空频域联合优化的通用对抗扰动生成方法

耿荣[1]，孙钦东[1,2]，曹晗[1,3]，王艳[1]

（1.西安理工大学网络计算与安全技术陕西省重点实验室，陕西 西安 710048；
2.西安交通大学网络空间安全学院，陕西 西安 710049；3.四川数字经济产业发展研究院，四川 成都 610036）

摘　要：通用对抗扰动（UAP）的空域信息直观表示了扰动的视觉特征，频域信息包含了扰动的结构和纹理，联合分析扰动的空域和频域信息，有助于理解 UAP 的生成机制及其对图像分类模型鲁棒性的影响。已有研究大多关注扰动空域信息的分布和变化，忽略了频率分量的作用，限制了 UAP 的泛化能力。针对此问题，提出一种空频域联合优化的图像 UAP 生成方法，使用对抗样本置信度损失、扰动空域距离损失和扰动频率引导损失，从空域和频域角度训练模型，生成具有高攻击性和迁移性的 UAP。其中，对抗样本置信度损失用于增强扰动的攻击性，扰动空域距离损失优化扰动的空域大小，扰动频率引导损失控制扰动中频率分量的比重。实验结果表明，UAP 的低频分量对攻击效果影响较大，在相同扰动空域内，低频分量越多，扰动攻击成功率越高；与基线方法对比，通过联合优化空域和频域生成的 UAP 具有较强的攻击性和迁移性，在生成速度方面也有显著的优势。

关键词：通用对抗扰动；空频域联合优化；对抗样本置信度；频率引导；频率分量

中图分类号：TP391　　**文献标志码：**A　　**DOI：**10.19678/j.issn.1000-3428.0069619

Generalized Method for Universal Adversarial Perturbations Using Joint Optimization in Spatial-Frequency Domains

GENG Rong[1], SUN Qindong[1,2], CAO Han[1,3], WANG Yan[1]

(1.Shaanxi Key Laboratory of Network Computing and Security, Xi′an University of Technology, Xi′an 710048, Shaanxi, China;
2.School of Cyber Science and Engineering, Xi′an Jiaotong University, Xi′an 710049, Shaanxi, China;
3.Sichuan Digital Economy Industry Development Research Institute, Chengdu 610036, Sichuan, China)

【Abstract】The spatial information of a Universal Adversarial Perturbation (UAP) intuitively represents the visual characteristics of perturbations, whereas the frequency domain information includes the structure and texture of perturbations. Joint analysis of the spatial and frequency domain information of perturbations helps understand the generation mechanism of UAP and its impact on the robustness of image classification models. Most existing studies have focused on the distribution and changes in perturbed spatial information, neglecting the role of frequency components and limiting the generalization ability of the UAP. To address this issue, a joint optimization method for image UAP generation in the spatial and frequency domains is proposed. This method utilizes the adversarial sample confidence loss, perturbation spatial distance loss, and perturbation frequency guidance loss to train the model from both spatial and frequency perspectives, generating a UAP with high attack and transferability. The adversarial sample confidence loss is used to enhance the aggressiveness of disturbances, disturbance spatial distance loss optimizes the spatial size of disturbances, and disturbance frequency guided loss controls the proportion of the frequency components in disturbances. The experimental results indicate that the low-frequency components of the UAP have a significant impact on attack effectiveness. Within the same perturbation space, the more low-frequency components, the higher the success rate of perturbation attacks. Compared with the baseline method, the UAP generated by jointly optimizing the spatial and frequency domains has strong aggressiveness and transferability. Moreover, it has significant advantages in terms of generation speed.

【Key words】Universal Adversarial Perturbations (UAP); joint optimization in spatial-frequency domains; adversarial sample confidence; frequency guided; frequency component

基金项目：国家自然科学基金（62272378）；四川省自然科学基金（2023NSFSC0502，2022NSFSC0554，2022NSFSC0549）；陕西省高校青年创新团队（2019-38）。

作者简介：耿荣，男，博士研究生，主研方向为机器学习、网络安全；孙钦东（通信作者），教授、博士、博士生导师；曹晗、王艳，博士研究生。

收稿日期：2024-03-19　　**修回日期：**2024-07-25　　**E-mail：**qdongsun@xjtu.edu.cn

0 引言

现有的主流深度学习模型容易被对抗样本干扰。文献[1]的研究表明,对原始图像添加小幅扰动,能够导致分类模型输出错误的分类结果,且人眼难以分辨原始图像与对抗样本图像之间的差别。在自动驾驶系统、安全监控系统、人脸识别系统等关键领域,对抗扰动具有难以预测的威胁[2],可能会导致交通事故、安全威胁、身份被盗窃等严重后果[3-4]。因此,需要研究对抗扰动生成方法及其特性,以“攻”促“防”,提高深度学习模型的鲁棒性。

在文献[1]首次提出图像对抗样本之后,研究学者对图像对抗扰动进行了广泛研究,前期工作主要集中在针对单张图像生成特定对抗扰动。随着研究的深入,研究学者发现同一扰动对于不同的图像和不同的分类模型仍具有攻击性,这表示对抗扰动具有通用性、迁移性等重要特性。文献[5]提出通用对抗扰动(UAP),其不依赖于特定图像,可以在多个样本上产生攻击效果,使它们被图像分类模型错误分类。

UAP 的空域信息关注原始像素坐标系中的扰动变化,直观地表示了扰动的形状、颜色等视觉特征。扰动的频域信息能够揭示扰动的模式、纹理和结构等,不同频率分量所蕴含的图像信息迥然不同[6]。仅依赖空域信息生成通用对抗扰动可能受限于视觉感知能力从而削弱其攻击效果,利用空频域的信息互补性[7]联合优化通用对抗扰动的空域和频域,有利于进一步分析其生成机制及对图像分类模型鲁棒性的影响[8]。现有的大多数通用对抗扰动生成算法关注扰动的空间分布和变化,未能充分考虑频率分量在扰动攻击中的作用,限制了通用对抗扰动的攻击能力。因此,亟需研究频率分量对通用对抗扰动攻击性的影响,联合空频域信息探索通用对抗扰动的生成方法,以解决现有工作的局限性。

为了研究通用对抗扰动的空域和频域信息对其攻击性的影响,本文从空域和频域角度联合优化通用对抗网络(UAN)以生成扰动,提出通用对抗扰动生成模型 SFUAN,使用损失函数从频域和空域角度训练模型,生成具有高攻击性和迁移性的通用对抗扰动。本文的主要贡献如下:

1)提出扰动频率引导损失函数,以不同的权重联合对抗样本置信度损失、扰动空域距离损失和扰动频率引导损失,从空域和频域角度联合优化扰动生成网络 SFUAN。

2)在扰动生成网络 SFUAN 的训练过程中,通过频率引导系数控制生成模型对扰动中高频与低频分量的关注程度,从而生成具有不同比例高频和低频成分的通用对抗扰动。

3)分别在目标攻击和非目标攻击场景下,研究对抗扰动的不同频率分量对其攻击性的影响。分析目标攻击场景下攻击成功率与扰动半径的关系。在非目标攻击场景下,将所提方法与 UAP、UAN、SGA 方法在攻击性和生成速度方面展开对比实验。

1 相关工作

在机器学习理论中,传统机器学习模型依赖于稳定性假设,即训练数据和测试数据具有相同的分布[9]。然而,当机器学习模型面临由攻击者蓄意生成的异常样本时,由于这些样本的特征分布与训练样本的特征分布差异较大,分类模型在识别时会产生错误的结果。根据对抗扰动攻击范围的不同,扰动被分为特定对抗扰动和通用对抗扰动。其中,特定对抗扰动是针对特定样本生成的对抗扰动,通用对抗扰动则不依赖于特定的输入样本,而是针对输入样本集和分类模型的结构进行设计的,能够对多个样本和模型产生攻击效果。

