基于纹理和颜色感知距离的对抗样本生成算法

徐 明*，蒋奔驰

(杭州电子科技大学网络空间安全学院 杭州 310016)

【摘要】理想的对抗样本不仅要成功欺骗机器学习分类器，同时还应不易被人类视觉感知到差异。传统的算法仅采用 L_p 范数衡量对抗样本扰动的大小，往往导致视距差异与感官不匹配等问题。该文提出了一种基于纹理和颜色感知距离的对抗样本生成算法 (Aho-λ)，其基本原理是尽可能地将扰动嵌入原始图像的高纹理区域，且基于颜色感知距离构建损失函数，从而降低原始图像和对抗样本之间的视距差异，最后利用自适应参数调节算法加快训练的收敛速度。在相近的 L_p 范数和可迁移性情形下，与 DDN 和 C&W 算法相比，该算法生成的对抗样本颜色感知距离更低，而且能以更少的迭代次数更快地生成对抗样本。

关 键 词 对抗样本； 自适应训练； 无感； 颜色感知距离

中图分类号 TP182 **文献标志码** A doi:10.12178/1001-0548.2021058

Adversarial Examples Generation Method Based on Texture and Perceptual Color Distance

XU Ming* and JIANG Ben-chi

(The School of Cyberspace, Hangzhou Dianzi University Hangzhou 310016)

Abstract Ideal adversarial examples should not only successfully deceive the machine learning classifier, but also should not easily be perceived by human vision. In the traditional algorithms, only the norm is adopted as a measurement index of the perturbation size of adversarial examples, which usually leads to the difference in the visibility range. In this paper, a method for adversarial examples generation based on the texture and perceptual color distance is developed. The main idea is to embed the perturbation into a high texture area of an image and optimize the perceptual color distance, so as to reduce the difference in the visibility range between the original image and adversarial example. Moreover, an automatic hyperparameter optimization method is employed to accelerate the convergence of backpropagation. Experimental evaluation shows that the proposed algorithm can obtain the smallest L_2 norm and perceptual color distance than other algorithms. Meanwhile, a smaller number of iterations was required to obtain adversarial examples

Key words adversarial examples; automatic hyperparameter optimization; imperceptible; perceptual color distance

近年来，深度学习在各个领域被广泛应用，其安全性备受关注，特别是对抗样本[1]带来了诸多潜在威胁。对抗样本是通过对原始图像添加刻意构造的微小扰动后，使特定的深度学习分类器以高置信度产生一个错误的分类输出。理想的对抗样本不仅能够欺骗机器学习分类器，且其差异应不易被人类视觉感知。

在目前的对抗样本生成算法中，为了保证添加扰动后图像篡改痕迹的不可见性，通常研究人员采用比较公认的标准，即在 RGB 颜色空间内满足一定的 L_p 范数约束，用 L_p 范数衡量对抗样本中扰动的大小。如 C&W[2]、FGSM[3] 及变种 (I-FGSM[4]、RFGSM[5])、Deepfool[6] 和 JSMA[7]。但范数距离与人类感官差异存在较大的偏差[8]，采用范数约束优化生成的对抗样本不可避免地会在图像平滑区域出现肉眼可见的异常纹理。

此外在基于迭代优化的对抗样本生成算法中，如 C&W[2]、DDN[9] 等算法的损失函数是由多个损

收稿日期：2021－03－01；修回日期：2021－05－06
基金项目：国家自然科学基金 (61702150，61803135)
作者简介：徐明 (1970－)，男，教授，博导，主要从事网络安全、数字取证及多媒体安全等方面的研究. E-mail：549614989@qq.com

失函数累加，通常引入超参数来表示每个损失之间的加权系数。损失函数中的超参数在图像风格转移[10]、图像超分辨率[11]及 GAN 等网络模型中都会涉及，通常采用遍历或者随机搜索的方式反复尝试，最终才能确定合适的超参数。

为了解决对抗样本平滑区域易出现异常纹理和超参数确认困难的问题，本文提出了一种超参数自适应调节算法 (Aho-λ)。该算法基于图像纹理和颜色感知距离，有效降低了对抗样本的视觉差异。训练过程中结合损失函数中超参数与攻击成功率和扰动距离之间的线性关系[2, 9]，进行动态调节超参数，有效避免了超参数的反复尝试，降低对抗样本扰动的同时也减少了算法的迭代次数。

