

# 基于 FGSM 的对抗样本生成算法

汤家军,王 忠

(中国人民解放军火箭军工程大学 基础部,陕西 西安 710025)

**摘 要:**近年来,深度学习算法在各个领域都取得了极大的成功,给人们的生活带来了极大便利。然而神经网络由于其固有特性,用于分类任务时,存在不稳定性,很多因素都影响着分类的准确性,尤其是对抗样本的干扰,通过给图片加上肉眼不可见的扰动,影响分类器的准确性,给神经网络带来了极大的威胁。通过对相关对抗样本的研究,该文提出一种基于白盒攻击的对抗样本生成算法 DCI-FGSM(Dynamic Change Iterative Fast Gradient Sign Method)。通过动态更新梯度及噪声幅值,可以防止模型陷入局部最优,提高了生成对抗样本的效率,使得模型的准确性下降。实验结果表明,在 MNIST 数据集分类的神经网络攻击上 DCI-FGSM 取得了显著的效果,与传统的对抗样本生成算法 FGSM 相比,将攻击成功率提高了 25%,具有更高的攻击效率。

**关键词:**深度学习;不稳定性;白盒攻击;对抗样本;对抗攻击

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2023)03-0105-05

doi:10.3969/j.issn.1673-629X.2023.03.016

## Adversarial Sample Generation Algorithm Based on FGSM

TANG Jia-jun, WANG Zhong

(PLA Rocket Force University of Engineering, Xi'an 710025, China)

**Abstract:** In recent years, deep learning algorithms have achieved great success in various fields and brought great convenience to people's life. However, due to its inherent characteristics, deep neural networks are unstable when used for classification tasks. Many factors affect the accuracy of classification, especially the interference against samples. By adding invisible disturbance to pictures, the accuracy of classifiers is affected, posing a great threat to deep neural networks. Based on the research of correlative adversarial samples, we propose a new adversarial sample generation algorithm DCI-FGSM based on white-box attacks. Through dynamic updating of gradient and noise amplitude, the model can be prevented from falling into local optimal, which improves the efficiency of generating adversarial samples and decreases the accuracy. Experimental results show that DCI-FGSM achieves a remarkable effect on the neural network attack of MNIST dataset classification. Compared with the traditional adversarial examples generation algorithm FGSM, DCI-FGSM improves the success rate of attack by 25% and has higher attack efficiency.

**Key words:** deep learning; instability; white-box attack; adversarial examples; adversarial attack

## 0 引言

随着统计学习和计算机硬件的不断发展,深度学习在各个领域都取得了飞速的发展,在有的方面的表现甚至已经超过了人类。无论是 AlphaGo 在围棋上的优异表现,还是机器翻译取得的巨大成功,都表明了深度学习存在着巨大的潜能。基于深度学习的各种应用也走进了大众的视野,比如人脸识别<sup>[1]</sup>、手语识别<sup>[2]</sup>、医学影像<sup>[3-4]</sup>等。但是深度学习还存在一定的不稳定性<sup>[5]</sup>,造成很多安全问题<sup>[6]</sup>。尤其是容易受到对抗样本的攻击,引起模型分类错误<sup>[7]</sup>。

前人提出了很多对抗攻击<sup>[8-9]</sup>与防御<sup>[10]</sup>的方法,但是还存在着很大的缺陷。该文对其中的一些缺陷进行了研究,提出了一种动态改变噪声幅值、随机添加扰动防止陷入局部最优解的算法,即 Dynamic Change Iterative Fast Gradient Sign Method (DCI-FGSM)。

## 1 相关工作

自从 Szegedy 等提出对抗样本<sup>[11]</sup>以来,学者们发现只需要添加细微的扰动就可以让模型产生错误的预测。近年来,越来越多的研究发现对抗样本不仅可以

收稿日期:2022-05-12

修回日期:2022-09-14

基金项目:国家自然科学基金委员会,青年科学基金项目(62103432)

作者简介:汤家军(1998-),男,硕士研究生,CCF 会员(K4636G),研究方向为智能算法可靠性、对抗样本生成;通讯作者:王 忠(1968-),男,教授,博士,CCF 高级会员(09160S),研究方向为嵌入式系统与嵌入式网络。

