0.01	38.1/**43.5**	37.4/32.7	**40.0**/41.0	35.0/35.5	37.4/31.2
	0.02	**44.7**/39.7	29.2/18.0	42.0/**40.7**	36.8/27.3	33.9/25.6
	0.03	**46.9/35.8**	19.6/10.9	43.9/32.6	30.7/15.0	26.3/16.9
	0.04	**43.4/30.0**	14.3/7.3	40.5/29.5	18.9/8.1	23.7/10.2
	Average	**43.3/37.3**	25.1/17.2	41.6/36.0	30.4/21.5	30.3/21.0
DeepFool	0.01	**49.1/46.4**	35.8/31.2	43.0/38.1	33.3/24.7	38.8/29.5
	0.02	**43.1/33.4**	23.0/17.9	40.9/30.6	26.4/18.9	31.1/21.6
	0.03	**38.5/26.7**	19.4/11.6	37.7/25.6	23.0/12.1	22.1/16.2
	0.04	**32.9/20.8**	15.4/7.5	18.9/14.1	16.3/8.6	16.6/11.0
	Average	**40.9/31.8**	23.4/17.1	35.1/27.1	24.8/16.1	27.1/19.6
Total average		**39.0/32.6**	23.8/18.0	36.3/30.1	26.2/18.4	27.2/19.6

从表2可以看出：在防御黑盒攻击实验中，所提算法在10个测试集上的表现优于对比算法；在灰盒攻击实验中，所提算法在11个测试集上取得了领先；而从总平均分类准确率来看，所提算法总体优于其他4种对比算法，总体领先第2名TVM算法2.7/2.5个百分点。在4种对比算法中，JPEG和TVM是与所提算法较为相似的算法，因为它们在图像处理时完全不依赖分类模型和任何攻击方的信息，但它们处理对抗扰动的基本原理均是通过较大幅度改变图像信息从而同时改变对抗扰动信息使其失效，这一改变会以损失原图像的细节信息为代价而降低处理后的图像分类准确率。所提算法对低秩图像的多尺度稀疏编码能够补充原图像细节信息，有效地解决了这一问题。表3展示了所提算法的消融实验结果，表中“NMF”表示只对图像进行NMF处理，“NMF+MSC”表示对图像先进行NMF处理再进行多尺度稀疏编码重构。实验结果表明，多尺度稀疏编码处理对于图像分类准确率的提升作用显著。另外两种对比算法PDWD和Comdefend在实施防御时依赖分类模型或攻击方的信息。PDWD在像素邻域替换阶段依赖于robust class activation map (R-CAM)[35]，而R-CAM的生成会受到攻击算法的影响。Comdefend则需要预先训练图像压缩和重构网络，这会受到训练时所使用的分类模型的影响。这些额外的信息依赖使这两种算法的泛化性受到影响，因此在本实验中取得较一般的分类准确率结果。观察各防御对单独某一种攻击算法的效果可以发现，当扰动增大时，4种对比算法除TVM以外，防御效果都显著变差。所提算法则在对3种攻击算法的防御中都取得了最高的平

表 3　NMF 和 NMF＋MSC 的 Top-1 分类准确率

Table 3　Top-1 classification accuracy of NMF and NMF＋MSC　　unit: %

Attack algorithm	Δ	NMF	NMF＋MSC	Variety
FGSM	0.01	31.4/32.0	36.4/34.0	5.0/2.0
	0.02	28.8/27.6	34.2/32.3	5.4/4.7
	0.03	28.6/21.3	31.1/26.0	2.5/4.7
	0.04	26.0/19.1	29.2/21.9	3.2/3.8
BIM	0.01	36.7/37.3	38.1/43.5	1.4/6.2
	0.02	39.3/33.9	44.7/39.7	5.4/5.8
	0.03	39.6/30.4	46.9/35.8	7.3/5.4
	0.04	37.1/23.3	43.4/30.0	6.3/6.7
DeepFool	0.01	41.2/36.9	49.1/46.4	7.9/9.5
	0.02	37.5/27.6	43.1/33.4	5.5/5.8
	0.03	31.9/22.0	38.5/26.7	6.6/4.7
	0.04	26.0/16.8	32.9/20.8	6.9/4.0