在特定对抗扰动的研究领域,文献[10]提出快速梯度符号法,通过利用神经网络模型的梯度信息来调整输入图像的像素值,使图像向梯度的反方向移动,从而生成能够被神经网络模型错误分类的对抗样本。文献[11]在快速梯度符号法的基础上进行改进,通过多次迭代地计算梯度并调整图像的像素值,提出了基本迭代法。通过多次迭代,基本迭代法可以逐步增加对抗样本的攻击强度,从而更加有效地欺骗深度学习模型。文献[12]提出了一种高效的对抗攻击方法,称为 DeepFool。DeepFool 方法的基本思想是将原始样本分类到一个错误的类别,在每次迭代中通过线性逼近确定样本的最短距离,并将输入样本移动到一个超平面上,该超平面用于分割样本的正确类别和错误类别,迭代过程不断重复,直到样本被识别为错误类别。文献[13]提出了一种基于优化的 C&W 攻击方法。该方法可以针对特定模型和数据集进行攻击,可以调节对抗样本的置信度,具有更强的攻击性和更高的精度。由于该方法基于优化的方式,对参数进行更新需要较长的时间,因此生成对抗样本的速度较慢。文献[14]通过优化算法提出 One-Pixel 方法,该方法使用差分进化算法对像素空间进行搜索,找到一个像素位置和颜色变化

的最优组合，通过微小的像素变化，最大化目标分类模型在分类对抗样本时的错误概率。

文献[12]对 DeepFool 方法进行了扩展研究，在国际计算机视觉与模式识别会议发表的文献中提出一个通用对抗扰动，可以在多个输入样本上产生攻击效果，使它们被深度学习模型错误分类。此外，通用对抗扰动也可以攻击不同的网络模型，并且往往能取得较好的攻击效果。文献[15]提出了快速特征欺骗法，该方法基于梯度下降进行迭代优化。该算法不依赖于目标数据分布的信息，仅使用模型提取的特征向量便可生成通用对抗扰动。该文发现，生成的通用对抗扰动不仅能够有效攻击使用相同数据训练的神经网络模型，还具备跨模型攻击的能力。文献[16]提出了 GD-UAP，该方法无须依赖数据，通过破坏神经网络多层提取的特征便能生成通用对抗扰动。相比于依赖数据生成的通用对抗扰动，该方法生成的扰动在攻击目标模型时更具有优势。文献[17]提出了 Cosine-UAP，将原始图像和对抗样本在深度神经网络中的 logit 输出视为高维向量，通过迭代优化这些向量之间的余弦相似度来生成通用对抗扰动。文献[18]提出了 FG-UAP，攻击产生神经崩塌现象的神经网络层，从而生成通用对抗扰动。文献[19]通过空间转换对深度神经网络发起通用对抗性攻击，生成不易察觉的对抗样本。在通用对抗扰动生成过程中，存在使用小批量随机梯度优化的梯度消失问题以及使用大批量优化的局部最优问题，针对这些问题，文献[20]提出随机梯度聚合方法 SGA，增强了通用对抗扰动的泛化能力。

为了探究影响深度神经网络鲁棒性的因素，研究学者从频率角度分析了对抗样本[21]。文献[22]发现，卷积神经网络仅能利用图像的高频对应部分作出正确的预测，而这些对应部分是人类无法感知的。他们通过研究图像中的高频部分，来探究卷积神经网络在准确性和鲁棒性方面的平衡关系。文献[23]证明，在自然训练模型中，对抗扰动主要集中在图像的高频部分[24]，而在经过对抗训练后，这些扰动则转移到了图像的低频部分。除了探索对抗性攻击的目标特征分布，一些研究还侧重于通过寻找不同频段的扰动来提高对抗性扰动的强度。文献[25]指出，在图像的低频部分，对抗方向可能会出现得更加密集，在低频部分寻找对抗扰动可以提高攻击的效率。文献[26]通过采用协方差矩阵自适应进化策略来学习低频域的对抗扰动分布，并通过将协方差矩阵设置为对角矩阵，减少多元高斯分布协方差矩阵需要更新的维数，从而降低攻击的计算成本。此外，低频扰动被证明对防御模型[27]非常有效[28]。另一方面，在中频段和高频段进行攻击可以平衡欺骗率和感知能力[29]。目前存在的观点并不统一，从频率角度分析对抗样本仍具有较大的挑战。

分析空域和频域信息对通用对抗扰动攻击性及迁移性的影响有助于解决现有研究的局限性，理解对抗扰动的生成机制。本文从空域和频域角度联合优化生成模型，在白盒攻击场景下生成通用对抗扰动，分析扰动的空域大小和频率分量对通用对抗扰动攻击性和迁移性的影响。

2 通用对抗扰动生成方法 SFUAN

2.1 问题描述

对于一个目标分类模型 f，输入样本集 $\mathcal{X}$ 中的样本为 x，目标分类模型正确预测样本 x 为类别 $f(x)=c$。寻找一个扰动向量 $\boldsymbol{\delta}$，使得 $x+\boldsymbol{\delta}$ 在视觉上与 x 保持相似，但 $f(x+\boldsymbol{\delta})\neq c$。该优化问题形式化表示为：

$$\begin{aligned}&\min L(x+\boldsymbol{\delta})\\&\text{s.t. } \|\boldsymbol{\delta}\|_p\leqslant\xi\\&\forall x\in\mathcal{X},\ \rho\geqslant 1-\tau\end{aligned}\tag{1}$$

式中：$L(\cdot)$为针对优化问题在不同的攻击场景下定义的损失函数；$\|\cdot\|_p$ 表示使用 l_p 范数对扰动的空域大小进行度量；ξ 对扰动的空域大小进行约束；τ（$0<\tau<1$）表示通用对抗扰动的攻击成功率限度；ρ 表示通用对抗扰动的攻击成功率。在目标攻击场景中，$\rho=\Pr(f(x+\boldsymbol{\delta})=c_{\text{target}})$且 $c_{\text{target}}\neq f(x)$，在非目标攻击场景中，$\rho=\Pr(f(x+\boldsymbol{\delta})\neq f(x))$，$\Pr(\cdot)$表示对应事件的概率。

针对该优化问题，通过最小化损失函数以寻找最优的扰动向量。在本文中，设计损失函数时考虑以下 3 个方面：

1）扰动攻击性。扰动攻击的目的是使得添加扰动后的图像被目标分类模型错误分类。在非目标攻击场景下，错误类别与正确类别不同即可。在目标攻击场景下，错误分类后的类别为预先指定的类别，且指定类别与正确类别不同。

2）扰动大小。添加扰动后的图像需保持与原始图像在视觉上的相似性，通过最小化扰动向量的范数使得添加的扰动尽可能不可见，从而保持图像添加扰动前后的视觉相似性。

3）扰动频率分量。为了分析通用对抗扰动的频率与其攻击性的关系，通过调控扰动中高频分量与低频分量的比重，对比各频率分量对扰动攻击成功率的影响。

2.2 模型结构

本文中通用对抗扰动生成网络 SFUAN 结构如图 1 所示(彩色效果见《计算机工程》官网 HTML 版,下同)。将服从正态分布 $\mathcal{N}(0,1)^{100}$ 的随机向量 $\boldsymbol{y}$ 输入到生成模型 G 中,生成模型通过学习随机向量到通用对抗扰动的映射关系,生成通用对抗扰动 $\boldsymbol{\delta}$。将通用对抗扰动 $\boldsymbol{\delta}$ 缩放后添加到原始样本 x 从而生成对抗样本,将其像素值裁剪至合理的范围内后输入到目标分类模型 f 中。目标模型是具有较高准确率的预训练深度学习分类模型 f,该模型能够对输入的图像进行识别,并给出该图像的预测概率向量。