1　相关工作

对抗样本的设计是为了产生与原始图像接近的篡改图像，不影响人类判断的前提下使深度学习模型受到明显的改变。对抗样本问题可以描述为：

$$\begin{aligned}&\min D(x'-x)\\&\text{s.t.}f(x')=l'\\&f(x)=l\\&l'\neq l\\&x\in[0,1]\end{aligned}\tag{1}$$

在距离 D 的约束下，使网络分类器的标签发生改变；距离 D 可以分为 L_p 范数和非 L_p 范数。

1.1　基于 L_p 范数的对抗样本生成算法

传统算法中，扰动大小通常用 L_p 范数来表示：

$$L_p=\sqrt[p]{\sum\|x_i\|^p}\tag{2}$$

常用的 L_p 范数包括 L_0、L_2 及 L_∞范数。L_0 范数表示非零元素的个数，L_0 范数限制可修改像素的数量，如 JSMA[7] 通过迭代的次数来限制 L_0 范数。另外，L_0 范数广泛应用于黑盒模型中，如单像素攻击[12] 仅通过修改某一个像素便可引起分类器的误判。L_2 范数也称欧式距离，使用在文献 [1, 4, 7] 中，是衡量对抗样本全局扰动大小的指标。L_∞范数使用在文献 [2, 4, 5] 中，表示向量中元素的最大值，相比于 L_2 范数，L_∞范数侧重于图像局部的修改限制，对 L_∞范数的约束是为了防止图像某一像素点扰动过大。

文献 [1] 首先提出范数约束扰动大小的方法，如式 (3) 所示，损失需要由超参数 λ 来调节，其中 J 是交叉熵，该文献提出了一种有效算法 L-BFGS 进行求解。

$$\text{minimize}\ \lambda\|\delta\|_2-J(x',y)\ \ \text{s.t.}\ x'\in[0,1]^n\tag{3}$$

C&W 算法[2] 对文献 [1] 算法进行了改进，如式 (4) 所示，通过引入 tanh 函数解决了图像的训练约束，把像素值约束在 [0,1] 之间，并且交叉熵用式 (9) 进行替代。

$$\begin{aligned}&\text{minimize}\|x'-x\|_2^2+\lambda f(x')\\&\text{where}\ f(x')\max(\max(\{Z(x')^i:i\neq t\}-Z(x')_t)-k)\\&\text{and}\ x'=\frac{1}{2}(\tanh(\operatorname{arctanh}(x)+\omega)+1)\end{aligned}\tag{4}$$

文献 [1] 和文献 [2] 两种算法都涉及超参数 λ 的选择，λ 维持 L_p 范数和交叉熵之间的平衡，λ 过大使对抗样本 L_p 范数过大，图像产生明显的扰动；λ 过小会导致对抗样本不能产生攻击效果。也有算法避开了超参数 λ 的选择，如 I-FGSM[4] 和 Deepfool[6]，这些算法通过降低迭代次数来降低 L_p 范数。

1.2　基于非 L_p 范数的对抗样本生成算法

目前，还没有一个统一的标准用来描述图像的感官差异。在文献 [13-14] 中，SSIM[15] 取代了 L_p 范数，其缺点是 SSIM 对图像微小的变化都较敏感，有时即使完全不同的图像的 SSIM 值却比相似图像要高[13]。还有研究用多重的 L_p 叠加来取代单一的范数约束[16]，虽然能够获得一定的效果，但是很难从根本上降低图像扰动的可见性，反而增加了训练的复杂程度。另外，文献 [17] 使用谐波 (Harmonic) 生成无边缘的平滑扰动来降低扰动的可见性；文献 [18-20] 将拉普拉斯平滑项 (Laplacian smoothing term) 和正则项 (regularization term) 引入损失函数中，以此来生成平滑的对抗样本。此外，很多研究将扰动尽可能地添加在图像的高纹理区域[21-24]，将纹理损失加入损失函数中降低平滑区域的噪声。文献 [21] 提出的 C-adv 方法通过改变背景的颜色和修改图像中物品的着色来获得有效的对抗样本。另外，文献 [25] 提出 PerC-C&W 和 PerC-AL 方法，使用 CIEDE2000[26] 取代了 L_p 范数约束，在反向传播中进行直接优化，使用这一标准作为图像质量的衡量指标，生成的对抗样本与原始图像视觉差异更小。