均分类准确率，说明所提算法的鲁棒性优于其他对比算法。由于图像的多尺度处理，所提算法具有较强的对扰动的自适应性。

灰盒攻击比黑盒攻击的防御难度大，故本实验中几乎所有灰盒攻击下的分类准确率都要低于黑盒攻击下的分类准确率。表 4 展示了攻击模式由黑盒转换为灰盒时各对应分类准确率的变化率，变化率的表达式为

$$R_{\text{change}}=\frac{P_{\text{h}}-P_{\text{b}}}{P_{\text{b}}},\tag{8}$$

式中：P_{h} 为灰盒准确率；P_{b} 为黑盒准确率。该变化率能够体现防御算法面对不同攻击模式时的鲁棒性，由于对灰盒攻击的防御难度更大，通常情况下变化率为负值，故变化率的算数值越大，说明对应的算法更鲁棒。

从表 4 可以看出，所提算法在 12 类测试集中的 6 类具有最大的变化率，TVM 其次，在其中 3 类取得最大变化率。PDWD 和 Comdefend 由于对额外信息的依赖，攻击方可以针对其依赖信息，削弱其防御效果。例如本实验的灰盒攻击模拟，攻击方知晓防御方的 VGG16 模型信息，也采用该模型作为攻击目标，这影响了 PDWD 中 R-CAM 的生成和 Comdefend 中的图像压缩和重构网络的效果。所以当攻击模式由黑盒攻击转变为灰盒攻击时，PDWD 和 Comdefend 的分类准确率大幅度减小，总平均分类准确率的变化率都接近 30%。得益于多尺度稀疏编码的处理和没有额外信息的依赖，所提算法在面对不同的攻击模式时具有最佳的鲁棒性。

表 4　黑盒攻击转变为灰盒攻击后对应 Top-1 分类准确率的变化率

Table 4　Change rate of Top-1 classification accuracy after a black box attack is converted to a gray box attack　　unit: %

Attack algorithm	Δ	Proposed algorithm	JPEG	TVM	PDWD	ComDefend
FGSM	0.01	−6.6	−6.2	−6.3	**−5.4**	−13.3
	0.02	**−5.5**	−33.3	−13.3	−27.2	−25.6
	0.03	**−16.4**	−28.0	−22.3	−37.9	−37.6
	0.04	−25.0	**−22.1**	−25.1	−39.2	−28.7
	Average	**−12.5**	−14.8	−16.1	−25.5	−24.9
BIM	0.01	**14.2**	−12.6	2.5	1.4	−16.6
	0.02	−11.2	−37.7	**−3.1**	−25.8	−24.5
	0.03	**−24.3**	−51.5	−25.7	−51.1	−35.7
	0.04	−30.9	−49.0	**−27.2**	−57.1	−57.0
	Average	−13.8	−31.5	**−13.5**	−29.3	−30.7
DeepFool	0.01	**−5.5**	−12.8	−12.0	−25.8	−23.9
	0.02	**−22.5**	−22.2	−25.2	−28.4	−30.5
	0.03	−30.6	−40.2	−32.1	−47.4	**−26.7**
	0.04	−36.7	−51.3	**−25.4**	47.2	−33.7
	Average	**−22.2**	−26.9	−22.8	−35.1	−27.7
Total average		**−16.4**	−24.4	−17.1	−29.8	−27.9

[图 4(a)]中，类别标签为“Indigo bird”的无扰动自然图像被 VGG16 模型以 26.1% 的置信度将其分类为“Indigo bird”；[图 4(b)]中，BIM 算法攻击后的扰动图像被 VGG16 模型以 26.4% 的置信度将其分

