Computer Science and Application 计算机科学与应用, 2023, 13(9), 1796-1805
Published Online September 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.139178

Hans汉斯

基于仿射变换和梯度细化的对抗样本生成方法

王　卓

广东工业大学计算机学院，广东　广州

收稿日期：2023年3月14日；录用日期：2023年9月21日；发布日期：2023年9月28日

摘　要

尽管白盒攻击已实现了较高的攻击成功率，但样本的过拟合现象，使得生成的对抗样本在攻击其它分类模型时成功率偏低。为缓解过拟合现象以提高对抗样本的迁移性，增加其在黑盒条件下的攻击成功率，本文提出了一种基于仿射变换和梯度细化的对抗样本生成方法AF-R-MI-FGSM。该方法不是仅使用原始图像生成对抗样本，而是在每次迭代时对输入图像进行随机的仿射变换来提高输入图像的多样性，利用数据增强技术来缓解对抗样本的过拟合现象，使得对抗样本更具有迁移性。由于引入图像随机变换导致噪声梯度随机性增加，影响攻击性能，本文提出了一种梯度细化的方式来缓解消极的梯度影响。此外，还通过使用集成模型来进一步提高样本的迁移性。并在ImageNet数据集上进行了实验，验证了本文方法的有效性，在黑盒条件下，与MI-FGSM相比，本文所提方法的单模型攻击的平均攻击成功率提升了14.3%，集成模型攻击的平均攻击成功率提升了22.1%。

关键词

对抗样本，黑盒攻击，图像仿射变换，梯度细化，可迁移性

Generating Adversarial Example Based on Affine Transformation and Gradient Refining

Zhuo Wang

School of Computer, Guangdong University of Technology, Guangzhou Guangdong

Received: Mar. 14th, 2023; accepted: Sep. 21st, 2023; published: Sep. 28th, 2023

Abstract

Although the white-box attack has achieved a high rate of attack success, the over-fitting pheno-

文章引用: 王卓. 基于仿射变换和梯度细化的对抗样本生成方法[J]. 计算机科学与应用, 2023, 13(9): 1796-1805.
DOI: 10.12677/csa.2023.139178

menon of samples makes the generated adversarial samples have a low success rate when attacking other classification models. Therefore, it is necessary to alleviate the over-fitting phenomenon to improve the migration of the adversarial samples, so as to enhance its attack performance under the condition of black-box. Therefore, it is necessary to improve the migration of adversarial samples to enhance their attack performance under black-box conditions. To solve this problem, this paper proposes a method of generating adversarial example based on affine transformation and gradient refining, AF-R-MI-FGSM. This method does not only use the original image to generate adversarial example, but performs random affine transformation on the input image at each iteration to improve the diversity of the input image, and uses data enhancement technology to alleviate the over-fitting phenomenon of adversarial example, so as to improve the attack success rate of adversarial example under black-box conditions. Because the introduction of image random transformation leads to the increase of noise gradient randomness and affects the attack performance, this paper proposes a gradient thinning method to alleviate the negative gradient effect. In addition, the migration of samples is improved by attacking the integration model. Experiments are carried out on ImageNet datasets to verify the significance of the proposed method. Compared with MI-FGSM, the average black-box attack success rate of AF-R-MI-FGSM in attacking a single model is increased by 14.3%, and the average black-box attack success rate of attack integration model is increased by 22.1%.

Keywords

Adversarial Example, Black-Box-Attack, Affine Transformation, Gradient Refining, Transferability

1. 引言

尽管深度神经网络(Deep Neural Networks, DNN)在各种图像任务中表现优良，例如在图像分类领域[1]，但研究表明对待识别的样本插入微小不可见扰动形成的对抗样本可导致神经网络的给出错误分类[2]。在已提出的对抗样本攻击方法中，主要分为白盒攻击和黑盒攻击。在白盒条件下，攻击者具备或者有能力获取神经网络模型结构、参数等先验知识，攻击者能生成对此模型更有针对性的对抗样本[3]，现已提出了大量白盒攻击方法，如 FGSM (Fast Gradient Sign Method) [4]、I-FGSM (Iterative Fast Gradient Sign Method) [5]和 MI-FGSM (Momentum Iterative Fast Gradient Sign Method) [6]，这些攻击方法需知道给定模型的梯度信息。与白盒攻击相反，黑盒攻击无法获取攻击模型的结构、参数等相关知识。

