计算机科学
COMPUTER SCIENCE

基于高频特征掩蔽的对抗攻击算法

王柳依, 周淳, 曾文强, 何星星, 孟华

引用本文

王柳依, 周淳, 曾文强, 何星星, 孟华. 基于高频特征掩蔽的对抗攻击算法[J]. 计算机科学, 2025, 52(10): 374-381.

WANG Liuyi, ZHOU Chun, ZENG Wenqiang, HE Xingxing, MENG Hua. High-frequency Feature Masking-based Adversarial Attack Algorithm [J]. Computer Science, 2025, 52(10): 374-381.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241000030

基于高频特征掩蔽的对抗攻击算法

王柳依[1]　周　淳[2]　曾文强[2]　何星星[2]　孟　华[2]
1 西南交通大学信息科学与技术学院　成都 611756
2 西南交通大学数学学院　成都 611756
(wangly202410@163.com)

摘　要　深度神经网络在图像识别领域取得广泛应用,但其结构复杂,容易受到对抗样本的攻击。构造人眼不可察觉的对抗样本,对测试网络的安全性有着重要的意义。现有针对图像的对抗样本生成方法通常是对原始样本进行微小扰动,而扰动通常用 l_p 范数距离进行约束,这种简单方案将所有像素点平等对待,每个点允许扰动的范围满足同样的约束,这限制了对抗样本的构造方式,使得扰动易被人眼察觉。而在现实应用中,人眼对不同颜色和区域的像素点扰动的敏感性亦不相同。针对这一特点,设计了一种基于观测敏感性的自适应扰动方案,为不同的像素点设计不同的扰动约束,从而提升对抗样本的鲁棒性。具体而言,该方法通过频谱分析将图像划分为高频和低频区域,并通过新的空间约束规范扰动,对高频不敏感区域增加更大的扰动,以提升对抗能力。基于 ImageNet-1K 和 CIFAR-10 数据集进行的一系列实验表明,新的对抗样本构造策略能与多种攻击方法相耦合,并在保障隐蔽性的前提下提升对抗性能。

关键词:深度神经网络;对抗样本;高频;扰动;鲁棒性
中图分类号　TP183

High-frequency Feature Masking-based Adversarial Attack Algorithm

WANG Liuyi[1],ZHOU Chun[2],ZENG Wenqiang[2],HE Xingxing[2] and MENG Hua[2]
1 School of Information Science and Technology,Southwest Jiaotong University,Chengdu 611756,China
2 School of Mathematics,Southwest Jiaotong University,Chengdu 611756,China

Abstract　Deep neural networks have achieved widespread application in the field of image recognition,however,their complex structures make them vulnerable to adversarial attacks. Constructing adversarial examples that are imperceptible to the human eye is crucial for evaluating the security of these networks. Existing adversarial example generation methods for images typically involve small perturbations to the original samples,constrained by l_p-norms. This simplistic approach treats all pixels equally,applying uniform constraints to the allowable perturbations at each pixel,which limits the flexibility of adversarial example generation and makes the perturbations more detectable to the human eye. In practical applications,human visual sensitivity varies across different colors and image regions. To address this issue,this paper proposes an adaptive perturbation scheme based on perceptual sensitivity,where different perturbation constraints are applied to different pixels,thereby enhancing the robustness of the adversarial examples. Specifically,this method employs spectral analysis to divide the image into high-frequency and low-frequency regions and applies novel spatial constraints to regulate perturbations. Larger perturbations are introduced in regions less sensitive to high-frequency changes,improving adversarial effectiveness. Extensive experiments conducted on the ImageNet-1K and CIFAR-10 datasets demonstrate that the proposed adversarial example generation strategy can be coupled with various attack me-thods,significantly enhancing adversarial performance while ensuring imperceptibility.

Keywords　Deep neural networks,Adversarial examples,High-frequency,Perturbations,Robustness

1　引言

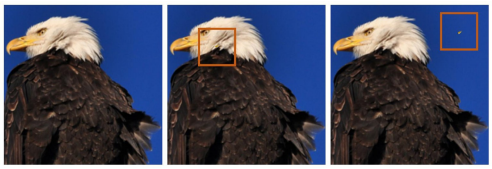
近年来,各种结构和功能各异的深度神经网络(Deep Neural Networks,DNN)在计算机视觉领域取得了显著的成功,涵盖了图像分类[1-3]、人脸识别[4-5]和自动驾驶[6-7]等多个领域。然而这些深度神经网络结构复杂,参数众多,且很多样本分布在分类边界附近,这使得深度神经网络容易受到对抗攻击。已有研究表明,仅需在图像上添加扰动即可欺骗最

到稿日期:2024-10-08　返修日期:2024-12-07
基金项目:中央高校基本科研业务费专项资金(2682024ZTPY041);四川省科技计划项目(2023YFH0066);成都市科技项目(2023-RK00-00080-ZF)
This work was supported by the Fundamental Research Funds for the Central Universities of Ministry of Education of China(2682024ZTPY041),Science and Technology Program of Sichuan Province(2023YFH0066) and Science and Technology Program of Chengdu(2023-RK00-00080-ZF).
通信作者:孟华(menghua@swjtu.edu.cn)

先进的神经网络分类器[8]。这揭示了 DNN 在应对对抗性攻击方面的脆弱性。

在图像分类任务中，对抗攻击的研究集中于通过向干净图像添加微小扰动来生成对抗样本，这些对抗样本会导致目标分类器输出错误的预测。其目标是保持扰动在人眼难以察觉的情况下实现对抗攻击的效果。同时，在对抗性扰动的研究中，扰动的 l_p 范数通常被视为不可感知性的良好保证。基于这一假设，一些防御方法被设计为在特定 l_p 界限下对对抗性扰动有效[9-12]。然而，有研究表明仅使用 l_p 范数度量扰动可能不足以评估视觉上的不可感知性[13]。

如图 1 所示，图 1(a)为原始样本，图 1(b)和图 1(c)分别在不同区域内施加了相同幅度、相同像素数量的扰动。图 1(b)的扰动位于复杂区域，而图 1(c)则位于光滑区域。结果显示，图 1(b)的扰动难以被人眼察觉，而图 1(c)的扰动在光滑背景中较为明显。这表明，扰动的可感知性不仅与扰动的大小有关，还受到图像内容的影响。因此，单纯依赖范数衡量对抗扰动的不可感知性存在局限性，应将人眼对不同区域的敏感度差异纳入考虑范围。