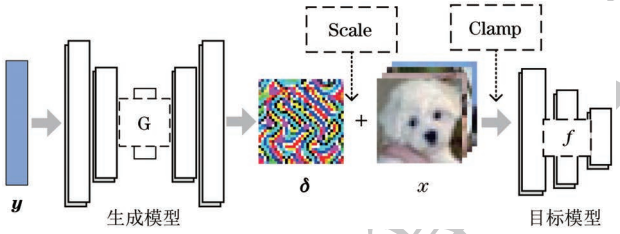

图 1 SFUAN 的结构

Fig. 1 The structure of SFUAN

在 SFUAN 网络模型中,生成模型需要学习输入的随机向量与通用对抗扰动之间的映射关系。生成模型对输入数据进行非线性变换处理,学习输入数据的特征,使用反向传播算法对非线性变换进行优化。随着网络深度的增加,使用更多的非线性变换可以学习到更高阶、更抽象的特征,从而增强对通用对抗扰动的学习能力。因此,为了提高生成模型的能力,在 SFUAN 网络模型中使用深层神经网络模型作为通用对抗扰动的生成模型。ResNet 模型通过引入残差结构实现了跨网络层的直接连接,从而将前一层的特征信息传递到当前层,有效解决了梯度消失和梯度爆炸问题。这种结构允许构建更深层次的网络结构,学习更复杂、更具攻击性的扰动模式,生成具有高攻击性和高迁移性的通用对抗扰动。在 SFUAN 模型中,生成模型基于 ResNet-101,其网络模型结构如表 1 所示。

ResNet-101 模型使用的标准化方法为批量归一化,该方法考虑整个批次的统计信息,更适用于边缘检测、图像分类等判别任务。但是,对于输入生成模型中的随机噪声向量,批量归一化会对其进行标准化处理,从而降低生成模型的多样性。在生成模型的任务中,需要考虑样本多样性和特征分布变化,实例归一化更有利于保留样本独特性,提高生成模型的泛化能力和生成扰动的多样性,因此,在 SFUAN 网络模型中,对 ResNet-101 模型进行调整,将其标准化方式调整为实例归一化,以提高生成扰动的质量和多样性。

表 1 ResNet-101 生成模型结构

Table 1 The structure of ResNet-101 generation model

网络层	输出大小	网络大小
conv1	112×112	7×7, 64, 步长为 2
conv2_x	56×56	3×3 最大池化,步长为 2 [1×1, 64 3×3, 64 1×1, 256] ×3
conv3_x	28×28	[1×1, 128 3×3, 128 1×1, 512] ×4
conv4_x	14×14	[1×1, 256 3×3, 256 1×1, 1 024] ×23
conv5_x	7×7	[1×1, 512 3×3, 512 1×1, 2 048] ×3
全连接层	1×1	平均池化 1 000-d 全连接层 Tanh

2.3 目标函数

本文在训练扰动生成网络 SFUAN 的过程中,使用对抗样本置信度损失优化扰动的攻击成功率,使用扰动空域距离损失最小化扰动大小,使用扰动频率引导损失控制扰动中高频分量与低频分量的强度。损失函数以不同的权重联合对抗样本置信度损失、扰动空域距离损失和扰动频率引导损失,在空域和频域优化扰动生成模型。

1)对抗样本置信度损失。

对于输入样本集 $\mathcal{X}$ 中的原始样本 x,目标分类模型 f 对原始样本 x 预测的类别标签为 c_0。在非目标攻击场景下,攻击成功的标志是使得对抗样本的预测类别与原始类别 c_0 不同,即该类别的置信度高于原始类别 c_0 的置信度。在非目标攻击场景中,对抗样本置信度损失如式(2)所示:

$$L_{\mathrm{nt}}=\ln[f(x_{\mathrm{adv}})]_{c_0}-\max_{i\neq c_0}\ln[f(x_{\mathrm{adv}})]_i \tag{2}$$

式中:x_{adv} 表示添加扰动后的对抗样本。

在目标攻击场景下,预测类别为预选的目标类别 c_{target},攻击成功的标志是使得对抗样本的预测类别 c_{target} 与原始类别 c_0 不同,目标类别 c_{target} 的置信度高于原始类别 c_0 的置信度。目标攻击场景中对抗样本置信度损失如式(3)所示:

$$L_{\mathrm{t}}=\max_{i\neq c_{\mathrm{target}}}\ln[f(x_{\mathrm{adv}})]_i-\ln[f(x_{\mathrm{adv}})]_{c_{\mathrm{target}}} \tag{3}$$

2)扰动空域距离损失。

在训练通用对抗扰动生成模型的过程中,不仅要提高扰动的攻击性,还要增强扰动的隐蔽性,使得攻击更加难以防范,在不引起防御者注意的情况下改变模型的预测结果。为了生成高质量的通用对抗扰动,对扰动的大小进行优化。在衡量两张图像之间的感知差异时,通常使用范数 l_p 进行度量。在本文生成模型的训练过程中,使用 l_∞ 范数计算扰动的空间域大小,用 l_∞ 范数度量扰动的最大像素值。l_∞ 范数定义如式(4)所示:

$$l_\infty=\|t\|_p=\max(|t_1|,|t_2|,\cdots,|t_i|) \tag{4}$$

式中:t 表示用于度量的图像;t_i 表示该图像的各个像素值。对抗扰动空域距离损失如式(5)所示:

$$L_{\mathrm{d}}=\|\boldsymbol{\delta}\|_\infty \tag{5}$$

3)扰动频率引导损失。

本文提出扰动频率引导损失,首先使用傅里叶变换将通用对抗扰动由空间域转为频率域,通过低通滤波和高通滤波提取扰动的高频分量和低频分量,在训练生成模型的过程中,利用频率引导系数控制生成模型对扰动高频和低频分量的关注程度,设置不同的引导系数以生成具有不同比例高频和低频成分的通用对抗扰动,从而对比扰动中各频率分量与扰动攻击性的关系。

使用傅里叶变换将图像从空间域变换为频率域。经傅里叶变换后,对频谱进行中心化处理,将低频率分量置于频谱中心。本文使用理想滤波技术对图像频率域进行滤波操作。其中,理想低通滤波器在以原点为圆心、以 D_0 为半径的圆内无衰减地通过所有的频率,而在圆外则阻止任何频率通过。理想低通滤波器形式化表达如下:

$$L(u,v)=\begin{cases}1, & D(u,v)\leqslant D_0\\ 0, & D(u,v)>D_0\end{cases} \tag{6}$$

式中:D_0 为正常数;$D(u,v)$是到频率域中点(u,v)的欧氏距离。$D(u,v)$的计算如式(7)所示:

$$D(u,v)=\sqrt{\left(u-\frac{P}{2}\right)^2+\left(v-\frac{Q}{2}\right)^2} \tag{7}$$

式中:P 和Q 表示图像的尺寸。

理想高通滤波器在以原点为圆心、以 D_0 为半径的圆外无衰减地通过所有的频率,而在圆内则阻止任何频率通过。理想高通滤波器形式化表达如下:

$$H(u,v)=\begin{cases}0, & D(u,v)\leqslant D_0\\ 1, & D(u,v)>D_0\end{cases} \tag{8}$$

使用理想高通滤波器和理想低通滤波器分别提取扰动的高频分量和低频分量,对提取的高频分量和低频分量取平均值,由频率引导系数控制生成模型对扰动高频和低频分量的关注程度,扰动频率引导损失如式(9)所示:

$$L_{\mathrm{f}}=\lambda\frac{\sum(\delta_{\mathrm{fct}}\times L(u,v))}{\sum L(u,v)}+(1-\lambda)\frac{\sum(\delta_{\mathrm{fct}}\times H(u,v))}{\sum H(u,v)} \tag{9}$$