对抗样本具有迁移性，即利用一个模型生成的对抗样本能成功地攻击另一个模型[7]，迁移性的存在能够使得攻击者在无法获得模型结构和参数的黑盒攻击中，使用已知模型生成对抗样本来完成攻击[8]。现有的基于梯度的攻击方法，在白盒攻击中，有着较高的攻击成功率，但在黑盒条件下，攻击成功率却不理想，上述问题被认为是对抗样本的过拟合现象[9]，导致其生成的对抗样本迁移性不足。若能提升对抗样本的迁移性，将会增加对抗样本的黑盒攻击成功率。

为提升对抗样本的迁移性，本文使用图像的仿射变换和梯度细化来对样本的生成过程进行优化。主要工作如下：

1) 使用图像的仿射变换来增加对抗样本的泛化能力，从而缓解过拟合问题，提高样本迁移性。

2) 使用梯度细化方法来消除图像随机变换所带来的消极噪声梯度，降低消极噪声梯度对对抗样本泛

化能力的影响。

3) 在 ImageNet 数据集上进行了单模型攻击和集成模型攻击实验，并与 I-FGSM [10]和 MI-FGSM [11]方法进行对比。

2. 相关工作

2.1. 对抗样本的生成方法

有研究表明，对抗样本能对深度神经网络进行攻击[10]。Szegedy 等人[11]指出，深度神经网络在对抗样本的攻击下是极其脆弱的，并提出了一种基于优化的 L-BFGS 方法。Goodfellow 等人提出了快速梯度符号法，此方法仅使用单次梯度计算来生成对抗样本，来减小生成对抗样本的计算代价。Alexey 等人将此方法扩展为了迭代版本，大大增加了白盒攻击的成功率，并表明生成的对抗样本能够存在于物理世界中。Donger 等人提出了一种基于动量的迭代方法，对梯度的更新方向和其收敛过程进行了优化，此方法还通过同时攻击一组网络来提高提高其迁移性[12]。

2.2. 对抗样本的防御方法

面对对抗样本的安全威胁，现已提出了许多方法来防御对抗样本。其中，对抗训练[8]是一种常用的方法，该方法提出在训练数据集中添加对抗样本来对模型进行训练，以此来提高模型的鲁棒性。Tramèr 等人[12]提出了一种集成对抗训练的方法，来弥补对抗训练过程中的不足，即训练其他的模型，并使用其传递过来的扰动对训练数据进行增强。Guo 等人[13] [14] [15]利用随机的图像变换，使得图像在保持关键视觉内容的同时消除对抗扰动。Prakash 等人[16]提出一种像素偏移和软小波去噪相结合的框架，以抵御对抗样本。

3. AF-R-MI-FGSM 方法

在本节中将对本文提出方法和基于梯度生成对抗样本算法 AF-R-MI-FGSM (Affine And Refine Momentum Iterative Fast Gradient Sign Method)进行详细描述。定义 x 为输入的原始干净样本，y^{true} 为样本的真实标签，θ 为模型参数，$L(x, y^{true}; \theta)$ 为损失函数。在对抗样本的生成过程中，需要使损失函数 $L(x, y^{true}; \theta)$ 最大化，以此来生成一个与原始样本 x 在人眼上难以区分的 x^{adv} 来欺骗深度学习模型，从而让模型给出一个错误的分类结果。对于扰动大小的限制，本文使用无穷范数来进行制约，即 $\left\| x^{adv} - x \right\|_\infty \le \varepsilon$。

3.1. 基于梯度生成方法

FGSM 是此类方法中最早提出的，利用梯度信息来生成对抗扰动，并对扰动使用无穷范数限制，其更新公式如下，

$$x^{adv} = x + \varepsilon \cdot sign(\nabla_x L(x, y^{true}; \theta)) \tag{1}$$

FGSM 沿着梯度的方向，仅使用单步添加扰动，来达到快速生成对抗样本的目的。但单步变化不够精准，导致其白盒成功率偏低。

I-FGSM 是在 FGSM 的基础上增加了迭代的过程，将 FGSM 中的梯度进行多步迭代计算，使其梯度变化更为精准，提升了方法的白盒成功率。其公式如下，

$$x_0^{adv} = x,$$

$$x_{t+1}^{adv} = Clip_x^{\varepsilon}\{x_t^{adv} + \alpha \cdot sign(\nabla_x L(x_t^{adv}, y^{true}; \theta))\} \tag{2}$$