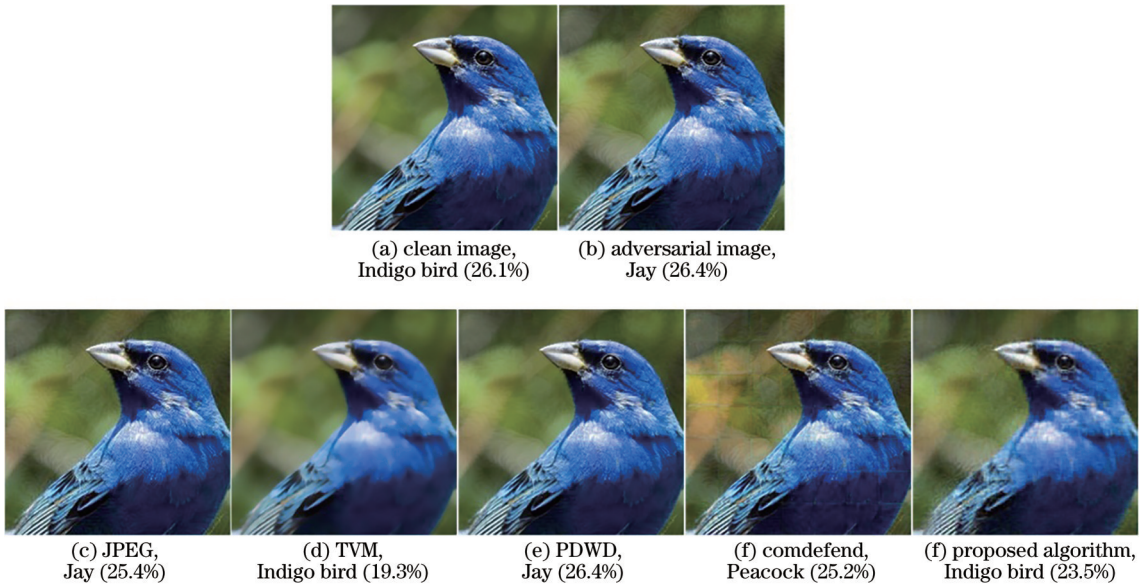

图 4　各算法的防御效果展示

Fig. 4　Defense effects of each algorithm

类为“Jay”;[图 4(c)～(g)]分别为 4 种对比算法及所提算法对扰动图像进行处理后的图像以及 VGG16 模型对其分类结果和置信度。结果表明,只有所提算法和 TVM 能够让该扰动图像在处理后被重新正确地分类为“Indigo bird”,且所提算法处理后的图像在该类别上的置信度比 TVM 高出 4.2 个百分点。

为了模拟真实场景下的随机攻击情形,额外增加了 ILCM[22]和 C&W[24]两种攻击算法,在 VGG19[20]、ResNet101[32]和 Inception V3[36]3 种分类模型上对所提算法进行了扩展实验,3 种分类模型在 Imagenet 测试集上的 Top-1 分类准确率分别为 72.4%、77.4% 和 77.3%。测试集为 FGSM、BIM、DeepFool、ILCM 和 C&W 5 种攻击算法分别攻击 VGG16 模型生成的扰动图像集,每种攻击算法生成 1000 张扰动图像(扰动大小从 0.01～0.04,每个级别各 250 张),所有测试集在 VGG16 模型中的分类准确率均为 0。5 个测试集经过所提算法预处理后,在 3 种分类模型上的分类准确率如表 5 所示。

从表 5 可以看出,所提算法在面对多种攻击算法以及应用于多种分类模型时,Top-1 分类准确率稳定在 40% 左右,进一步体现了所提算法的鲁棒性。

表 5　所提算法在扩展实验中的 Top-1 分类准确率

Table 5　Top-1 classification accuracy of proposed algorithm in extended experiments　　unit: %