(a) (b) (c)

图 1 不同像素位置的扰动对人类感知的影响

Fig. 1 Impact of perturbations at different pixel locations on human perception

Luo 等[14]提出通过计算像素的方差来衡量扰动承载能力，使高方差像素点容纳更大的扰动，从而降低感知度。然而，该方法局限于局部信息，未考虑图像的全局特征。针对上述问题，本文提出了一种基于梯度的自适应分频段攻击方法。该方法通过图像的频域信息，在全局范围内识别出高频(复杂)区域，将较大扰动隐藏于人眼不敏感的高频细节中，从而显著提升扰动的不可感知性。具体而言，首先利用图像的频谱特征将高频和低频区域分离，得到高频重构图像；然后对高频重构图像进行高斯函数映射，生成高频权重掩码矩阵；最后，在传统梯度攻击的基础上，使用高频权重掩码矩阵对扰动施加额外的空间约束，从而抑制图像低频区域的扰动，降低扰动的感知度，同时在高频区域内嵌入较大扰动，以提高对抗样本的鲁棒性。本文的主要贡献总结如下：

1)提出了一种自适应高频权重矩阵生成算法，并通过在频率域中有效分离高频与低频分量，利用重构的高频区域生成自适应权重矩阵。

2)利用自适应高频权重矩阵对扰动进行空间约束，通过逐像素加权来抑制低频区域的扰动强度，与基于梯度的攻击方法(如 IFGSM，MIM 等)结合，能够快速生成在视觉上难以察觉且具鲁棒性的对抗样本。

3)在 ImageNet-1K 和 CIFAR-10 数据集上的实验结果验证了所提方法的有效性。结果表明，结合该算法的对抗样本在感知相似度、鲁棒性测试及面对防御措施时均表现优异，展示了其生成自然且具备鲁棒性对抗样本的潜力。

2 相关工作

本文聚焦于图像分类中的对抗攻击问题，重点研究对抗样本的攻击效果和扰动的不可感知性能。与本文研究密切相关的工作包括：基于梯度的攻击方法、难以察觉的对抗样本生成方法和人类视觉识别分辨能力的研究。

2.1 基于梯度的对抗攻击方法

基于梯度的对抗攻击方法，主要通过计算和利用目标模型的梯度来生成对抗样本。对抗样本指在原始数据上添加精心设计的、在视觉上几乎不可察觉的细微扰动所形成的样本，这类样本会导致训练好的模型给出错误的分类输出[15]。

Goodfellow 等[16]提出的快速梯度符号法(Fast Gradient Sign Attack，FGSM)是基于梯度的攻击，可以快速生成对抗样本。FGSM 利用目标模型的损失函数，计算输入图像的梯度，再利用梯度信息让输入图像向梯度的反方向移动，并使用 l_p 范数直接约束单个像素的扰动范围以生成对抗样本。迭代快速梯度符号法(I-FGSM)[17]是 FGSM 的一种改进版本，通过多次迭代应用 FGSM 并减小每次迭代的步长 α 来增加对抗扰动的效果。每次迭代中，对对抗样本进行逐步更新，并对更新后的结果进行裁剪，以确保扰动处于约束范围 ϵ 内。对抗样本的更新式如下：

$$x_{\text{adv}}^{(0)} = x \tag{1}$$

$$x_{\text{adv}}^{(k+1)} = clip_{x,\epsilon}\{x_{\text{adv}}^{(k)} + \alpha \cdot \text{sign}(\nabla_x J(\theta, x_{\text{adv}}^{(k)}, y))\} \tag{2}$$

其中，$x_{\text{adv}}^{(k)}$ 是第 k 次迭代后的对抗样本。研究表明，随着 ϵ 值的增加，对抗样本的成功率提高，且在物理环境中的鲁棒性也更强，但同时扰动也更容易被感知。

Carlini&Wagner(C&W)[18]攻击是一种基于优化的对抗攻击方法，旨在生成具有较高攻击成功率和较低可察觉性的对抗样本。与 FGSM 和 I-FGSM 不同，C&W 攻击是通过优化目标函数来生成对抗样本，使得对抗样本在成功地欺骗目标分类器的同时，尽可能地减小扰动的感知性。其基本思想是通过优化以下目标函数来生成对抗样本：

$$\underset{\delta}{\text{minimize}} \| \delta \|_p + c \cdot f(x+\delta) \tag{3}$$

其中，$\| \delta \|_p$ 表示扰动的 l_p 范数，通常取 l_2 范数；c 是用于平衡范数和 f 函数的权重参数。$f(x)$ 用于衡量攻击是否成功，其定义如下：

$$f(x) = \max(\max\{Z(x)_i : i \neq t\} - Z(x)_t, -k) \tag{4}$$

其中，$Z(x)$ 是分类器的 logit 层输出；t 是目标类别；k 是非负常数，称为信心度量，用于控制对抗样本的置信度。通过最小化 $f(x)$，使得分类器对目标类别 t 的置信度最高，从而成功实现对抗攻击。C&W 攻击通过梯度下降等方法，迭代地调整 δ，使得目标函数逐步减小。该攻击通常能够达到较高的攻击成功率，并且由于优化过程考虑了扰动的感知性，生成的对抗样本在视觉上更加接近原始图像，相比于 I-FGSM 等方法，其扰动较难被人眼察觉。

2.2 难以察觉的对抗攻击

Zhao 等[19]提出 PerC-AL 攻击方法，该方法根据 CIEDE2000[20]测量的感知颜色距离来优化对抗性扰动，能够在人眼不易察觉的情况下，在 RGB 空间中隐藏大扰动，其鲁