2　自适应无感对抗样本生成算法

基于纹理度筛选和颜色感知距离 CIEDE2000，本文提出了对抗样本生成算法 Aho-λ，能够在训练过程中自适应调节超参数，算法流程如图 1 所示。

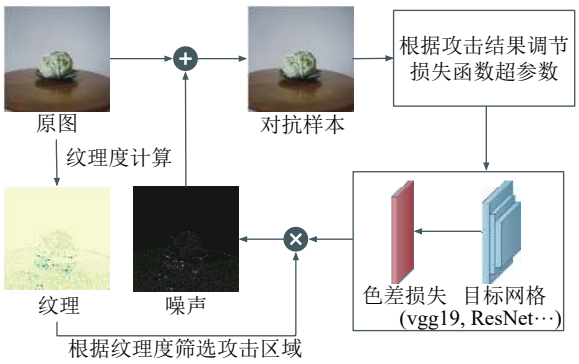


图 1 无感对抗样本算法流程图

2.1 CIEDE2000 颜色感知距离

本文将 CIEDE2000 标准引入损失函数中，取代原先的 L_p 范数。CIEDE2000 计算公式为：

$$\Delta E_{00}=\sqrt{\left(\frac{\Delta L'}{k_L S_L}\right)^2+\left(\frac{\Delta C'}{k_C S_C}\right)^2+\left(\frac{\Delta H'}{k_H S_H}\right)^2+\Delta R}$$
$$\Delta R=R_T\left(\frac{\Delta C'}{k_C S_C}\right)\left(\frac{\Delta H'}{k_H S_H}\right) \tag{5}$$

式中，L、C、H 分别代表了图像的亮度 (lightness)、色度 (chroma)、明度 (hue)。参数参考文献 [26]。通过文献 [26] 的研究证明，这一标准比范数距离更符合人类肉眼对于颜色的感知。

2.2 图像纹理度

图像纹理度是用来表述图像每个像素点位置的纹理程度大小的指标。在文献 [10-11] 研究中，对抗样本的扰动应尽可能添加在图像的平滑区域，文献 [11] 首次在对抗样本训练中引入了纹理度损失，如式 (6)：

$$D(X^*,X)=\sum_{i=1}^{N}\varepsilon_i\times \mathrm{Sen}(x_i) \tag{6}$$

式中，Sen 是每个像素点的敏感度因子 (为每个点和周围像素点之间的方差的倒数)，如式 (7)：

$$\mathrm{Sen}(x_i)=1/\mathrm{SD}(x_i)$$
$$\mathrm{SD}(x_i)=\sqrt{\frac{\sum_{x_k\in S_i}(x_k-\mu)^2}{n^2}} \tag{7}$$

式中，n 表示计算方差和均值的分块大小。

本文选取了方差 SD 来表示图像纹理度，作为筛选图像扰动区域的量化指标，n 取值 3，并在迭代的过程直接过滤掉低纹理区域，降低计算成本。

2.3 自适应训练算法

首先本文将为式 (8) 定义对抗样本训练的损失，

$$\text{minimize}\|\Delta E_{00}(x,x')\|+\lambda f(x') \tag{8}$$

式中，ΔE_{00} 表示颜色感知距离；f 表示攻击目标网络的损失函数，超参数 λ 调节 ΔE_{00} 和 f 之间的比率，对于不同的图像 λ 的取值不同。

本文设计的 Aho-λ 算法，能够适应不同的图像和网络模型，通过训练得到一个相对较优的参数 λ 来降低加入的扰动大小。算法使用的 $f(x)$ 如式 (9) 所示，文献 [2] 已经实验证明了 $f(x)$ 能够有效代替交叉熵，其中参数 k 用来描述模型中最大概率的预测项和次预测项目之间的距离大小，能够有效反映生成对抗样本的置信度。