Model	FGSM	BIM	DeepFool	ICLM	C&W	Average
VGG19	30.1	38.2	33.0	40.1	31.6	34.6
ResNet101	34.9	44.8	41.2	43.7	39.4	40.8
Inception V3	35.4	43.1	42.9	42.8	41.7	41.2
Average	33.5	42.0	39.0	42.2	37.6	**38.9**

5　结　　论

提出了一种基于图像低秩降维和多尺度稀疏编码重构的图像扰动防御算法。所提算法采用非负矩阵分解获取图像的低秩表示,再用稀疏编码对图像的拉普拉斯金字塔的每一层进行重构,处理完成后再将其输入分类模型。与其他防御算法相比,所提算法针对图像的不同尺度进行了差别处理,使图像在稀疏编码重构过滤残余扰动的同时更加接近原始自然图像,在攻击算法及扰动大小发生变化时,防御性能更加稳定。实验结果表明,不论是面对黑盒攻击还是灰盒攻击,所提算法都具有更好的防御效果,且在攻击算法、对抗扰动大小发生改变时,所提算法也具有比其他 4 种防御算法更好的 Top-1 分类准确率,鲁棒性更佳。

参　考　文　献

[1] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[EB/OL]. (2013-12-21) [2021-05-07]. https://arxiv. org/abs/1312.6199.

[2] Cisse M M, Adi Y, Neverova N, et al. Houdini: fooling deep structured visual and speech recognition models with adversarial examples[C]//Advances in Neural Information Processing Systems, December 4-9, 2017, Long Beach, CA, USA. [S. l.: s. n.],

2017: 6977-6987.

[3] Kurakin A, Goodfellow I J, Bengio S. Adversarial machine learning at scale[EB/OL]. (2016-11-04) [2021-06-04]. https://arxiv.org/abs/1611.01236.

[4] 王嘉欣, 雷志春. 一种基于特征融合的卷积神经网络人脸识别算法[J]. 激光与光电子学进展, 2020, 57(10): 101508.
Wang J X, Lei Z C. A convolutional neural network based on feature fusion for face recognition[J]. Laser & Optoelectronics Progress, 2020, 57(10): 101508.

[5] 李智唯, 曹慧, 杨锋, 等. 基于卷积神经网络的脑肿瘤分割研究进展[J]. 激光与光电子学进展, 2021, 58(24): 2400003.
Li Z W, Cao H, Yang F, et al. Research progress of brain tumor segmentation based on convolutional neural network[J]. Laser & Optoelectronics Progress, 2021, 58(24): 2400003.

[6] 周苏, 吴迪, 金杰. 基于卷积神经网络的车道线实例分割算法[J]. 激光与光电子学进展, 2021, 58(8): 0815007.
Zhou S, Wu D, Jin J. Lane instance segmentation algorithm based on convolutional neural network[J]. Laser & Optoelectronics Progress, 2021, 58(8): 0815007.

[7] Guo C, Rana M, Cisse M, et al. Countering adversarial images using input transformations[EB/OL]. (2017-10-31) [2021-05-08]. https://arxiv. org/abs/1711.00117.

[8] Shaham U, Garritano J, Yamada Y, et al. Defending against adversarial images using basis functions transformations[EB/OL]. (2018-03-28) [2021-05-04]. https://arxiv.org/abs/1803.10840.

[9] Dziugaite G K, Ghahramani Z, Roy D M. A study of the effect of JPG compression on adversarial images [EB/OL]. (2016-08-02) [2021-05-01]. https://arxiv.org/abs/1608.00853.

[10] Jia X J, Wei X X, Cao X C, et al. ComDefend: an efficient image compression model to defend adversarial examples[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 6077-6085.

[11] Xu W L, Evans D, Qi Y J. Feature squeezing: detecting adversarial examples in deep neural networks[C]//25th Annual Network and Distributed System Security Symposium, NDSS 2018, February 18-21, 2018, San Diego, California, USA. Reston: Internet Society, 2018.

[12] Prakash A, Moran N, Garber S, et al. Deflecting adversarial attacks with pixel deflection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 8571-8580.