棒性和不可感知性均优于传统方法。但 PerC-AL 方法的色彩空间转化处理和基于迭代的优化方式，导致其时间消耗巨大，难以在实际应用中推广。Luo 等[21]提出的 SSAH 方法通过攻击模型特征层的语义相似性，不再仅依赖于分类层的决策。该方法引入低频约束，将扰动限制在高频区域，以保持视觉相似性。然而，SSAH 依赖于在输入样本批次中寻找不相似样本来构建对抗损失，导致计算复杂度和内存需求大幅增加，限制了其实用性。Liu 等[22]提出了一种稀疏低频攻击方法，该方法通过梯度掩码定位图像语义区域，并在区域内自适应生成低频扰动，提升对抗样本的视觉相似性。Zhang 等[23]将扰动投影到原始图像的最小可觉察差异(JND)空间中，并通过视觉系数调整扰动方向，以保持对抗样本的视觉相似性。Li 等[24]通过 K-means 聚类选择对人类视觉系统不敏感的频率分量，引入损失限制扰动的频率分布，生成难以感知的对抗样本。

2.3 人类视觉系统

人类通过视觉系统获取外界的图像信息，当光辐射刺激人眼时，会引起生理和心理上的变化，这种感知就是视觉。HVS(Human Vision System)作为一种图像处理系统，其认知是非均匀且非线性的。人类视觉系统的主要特性体现在 3 个方面：亮度、图像类型和频域特性。在不同局部特性区域，允许改变的信号强度各不相同。一般来说，人眼对高亮度区域附加噪声的敏感性较低。从图像类型来看，平滑区域比纹理密集区域更敏感，在频域中，频率越高，人眼的分辨能力越弱，即人眼对高频内容的敏感性较低。因此，图像的高频部分和纹理密集区域可以嵌入更多信息而不易被察觉。

一些研究模型从频域角度分析了深度神经网络。Wang 等[25]注意到，深度神经网络能够捕捉到人类几乎无法察觉的图像高频成分。Dong 等[26]发现，自然训练的模型对高频的加性扰动高度敏感，而高斯数据增强和对抗性训练都可以显著提高对高频噪声的鲁棒性。Subramanian 等[27]则发现，神经网络分类器覆盖的频率通道比人类宽 2～4 倍，这意味着过于高频和低频的噪声会损害网络性能，但对人类视觉几乎没有影响。

本文综合上述 3 个方面的研究，提出一种新的对抗攻击方法。该方法不仅能够快速生成对抗样本，还能确保这些扰动在视觉上难以被察觉，同时提高攻击的隐蔽性和鲁棒性。

3 基于高频特征掩蔽的对抗攻击算法

现有基于 l_p 范数约束的攻击方法旨在对整个图像添加扰动，但在面对一些防御方法和更为鲁棒的防御模型时，通常需要增大扰动幅度，这显著降低了扰动的不可感知性。尤其是在光滑区域，人眼对该区域的扰动更为敏感，而复杂区域在相似的感知程度下能够嵌入更多信息。基于此，提出了一种基于图像频谱特征的自适应高频权重矩阵生成方法。

图 2 展示了通过提取图像频谱中不同频率分量重构的高频区域和低频区域。图 2(a)为原始图像，图 2(b)为通过提取高频分量重构得到的高频成分图像，图 2(c)为低频分量重构的低频图像。高频重构图像突出了图像中细节和纹理丰富的区域，而低频图像则包含了图像的整体轮廓和光滑区域。基于图像频谱特征分离高频和低频，并通过高频重构图像生成自适应高频权重矩阵，作为每个像素点属于高频区域的权重，用于衡量该像素点能够嵌入的扰动量。将该方法与其他基于梯度的攻击方法相结合，能够在高频区域隐藏较大的扰动，从而生成对人类视觉系统更友好的对抗样本。如此不仅保留了传统方法快速生成对抗样本的优势，还进一步考虑了人眼的感知特性，有效提升了对抗样本的隐蔽性和鲁棒性。

(a) (b) (c)

图 2 利用频谱信息分离高频和低频区域

Fig. 2 Spectral information is leveraged to separate high-frequency and low-frequency regions

3.1 离散余弦变换

离散余弦变换(Discrete Cosine Transform，DCT)[28]是一种将图像数据从空间域转换到频率域的线性变换。DCT 将图像中的像素值表示为一组频率分量，使得高频分量(快速变化部分)和低频分量(缓慢变化部分)的分离度更高。对于大小为 $N\times N$ 的图像 x，其离散余弦变换定义如下：

$$X_{k_1,k_2}=\alpha(k_1)\alpha(k_2)\sum_{n_1=0}^{N-1}\sum_{n_1=0}^{N-1}x_{n_1,n_2}\cos\left[\frac{\pi}{N}\left(n_1+\frac{1}{2}\right)k_1\right]\cos\left[\frac{\pi}{N}\left(n_2+\frac{1}{2}\right)k_2\right] \tag{5}$$

$$\alpha(k_i)=\begin{cases}\sqrt{\dfrac{1}{N}}, & k_i=0\\ \sqrt{\dfrac{2}{N}}, & k_i\neq 0\end{cases},i=1,2 \tag{6}$$

离散余弦逆变换(Inverse Discrete Cosine Transform，IDCT)用于将图像从频域转换回空间域。其定义如下：

$$x_{n_1,n_2}=\sum_{k_1=0}^{N-1}\sum_{k_2=0}^{N-1}\alpha(k_1)\alpha(k_2)X_{k_1,k_2}\cos\left[\frac{\pi}{N}\left(n_1+\frac{1}{2}\right)k_1\right]\cos\left[\frac{\pi}{N}\left(n_2+\frac{1}{2}\right)k_2\right] \tag{7}$$

其中，X_{k_1,k_2} 为频域中的 DCT 系数，k_1，k_2 是频率索引，$0\leqslant k_i\leqslant N-1$，$i=1,2$；$x_{n_1,n_2}$ 为空间域中(n_1,n_2)处像素值，$0\leqslant n_i\leqslant N-1$，$i=1,2$。

在使用 DCT(记作 $D(\cdot)$)对图像进行转换时，可用矩阵乘法简述上述过程：

$$D(x)=\boldsymbol{C}x\boldsymbol{C}^{\mathrm{T}} \tag{8}$$

IDCT(记作 $D_I(\cdot)$)简化为：

$$D_I(D(x))=\boldsymbol{C}^{\mathrm{T}}D(x)\boldsymbol{C}=x \tag{9}$$

其中，矩阵 $\boldsymbol{C}$ 是 DCT 基函数矩阵且为正交矩阵，因此 $\boldsymbol{C}\boldsymbol{C}^{\mathrm{T}}$ 等于单位矩阵 $\mathbf{E}$。DCT 和 IDCT 都是无损变换。