其中，$Clip_x^{\varepsilon}$ 函数是将对抗样本限制在原始样本 x 的 ε 领域内，α 为步长，$T(0,1,2,\ldots,t)$ 为迭代次数，

$\alpha = \varepsilon / T$ 。

I-FGSM 使用了迭代来逐步生成扰动，使其梯度变换更为平稳，但由于迭代的次数增加，生成的对抗样本会出现过拟合现象，导致其迁移性降低。

MI-FGSM 是在迭代的过程中使用动量，此方法能够稳定梯度的更新方向并且能避免局部的极大值问题，其公式如下，

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x L(x_t^{adv}, y^{true}; \theta)}{\left\| \nabla_x L(x_t^{adv}, y^{true}; \theta) \right\|_1}$$

$$x_{t+1}^{adv} = Clip_x^{\varepsilon}\{x_t^{adv} + \alpha \cdot sign(g_{t+1})\} \tag{3}$$

其中， μ 是动量的衰减因子， g_t 是迭代时的累积梯度。

MI-FGSM 使用动量来解决局部最优值的问题，来提高对抗样本的迁移性。

为了进一步提升对抗样本的迁移性，增加其黑盒攻击成功率，本文提出了一种基于仿射变换和梯度细化的生成方法，来缓解对抗样本的过拟合现象，提高对抗样本的迁移性。

3.2. 基于仿射变换和梯度细化生成方法

在深度神经网络训练中，数据增强是一种常用的用于解决过拟合问题的方式，本文采用一种仿射变换的数据增强方式来缓解生成对抗样本中的过拟合现象，AF-MI-FGSM 方法。图像的几何变换有很多种，其中包括一些基本变换(如平移、旋转、裁剪等)，图像的仿射变换就是将基本变换进行组合，来实现一种更加复杂的图形变换。本文则采用平移和旋转的组合来对原始样本进行变换。该方法在每次迭代的过程中，对原始样本以概率 p 进行随机的仿射变换，以此增加对抗样本的迁移性，其变换公式如下，

$$AF(x^{adv}; p)\begin{cases} AF(x), \text{with probability } p \\ x, \text{with probability } 1-p \end{cases} \tag{4}$$

公式(4)使用变换函数 $AF(\cdot)$ 对原始样本进行随机的仿射变换，即对图像进行随机度数的旋转，在旋转的基础上对图像进行水平和垂直方向上的平移变换。转换概率 p 控制着原始样本转换的比例，以此方式来平衡生成样本的白盒和黑盒攻击成功率。

由于图像的变换的随机性，导致其生成的噪声梯度具有随机性。这种随机的噪声梯度所生成的扰动，将其加在干净的样本中，能够增加对抗样本的迁移性。然而与此同时，随机梯度中也存在部分噪声梯度，会抑制对抗样本的迁移性。为了缓解随机性带来的消极影响，同时保证对抗样本的迁移性，本文使用梯度细化方法来缓解消极梯度影响，AF-R-MI-FGSM。即相同图像经随机变换后所产生的 n 次噪声梯度，使用 n 次噪声梯度的均值来生成扰动，具体公式如下：

$$g_i = \nabla_x L(AF(x_n^{adv}; p), y^{true}; \theta), i = 1, 2, \ldots, n \tag{5}$$

$$g = \frac{1}{n}\sum_{i=1}^{n} g_i \tag{6}$$

$$x_{n+1}^{adv} = x_n^{adv} + \alpha \cdot sign(g)\,. \tag{7}$$

公式(5)对图像 x_n^{adv} 计算 n 次噪声梯度，由于图像输入模型采用的是随机变换处理，所以图像多次计算的噪声梯度是不同的。通过对多个噪声梯度求平均梯度，可以抵消部份的消极梯度，突出有效的梯度信息，如公式(6)所示。最后利用公式(7)来更新噪声，生成对抗样本。

3.3. 算法描述

上述的攻击方法属于 FGSM 家族中的一员，AF-R-MI-FGSM 通过调整不同的参数设置，可以转换为

FGSM 家族中的其他方法，当 $n=1$，即仅使用单次梯度信息进行迭代时，AF-R-MI-FGSM 退化为 AF-MI-FGSM。当转换的概率 $p=0$ 时，即不对图像进行处理，仅使用原始样本生成对抗样本，方法则退化为 MI-FGSM。具体转换关系如图 1 所示。