$$f(x')=\max(\max(\{z(x')_i:i\neq t\}-z(x')_t),-k) \tag{9}$$

Aho-λ 如算法 1 所示，使用式 (7) 将所有像素点的纹理度进行计算及排序，把纹理度较低的点按照一定百分比进行过滤。依据对抗样本攻击成功与否，在迭代过程中动态调节超参数 λ 的大小。超参数 λ 值越大越能保证对抗样本攻击的成功率，但是为了有适当的感知距离且不易被肉眼察觉，对于每一张图像需要一个适合的 λ 值来权衡感知距离和攻击成功率之间的关系。结合每次迭代的对抗样本攻击结果，参考机器学习训练中的优化算法对 λ 进行自适应的调节，其中衰减率 θ 满足 $0<\theta<1$，目的是为了让 λ 最终稳定在某一范围内，随着迭代的不断进行，λ 的变化率逐渐减小。

算法 1：自适应训练算法

输入：

x：原始图像；t：原始标签

k：超参数；θ：超参数变化率；η：学习率

输出：

$\hat{x}$：对抗样本

初始化 $\hat{x}\leftarrow x$

初始化损失函数 $\mathrm{Loss}=\|\Delta E_{00}(x,\hat{x})\|+\lambda f(\hat{x})$

计算图像纹理度 x'，将高纹理区域置为 1，低纹理区域置为 0

for $i\leftarrow 1$ to n do

  if $\hat{x}$ 是对抗样本 then

    $\lambda\leftarrow\lambda(1+\theta^n)$

  else

    $\lambda\leftarrow\lambda(1-\theta^n)$

  end if

  $\mathrm{Loss}=\mathrm{Loss}_{\mathrm{same}}+\lambda\times\mathrm{Loss}_{\mathrm{target}}$

  $g\leftarrow-\nabla_{\hat{x}}J(\mathrm{Loss})$

  $\delta\leftarrow\eta\times g\times x'$

  $\hat{x}\leftarrow\mathrm{clip}(\hat{x}+\delta,0,1)$

end for

return $\hat{x}$

3 实验

3.1 实验设计

数据集与网络：选用 NIPS 2017 对抗样本攻防比赛[26] 所采用的数据集 ImageNet-Compatible dataset，共包含 6 000 张图像，属于 ImageNet 1 000 种标签类，Inception V3 [27] 具有较高的识别率。因此，本文将 Inception V3 作为目标网络进行攻击产生对抗样本，最后将获取到的图像直接缩放到指定大小，图像的长宽为 299*299。

实验对比的算法：与 L_p 范数的 I-FGSM[4]、C&W[2] 和 DDN[9] 算法，及感官距离 CIEDE2000 的 PerC-AL[25] 算法进行对比。比较对抗样本的攻击成功率、L_p 范数和感官距离。

实验建立与参数选择：I-FGSM 每次迭代的步长设置为 η=1/255，直到攻击成功后停止。C&W 算法中，学习率设置为 0.005，超参数 λ 使用 [0.01,0.1,1,10,100] 进行选择，生成的对抗样本中扰动最小的图像作为最终结果。PerC-C&W 和 PerC-AL 算法的学习率为 0.001，DDN 的单步步长为 0.01。DDN、PerC-AL 和本文的 Aho-λ 算法，迭代次数设置为 [100,300,1 000] 分别进行比较。

3.2 局部像素修改实验

本文提出的方法能够有效地对添加扰动的区域进行筛选和修改限制。如图 2 所示，图像第一行为不同修改比率下生成的对抗样本；第二行为对抗样本修改像素点的位置，使用 255 替换原来的颜色，图像中白色部分表示修改像素点对应的 RGB 三个颜色通道都被修改过。根据本文提出的算法，将纹理度排序并筛选出一定比率的可攻击区域，实验结果如表 1 所示。可以看出仅需修改图像中的少量点就能得到较高的攻击成功率，当修改区域大于 70% 时能保证图像 100% 的攻击成功率；对抗样本 L_2 范数随着修改点的减少呈现下降的趋势，从 L_∞范数可以看出单个像素点的最大扰动随之提高。另外，当修改像素在 70% 时，颜色感知距离 C_2 具有最低值 56.54。因此，本文后续的实验都采用 70% 的像素修改作为实验参数。