在图像处理领域，除了 DCT，还有其他方法可以将图像从空间域变换到频域，例如离散傅里叶变换(Discrete Fourier Transform，DFT)和离散小波变换(Discrete Wavelet Transform，DWT)。相较于其他变换，DCT 具有一个显著的优点，即能够将图像的能量集中在低频系数上。这意味着大多数信

息在变换后被集中到矩阵的左上角，而高频分量则主要集中在右下角。利用这一特性，通过掩蔽 DCT 变换后矩阵左上角的低频分量，仅用剩余分量重构图像，来自适应地分离原始图像中的高频部分和低频部分。

如图 2 所示结果中，低频重构图像由 DCT 变换后的矩阵左上角的低频分量重构，高频重构图像则由剩余分量重构。具体而言，复杂区域和光滑区域分别对应高频和低频部分。本文方法通过掩蔽特定的低频分量，用剩余分量重构高频图像，能够更加清晰地划分复杂区域和光滑区域，进而构建高频权重矩阵。

3.2 自适应高频权重矩阵生成算法

大多数基于梯度的攻击方法，如 FGSM 和 PGD 等，在生成对抗样本时普遍利用 l_p 范数约束对抗样本与原始样本之间的差异，然而这些方法可能会产生两个问题。

1）l_p 范数约束只限制扰动的大小，并且对所有像素点的扰动上限是一致的，忽略了扰动的空间分布，这通常会导致光滑的背景区域出现较为明显的扰动。

2）C&W 等方法使用 l_2 范数作为正则化器，旨在引导模型产生错误分类的同时最小化扰动的 l_2 范数，使得扰动在所有像素中相对均匀地分布。这在一定程度上降低了扰动的视觉感知，但也减小了扰动的强度。在面对某些对抗性防御方法（如 JPEG 压缩和高斯模糊）时，这种弱扰动容易被破坏，从而削弱其对抗效果。

同时，研究发现，同量级的扰动被添加到图像的不同区域时，在感知上的影响是不同的。如图 1 所示，与光滑区域相比，复杂区域通常包含丰富的细节和色彩，在此区域添加新的扰动对人眼的感知相似性影响较小。基于上述问题，提出了一种自适应高频权重矩阵生成算法，将自适应高频权重矩阵与基于梯度的攻击方法结合，能够快速生成难以感知且具有鲁棒性的对抗样本。该方法首先对原始图像进行离散余弦变换，得到频谱信息；然后将提取的高频分量重构，并生成自适应高频权重矩阵。结合基于梯度的攻击方法，利用自适应高频权重矩阵对扰动进行空间约束，利用复杂区域掩盖大扰动，可以生成感知上真实的对抗样本。图 3 显示了该方法的框架，其中包括两个主要步骤：生成自适应高频权重矩阵和生成空间约束的对抗样本。

自适应高频权重矩阵的生成过程可分为两个步骤：高频分量提取和高斯函数映射。首先，对输入图像 x 进行离散余弦变换，将图像从空间域转换到频率域。其次，使用频谱掩蔽矩阵 $\boldsymbol{M}$ 对频率分量进行筛选，仅保留高频分量。掩蔽矩阵 $\boldsymbol{M}$ 是与输入图像 x 形状相同的二值矩阵，$\boldsymbol{M}\in\{0,1\}^{c\times h\times w}$。此操作通过掩蔽低频分量，保留高频分量，从而实现了对图像中复杂细节信息的聚焦。然后对提取后的高频分量进行逆离散余弦变换（IDCT），将其转换回空间域，得到高频重构图像：

$$X_h = D_I(D(x)\odot \boldsymbol{M}) \tag{10}$$

其中，$D(x)$ 和 $D_I(x)$ 均对图像的 3 个通道分别进行操作，X_h 是通过高频分量进行 IDCT 后得到的高频重构图像。该图像包含了输入图像中的高频细节和边缘信息，是高频成分在空间域的表示。利用高斯函数对高频重构图像 X_h 进行高斯映射，生成高频权重矩阵：

$$\boldsymbol{M}_h = G(X_h) = \exp\left(-\frac{(X_h-\mu)^2}{2\sigma^2}\right) \tag{11}$$

其中，μ 和 σ 分别为高频重构图像 X_h 的均值和标准差；$G(\cdot)$ 为高斯映射函数。高频权重矩阵的每个元素代表该像素点属于高频区域的程度，用于衡量该像素点能够嵌入的扰动量。

在生成对抗样本的过程中，将高频权重矩阵与基于梯度的攻击方法（如快速梯度符号法 FGSM）结合使用。通过 FGSM 方法计算扰动，再使用高频权重矩阵对其进行加权，将加权后的扰动添加到原始输入图像中，生成对抗样本：

$$\boldsymbol{M}_h = G(D_I(D(x)\odot \boldsymbol{M}),\mu,\sigma) \tag{12}$$

$$x_{\text{adv}} = clip_{x,\epsilon}\{x+\alpha\cdot \text{sign}(\nabla_x J(\theta,x,y))\cdot \boldsymbol{M}_h\} \tag{13}$$

此过程通过在复杂区域嵌入更多扰动，在光滑区域嵌入较少扰动，能够在保持视觉质量的同时提高攻击的鲁棒性。

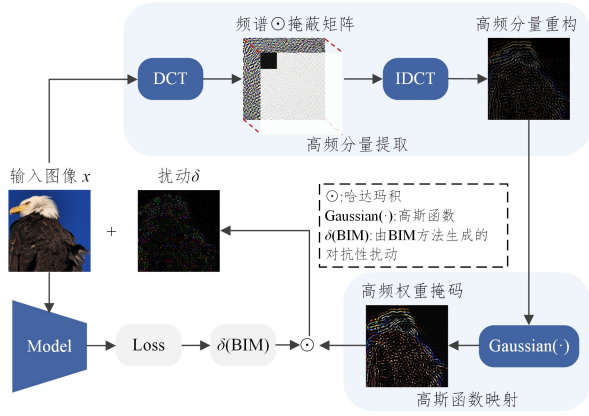


图 3 生成自适应高频权重矩阵约束对抗扰动流程图

Fig. 3 Flowchart of generating adaptive high-frequency weighting matrix to constrain adversarial perturbation