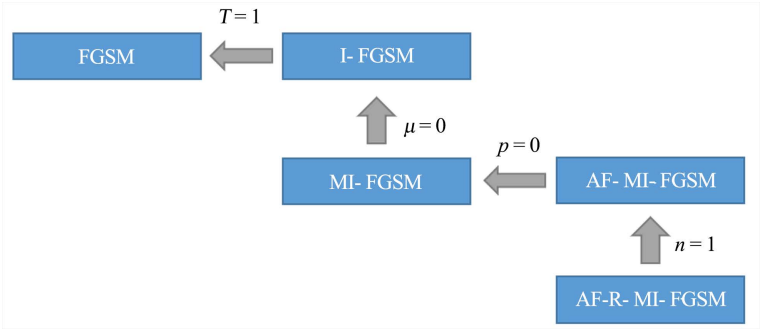

Figure 1. Conversion relationships between different methods
图 1. 不同方法之间的转换关系

算法伪代码如下：

算法 1：AF-R-MI-FGSM (单个模型)

输入：原始样本 x 以及其对应的真实标签 y^{true}，神经网络 f 和损失函数 L，扰动大小 ε，迭代次数 T，衰减因子 μ，每张图像的平均梯度次数 n。

输出：对抗样本 x^{adv}

1) $\alpha=\epsilon/T$

2) $x_0^{adv}=x$

3) for t=0 to T do

4) 通过公式(4) $AF(x^{adv};p)$ 求取 x_t^{adv}

5) 通过公式(5)计算模型的损失函数并得到梯度 g_i

6) 对同一张图片进行 n 次梯度计算，$g_{1,2,3,\ldots,n}$

7) 通过公式(6)对上述 n 次梯度求均值，得到梯度 g

8) 通过公式(7)更新 x_{t+1}^{adv}

9) 返回对抗样本 $x^{adv}=x_T^{adv}$。

3.4. 集成模型攻击

Liu 等人认为，使用集成模型能够生成迁移性更强的对抗样本。由于对抗样本对多个深度神经网络保持对抗性，那么其更能够迁移到其它的分类网络中。因此，本节使用 AF-R-MI-FGSM 方法攻击集成网络来进一步提高样本的迁移性。

本文遵循文献[6]的策略，使用 AF-R-MI-FGSM 方法同时攻击多个网络，将多个网络的逻辑激活融合在一起，称此为逻辑集成。具体融合方式如下：

$$l(x;\theta_1,\ldots,\theta_K)=\sum_{k=1}^{K}\omega_k l_k(x;\theta_k) \tag{8}$$

其中 $l_k(x;\theta_k)$ 表示参数为 θ_k 的第 K 个网络的逻辑输出值，ω_k 表示多个网络的集成权重，并满足 $\sum_{k=1}^{K}\omega_k=1$ 且 $\omega_k\geq 0$。

算法伪代码如下：

算法 2：AF-R-MI-FGSM (集成模型)

输入：原始样本 x 以及其对应的真实标签 y^{true}，K 个神经网络 $f(f_1,f_2,\ldots,f_k)$，其对应的网络逻辑值 $l(l_1,l_2,\ldots,l_k)$ 和网络集成权重 $\omega(\omega_1,\omega_2,\ldots,w_k)$，扰动大小 ε，迭代次数 T，衰减因子 μ，每张图像的平均梯度次数 n。

输出：对抗样本 x^{adv}

1) $\alpha=\epsilon/T$

2) $x_0^{adv}=x$

3) for t=0 to T do

4) 通过公式(4) $AF(x^{adv};p)$ 求取 x_t^{adv}

5) 将 x_t^{adv} 输入到每个网络，并得到其对应的逻辑值 $l_k\left(x_t^{adv}\right),k=1,2,\cdots,K$

6) 计算集成的逻辑值 $l\left(x_t^{adv}\right)=\sum_{k=1}^{K}w_k l_k\left(x_t^{adv}\right)$

7) 根据集成逻辑值求取梯度，得到梯度 g_i

8) 对同一张图片进行 n 次梯度计算，$g_{1,2,3,\ldots,n}$

9) 通过公式(6)对上述 n 次梯度求均值，得到梯度 g

10) 通过公式(7)更新 x_{t+1}^{adv}

11) 返回对抗样本 $x^{adv}=x_T^{adv}$。

4. 实验与结果分析

本节将对上述方法进行实验来验证其有效性，在单个模型和集成模型上都做了大量的实验来进行对比。在实验中，本文所提方法将与 FGSM 系列方法中的 I-FGSM 和 MI-FGSM 进行对比。