[13] Sun B, Tsai N H, Liu F C, et al. Adversarial defense by stratified convolutional sparse coding[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 11439-11448.

[14] Zheng S, Song Y, Leung T, et al. Improving the robustness of deep neural networks via stability training[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 4480-4488.

[15] Metzen J H, Genewein T, Fischer V, et al. On detecting adversarial perturbations[C]//5th International Conference on Learning Representations, April 24-26, 2017, Toulon, France. [S.l.: s.n.], 2017.

[16] Zantedeschi V, Nicolae M I, Rawat A. Efficient defenses against adversarial attacks[C]//AISec '17: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, November 3, 2017, Dallas, TX, USA. New York: ACM Press, 2017: 39-49.

[17] Ross A S, Doshi-Velez F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients[C]// Proceedings of 2018 AAAI Conference on Artificial Intelligence, February 2-7, 2018, New Orleans, Louisiana, USA. Menlo Park: AAAI Press, 2017: 1660-1669.

[18] Samangouei P, Kabkab M, Chellappa R. Defense-GAN: protecting classifiers against adversarial attacks using generative models[C]//6th International Conference on Learning Representations, April 30-May 3, 2018, Vancouver, BC, Canada. [S.l.: s.n.], 2018.

[19] Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization[J]. Nature, 1999, 401(6755): 788-791.

[20] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]//3rd International Conference on Learning Representations, May 7-9, 2015, San Diego, CA, USA. [S.l.: s.n.],

2015.

[21] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[C]//3rd International Conference on Learning Representations, ICLR 2015-Conference Track Proceedings, May 7-9, 2015, San Diego, CA, USA. [S.l.: s.n.], 2015.

[22] Kurakin A, Goodfellow I J, Bengio S. Adversarial examples in the physical world[EB/OL]. (2016-07-08) [2021-05-04]. https://arxiv. org/abs/1607.02533v4.

[23] Moosavi-Dezfooli S M, Fawzi A, Frossard P. DeepFool: a simple and accurate method to fool deep neural networks[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 2574-2582.

[24] Carlini N, Wagner D. Towards evaluating the robustness of neural networks[C]//2017 IEEE Symposium on Security and Privacy, May 22-26, 2017, San Jose, CA, USA. New York: IEEE Press, 2017: 39-57.

[25] Papernot N, McDaniel P, Jha S, et al. The limitations of deep learning in adversarial settings [C]//2016 IEEE European Symposium on Security and Privacy (EuroS&P), March 21-24, 2016, Saarbruecken, Germany. New York: IEEE Press, 2016: 372-387.

[26] Xiao C W, Li B, Zhu J Y, et al. Generating adversarial examples with adversarial networks[C]// Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, July 13-19, 2018. Stockholm, Sweden. Menlo Park: International Joint Conferences on Artificial Intelligence Organization, 2018: 3905-3911.

[27] 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.

[28] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal adversarial perturbations[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 86-94.

[29] Xie C H, Wang J Y, Zhang Z S, et al. Mitigating adversarial effects through randomization[C]//6th International Conference on Learning Representations, April 30-May 3, 2018, Vancouver, BC, Canada. [S.l.: s.n.], 2018.

[30] Aharon M, Elad M, Bruckstein A. K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation[J]. IEEE Transactions on Signal Processing, 2006, 54(11): 4311-4322.

[31] Pati Y C, Rezaiifar R, Krishnaprasad P S. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition[C]//Proceedings of 27th Asilomar Conference on Signals, Systems and Computers, November 1-3, 1993, Pacific Grove, CA, USA. New York: IEEE Press, 1993: 40-44.

[32] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[33] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[34] Papernot N, Faghri F, Carlini N, et al. Technical report on the CleverHans v2.1.0 adversarial examples library[EB/OL]. (2016-10-03) [2021-06-02]. https://arxiv.org/abs/1610.00768.

[35] Zhou B L, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 2921-2929.

[36] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 2818-2826.