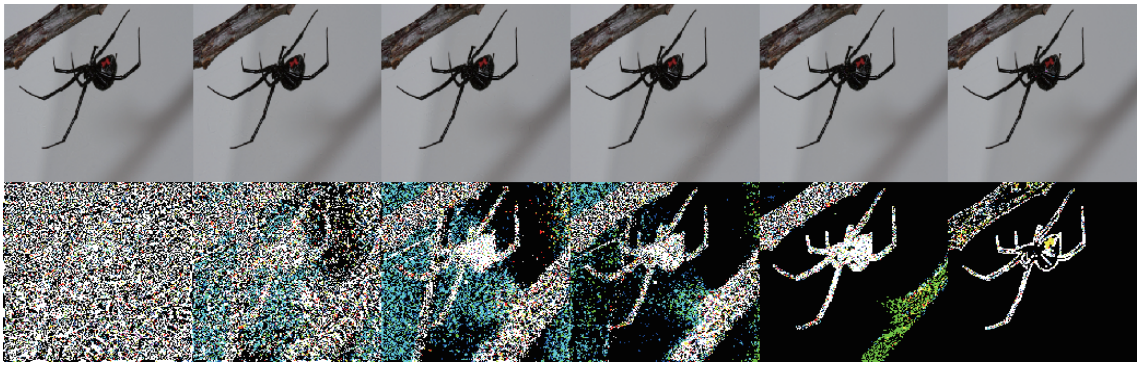


图 2　Aho-λ 不同修改比率得到的对抗样本效果

表 1　图像修改比率、攻击成功率及扰动距离对比

修改比率/%	成功率/%	扰动距离		
		$\overline{L_2}$	$\overline{L_\infty}$	$\overline{C_2}$
10	78.3	0.51	25.90	123.31
20	80.4	0.48	24.73	120.76
30	85.3	0.52	20.23	102.65
40	92.3	0.69	19.66	75.78
50	95.1	0.74	19.48	68.42
60	96.2	0.82	18.65	90.75
70	100	0.79	16.78	56.54
80	100	1.03	17.03	67.52
90	100	1.11	17.25	82.48
100	100	1.30	16.89	85.60

3.3 对抗样本质量实验

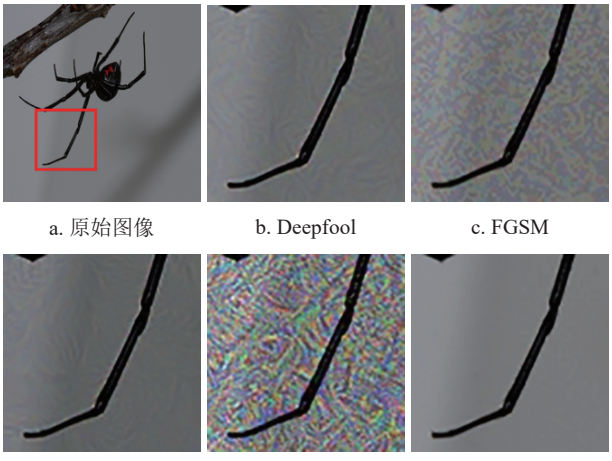
如表 2 所示，在不考虑感官距离的 DDN、C&W 以及 I-FGSM 三种算法中，DDN 算法能够获得最低的 L_2 范数，略优于 C&W 算法；在 I-FGSM 算法中，L_∞取决于迭代的次数，且每次迭代具有固定步长，虽然 I-FGSM 能够获得较低的 L_∞范数，但在 L_2 和 C_2 上都不如其他算法优越。在结合感官距离的算法 PerC-AL 和 Aho-λ 中，本文提出的 Aho-λ 算法能够达到和 DDN 算法基本相同的 L_2 范数，并具有比 PerC-AL 算法更小的颜色感知距离 C_2。与 DDN 和 PerC-AL 算法相比，Aho-λ 算法在

300 次和 1 000 次迭代过程中生成的两组对抗样本差异更小，这说明本文提出的 Aho-λ 算法能够更快地收敛，在 300 次左右就能够达到最佳的攻击效果。

表 2　对抗样本质量对比

方法	迭代次数	成功率/%	扰动距离		
			$\overline{L_2}$	$\overline{L_\infty}$	$\overline{C_2}$
I-FGSM	–	100.0	2.51	1.59	317.96
C&W	1 000	100.0	1.09	8.20	132.86
	100	100.0	1.00	7.84	136.11
DDN	300	100.0	0.88	7.58	120.12
	1 000	100.0	0.82	7.62	111.65
	100	100.0	1.30	11.98	69.49
PerC-AL	300	100.0	1.17	13.97	61.21
	1 000	100.0	1.13	17.04	57.10
	100	100.0	1.21	13.89	70.75
Aho-λ	300	100.0	0.81	16.78	56.55
	1 000	100.0	0.81	16.98	56.52