Table 1. The success rate (%) of single model attack
表 1. 单模型攻击的成功率(%)

模型	攻击方法	Inc-v3	Inc-v4	IncRes-v2	Res-101	Adv-Inc-v3	IncRes-v2-ens	平均值
Inc-v3	I-FGSM	**99.9***	21.8	16.1	16.5	13.6	5.6	28.9
	AF-R-I-FGSM	**99.9***	54.2	40.3	33.4	19.2	8.7	42.6
	MI-FGSM	**99.9***	42.1	36.8	37.4	22.1	9.5	41.3
	AF-MI-FGSM	**99.9***	58.2	51.1	49.3	26.6	12.8	49.7
	AF-R-MI-FGSM	**99.9***	**70.8**	**61.6**	**58.5**	**32.8**	**18.2**	**56.9**
Inc-v4	I-FGSM	19.9	**99.8***	14.9	17.2	12.6	5.2	28.2
	AF-R-I-FGSM	47.6	99.7*	37.3	32.2	18.4	8.4	40.6
	MI-FGSM	44.6	**99.8***	38.3	39.6	22.2	10.8	42.5
	AF-MI-FGSM	60.2	99.7	52.1	51.1	26.3	14.1	50.6
	AF-R-MI-FGSM	**70.1**	99.4*	**64.1**	**61.2**	**31.9**	**19.2**	**57.6**
IncRes-v2	I-FGSM	17.9	19.3	**98.8***	16.2	14.5	6.9	28.9
	AF-R-I-FGSM	44.2	48.7	98.5*	31.9	21.6	13.5	43.1
	MI-FGSM	45.1	43.1	98.2*	39.6	26.4	15.1	44.6
	AF-MI-FGSM	54.3	54.8	98.4	46.3	31.1	21.8	51.1
	AF-R-MI-FGSM	**66.3**	**66.9**	97.7*	**56.1**	**37.8**	**30.4**	**59.2**

Continued

Res-101	I-FGSM	10.4	13.8	10.3	99.5*	12.3	5.3	25.3
	AF-R-I-FGSM	30.2	37.1	27.2	99.1*	17.5	8.3	36.6
	MI-FGSM	26.6	31.7	22.7	**99.7***	20.8	8.1	34.9
	AF-MI-FGSM	32.3	38.9	32.2	98.8	21.6	9.1	38.8
	AF-R-MI-FGSM	**48.8**	**53.2**	**44.5**	98.4*	**25.2**	**13.5**	**47.2**

Table 2. The success rate (%) of integrated model attack
表 2. 集成模型攻击的成功率对比(%)

攻击方法	Inc-v3*	Inc-v4*	IncRes-v2*	Res-101*	Adv-Inc-v3	IncRes-v2-ens	平均值
I-FGSM	**99.9**	98.6	93.7	95.9	18.2	9.1	69.2
AF-R-I-FGSM	99.8	98.6	**95.9**	**96.5**	31.4	22.3	74.1
MI-FGSM	99.8	**99.1**	93.3	95.9	31.9	18.9	73.2
AF-R-MI-FGSM	99.2	97.9	93.5	94.3	**52.2**	**42.7**	**79.9**

4.1. 实验设置

数据集。本文使用 ImageNet 数据集，从其验证集中随机选取 1000 张图片，共选取了 1000 个类别，即每一个类别选取了一张图片。所选取的图片都能够被本文所使用的分类网络正确分类。所有的图片大小都预先调整为 299 × 299 × 3。

网络模型。在本文中，一共选取了 6 个网络模型，包含四个正常训练的网络模型，Inception-v3 (Inc-v3) [17]，Inception-v4 (Inc-v4) [18]，Inception-Resnet-v2 (IncRes-v2) [19]和 Resnet-v2-101 (Res-101) [19]和两个经过了对抗训练的网络模型，adv-Inception-v3 (adv-Inc-v3)和 ens-adv-Inception-Resnet-v2 (IncRes-v2-ens)。

实验细节。本文超参数的选择按照文献[6]中的设定，迭代次数 $T=10$，扰动量大小为 $\varepsilon=16$，步长为 $\alpha=1.6$。对于 MI-FGSM，其衰减因子 $\mu=1.0$，对于随机变换函数 $AF(x^{adv};p)$，其转换概率 $p=0.5$。对于公式(5)中对图像求 $n=9$ 次梯度，最后取平均梯度进行计算。