式中:δ_{fct} 表示经过中心化处理的扰动频谱信息;λ 为频率引导系数。在生成模型训练过程中,当 $\lambda=0$ 时,引导生成模型减少扰动的高频分量,增强低频分量的比重,称为“高频优化”;当 $\lambda=1$ 时,引导生成模型减少扰动的低频分量,增强高频分量的比重,称为“低频优化”。

综上所述,在目标攻击场景下,生成模型的损失函数如式(10)所示:

$$L_{\mathrm{Gt}}=\beta_1L_{\mathrm{t}}+\beta_2L_{\mathrm{d}}+\beta_3L_{\mathrm{f}} \tag{10}$$

式中:β_1、β_2、β_3 为各损失的权重。在非目标攻击场景下,生成模型的损失函数如式(11)所示:

$$L_{\mathrm{Gnt}}=\beta_1L_{\mathrm{nt}}+\beta_2L_{\mathrm{d}}+\beta_3L_{\mathrm{f}} \tag{11}$$

2.4 生成过程

本文联合对抗样本置信度损失、扰动空域距离损失和扰动频率引导损失,从空域和频域角度联合优化扰动生成模型。通用对抗扰动生成模型的训练过程如图 2 所示。

将符合正态分布 $\mathcal{N}(0,1)^{100}$ 的随机噪声向量 $\boldsymbol{y}$ 输入到生成模型 G 中,生成模型通过学习随机噪声到通用对抗扰动的映射关系生成通用对抗扰动 $\boldsymbol{\delta}$,将 $\boldsymbol{\delta}$ 与标量 $\alpha\in\left(0,\frac{\varepsilon}{\|\boldsymbol{\delta}\|_p}\right]$ 相乘,对扰动中的像素上限值进行限制,其中,ε 表示允许的最大扰动阈值,$p=\infty$。将经过缩放的扰动 $\boldsymbol{\delta}'=\alpha\times\boldsymbol{\delta}$ 与输入样本集 $\mathcal{X}$ 中的样本 x 相加得到对抗样本,将其像素值裁剪至合理的范围内得到对抗样本 $x_{\mathrm{adv}}=\mathrm{Clamp}(x+\boldsymbol{\delta}')$,将其输入到目标分类模型 f 后能够得到该对抗样本的概率向量。通过概率向量计算在目标攻击和非目标攻击场景下对抗样本 x_{adv} 的置信度损失 L_1。对通用对抗扰动 $\boldsymbol{\delta}'$ 进行傅里叶变换及中心化变换,将扰动从空域转换到频域,使用理想滤波技术对扰动频率域进行滤波操作后计算其扰动频率引导损失 L_2。使用 l_∞ 范数计算通用对抗扰动 $\boldsymbol{\delta}'$ 的空域距离损失 L_3。通过不同的权重联合置信度损失 L_1、扰动频率引导损失 L_2 和扰动空域距离损失 L_3 作为损失函数,训练通用对抗扰动生成模型。

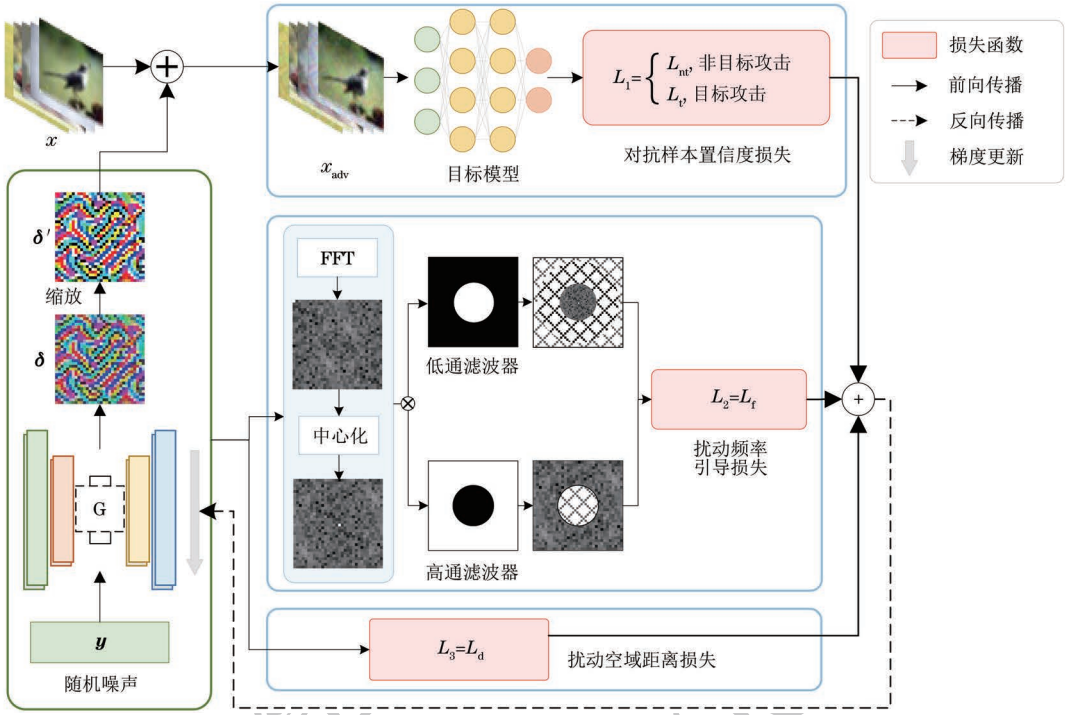

图 2 通用对抗扰动生成模型的训练过程

Fig. 2 The training process of UAP generation model

随着训练的迭代进行，生成模型通过置信度损失 L_1 逐渐学习到如何生成攻击性更强的通用对抗扰动，通过扰动频率引导损失 L_2 控制生成的通用对抗扰动中高频分量和低频分量的比重，通过扰动空域距离损失 L_3 优化扰动的空域大小，生成更加微小的通用对抗扰动。

3 实验结果与分析

3.1 实验设置与数据集

本文提出的通用对抗扰动生成方法基于 PyTorch 框架实现，实验的软硬件配置如表 2 所示。本文使用 CIFAR-10 数据集进行通用对抗扰动的生成实验。CIFAR-10 数据集由 60 000 张 32×32 像素的彩色图片组成，包括 Airplane、Automobile、Bird、Cat、Deer、Dog、Frog、Horse、Ship、Truck 共 10 个类别，每个类别有 6 000 张图像，是广泛用于机器学习和计算机视觉的公开数据集。该数据集的训练集和测试集分别包含 50 000 张和 10 000 张图像。

在非目标攻击实验和目标攻击实验中，设置 β_1、β_2、β_3 分别为 1、4、2 以控制损失函数中各部分损失的权重。消融实验中设置 $\beta_3=0$，生成模型将不直接改变扰动中高频分量和低频分量的强度，称为“无频率优化”。在实验中，使用 l_∞ 范数计算通用对抗扰动的空间域大小，限制扰动空间域大小为 0.04，实验迭代次数为 20 次，批次大小为 64，理想滤波器的 $D_0=11$。

表 2 软硬件配置信息

Table 2 Software and hardware configuration information

实验环境	相关配置
操作系统	Ubuntu 16.04
内存大小/GB	32
CPU	Intel® E5-2620 2.10 GHz
GPU	NVIDIA TITAN Xp 12 GB
Python	3.6.13
PyTorch	1.10.0+cu113
torchvision	0.11.0+cu113