本文算法与其他几种算法生成的对抗样本局部细节对比如图 3 所示。图 3a 是原始图像，放大方块的选中区域后本文的算法并未出现异常纹理，与原图基本一致。其余几种方法都出现了可见的异常纹理。



a. 原始图像　b. Deepfool　c. FGSM

d. I-FGSM　e. RFGSM　f. Aho-λ

图 3　5 种算法生成的对抗样本

3.4　置信度与鲁棒性实验

置信度 k 值变化会影响对抗样本的质量，能有效地保证对抗样本输出的标签与其他标签之间的差异，如式 (9) 所示。在不同置信度 k 下，本文算法 L_p 范数和颜色感知距离变化如图 4 所示，置信度 k 值越大，扰动距离 L_2、L_∞和 C_2 不断增大。

图片的有损压缩通常也被当作是防御对抗样本攻击的有效手段。本实验选取了与文献 [20, 28] 相同的 JPEG 和 Bit Depth 压缩方法。在不同质量因子下，JPEG 压缩后的对抗样本保持原有攻击效果的比率，如图 5 所示。在常用的质量因子大于 70 时，本文提出的算法具有一定的抗 JPEG 压缩能力；但当质量因子小于 60 时，压缩图像越来越模糊，对抗样本的攻击成功率降低。Bit Depth 压缩下也表现出了近似的效果，如图 6 所示，原来图像颜色由 8 位压缩到 4 位以上时，对抗样本表现出较出色的抗压缩能力。同时，图 5 和图 6 表明，由于置信度增加，需要在图像中嵌入扰动变大，对抗样本的鲁棒性也随之提高。