找到模型的缺陷<sup>[12]</sup>,还能使模型更具有鲁棒性<sup>[13]</sup>,同时这种方法还可以适用于不同的模型。

为了攻击神经网络模型,Szegedy等人<sup>[11]</sup>提出有边界约束的L-BFGS算法,其收敛速度快、内存开销小,但步长计算精度不高;Goodfellow等<sup>[14]</sup>提出基于梯度方向的FGSM算法,可以有效生成对抗样本对模型进行白盒测试,但其只对梯度方向上进行一次更新,生成的对抗样本效果一般;Kurakin等人<sup>[15]</sup>基于FGSM,提出了I-FGSM算法,可以在梯度方向上进行多次迭代,但是可能会陷入梯度的局部最大值。所以Dong等<sup>[16]</sup>为了增强对抗攻击,提出了一类基于动量的迭代算法,可以摆脱局部最大值,但其只是盲目的去摆脱,存在着一定的局限性。Jia等人为了通过梯度的均方根对学习率进行约束,提出了AdaDelta-Nesterov动量方法<sup>[17]</sup>,这种方法甚至不需要提前设定学习率,同时减少了无用的迭代过程。Wang等人<sup>[18]</sup>为了欺骗模型,利用风格迁移使得对抗样本人眼不易区分,这种方法虽然具有一定的效率,但是在风格迁移的过程中可能会丢失图片本来的特征。Lin等人<sup>[19]</sup>提出一种防御量化的方法来防御对抗攻击,同时保持攻击的效率,在对于经过对抗训练的模型上取得了显著的成果。

## 2 对抗攻击算法

为便于介绍,现在对算法涉及到的一些参数进行说明。将原始样本定义为 $x$ ,与其对应的真实标记为 $y$ , $f(x, \theta)$ 为分类器,可以预测出正确的结果,其中 $\theta$ 为网络参数,将 $J(\theta, x, y)$ 定义为分类器 $f$ 的损失函数,通常为交叉熵损失函数。将添加扰动后的对抗样本定义为 $\tilde{x} = x + \eta$ ,其中 $\eta$ 即为扰动, $x$ 与 $\tilde{x}$ 在视觉上没有任何区别,但是 $\tilde{x}$ 会导致神经网络出现错误的分类。定义一个阈值 $\varepsilon$ ,当 $\|\eta\|_{\infty} < \varepsilon$ 时,分类器的性能不会受到影响,但是会产生错误的预测,即 $f(x, \theta) \neq f(\tilde{x}, \theta)$ 。当经过一次线性梯度方向后,权重和对抗样本的点积为: $\omega^T \tilde{x} = \omega^T(x + \eta) = \omega^T x + \omega^T \eta$ ,对抗样本产生的原因即为 $\omega^T \eta$ ,可以最大化 $\eta = \text{sign}(\omega)$ 使得扰动最大。添加的扰动值为: $\eta = \varepsilon \text{sign}(\nabla_x J(\theta, x, y))$ 。

### 2.1 FGSM 算法

Goodfellow等人<sup>[14]</sup>首次提出对抗训练,即用对抗样本训练神经网络,使得网络更具鲁棒性。他们提出FGSM(Fast Gradient Sign Method)方法,认为高维度的线性模型就可以产生对抗样本。他们在梯度上进一步探索去最大化损失:

$$\tilde{x} = x + \varepsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \quad (1)$$

其中, $\nabla_x J(\theta, x, y)$ 表示损失函数的梯度, $\text{sign}(\cdot)$ 表示sign函数。

### 2.2 I-FGSM 算法

FGSM算法在梯度方向上做攻击,速度快,但是其只能在梯度方向上做一次攻击。在FGSM算法的基础上,提出迭代式的FGSM算法,即I-FGSM(Iterative Fast Gradient Sign Method)算法<sup>[15]</sup>。I-FGSM算法沿着梯度的方向进行多步扰动,并且每一次都重新计算扰动的方向。

算法流程为:

$$\tilde{x}_0 = x \quad (2)$$

$$\tilde{x}_{i+1} = \text{Clip}_{x, \varepsilon} \{ \tilde{x}_i + \alpha \text{sign}(\nabla_{x\tilde{x}} J(\theta, \tilde{x}_i, y_{\text{true}})) \} \quad (3)$$