评价指标。在对抗样本的攻击领域中，采取对抗样本的攻击成功率(Attack success rate, ASR)来评价攻击算法的好坏，即被攻击模型的分类错误率。但在不同的攻击算法中，ASR 有着不同的含义，本文采取无目标攻击的方式，即所生成的对抗样本能够使得分类网络给出一个与真实标签不一致的结果即可，无需使其分类为指定类别。因此，本文使用的 ASR 可被定义为：

$$\text{ASR}=\frac{\text{样本被错误分类的数量}}{\text{样本的总数量}}\times 100\% \tag{9}$$

4.2. 单模型攻击

本节中，使用单个深度神经网络来生成对抗样本，对分类模型进行攻击。实验使用 I-FGSM、AF-R-I-FGSM 以及 MI-FGSM、AF-R-MI-FGSM 方法，分别在在四个正常训练的网络模型上生成对抗样本，并在所有的 6 个网络模型(包括四个正常训练的网络模型和两个经过了对抗训练的网络模型)上进行攻击。实验结果如表 1 所示。

表 1 中的结果表明，AF-R-I-FGSM 和 AF-R-MI-FGSM 方法在黑盒设定下的攻击成功率高于其它攻击方法并且在白盒设定下也保持着较高的攻击成功率。例如，使用 Inc-v3 网络模型，来生成对抗样本，攻击正常训练的网络模型 Inc-v4 时，AF-R-MI-FGSM 方法的黑盒攻击成功率达到了 70.8%，而 I-FGSM 和 MI-FGSM 仅有 21.8%和 42.1%。攻击对抗训练的网络时也表现出了良好的性能，在 IncRes-v2 网络上使

用 MI-FGSM 和 AF-R-MI-FGSM 方法生成的对抗样本进行黑盒攻击，其平均攻击成功率分别为 33.8%和 51.5%。与 MI-FGSM 相比，AF-R-MI-FGSM 的整体平均黑盒攻击成功率提升了 14.3%。说明了将仿射变换和梯度细化引入到对抗样本的生成过程中，能有效的增强样本的迁移性，从而提升其在黑盒条件下的攻击性能。

在仅使用仿射变换，即 AF-MI-FGSM 方法，生成的对抗样本的攻击成功率均低于使用了梯度细化的攻击方法，AF-R-MI-FGSM。如表 1 结果所示，AF-MI-FGSM 方法的平均黑盒攻击成功率为 37.2%，而 AF-R-MI-FGSM 方法的平均黑盒攻击成功率为 46.6%，提高了 9.4%。证明了使用梯度细化能够在一定程度上缓解因随机变换所产生的消极梯度影响。

4.3. 集成模型攻击

表 1 的结果表明 AF-R-MI-FGSM 方法能提高对抗样本的迁移性，但还能够使用集成网络来生成对抗样本，进一步提高其黑盒攻击成功率。实验遵循文献[15]的设定，通过集成多个网络的逻辑值，来生成对抗样本。本实验集成四个正常训练的网络模型，使用 I-FGSM、AF-R-I-FGSM 以及 MI-FGSM、AF-R-MI-FGSM 方法来生成对抗样本，并在所有的 6 个网络模型上进行攻击测试。在实验中，迭代次数 $T=10$，扰动大小 $\varepsilon=16$，集成模型使用相等的集成权重，$\omega_k=1/4$。其实验结果如表 2 所示。

表 2 的结果表明，使用 AF-R-MI-FGSM 方法在集成模型上生成的对抗样本，其黑盒成功率更高。例如，AF-R-MI-FGSM 方法生成的对抗样本攻击经过对抗训练的网络时，其平均黑盒攻击成功率为 47.5%，而 MI-FGSM 仅有 25.4%，提高了 22.1%。同时，在白盒攻击方面，AF-R-MI-FGSM 也保持着较高的成功率，均在 90%以上。

4.4. 超参数研究

4.4.1. 转换概率 p 对攻击成功率的影响

本节研究不同转换概率对 AF-R-MI-FGSM 方法的影响，使用 AF-R-MI-FGSM 方法在四个正常训练的网络的集成来生成对抗样本，并在所有的 6 个网络上进行测试，以此来评估不同转换概率对于攻击成功率的影响。转换概率 p 从 0 以 0.1 步长增加到 1，其他超参数设置，迭代次数 $T=10$，扰动大小 $\varepsilon=16$，动量衰减因子 $\mu=1.0$。