3.3 攻击算法

自适应高频权重矩阵生成算法可以与基于梯度的攻击相结合。为了展示这一算法的有效性，以迭代快速梯度符号法（I-FGSM）为例，详细描述了生成扰动后逐像素与高频权重矩阵进行加权生成对抗样本的过程。该过程被称为高频矩阵加权的 I-FGSM（HFM-I-FGSM），其步骤如算法 1 所示。

算法 1 HFM-I-FGSM

输入：输入图像 x、分类器 f(x,θ)、迭代轮次 N、扰动上限ϵ

输出：对抗样本x_{adv}

1. 使用 3.2 节所述算法生成输入图像 x 的自适应高频权重矩阵 $\mathbf{M}_h$
2. 初始化 $x_0=x,\delta_0=0,\alpha=\epsilon/N$
3. for i=0 to N−1 do
4. 计算梯度 $g_i=\nabla_{x_i}J(\theta,x_i,y)$
5. 计算扰动 $\delta_i=\alpha\cdot\text{sign}(g_i)$
6. 更新扰动 $\delta_i'=\delta_i\cdot\mathbf{M}_h$
7. 更新 $x_{i+1}:x_{i+1}=\text{clip}_{x,\epsilon}\{x_i+\delta_i'\}$
8. 更新 $x_{i+1}:x_{i+1}=\text{clip}(x_{i+1},0,1)$
9. end for
10. return $x_{adv}=x_N$

4 实验结果与分析

4.1 实验设置

为验证所提方法的有效性，将自适应高频权重矩阵生成

算法与基于梯度的攻击算法集成，并在两个广泛使用的数据集 ImageNet-1K[29] 和 CIFAR-10[30] 上进行实验。ILSVRC2012 数据集为 ImageNet 数据集的子数据集，共包含 1 000 个类别，图像分辨率为 299×299×3。CIFAR-10 数据集共包含 10 个类别，图像分辨率为 32×32×3。

实验使用了从 ILSVRC2012 验证数据集中选取的 1 000 张图像，这些图像包括所有类别，图像选择方法参考了 Wang 等[31]的工作。从 CIFAR-10 训练集中选择一个批次的 10 000 张彩色图像，几乎所有图像都被目标分类器正确分类。选择几种常用的梯度攻击方法（I-FGSM，PGD，MIM 和 C&W）与自适应高频权重生成算法（HFM）相结合，形成 HFM-I-FGSM，HFM-PGD，HFM-MIM 和 HFM-C&W 方法。此外，还选取了两种专注于生成难以感知对抗样本的算法 PerC-AL 和 SSAH 进行对比分析。

如表 1 所列，这些方法通过不同策略提高了对抗样本的隐蔽性。

表 1 攻击算法

Table 1 Attack algorithms

方法	年份	类型	描述
I-FGSM	2016	白盒	经典的迭代攻击方法，通过多次迭代小步扰动逐步逼近目标
PGD	2018	白盒	基于投影梯度下降的迭代攻击方法，迭代后将扰动投影回合法空间，保证扰动幅度不超过预设上限
MIM	2018	白盒	在 I-FGSM 的基础上引入了动量项，以更好地捕捉优化方向
C&W	2017	白盒	基于优化的攻击方法，通过优化目标函数生成对抗样本，具有较高的攻击成功率和较低的可察觉性
PerC-AL	2021	白盒	基于感知损失的攻击方法，自适应调整扰动以确保在颜色和亮度上的不可见性
SSAH	2023	白盒	攻击语义特征相似性，使用高频约束生成跨数据集的高质量对抗样本

4.2 评价指标及模型

为对性能进行评估和比较，实验分别选择了扰动的 l_2 和 l_∞ 范数以及攻击成功率 ASR 和 LPIPS[32] 来评估对抗样本。LPIPS 用于衡量样本的视觉质量和不可感知性，传统的 l_2 范数度量和最大扰动强度 l_∞ 用于衡量扰动的强度。LPIPS 是一种衡量图像感知相似度的指标，旨在更好地模拟人类视觉系统对图像差异的感知。在实验中，使用 ResNet50 模型作为 ImageNet-1K 数据集的白盒模型，使用 ResNet20 模型作为 CIFAR-10 数据集的白盒模型。

4.3 实验结果分析

在白盒攻击场景下，对不同的对抗样本生成方法进行了评估，其中目标模型是完全可访问的。实验使用了 ImageNet-1K 和 CIFAR-10 数据集，选择将梯度攻击方法（I-FGSM，PGD，MIM 和 C&W）与自适应高频权重生成算法（HFM）结合，每组方法分别在相同范数限制下生成对抗样本。I-FGSM，PGD 和 MIM 均采用 l_∞ 范数约束，扰动上限设定为 $\epsilon=8/255$，迭代步长为 $\alpha=1/255$。

实验结果如表 2 所列，与基线相比，HFM-I-FGSM 和 HFM-PGD 的 LPIPS 从 0.016 38 和 0.028 35 降至 0.001 32 和 0.001 91。HFM-C&W 在扰动范数略有增加的情况下，LPIPS 仍从 0.000 14 降至 0.000 12。一些与 HFM 结合的梯度攻击如 HFM-I-FGSM 等，在感知相似度上的表现也比 PerC-AL 和 SSAH 更好。这表明，HFM 通过将扰动隐蔽于高频区域，有效提升了对抗样本的视觉隐蔽性。此外，HFM-PGD 等还继承了 PGD 方法生成效率高的特点，在运行速度上明显优于 PerC-AL 和 SSAH，这使得 HFM 方法在生成难以察觉的对抗样本时具备更高的实用性。