3.2 非目标攻击实验及分析

本节展示在 CIFAR-10 数据集上使用 SFUAN 进行非目标攻击的实验结果。在非目标攻击场景下，分别从不同频域分量对扰动攻击性的影响、不同频域分量对扰动迁移性的影响、对比实验、模型复杂度这 4 个方面展开实验分析。

1）频域分量对扰动非目标攻击性的影响。

本节分别利用高频优化、低频优化和无频率优化训练生成模型，对比不同频率分量下扰动非目标攻击成功率。在训练实验中，目标模型为在 CIFAR-10 数据集上经过训练的 GoogLeNet 模型。使用 SFUAN 模型在 CIFAR-10 数据集的训练集上生成通用对抗扰动，分别保留其全部频率分量、高频分量、低频分量，在测试集上进行攻击实验。表 3 展示了使用本文方法在 3 种频率引导条件下的非目标攻击成功率（最优结果加粗标注，下同）。

表 3 不同频率优化的非目标攻击成功率对比

Table 3 Comparison of success rates of non-target attacks optimized at different frequencies %

频率优化	训练集攻击成功率	测试集攻击成功率		
		全部保留	保留高频	保留低频
高频优化	**86.71**	**84.49**	14.18	**63.55**
低频优化	80.67	80.50	**32.68**	17.69
无频率优化	82.68	82.13	26.68	28.11

分析表 3 中的数据可以发现，高频优化训练生成的通用对抗扰动具有更高的攻击成功率，对比低频优化训练生成的扰动，在训练集和测试集上的攻击成功率分别提高了约 6 和 4 百分点，对比无频率优化训练生成的扰动，在训练集和测试集上的攻击成功率分别提高了约 4 和 2 百分点。低频优化训练生成的扰动在训练集和测试集上的攻击成功率最低。由于高频优化条件的限制，在通用对抗扰动的生成过程中，生成模型偏向于在低频区域生成扰动，生成的扰动中低频分量比重较高，保留低频分量作为对抗扰动时的攻击成功率高于仅保留高频分量。反之，低频优化条件下生成的扰动中低频分量比重较低，保留低频分量作为对抗扰动时的攻击成功率低于仅保留高频分量。无频率优化条件下保留低频分量时的攻击成功率略高于仅保留高频分量的对抗扰动。据此可知，本文方法使用的高频优化和低频优化策略是有效的，且通用对抗扰动的低频分量对其非目标攻击影响更大，低频分量越多，其攻击成功率越高。

2）频率分量对非目标攻击扰动迁移性的影响。

为了分析不同频域分量对非目标攻击场景下通用对抗扰动的迁移性的影响，分别选择 GoogLeNet、DenseNet-121、VGG-19 和 ResNet-101 作为目标模型，使用本文提出的通用对抗扰动生成方法在高频优化、低频优化和无频率优化条件下生成通用对抗扰动。在 CIFAR-10 测试集上，使用生成的通用对抗扰动分别攻击 4 个模型。实验中使用的 4 个模型的网络结构和深度各不相同，且在图像分类任务中表现出极高的准确率。针对本文提出的通用对抗扰动生成算法，分析非目标攻击场景下不同频率优化对扰动迁移性的影响，对 CIFAR-10 数据集的 50 000 张训练集生成通用对抗扰动，生成的扰动在 10 000 张测试集上的攻击成功率如表 4 所示。

表 4 不同频率优化的非目标攻击迁移性对比

Table 4 Comparison of non-target attacks transferability optimization at different frequencies %

频率优化	目标模型	攻击模型的攻击成功率				
		GoogLeNet	DenseNet-121	VGG-19	ResNet-101	平均
高频优化	GoogLeNet	84.49	69.65	54.45	29.60	59.55
	DenseNet-121	81.19	84.28	65.73	64.95	74.04
	VGG-19	79.74	74.58	80.10	34.04	67.12
	ResNet-101	82.94	67.28	47.54	88.22	71.50
低频优化	GoogLeNet	80.50	64.20	47.59	29.16	55.36
	DenseNet-121	66.27	81.36	40.80	61.75	62.55
	VGG-19	75.56	67.03	76.39	28.76	61.94
	ResNet-101	77.17	62.83	43.55	85.69	67.31
无频率优化	GoogLeNet	82.13	64.53	51.17	27.49	56.33
	DenseNet-121	76.25	84.24	55.98	62.07	69.64
	VGG-19	73.67	65.97	76.28	32.77	62.17
	ResNet-101	74.27	64.15	45.69	87.61	67.93

根据表 4 所示，相较于在低频优化和无频率优化条件下生成的通用对抗扰动，在高频优化条件下生成的非目标通用对抗扰动迁移性攻击具有更高的攻击成功率，按照平均值计算，高频优化条件下生成的通用对抗扰动迁移性攻击成功率比低频优化条件下高 4～10 百分点，比无频率优化条件下高 3～5 百分点。在低频优化和无频率优化条件下生成的通用对抗扰动迁移性攻击成功率则不稳定，使用目标模型 VGG-19、GoogLeNet 和 DenseNet-121 生成的通用对抗扰动攻击 ResNet-101 模型时，在低频优化条件下生成的通用对抗扰动迁移性攻击成功率高于无频率优化条件，目标模型为 GoogLeNet 和 DenseNet-121 攻击 VGG-19、GoogLeNet 和 DenseNet-121 模型时则相反。通过对实验数据进行分析可知，对于非目标攻击而言，通用对抗扰动的低频分量对其迁移性攻击成功率影响较大，低频分量越多，其迁移性

攻击成功率越高。

3)对比实验分析。

上述实验结果表明,在高频优化条件下生成的非目标通用对抗扰动具有较好的攻击性和迁移性。为了验证本文提出的通用对抗扰动生成方法的有效性,选择 ResNet-101、DenseNet-121 和 VGG-19 模型作为目标模型和攻击模型,使用本文方法高频优化训练生成通用对抗扰动,并与 UAN、UAP、SGA 攻击算法生成的扰动进行对比。使用各模型在 CIFAR-10 训练集上生成通用对抗扰动,在 CIFAR-10 测试集上的攻击成功率及在训练过程中每个 Epoch 所用的时间如表 5 所示。

表 5 对比实验结果

Table 5 Comparative experimental results

攻击方法	数据集	ResNet-101		DenseNet-121		VGG-19	
		攻击成功率/%	训练时间/s	攻击成功率/%	训练时间/s	攻击成功率/%	训练时间/s
UAP	训练集	76.30	2 943	68.40	2 764	59.90	924
	测试集	76.00		67.90		57.20	
UAN	训练集	83.20	**151**	75.30	**127**	64.90	**64**
	测试集	85.10		75.00		66.60	
SGA	训练集	83.79	634	**85.93**	560	79.05	118
	测试集	79.63		**86.16**		79.57	
SFUAN	训练集	**87.67**	181	85.01	154	**80.87**	96
	测试集	**88.22**		84.28		**80.10**	

从表 5 可以看出,相较于 UAN、UAP 攻击算法,使用本文方法在高频优化条件下生成的通用对抗扰动在训练集和测试集上均具有更强的攻击性。对 ResNet-101 模型展开攻击实验时,本文方法较 UAN 方法提升幅度较小,较 UAP 和 SGA 方法提升约 4 百分点。依据对 DenseNet-121 目标模型的攻击成功率,本文方法较 UAN 方法提升 10 百分点左右,较 UAP 方法提升 15 百分点左右,较 SGA 方法低约 1 百分点。对 VGG-19 目标模型进行攻击时,本文方法较 UAN 和 UAP 方法提升幅度较大,约为 15～23 百分点,较 SGA 方法提升约 1 百分点。由于 UAP 和 SGA 方法根据目标模型的判别结果,通过梯度更新生成通用对抗扰动,因此其每个 Epoch 所需的训练时间较长。而本文方法对比 UAN 方法增加了频域优化部分,因此增加了每个 Epoch 所需的训练时间。综合分析攻击成功率和每个 Epoch 所需的训练时间,本文方法是可行且有效的,并具有明显的优势。