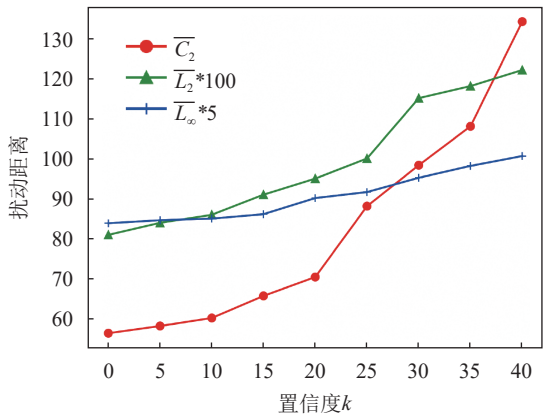


图 4　不同置信度下 3 种扰动距离的大小

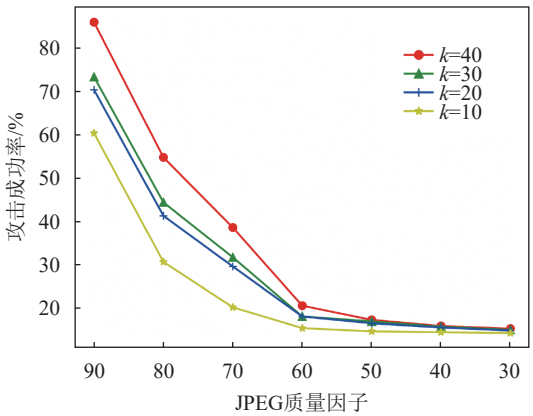


图 5　不同 JPEG 质量因子下对抗样本的成功率

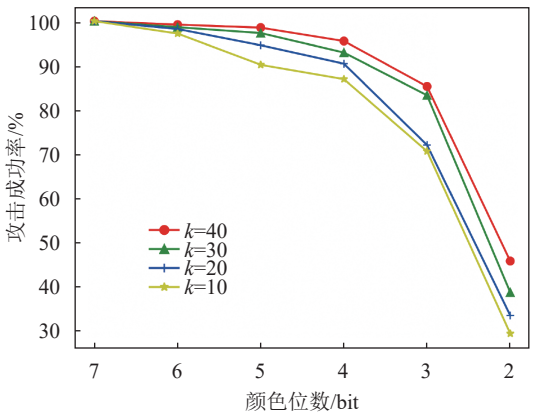


图 6　Bit Depth 压缩下对抗样本的成功率

3.5　迁移性实验

现有的许多研究表明[2, 3, 23]，对于不同网络模型使用相同的对抗样本可能达到同样的攻击效果，即对抗样本具有一定的迁移性。本文选取 ImageNet-Compatible dataset 数据集作为实验对象，置信度 k 选择 20 和 40，分别在 Google net[27]、Vgg-16[29] 和 ResNet-152[30] 网络模型上进行迁移性实验，实验结果如表 3 所示。表 3 中的数据表示在不同置信度 k 下，不同算法生成的对抗样本在另外两种网络模型中具有相同的分类结果的比率。另外，在所有算法中 I-FGSM 迁移性最好，其加入的 L_2 和 C_2 颜色感知距离都是最大的；Aho-λ 算法虽然具有一定的迁移性，但是迁移性不高，原因可能有以下两点：首先 Aho-λ 算法加入的扰动是所有算法中最小的；其次可能是不同的网络模型对于图像纹理区域的改变比较敏感。因此，Aho-λ 算法在不同网络模型中的判别结果一致性不高。

表 3　不同神经网络模型中的迁移性

方法	GoogLeNet		Vgg-16		ResNet-152	
	k=20	k=40	k=20	k=40	k=20	k=40
I-FGSM	3.4	5.3	6.5	11.9	7.5	9.9
C&W	1.8	2.8	3.9	5.9	4.5	5.1
DDN	1.0	2.0	4.5	6.7	4.3	5.1
PerC-C&W	2.2	3.9	4.3	8.1	5.5	6.5
PerC-AL	1.6	3.4	5.1	7.9	5.3	7.3
Aho-λ	1.2	2.3	4.0	7.0	4.5	5.6

4　结 束 语

本文结合纹理度筛选与颜色感知距离 CIEDE2000 作为图像损失函数，设计了一种能够自适应调节超参数的算法 Aho-λ，生成的图像具有更小的颜色感知距离和更快的收敛。在 JPEG 和 Big Depth 压缩下具有良好的鲁棒性，且对抗样本在多种网络模型下具备一定的迁移能力。

使用 CIEDE2000 标准作为人类感知距离，在一定程度上降低了对抗样本在视觉上的可见度，但在图像平滑区域依然存在一定的可感知性，未来希望找到一种更符合人类感官的新标准引入训练损失中；同时也希望找到一种能够定量区分图像修改区域的方法，进一步完善纹理度筛选。

参　考　文　献

[1] SZEGEDY C, ZAREMBA W, SUTSKEVER I. Intriguing properties of neural networks[C]//The 2nd International Conference on Learning Representations. [S.l.]: 2014, 4: 3861-3864.

[2] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Symposium on Security and Privacy. San Jose: IEEE, 2017, 5: 39-57.

[3] GOODFELLOW I, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//The 3rd International Conference on Learning Representations. San Diego: [s.n.], 2015, 5: 1353-1362.

[4] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world[C]//The 5th International Conference on Learning Representations. Toulon: [s.n.], 2017, 4: 1238-1249.

[5] TRAMER F, KURAKI A, PAPERNOT N, et al. Ensemble adversarial training: attacks and defenses[C]//The 6th International Conference on Learning Representations. Vancouver: IEEE, 2018, 5: 131-138.

[6] MOOSAVI-DEZFOOLI S, FAWZI A, FROSSARD P. Deepfool: A simple and accurate method to fool deep neural networks[C]//Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016, 6: 2574-2582.

[7] PAPERNOT N, MCDANIEL P, JHA S. The limitations of deep learning in adversarial settings[C]//European Symposium on Security and Privacy. [S.l.]: IEEE, 2016, 3: 372-387.

[8] SHARIF M, BAUER L, REITER M. On the suitability of L_p-norms for creating and preventing adversarial examples[C]//2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops. [S.l.]: IEEE, 2018, 6: 1605-1613.