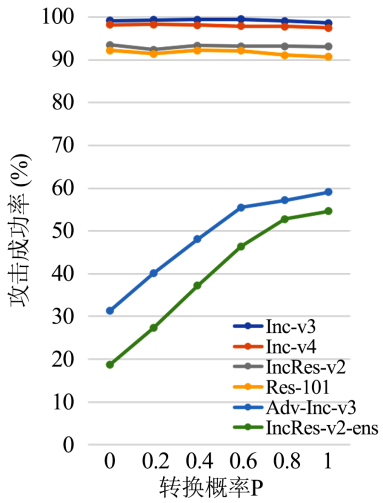

Figure 2. The influence of different conversion probabilities on the success rate of the attack
图 2. 不同转换概率与攻击成功率的关系曲线图

实验结果如图 2 所示，其中 Inc-v3、Inc-v4、IncRes-v2、Res-101 表示，AF-R-MI-FGSM 方法在正常训练的网络模型上的白盒攻击成功率。adv-Inc-v3 和 IncRes-v2-ens 则表示，在使用了对抗训练的网络模型上，进行黑盒攻击，其成功率的变化趋势。实验结果表明，随着转换概率 p 的增大，其白盒成功率是逐步降低的，但趋势平稳且还保持着较高的成功率。其黑盒的攻击成功率随着转换概率 p 的增加而不断地增加。

4.4.2. 扰动大小ε对攻击成功率的影响

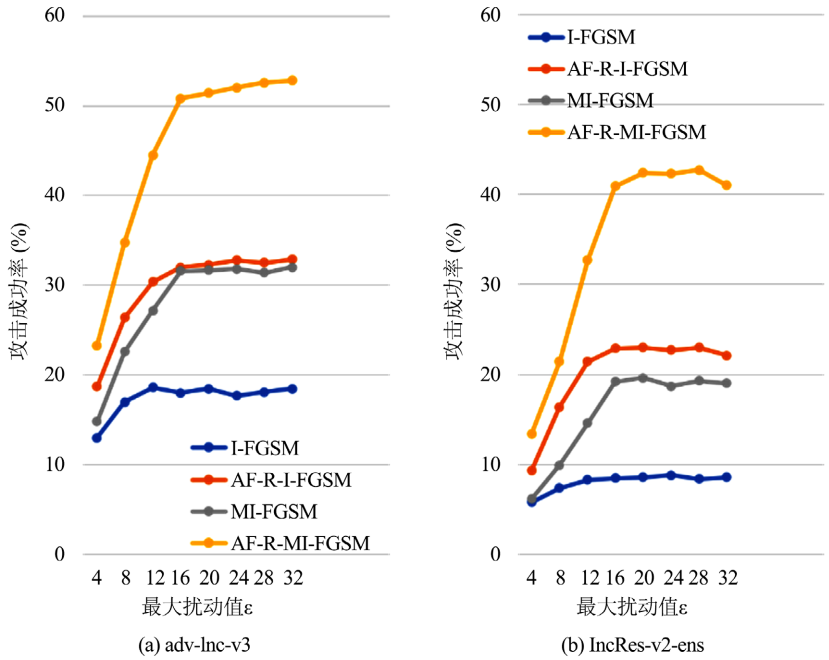

Figure 3. Graph of the relationship between different disturbance magnitude and attack success rate
图 3. 不同扰动大小与攻击成功率的关系曲线图

本节研究不同扰动大小对各方法的攻击成功率的影响，实验使用 I-FGSM、AF-R-I-FGSM 以及 MI-FGSM、AF-R-MI-FGSM 四种方法，将四个正常训练网络进行权重相等的集成，以此来生成对抗样本，并分别在两个对抗训练的网络上进行黑盒测试。本实验所设置的转换概率为 $p=0.5$，扰动大小 ε 从 4 以 4 步长增加到 32。图 3 中的(a) (b)分别为攻击对抗训练网络 adv-Inc-v3 和 IncRes-v2-ens，其中实线表示其黑盒攻击成功率的变化。实验结果表明，攻击成功率开始会随着扰动的增大而增大，当到达一个峰值后成功率其成功率会趋于平稳。