其中, $\text{Clip}_{x, \varepsilon}(\cdot)$ 函数将对抗样本 $\tilde{x}$ 中的像素值进行裁剪,使其不超过原始样本的无穷范数边界, $\alpha$ 为迭代的步长。

### 2.3 MI-FGSM 算法

I-FGSM在每次迭代中将对抗样本沿梯度的方向贪婪地移动,很容易让模型陷入局部最大值,并且过拟合,可能并不适用于所有的模型。因此,Dong等人<sup>[12]</sup>提出基于动量的方法MI-FGSM(Momentum Iterative Fast Gradient Sign Method):

$$g_{i+1} = \mu \cdot g_i + \frac{\nabla_{\tilde{x}} J(\tilde{x}_i, y, \theta)}{\|\nabla_{\tilde{x}} J(\tilde{x}_i, y, \theta)\|_1} \quad (4)$$

$$\tilde{x}_{i+1} = \tilde{x}_i + \alpha \cdot \text{sign}(g_{i+1}) \quad (5)$$

$g_0 = 0$ 为初始动量且 $\mu$ 为衰减因子。

## 3 DCI-FGSM 算法

在现实世界中存在着很多的对抗样本<sup>[14]</sup>,让人们更加关注深层神经网络的稳定性,学者们提出了大量的对抗攻击算法去对抗各个领域的模型(如:人脸识别、目标检测、语义分割等)。该文仅针对白盒测试中基于梯度的图像对抗样本算法做出分析。

FGSM通过给原始样本加扰动得到对抗样本 $\tilde{x} = x + \eta$ ,只在梯度方向进行一次更新,生成的对抗样本表现一般;与FGSM相比,I-FGSM进行了多次迭代,但其只向梯度的一个方向不断进行探索,可能会陷入局部最优的局面。MI-FGSM首次将动量引入对抗样本生成,提高了攻击的成功率。为了动态更新梯度的大小和方向,使得每次梯度更新的步长动态改变,由此提出了DCI-FGSM算法。图1对不同的对抗样本生成算法的关系进行了介绍,当 $\alpha = 0$ 时,DCI-FGSM算法退化为MI-FGSM算法;当 $\mu = 0$ 时,MI-FGSM算法退化为I-FGSM算法;当 $T = 1$ 时,I-FGSM算法退化为FGSM算法。

算法:DCI-FGSM。

输入:分类器 $f$ ,损失函数 $J$ ,真实样本 $x$ 及其标签 $y$ ,扰动幅度 $\varepsilon$ ,迭代轮次 $T$ 。

输出:对抗样本 $\tilde{x}$ 。

- (a)  $x^{\text{adv}} = x$ ;
- (b) 获得初始梯度  $g_0 = \nabla_x J(\theta, x, y)$
- (c) for  $t = 0$  to  $T - 1$  do
- (d)  $\alpha = \varepsilon / t + 1$ ;
- (e) 更新梯度  $g_{t+1} = \mu * \nabla_x J(\theta, x_t^{\text{adv}}, y) / \|g_t\|_\infty$ ;
- (f) 获得对抗样本  $\tilde{x}_{t+1} = x_t + (\varepsilon - \alpha) \text{sign}(g_{t+1})$ ;
- (g) end for
- (h) return  $x^{\text{adv}} = \tilde{x}$

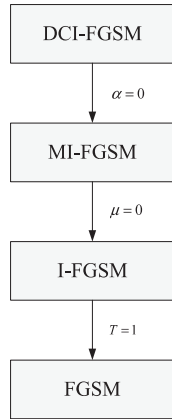


图1 各种攻击算法关系

根据前人的研究,  $\alpha$  最好的取值为  $\varepsilon/T$ ,  $T$  是总的迭代次数, 在迭代的过程中,  $\alpha$  始终保持不变可能会导致模型跳过最优值。受此启发, 在 DCI-FGSM 算法中,  $\alpha$  不再这样取值, 而是根据每次迭代的次数去取值, 这样做的好处是不至于让每次迭代的步长过大, 能够有效找到原始数据的边界。