[9] RONY J, HAFEMANN L, OLIVEIRA L, et al. Decoupling direction and norm for efficient gradient-based L_2 adversarial attacks and defenses[C]//Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019, 6: 4322-4330.

[10] GATYS L, ECKER A, BETHGE M. A neural algorithm of artistic style[EB/OL]. [2015-09-02]. https://arxiv.-org/abs/1508.06576.

[11] YOON Y, JEON H, YOO D, et al. Learning a deep convolutional network for light-field image super-resolution[C]//International Conference on Computer Vision Workshop. [S.l.]: IEEE, 2015, 12: 57-65.

[12] SU Jia-wei, VARGAS D, SAKURAI K. One pixel attack for fooling deep neural networks[J]. Trans Evol Comput, 2019, 23(5): 828-841.

[13] WANG Z, BOVIK A, SHEIKH H, et al. Image quality assessment: from error visibility to structural similarity[J]. Trans Image Process, 2004, 13(4): 600-612.

[14] ROZSA A, RUDD E, BOULT T. Adversarial diversity and hard positive generation[C]//Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016, 6: 410-417.

[15] KANBAK C, MOOSAVI-DEZFOOLI S, FROSSARD P. Geometric robustness of deep networks: Analysis and improvement[C]//Conference on Computer Vision and Pattern Recognition. San Diego: IEEE, 2018, 6: 4441-4449.

[16] ENGSTROM L, TSIPRAS D, SCHMIDT L, et al.

Exploring the landscape of spatial robustness[EB/OL]. [2019-09-16]. https://arxiv.org/abs/1712.02779.

[17] HENG Wen, ZHOU Shu-chang, JIANG Ting-ting. Harmonic adversarial attack method[EB/OL]. [2018-08-08]. https://arxiv.org/abs/1807.10590.

[18] LI Yi-jun, LIU Ming-yu, YANG Ming-hsuan, et al. A closed-form solution to photorealistic image stylization[J]. Springer Journal of Computer Vision, 2018, 5: 468-483.

[19] PUY G, PEREZ P. A flexible convolutional solver with application to photorealistic style transfer[EB/OL]. [2018-06-13]. https://arxiv.org/abs/1806.05285.

[20] ZHANG Han-wei, AVRITHIS Y, FURON T, et al. Smooth adversarial examples[J]. EURASIP Journal on Information Security, 2020(1): 15-24.

[21] BHATTAD A, CHONG Min-jin, LIANG Kai-zhao. Unrestricted adversarial examples via semantic manipulation[EB/OL]. [2019-03-20]. https://arxiv.org/abs/1904.06347.

[22] LUO Bo, LIU Yan-nan, WEI Ling-xiao, et al. Towards imperceptible and robust adversarial example attacks against neural networks[C]//The 30th Innovative Applications of Artificial Intelligence. Louisiana: IEEE, 2018, 2: 1652-1659.

[23] CROCE F, HEIN M. Sparse and imperceivable adversarial attacks[C]//International Conference on Computer Vision. Seoul: IEEE, 2019, 10: 4728-4731.

[24] KURAKIN A, GOODFELLOW I, BENGIO S, et al. Adversarial attacks and defences competition[EB/OL]. (2018-03-31). https://arxiv.org/abs/1804.00097.

[25] ZHAO Zheng-yu, LIU Zhuo-ran, LARSON M. Towards large yet imperceptible adversarial image perturbations with perceptual color distance[C]//Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020, 6: 1036-1045.

[26] RONNIER L, CUI Gui-hua, BRYAN R. The development of the CIE 2000 colour-difference formula: CIEDE-2000[J]. Color Research, 2001, 26(5): 340-350.

[27] SZEGEDY C, VANHOUCKE V, IOFFE S. Rethinking the inception architecture for computer vision[C]// Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016, 6: 2818-2826.

[28] GUO C, RANA M, CISSE M, et al. Countering adversarial images using input transformations[EB/OL]. [2018-01-25]. https://arxiv.org/abs/1711.00117.

[29] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//The 3rd International Conference on Learning Representations. San Diego: IEEE, 2016, 6: 2818-2826.

[30] HE Kai-ming, ZHANG Xiang-yu, SUN Jian. Deep residual learning for image recognition[C]//Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016, 6: 770-778.

编 辑 蒋 晓