4)模型复杂度分析。

为了分析所提模型的复杂度,使用 THOP(Torch-OpCounter)工具统计了所提模型的计算量 FLOPs 和参数量 Params,并分别替换主干网络为 ResNet-18、ResNet-34、ResNet-50、ResNet-152 后进行对比分析。选择 ResNet-101、DenseNet-121 和 VGG-19 模型作为目标模型和攻击模型,使用本文方法高频优化训练生成通用对抗扰动。在实验中,限制相对扰动大小为 0.04,实验迭代次数为 20 次,批次大小为 64。使用 CIFAR-10 训练集生成通用对抗扰动,在 CIFAR-10 测试集上的攻击成功率、训练过程中每个 Epoch 所需的时间及模型的计算量 FLOPs 和参数量 Params 如表 6 所示。

表 6 模型性能对比

Table 6 Model performance comparison

主干网络	数据集	ResNet-101		DenseNet-121		VGG-19		模型复杂度	
		攻击成功率/%	训练时间/s	攻击成功率/%	训练时间/s	攻击成功率/%	训练时间/s	FLOPs	Params
ResNet-18	训练集	83.63	**131**	81.71	**105**	74.63	**40**	**616 538 624**	**12 800 512**
	测试集	82.29		82.21		77.98			
ResNet-34	训练集	84.20	142	82.18	114	77.32	54	1 222 156 800	22 908 672
	测试集	85.10		82.65		76.62			
ResNet-50	训练集	85.26	156	83.98	127	78.24	70	1 374 947 328	29 850 624
	测试集	84.63		84.25		79.51			
ResNet-101	训练集	87.67	181	85.01	154	80.87	96	2 593 785 856	48 842 752
	测试集	**88.22**		84.28		80.10			
ResNet-152	训练集	**88.12**	248	**85.24**	209	**81.14**	147	3 814 197 248	64 486 400
	测试集	88.07		**84.40**		**80.97**			

根据表 6 所示，生成的通用对抗扰动攻击成功率与模型的计算量和参数量呈正相关。对比使用 ResNet-18、ResNet-34、ResNet-50 为主干网络的实验结果可知，使用 ResNet-101 作为主干网络生成的通用对抗扰动攻击成功率提升较大，使用 ResNet-101 和 ResNet-152 作为主干网络生成的通用对抗扰动攻击成功率差异较小。分析每个 Epoch 所需的训练时间，当选择 ResNet-101 作为主干网络时，相较于使用 ResNet-18，每个 Epoch 的训练时间大约增加了 50 s，并且攻击成功率提升了约 4 百分点。然而，选择 ResNet-152 为主干网络时，相较于使用 ResNet-101，虽然每个 Epoch 的训练时间同样增加了约 50 s，但攻击成功率的提升相对有限，约为 0.5～1 百分点。综合考虑模型的计算量 FLOPs 和参数量 Params、扰动生成速度及扰动攻击成功率，本文所提方法使用 ResNet-101 作为主干网络是合理的。

3.3 目标攻击实验及分析

本节将展示在 CIFAR-10 数据集上使用本文方法进行目标攻击的实验结果及相关分析。在目标攻击场景下，分别从频率分量对扰动目标攻击性的影响、频率分量对扰动目标攻击迁移性的影响、扰动空间大小对扰动目标攻击性的影响 3 个方面展开实验和分析。

1）频率分量对扰动目标攻击性的影响。

本节对比了不同频率优化下扰动目标攻击成功率。在训练实验中，目标模型为在 CIFAR-10 数据集上经过训练的 GoogLeNet 模型，使用 CIFAR-10 训练集在高频优化、低频优化和无频率优化条件下生成通用对抗扰动。在实验中使用 CIFAR-10 数据集的训练集生成通用对抗扰动的攻击成功率和在测试集上的攻击成功率如表 7 所示。

表 7 不同频率优化的扰动目标攻击成功率

Table 7 Success rate of perturbation target attacks optimized at different frequencies %

类别	高频优化		低频优化		无频率优化	
	训练集攻击成功率	测试集攻击成功率	训练集攻击成功率	测试集攻击成功率	训练集攻击成功率	测试集攻击成功率
Airplane	**90.50**	**90.10**	88.98	82.35	88.76	89.17
Automobile	**90.81**	84.59	89.03	**88.92**	87.57	87.43
Bird	**95.28**	**95.13**	87.41	87.67	90.10	90.58
Cat	**94.63**	**93.86**	93.49	93.34	93.64	93.75
Deer	**90.59**	**90.20**	72.00	72.06	73.31	73.46
Dog	**78.88**	**77.88**	77.87	77.70	76.93	76.84
Frog	**95.91**	**95.08**	92.95	93.05	90.59	90.56
Horse	**85.48**	**85.12**	82.99	77.67	84.44	84.62
Ship	92.22	91.15	**92.60**	**92.71**	90.95	91.15
Truck	**90.39**	**89.91**	89.41	89.15	87.49	82.37
平均	**90.47**	**89.30**	86.67	85.46	86.38	85.99

通过实验发现，在 10 个类别中，高频优化条件下生成的通用对抗扰动在训练集上攻击成功率最高的类别有 Airplane、Automobile、Bird、Cat、Deer、Dog、Frog、Horse、Truck，共 9 种类别，低频优化条件下生成的通用对抗扰动在训练集上攻击成功率最高的类别有 Ship 类别，共 1 种类别，无频率优化条件下生成的通用对抗扰动在训练集上攻击成功率没有达到最高的类别。在测试集中，高频优化条件下生成的通用对抗扰动攻击成功率最高的类别有 Airplane、Bird、Cat、Deer、Dog、Frog、Horse、Truck，共 8 种类别，低频优化条件下生成的通用对抗扰动攻击成功率最高的类别有 Automobile、Ship 类别，共 2 种类别，无频率优化条件下生成的通用对抗扰动攻击成功率没有达到最高的类别。对 3 种优化条件下生成的通用对抗扰动在 10 个类别训练集和测试集上的攻击成功率分别求平均值，高频优化条件下生成的通用对抗扰动攻击成功率最高，在训练集和测试集上的攻击成功率分别为 90.47% 和 89.30%。