5. 结束语

基于仿射变换和梯度细化的对抗样本生成方法，使用数据增强技术来缓解模型的过拟合问题，使用梯度细化来解决随机变换带来的消极梯度影响，来提高对抗样本的迁移性。此外，通过使用集成模型进一步提升了对抗样本的迁移性。与 MI-FGSM 方法相比，AF-R-MI-FGSM 方法单模型攻击的平均黑盒攻击成功率提升了 14.3%，集成模型攻击的平均黑盒攻击成功率提升了 22.1%。本文还对不同的转换概率进行实验研究，给出了最优参数建议。使用了梯度细化来消除随机变换带来的影响，对于非基于梯度的攻击方法不适用，寻找一种不需要梯度信息来消除随机变换影响的方法是今后探索的方向之一。

基金项目

广州市重点领域研发项目(202007010004)。

参考文献

[1] Krizhevsky, A., Sutskever, I. and Hinton, G. (2012) ImageNet Classification with Deep Convolutional Neural Networks. 26*th Annual Conference on Neural Information Processing Systems* 2012, Lake Tahoe, 3-6 December 2012, 110-117.

[2] 陈晓楠, 胡建敏, 张本俊, 等. 基于模型间迁移性的黑盒对抗攻击起点提升方法[J]. 计算机工程, 2021, 47(8): 162-169.

[3] Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A. (2017) Towards Deep Learning Models Resistant to Adversarial Attacks.

[4] Goodfellow, I.J., Shlens, J. and Szegedy, C. (2018) Explaining and Harnessing Adversarial Examples. https://arxiv.org/abs/1412.6572

[5] Kurakin, A., Goodfellow, I. and Bengio, S. (2017) Adversarial Examples in the Physical World. https://arxiv.org/abs/1607.02533v4

[6] Dong, Y.P., Liao, F.Z., *et al*. (2018) Boosting Adversarial Attacks with Momentum. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 9185-9193. https://doi.org/10.1109/CVPR.2018.00957

[7] Liu, Y.P., Chen, X.Y., Liu, C. and Song, D. (2021) Delving into Transferable Adversarial Examples and Black-Box Attacks. https://arxiv.org/abs/1611.02770

[8] Carlini, N. and Wagner, D. (2017) Towards Evaluating the Robustness of Neural Networks. 2017 *IEEE Symposium on Security and Privacy* (*SP*), San Jose, 22-26 May 2017, 39-57. https://doi.org/10.1109/SP.2017.49

[9] Wang, X.S., He, X.R., Wang, J.D., *et al*. (2021) Admix: Enhancing the Transferability of Adversarial Attacks. http://arxiv.org/abs/2102.00436V3

[10] Tramer, F., Kurakin, A., Papernot, N., *et al*. (2020) Ensemble Adversarial Training: Attacks and Defenses. https://arxiv.org/abs/1705.07204v5

[11] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R. (2014) Intriguing Properties of Neural Networks. *International Conference on Learning Representations*.

[12] Wang, G., Yan, H., Guo, Y., *et al*. (2021) Improving Adversarial Transferability with Gradient Refining. Computer Vision and Pattern Recognition.

[13] Guo, C., Rana, M., Cisse, M., *et al*. (2018) Countering Adversarial Images Using Input Transformations. https://arxiv.org/abs/1711.00117v3

[14] Xie, C., Wang, J., Zhang, Z., *et al*. (2017) Mitigating Adversarial Effects through Randomization. *ICLR* 2018 *Conference Track*. 6*th International Conference on Learning Representations*, Vancouver, 30 April-3 May 2018, 1-16.

[15] Prakash, A., Moran, N., Garber, S., *et al*. (2018) Deflecting Adversarial Attacks with Pixel Deflection. 2018 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 8571-8580. https://doi.org/10.1109/CVPR.2018.00894

[16] 杨博, 张恒巍, 李哲铭, 等. 基于图像翻转变换的对抗样本生成算法[J]. 计算机应用, 2022, 42(8): 2319-2325.

[17] Szegedy, C., Vanhoucke, V., Loffe, S., *et al*. (2016) Rethinking the Inception Architecture for Computer Vision. *Proceedings of* 2016 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 2818-2826. https://doi.org/10.1109/CVPR.2016.308

[18] Szegedy, C., Loffe, S., Vanhoucke, V., *et al*. (2017) Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, **31**, 4278-4284. https://doi.org/10.1609/aaai.v31i1.11231

[19] He, K.M., Zhang, X.Y., Ren, S.Q., *et al*. (2016) Identity Mappings in Deep Residual Networks. *Proceedings of the* 2016 *European Conference on Computer Vision*, Amsterdam, 11-14 October 2016, 630-645. https://doi.org/10.1007/978-3-319-46493-0_38