表 2 白盒攻击效果比较

Table 2 Comparison of white-box attack effectiveness

Dataset	Attack	*RunTime*/s ↓	*Iteration*	*ASR*/% ↑	l_2 ↓	l_∞ ↓	*LPIPS* ↓
Image Net-1K	PerC-AL	14 668	1 000	99.3	1.82	0.110 0	0.003 10
	SSAH	975	150	99.7	2.23	0.002 7	0.002 33
	I-FGSM	253	10	100	4.29	0.030 0	0.016 38
	HFM-I-FGSM	274	10	100	2.04	0.030 0	0.001 32
	PGD	826	40	100	7.32	0.030 0	0.028 35
	HFM-PGD	776	40	100	2.69	0.030 0	0.001 91
	MIM	260	10	100	8.80	0.030 0	0.095 43
	HFM-MIM	244	10	100	2.38	0.030 0	0.001 83
	C&W	12 616	1 000	100	0.39	0.004 2	0.000 14
	HFM-C&W	12 871	1 000	100	0.39	0.004 8	0.000 12
CIFAR-10	PerC-AL	3 248	1 000	97.3	0.90	0.180 0	0.000 51
	SSAH	436	150	99.9	0.47	0.020 0	0.000 10
	I-FGSM	33	10	100	0.82	0.030 0	0.000 37
	HFM-I-FGSM	33	10	100	0.81	0.040 0	0.000 35
	PGD	90	40	100	1.15	0.030 0	0.000 68
	HFM-PGD	92	40	100	0.80	0.030 0	0.000 32
	MIM	34	10	99.8	0.63	0.030 0	0.000 20
	HFM-MIM	34	10	99.8	0.61	0.030 0	0.000 19
	C&W	5 032	1 000	100	0.09	0.010 0	5.7×10^{-6}
	HFM-C&W	4 949	1 000	100	0.09	0.010 0	5.6×10^{-6}

将自适应高频权重矩阵生成算法与基于梯度的攻击方法结合，在对抗样本生成过程中通过结合原始图像的高频空间信息和梯度信息，使扰动集中在图像的高频区域，在相同的攻击成功率下，生成的对抗样本更加隐蔽。为了更加直观地展示攻击效果，将 I-FGSM 和 C&W 与自适应高频权重矩阵生成算法结合生成的对抗样本进行了可视化，结果如图 4 所示。综上所述，结合了自适应高频权重生成算法的攻击方法在保持高攻击成功率的同时，显著降低了对抗样本的感知损失，并且在某些情况下，即使扰动幅度增大，视觉隐蔽性仍然得到了增强。

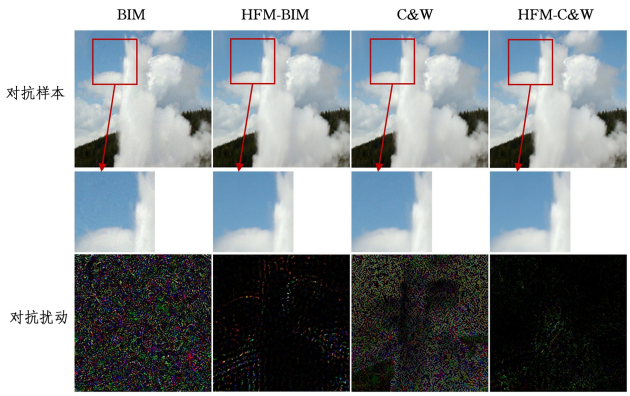


图 4 两种攻击方法结合高频权重矩阵的对比

Fig. 4 Comparison combining two attack methods withhigh-frequency weight matrices

4.4 鲁棒性评价

本节评估了不同攻击方法在 ImageNet-1K 数据集上生成的对抗样本的鲁棒性。实验选择了两种防御方法进行评估，分别为特征压缩[33]（Feature Squeeze，位深设置为 6）和 JPEG 压缩[34]（压缩质量设置为 75）。特征压缩通过降低模型输入的分辨率或者颜色深度来减小输入空间的复杂性，从而使得对抗样本难以生效。此方法假设对抗攻击通常依赖于输入中的微小扰动，而这些扰动在压缩过程中会被削弱或消除。JPEG 压缩防御通过对输入图像进行 JPEG 压缩和解压缩操作，来削弱或消除对抗样本中的微小扰动。实验中，对每一组对比方法生成的对抗样本在相同的感知相似度（LPIPS）下进行比较。表 3 列出了在无防御、特征压缩以及 JPEG 压缩防御下，不同攻击方法的攻击成功率（ASR）。

表 3 不同攻击方法的攻击成功率

Table 3 ASR of different attack methods

（%）

Attack	No Defence	Feature Squeeze	JPEG
PerC-AL	99.3	89.4	80.3
SSAH	99.9	43.7	39.5
I-FGSM	100.0	85.7	74.9
HFM-I-FGSM	100.0	86.1	78.6
PGD	100.0	85.9	75.7
HFM-PGD	100.0	87.4	78.0
MIM	100.0	99.7	98.9
HFM-MIM	100.0	99.7	99.7
C&W	100.0	31.5	34.1
HFM-C&W	100.0	31.8	34.6

在 JPEG 压缩防御下，HFM-I-FGSM 和 HFM-PGD 的攻击成功率分别提升至 78.6%和 78.0%，相比未使用 HFM 的 I-FGSM 和 PGD 有显著提升。HFM-MIM 在所有防御情况下均保持接近 100%的攻击成功率。而一些方法在防御上的表现不如 PerC-AL，可能是因为 PerC-AL 在生成对抗样本时使用了较大的扰动强度，从而提高了攻击成功率。结合 HFM 算法的攻击方法不仅能提升扰动的隐蔽性，还能显著增强对抗样本在压缩防御下的有效性，能够在保证视觉相似度的前提下生成更强的扰动。

自适应高频权重生成算法通过高频权重矩阵优化对抗扰动，使得扰动更加隐蔽且鲁棒。传统方法在结合了 HFM 方法后，能够在保证视觉感知效果的同时，生成更有效的对抗样本，提高了对抗攻击的整体效果。