图2为 DCI-FGSM 算法的基本流程, 包含五个步骤。通过卷积神经网络获得扰动, 不断迭代更新梯度获取对抗扰动, 基于原始图像获取对抗样本。

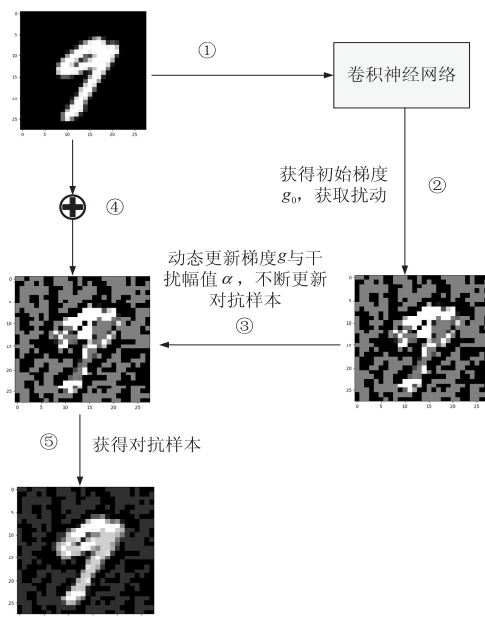


图2 DCI-FGSM 攻击算法流程

图2中, ①表示通过卷积神经网络训练原始样本; ②表示获取初始扰动, 代表算法中的第2步; ③表示迭

代不断更新对抗样本, 代表算法中的循环; ④⑤表示将原始样本加上对抗扰动获取对抗样本。

DCI-FGSM 算法总体公式为:

$$x_0^{\text{adv}} = x \quad (6)$$

$$x_{t+1}^{\text{adv}} = \text{Clip}_{x, \varepsilon} \{ x_t^{\text{adv}} + (\varepsilon - \alpha) \text{sign}(g_{t+1}) \}, \text{ where}$$

$$g_{t+1} = \alpha * \nabla_x J(\theta, x_t^{\text{adv}}, y) / \|g_0\|_\infty \quad (7)$$

## 4 实验测试

在本节, 通过实验对比了所提方法的效率, 以验证 DCI-FGSM 的优势。

### 4.1 实验设置

(1) 数据集。采用的是 MNIST 数据集, 由 60 000 个训练样本和 10 000 个测试样本组成, 每个样本都是一张  $28 * 28$  像素的灰度手写数字图片。

(2) 网络模型。实验采用的网络模型为简单神经网络模型, 网络结构如图3所示。

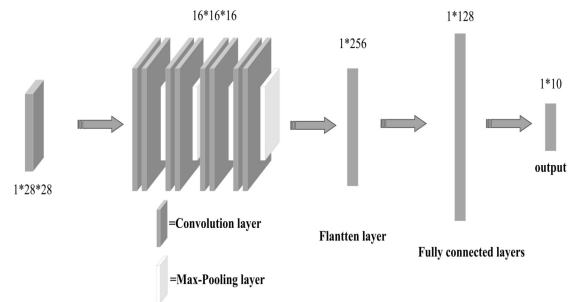


图3 简单神经网络模型结构

(3) 实验设置。对于超参数, 遵循文献[12]中的设定, 迭代轮次  $T = 10$ , 步长初始为 1.6, 对于 MI-FGSM 的衰减因子  $= 1.0$ 。实验在 GTX1080Ti 显卡上运行计算。

(4) 评价指标。通过模型的准确率以及模型的损失值判断生成的对抗样本的效果。准确率指简单神经网络模型对于正确分类图片对图片总数的占比, 准确率越低, 对抗样本的效果越好; 损失值越高, 对抗样本的效果越好。

### 4.2 实验结果

在本节, 采用算法对模型进行了攻击, 并对不同的算法进行比较。

图4、图5为神经网络在进行 10 个 epoch 训练下的准确率和损失值, 可以看到损失已经降到极低且准确率很高, 基本达到百分之百。本节采用 FGSM、I-FGSM、MI-FGSM、DCI-FGSM 算法对此神经网络进行攻击, 生成对抗样本, 结果见表1, 其中模型准确率是攻击前后模型对于图片的分类准确率, 损失是攻击前后模型的损失值。在图6中最左边为原始图像, 中间的为噪声, 右边的为原始图像加上噪声后的图片。可以看到, 原始图像与对抗样本无明显区别, 但是神经