图 3 展示了高频优化条件下生成的所有类别的目标攻击通用对抗扰动及攻击效果。每一列图像的原始类别相同，在下方显示图像的原始类别；

每一行图像的目标攻击类别相同，即攻击模型对同一行图像的预测类别相同，在左侧显示目标攻击类别；在最后一列展示各目标攻击类别的通用对抗扰动。

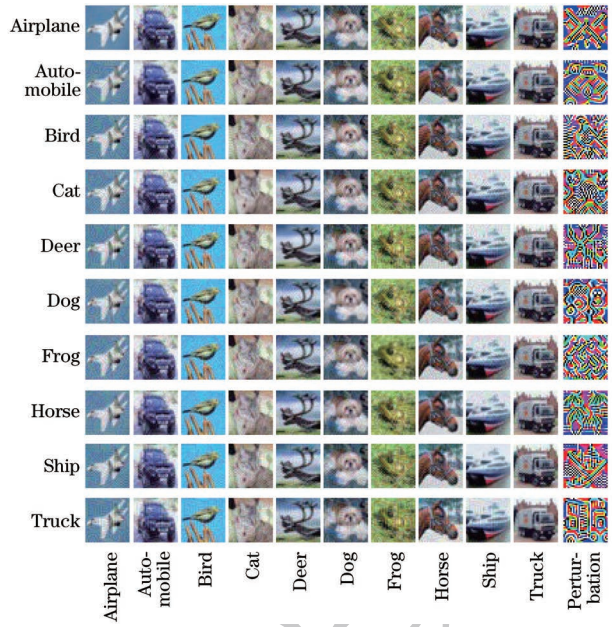

图 3 高频优化的扰动目标攻击效果

Fig. 3 The effect of high-frequency optimized perturbation target attack

2）频率分量对目标攻击扰动迁移性的影响。

为了分析在目标攻击场景下不同频率分量对扰动迁移性的影响，分别选择 GoogLeNet、DenseNet-121 和 VGG-19 作为目标模型，在高频优化、低频优化和无频率优化条件下生成目标攻击类别为 Airplane 的通用对抗扰动。在 CIFAR-10 测试集上，分别攻击 GoogLeNet、DenseNet-121 和 VGG-19 模型。使用 CIFAR-10 数据集的 45 000 张训练集（不包含目标攻击类别为 Airplane 的训练图像）生成通用对抗扰动，对于 GoogLeNet、DenseNet-121 和 VGG-19 模型，生成的扰动在 CIFAR-10 数据集的 9 000 张测试集（不包含目标攻击类别为 Airplane 的测试图像）上的攻击成功率如表 8 所示。

根据表 8 所示，相比于在低频优化和无频率优化条件下生成的目标攻击类别为 Airplane 的通用对抗扰动，在高频优化条件下生成的通用对抗扰动具有更高的迁移性攻击成功率，且在低频优化条件下生成的通用对抗扰动迁移性攻击成功率最低。实验数据表明，对于图像通用对抗扰动目标攻击，扰动的低频分量对其迁移性攻击成功率影响较大，保留更多的低频分量能够使通用对抗扰动具有更高的迁移性攻击成功率。

表 8 不同频率优化的扰动目标攻击迁移性对比

Table 8 Comparison of transferability of perturbation target attacks optimized at different frequencies %

频率优化	目标模型	攻击模型的攻击成功率		
		GoogLeNet	DenseNet-121	VGG-19
高频优化	GoogLeNet	**90.10**	**70.57**	**42.53**
	DenseNet-121	**49.88**	**91.88**	**46.77**
	VGG-19	**52.99**	**77.23**	**86.72**
低频优化	GoogLeNet	82.35	55.07	28.96
	DenseNet-121	40.45	88.08	41.27
	VGG-19	17.73	47.56	72.18
无频率优化	GoogLeNet	89.17	68.73	34.34
	DenseNet-121	44.13	89.50	43.28
	VGG-19	30.24	63.47	78.01

3）攻击成功率与扰动空域大小关系分析。

在训练生成模型的实验中，限制通用对抗扰动的扰动空域大小为 0.04。为了研究对抗扰动的大小与通用对抗扰动攻击成功率的关系，设置不同的相对扰动大小，在 CIFAR-10 数据集的测试集上进行攻击实验。图 4 分别展示了在高频优化、低频优化和无频率优化条件下生成的 CIFAR-10 数据集中所有类别的通用对抗扰动在测试集上的攻击成功率与扰动空域大小之间的关系。

如图 4 所示，在扰动空域大小相同的情况下，高频优化条件下生成的通用对抗扰动攻击成功率相比于低频优化和无频率优化条件下生成的扰动攻击成功率较高，部分类别则持平。当扰动空域大小限制在 0.03 时，使用高频优化生成的大多数类别的通用对抗扰动的目标攻击成功率超过 60%，低频优化和无频率优化条件下生成的通用对抗扰动攻击成功率则低于或在 60%附近。使用高频优化生成的目标攻击通用对抗扰动能够在更小的扰动空域大小限制下达到 100%的攻击成功率，如 Airplane、Bird、Deer、Frog 等类别。随着扰动空域大小的逐渐增加，3 种频率优化条件下生成的通用对抗扰动攻击成功率最终均能达到 100%。

图 5 展示了各攻击类别在扰动空域大小为 0、0.02、0.04、0.06、0.08 和 0.10 时的攻击效果，各列下方显示了目标攻击的类别。

综上所述，在扰动空域大小相同的情况下，高频优化生成的通用对抗扰动具有更高的攻击成功率。通用对抗扰动较小时仍有一定的概率攻击成功，这表明通用对抗扰动具有强大的攻击性，增强神经网络模型的鲁棒性刻不容缓。