4.5 高频与低频权重矩阵对比分析

仿照高频权重矩阵生成的方式，生成自适应低频权重矩阵（LFM），其中每个元素表示该像素点属于低频区域的程度。表 4 对比了 LFM 与 HFM 分别结合 I-FGSM，PGD，MIM 和 C&W 方法生成的对抗样本在攻击成功率（ASR）、l_2 范数、l_∞ 范数和感知相似度（LPIPS）下的表现。

表 4 LFM 和 HFM 分别与攻击方法结合的对比

Table 4 Comparison of attack methods combined with IFM and HFM

Attack	$ASR/\%\uparrow$	$l_2\downarrow$	$l_\infty\downarrow$	$LPIPS\downarrow$
LFM-I-FGSM	97.2	1.96	0.0300	0.00240
HFM-I-FGSM	100.0	2.04	0.0300	0.00130
LFM-PGD	97.5	2.66	0.0300	0.00400
HFM-PGD	100.0	2.69	0.0300	0.00190
LFM-MIM	91.4	2.33	0.0300	0.00710
HFM-MIM	100.0	2.38	0.0300	0.00180
LFM-C&W	100.0	0.50	0.0053	0.00017
HFM-C&W	100.0	0.39	0.0048	0.00012

从表 4 可以看出，与 HFM 结合后的方法在保持相似的扰动强度（即 l_2 和 l_∞ 范数基本一致）的前提下，通常表现出更小的 LPIPS 和更高的攻击成功率。HFM-I-FGSM 的攻击成功率为 100%，其 LPIPS 为 0.0013，明显低于 LFM-I-FGSM 的 LPIPS 值（0.0024）；HFM-PGD 和 HFM-MIM 也分别展现了更好的性能，其 LPIPS 值均远低于相应的 LFM 版本。这表明高频权重矩阵能更有效地隐藏扰动，并且更符合人眼对图像的感知，使得对抗样本在人类视觉上更加自然。另一方面，与 LFM 结合生成的对抗样本在攻击成功率上的表现不及 HFM 结合的方法，特别是在 MIM 攻击中，LFM-MIM 方法的攻击成功率下降至 91.4%，而 HFM-MIM 则能够保持 100%的攻击成功率，表明高频扰动对 DNN 的攻击更有效。而 Wang 等的研究表明，与人类识别时更关注低频成分不同，DNN 在识别时更依赖图像的高频细节部分，而与 LFM 结合后，大部分扰动被限制在低频成分内，这对 DNN 识别任务造成的影响更小，导致攻击效果降低。因此，高频扰动的隐蔽性和攻击性均优于低频扰动。

总体而言，与 HFM 结合的攻击方法能够在更高的攻击成功率下实现更好的感知隐蔽性和较小的扰动感知程度；而与 LFM 结合的攻击方法虽然在部分指标上表现出一定效果，但其总体攻击效果较弱。

4.6 基于高斯噪声的感知性评价

为验证自适应高频权重矩阵对隐藏扰动的有效性，自适应生成 ImageNet-1K 数据集图像的高频权重矩阵 HFM 与低

频权重矩阵 LFM。在实验过程中，首先生成服从标准正态分布 N(0,1)的高斯噪声 N，并将其绝对值限制在预设的 ε 内。缩放 LFM，使其 l_2 范数与 HFM 一致，保证在与同样大小的扰动相乘后的强度也相同。其次，将生成的高斯噪声 N 分别与 HFM 和 LFM 点乘，得到高频区域的噪声和低频区域的噪声。然后，将这两种区域的噪声分别添加到输入图像上，生成具有不同区域扰动的图像。最后，分别测试高频和低频区域对图像感知质量的影响，结果如图 5 所示。可以看出，随着扰动强度的增大，具有高频区域扰动的图像的 LPIPS 得分增长比具有低频区域的更缓慢，且整体得分比低频区域小。这表明在相同扰动强度下，高频区域能够容纳更多、更大的噪声，进一步验证了自适应扰动算法在噪声隐藏能力方面的有效性。

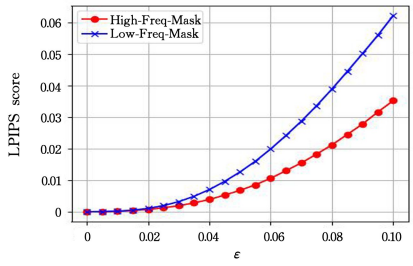


图 5 高频区域与低频区域噪声图像感知性的对比

Fig. 5 Comparison of perceptibility of noise in high-frequency and low-frequency regions

4.7 可视化分析

为证明结合自适应高频权重矩阵方法的对抗样本在攻击成功率和鲁棒性方面的有效性，采用 Grad-CAM 方法[35]分别对 I-FGSM，HFM-I-FGSM 和 LFM-I-FGSM 方法所生成的对抗样本进行可视化，分析模型对这些样本的注意力区域，结果如图 6 所示。

从图 6 中可以看出，采用 I-FGSM 方法生成的对抗样本会导致模型的注意力发生偏移，不再聚焦于原始目标物体；而结合了 HFM 后，扰动的形状发生了变化，增强了模型对目标物体的注意力，同时扩展了注意力的覆盖范围；相比之下，LFM 结合的对抗样本并未使模型的注意力有效集中于物体本身，且注意力范围小于 I-FGSM 生成的样本。这一结果表明，引入 HFM 改变扰动的整体形状后，对抗样本的攻击效果进一步增强，使其对模型的迷惑性更强。

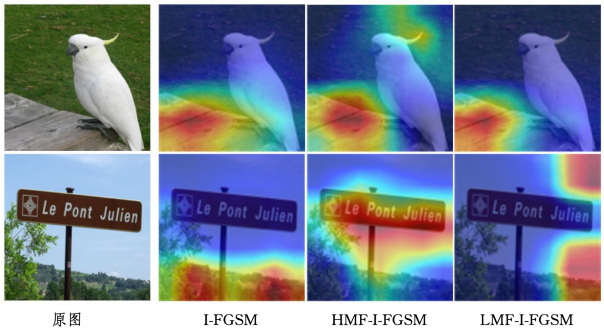


图 6 不同方法生成的对抗样本的可视化

Fig. 6 Visualization of adversarial samples generated by different methods

结束语 本文提出了一种基于梯度的自适应分频段攻击方法。该方法充分考虑了图像的频谱特性及人类视觉系统的感知特性，将较大幅度的扰动集中于高频区域，从而提高对抗样本的隐蔽性和有效性。该方法尽管在白盒攻击场景下展现了较强的不可感知性和鲁棒性，但在黑盒场景下表现出一定的局限性：自适应分频段攻击可能存在依赖源模型特征的风险；此外，该方法未能充分结合频率特征设计提升可迁移性的策略，导致对抗样本在黑盒场景下的可迁移性受限。未来工作将延续本文思路，进一步探索结合图像的频率特征提升对抗样本可迁移性的方法，以在保持对抗样本不可感知性的同时，实现更强的跨模型攻击能力。

参 考 文 献

[1] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[2] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4700-4708.