网络却会对图像进行错误的分类。

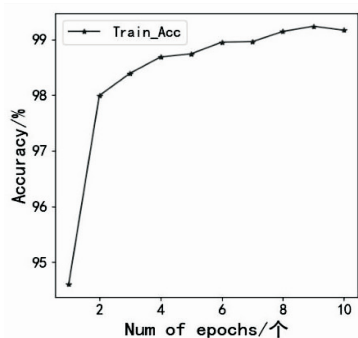


图 4 模型准确率随 epoch 变化

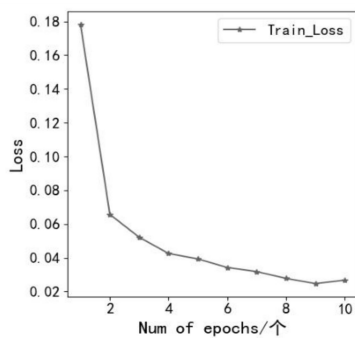
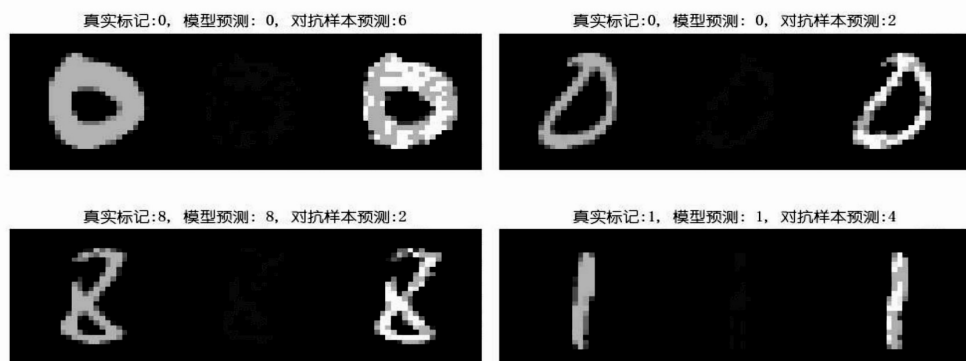


图 5 模型损失随 epoch 变化



(左:真实样本,中:噪声,右:对抗样本)

图 6 真实标记以及对抗样本标记

表 1 中的结果显示,DCI-FGSM 能够提高白盒攻击下的成功率,原始模型的准确率为 98.25%,经过 DCI-FGSM 算法的攻击,其准确率降至 4.5%,且相比

较于 FGSM、I-FGSM 和 MI-FGSM 算法,DCI-FGSM 具有更高的攻击效率,使得模型的损失值更大。

表 1 攻击自定模型的准确率与损失

攻击算法	模型准确率/%	攻击成功率/%	模型损失
原始模型	98.25	0	0.000 2
FGSM	7.79	45.3	13.421 1
I-FGSM	6.21	51.6	14.558 1
MI-FGSM	5.24	61.3	15.445 4
DCI-FGSM(ours)	4.50	69.7	16.388 8

研究设置了不同的扰动值  $\epsilon$ , 针对正常训练的神经网络模型,生成对抗样本。图 7、图 8 展示了在不同的  $\epsilon$  下各种攻击算法的成功率。从图中可以看到,随着扰动值  $\epsilon$  的增大,各个算法的攻击成功率也在升高,

但是 DCI-FGSM 算法的变化趋势与其他不一样,DCI-FGSM 能够使模型的准确率以更快的速度降到更低的值,使得模型的损失值以更快的速度提升至更高的值。足以表明 DCI-FGSM 的优越性。