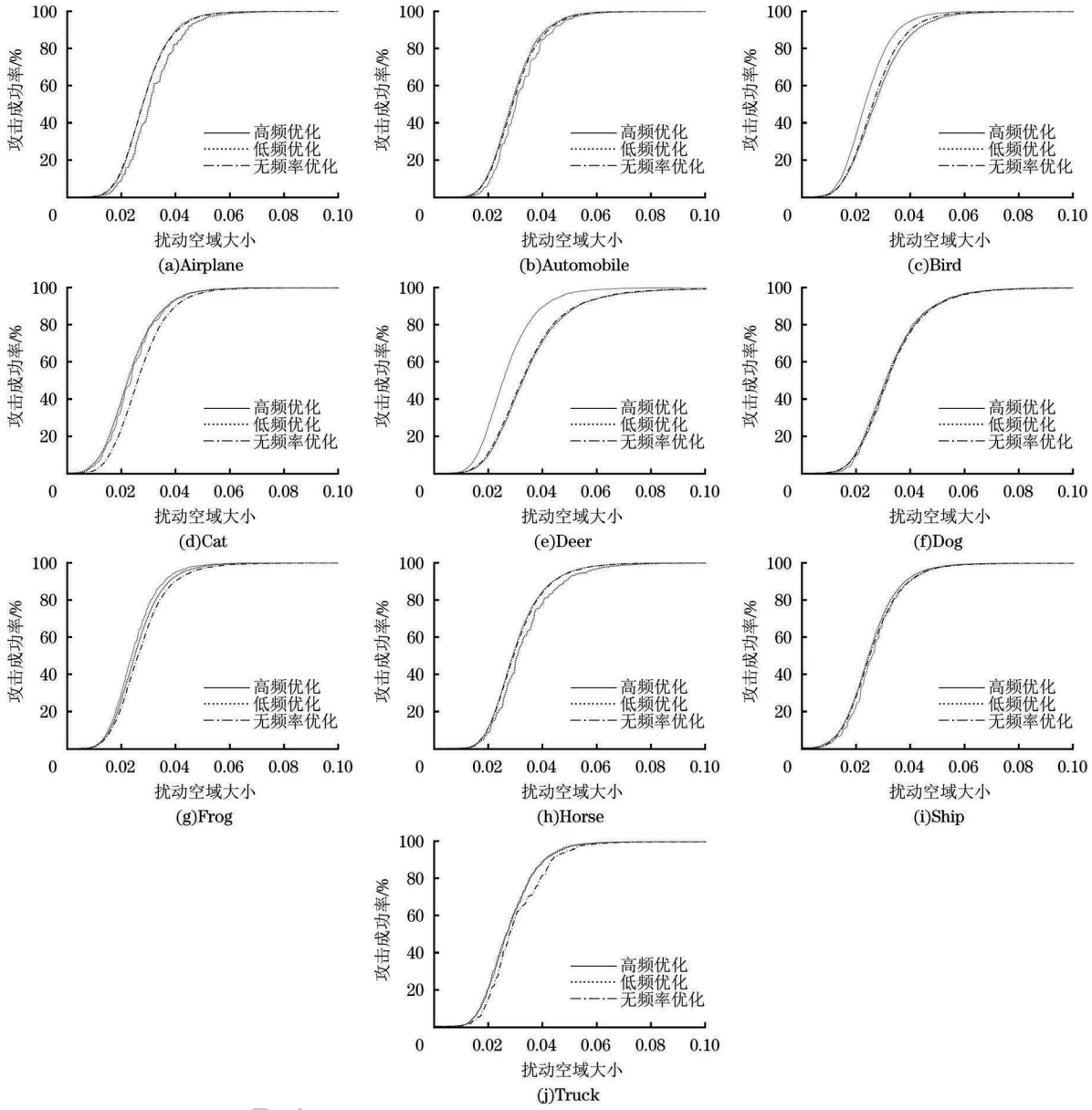

图 4 攻击成功率与扰动空域大小的关系

Fig.4 The relationship between attack success rate and disturbance airspace size

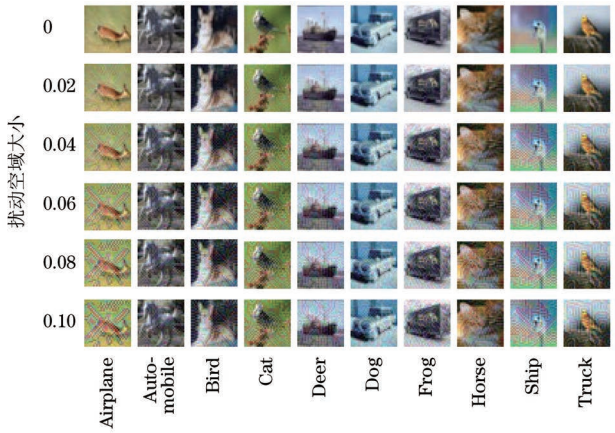

图 5 不同扰动空域大小下的目标攻击效果

Fig.5 Target attack effectiveness under different disturbance airspace sizes

4 结束语

本文提出空频域联合优化的通用对抗扰动生成方法，分析了扰动的频域分量和空域大小对通用对抗扰动攻击性的影响。该方法使用对抗样本置信度损失优化扰动的攻击性，使用频率引导系数控制生成模型对扰动高频和低频分量的关注程度，使用扰动空域距离损失优化扰动的空域大小，从空域和频域角度联合优化扰动生成模型。实验结果表明，通用对抗扰动的低频分量对扰动的攻击性影响较大，低频分量越多，其攻击性越强，且较小的通用对抗扰动仍具有攻击性。为了构建安全可靠的神经网络模型应用环境，对于通用对抗扰动存在的原因仍需进行深入研究。未来将从深度模型可解释性角度研究

通用对抗扰动的生成机制及其对图像分类模型鲁棒性的影响。

参考文献

[1] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2015: 1-9.

[2] REN H L, HUANG T, YAN H Y. Adversarial examples: attacks and defenses in the physical world[J]. International Journal of Machine Learning and Cybernetics, 2021, 12(11): 3325-3336.

[3] 冯博, 刘万平, 南海. 结合最大内接圆的图像对抗样本生成算法[J]. 小型微型计算机系统, 2024, 45(6): 1436-1443.
FENG B, LIU W P, NAN H. Image adversarial examples generation algorithm combined with maximum inscribed circle[J]. Journal of Chinese Computer Systems, 2024, 45(6): 1436-1443. (in Chinese)

[4] TANG L, YE D, LV Y, et al. Once and for all: universal transferable adversarial perturbation against deep hashing-based facial image retrieval[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI Press, 2024: 5136-5144.

[5] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2017: 86-94.

[6] XU K, QIN M H, SUN F, et al. Learning in the frequency domain[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2020: 1737-1746.

[7] JIA Z Y, LIN Y F, CAI X Y, et al. SST-EmotionNet: spatial-spectral-temporal based attention 3D dense network for EEG emotion recognition[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York, USA: ACM Press, 2020: 2909-2917.

[8] SONG X F, XU D H, PENG C, et al. A two-stage frequency-domain generation algorithm based on differential evolution for black-box adversarial samples[J]. Expert Systems with Applications, 2024, 249: 123741.

[9] 陈宇飞, 沈超, 王骞, 等. 人工智能系统安全与隐私风险[J]. 计算机研究与发展, 2019, 56(10): 2135-2150.
CHEN Y F, SHEN C, WANG Q, et al. Security and privacy risks in artificial intelligence systems[J]. Journal of Computer Research and Development, 2019, 56(10): 2135-2150. (in Chinese)

[10] WANG D H, YAO W, JIANG T S, et al. Improving transferability of universal adversarial perturbation with feature disruption[J]. IEEE Transactions on Image Processing, 2024, 33: 722-737.

[11] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[EB/OL]. [2024-02-05]. https://arxiv.org/abs/1607.02533v3.

[12] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. DeepFool: a simple and accurate method to fool deep neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2016: 2574-2582.

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

[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] MOPURI K R, GARG U, VENKATESH BABU R. Fast feature fool: a data independent approach to universal adversarial perturbations[EB/OL]. [2024-02-05]. https://arxiv.org/pdf/1707.05572.

[16] MOPURI K R, GANESHAN A, BABU R V. Generalizable data-free objective for crafting universal adversarial perturbations[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(10): 2452-2465.

[17] ZHANG C N, BENZ P, KARJAUV A, et al. Data-free universal adversarial perturbation and black-box attack[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Washington D.C., USA: IEEE Press, 2021: 7848-7857.

[18] YE Z X, CHENG X W, HUANG X L. FG-UAP: feature-gathering universal adversarial perturbation[C]//Proceedings of the International Joint Conference on Neural Networks (IJCNN). Washington D.C., USA: IEEE Press, 2023: 1-8.

[19] ZHANG Y H, RUAN W J, WANG F, et al. Generalizing universal adversarial perturbations for deep neural networks[J]. Machine Learning, 2023, 112(5): 1597-1626.

[20] LIU X N, ZHONG Y Y, ZHANG Y H, et al. Enhancing generalization of universal adversarial perturbation through gradient aggregation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Washington D.C., USA: IEEE Press, 2023: 4412-4421.

[21] LIU Y, LI C, WANG Z C, et al. Transferable adversarial attack based on sensitive perturbation analysis in frequency domain[J]. Information Sciences, 2024, 678: 120971.

[22] WANG H H, WU X D, HUANG Z Y, 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 (CVPR). Washington D.C., USA: IEEE Press, 2020: 8681-8691.

[23] YIN D, LOPES R G, SHLENS J, et al. A Fourier perspective on model robustness in computer vision[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Washington D.C., USA: IEEE Press, 2019: 13276-13286.

[24] WANG Y, SUN Q D, RONG D Z, et al. Multi-domain awareness for compressed deepfake videos detection over social networks guided by common mechanisms between artifacts[J]. Computer Vision and Image Understanding, 2024, 247: 104072.

[25] GUO C, FRANK J S, WEINBERGER K Q. Low frequency adversarial perturbation[EB/OL]. [2024-02-05]. https://arxiv.org/abs/1809.08758.

[26] CAO H, SUN Q D, LI Y Q, et al. Efficient history-driven adversarial perturbation distribution learning in low frequency domain[J]. ACM Transactions on Privacy and Security, 2024, 27(1): 1-25.

[27] WENG J J, LUO Z M, LIN D Z, et al. Comparative evaluation of recent universal adversarial perturbations in image classification[J]. Computers & Security, 2024, 136: 103576.

[28] SHARMA Y, DING G W, BRUBAKER M A. On the effectiveness of low frequency perturbations[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. New York, USA: ACM Press, 2019: 3389-3396.

[29] DENG Y P, KARAM L J. Frequency-tuned universal adversarial attacks on texture recognition[J]. IEEE Transactions on Image Processing, 2022, 31: 5856-5868.

文字编辑 吴云芳
栏目编辑 宋 圆