[3] WU W, SU Y, LYU M R, et al. Improving the transferability of adversarial samples with adversarial transformations [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 9024-9033.

[4] TAIGMAN Y, YANG M, RANZATO M A, et al. DeepFace: Closing the gap to human-level performance in face verification [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 1701-1708.

[5] WANG H, WANG Y, ZHOU Z, et al. CosFace: Large margin cosine loss for deep face recognition [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5265-5274.

[6] LIU A, LIU X, FAN J, et al. Perceptual-sensitive GAN for generating adversarial patches [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 1028-1035.

[7] SALLAB A E L, ABDOU M, PEROT E, et al. Deep reinforcement learning framework for autonomous driving [J]. Electronic Imaging, 2017, 29: 70-76.

[8] AKHTAR N, MIAN A. Threat of adversarial attacks on deep learning in computer vision: A survey [J]. IEEE Access, 2018, 6: 14410-14430.

[9] COHEN J, ROSENFELD E, KOLTER Z. Certified adversarial robustness via randomized smoothing [C]// International Conference on Machine Learning. PMLR, 2019: 1310-1320.

[10] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks [C]// International Conference on Learning Representations. 2018.

[11] TRAMÈR F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: Attacks and defenses [C]// International Conference on Learning Representations. 2018.

[12] WONG E, KOLTER Z. Provable defenses against adversarial examples via the convex outer adversarial polytope [C]// International Conference on Machine Learning. PMLR, 2018: 5286-5295.

[13] SHARIF M，BAUER L，REITER M K. On the suitability of lp-norms for creating and preventing adversarial examples [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern RecognitionWorkshops. 2018：1605-1613.

[14] LUO B，LIU Y，WEI L，et al. Towards imperceptible and robust adversarial example attacks against neural networks [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018.

[15] AKHTAR N，MIAN A. Threat of adversarial attacks on deep learning in computer vision：A survey [J]. IEEE Access，2018，6：14410-14430.

[16] GOODFELLOW I J，SHLENS J，SZEGEDY C. Explaining and harnessing adversarial examples[C]// International Conference on Learning Representations(Poster). 2015.

[17] KURAKIN A，GOODFELLOW I J，BENGIO S. Adversarial examples in the physical world [M]// Artificial Intelligence Safety and Security. Chapman and Hall/CRC，2018：99-112.

[18] CARLINI N，WAGNER D. Towards evaluating the robustness of neural networks [C]// 2017 IEEE Symposium on Security and Privacy(SP). IEEE，2017：39-57.

[19] ZHAO Z，LIU Z，LARSON M. Towards large yet imperceptible adversarial image perturbations with perceptual color distance [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：1039-1048.

[20] LUO M R，CUI G，RIGG B. The development of the CIE 2000 colour-difference formula：CIEDE2000 [J]. Color Research & Application，2001，26(5)：340-350.

[21] LUO C，LIN Q，XIE W，et al. Frequency-driven imperceptible adversarial attack on semantic similarity [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022：15315-15324.

[22] LIU J，LU B，XIONG M，et al. Low frequency sparse adversarial attack[J]. Computers & Security，2023，132：103379.

[23] ZHANG Y，TAN Y，SUN H，et al. Improving the invisibility of adversarial examples with perceptually adaptive perturbation [J]. Information Sciences，2023，635：126-137.

[24] LI C，LIU Y，ZHANG X，et al. Exploiting Frequency Characteristics for Boosting the Invisibility of Adversarial Attacks[J]. Applied Sciences，2024，14(8)：3315.

[25] WANG H，WU X，HUANG Z，et al. High-frequency component helps explain the generalization of convolutional neural networks [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：8684-8694.

[26] YIN D，GONTIJO LOPES R，SHLENS J，et al. A Fourier perspective on model robustness in computer vision [C]// Proceedings of the 33rd Conference on Neural Information Processing Systems. 2019.

[27] SUBRAMANIAN A，SIZIKOVA E，MAJAJ N，et al. Spatial-frequency channels，shape bias，and adversarial robustness [C]// NeurIPS 2023. 2023.

[28] AHMED N，NATARAJAN T，RAO K R. Discrete cosine transform [J]. IEEETransactions on Computers，1974，c-23(1)：90-93.

[29] RUSSAKOVSKY O，DENG J，SU H，et al. ImageNet large scale visual recognition challenge [J]. International Journal of Computer Vision，2015，115：211-252.

[30] KRIZHEVSKY A. Learning multiple layers of features from tiny images [D]. Toronto：University of Toronto，2009.

[31] WANG X，HE K. Enhancing the transferability of adversarial attacks through variance tuning [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021：1924-1933.

[32] ZHANG R，ISOLA P，EFROS A A，et al. The unreasonable effectiveness of deep features as a perceptual metric [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：586-595.

[33] XU W，EVANS D，QI Y. Feature squeezing：Detecting adversarial examples in deep neural networks [C]// Proceedings of the 2018 Network and Distributed System Security Symposium. Internet Society. 2018.

[34] DAS N，SHANBHOGUE M，CHEN S T，et al. Shield：Fast，practical defense and vaccination for deep learning using JPEG compression [C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018：196-204.

[35] 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，2020，128：336-359.

WANG Liuyi，born in 2002，postgraduate. Her main research interests include artificial intelligence and adversarial examples.

MENG Hua，born in 1982，Ph.D，associate professor. His research interests include interpretability in deep learning，topological data analysis and knowledge representation and reasoning.

（责任编辑：柯颖）