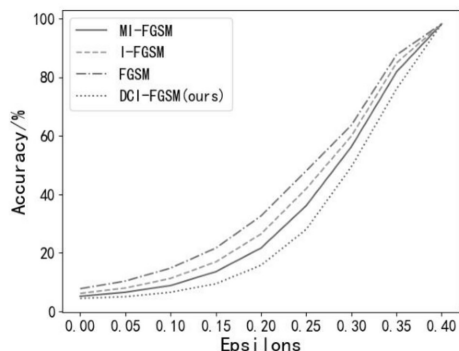


图 7 模型准确率变化

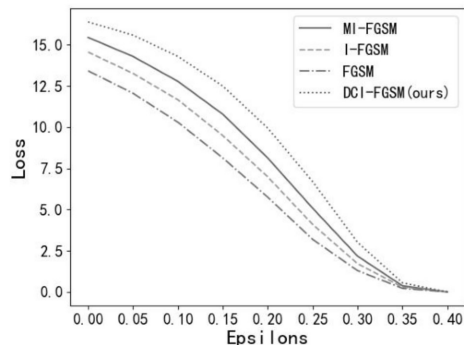


图 8 模型损失变化



## 5 结束语

该文提出一种基于白盒攻击的对抗样本生成算法,通过动态改变扰动的值,并且通过前一步梯度去更正梯度前进的方向,防止模型陷入局部最优,从而生成更有效的对抗样本。实验结果表明,与传统的对抗样本生成算法相比,DCI-FGSM 在白盒攻击上具有更强大的效率。目前 DCI-FGSM 算法仅针对白盒攻击,其在黑盒攻击上是否有更高的效率,还需要进一步研究。

### 参考文献:

- [1] DHAR P, GLEASON J, ROY A, et al. PASS: protected attribute suppression system for mitigating bias in face recognition [C]//IEEE/CVF international conference on computer vision (ICCV). Montreal: IEEE, 2021: 15067–15076.
- [2] SAUNDERS B, CAMGOZ N C, BOWDEN R. Mixed Sign language production via a mixture of motion primitives [C]//IEEE/CVF international conference on computer vision (ICCV). Montreal: IEEE, 2021: 1899–1909.
- [3] ZHOU H Y, LU C, YANG S, et al. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts [C]//IEEE/CVF international conference on computer vision (ICCV). Montreal: IEEE, 2021: 3479–3489.
- [4] 覃 延. 神经网络模型和多元线性回归预测肾结石 CT 值的比较[J]. 影像研究与医学应用, 2020, 4(6): 26–28.
- [5] PONS L, OZKAYA I. Priority quality attributes for engineering AI-enabled systems [C]//AAAI FSS-19: artificial intelligence in government and public sector. Hawaii: AAAI, 2019.
- [6] 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.
- [7] CHEN X N, HU J M, ZHANG B J, et al. Blackbox attack adversarial starting point promotion method based on mobility between models[J]. Computer Engineering, 2021, 47(8): 162–169.
- [8] LIN J, SONG C, HE K, et al. Nesterov accelerated gradient and scale invariance for adversarial attacks [C]//International conference on learning representations. [s. l.]: ICLR, 2020.
- [9] 姜 妍, 张立国. 面向深度学习模型的对抗攻击与防御方法综述[J]. 计算机工程, 2021, 47(1): 1–11.
- [10] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks [C]//6th international conference on learning representations. Vancouver: ICLR, 2018.
- [11] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks [C]//International conference on learning representations. Banff: ICLR, 2014.
- [12] ATHALYE A, CARLINI N, WAGNER D. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples [C]//International conference on machine learning. [s. l.]: [s. n.], 2018: 436–448.
- [13] ZHANG H, YU Y, JIAO J, et al. Theoretically principled trade-off between robustness and accuracy [C]//International conference on machine learning. California: ICML, 2019: 12907–12929.
- [14] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples [C]//International conference on learning representations. San Diego: [s. n.], 2015.
- [15] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world [C]//International conference on learning representations. Toulon: ICLR, 2017.
- [16] DONG Y, LIAO F, PANG T, et al. Boosting adversarial attacks with momentum [C]//IEEE computer society conference on computer vision and pattern recognition. Salt Lake City: IEEE, 2018: 9185–9193.
- [17] JIA X B, SHI J S. Ada\_Nesterov momentum algorithm—the nesterov momentum algorithm with adaptive learning rate [J]. Computer Science and Application, 2019, 9: 351–358.
- [18] WANG Lu, ZENG Guohui, HUANG Bo. Implementation of style transfer algorithm based on deep learning [J]. Intelligent Computer and Application, 2020, 10(2): 57–60.
- [19] LIN J, GAN C, HAN S. Defensive quantization: when efficiency meets robustness [C]//International conference on learning representations. New Orleans: ICLR, 2019.