基于多尺度特征融合的对抗循环生成网络防御模型

黄楷铭，贾世杰*，陈洁

大连交通大学自动化与电气工程学院，辽宁大连 116028

摘要： 深度神经网络目前在诸多领域均得到较好应用，然而关于深度模型的安全问题日渐突出，最近研究表明通过对抗样本会使神经网络模型输出错误的分类。因此对抗样本是深度神经网络继续推进发展的必须克服的一大障碍。设计一个高效且强大的，并且能够防御多种攻击算法的强鲁棒性防御模型是目前的主要任务，本文将生成对抗网络（Generative adversarial networks, GAN）和现有的攻击算法结合，提出一种基于多尺度特征融合循环对抗生成网络的防御模型。首先利用攻击算法生成的对抗样本作为 GAN 的训练样本，同时利用 CycleGan 独有的网络结构来使重构的图像更加接近于干净图像并清除潜在的扰动。在生成器端本文采用多特征融合 TernausNet 结构保证在扰动除去的过程中尽可能还原图像的特征信息，在判别器端加入注意力机制增大感受野，建立全局的依赖关系，训练出一个更鲁棒性的判别器来帮助 GAN 训练通过在 CIFAR-10 和 ImageNet 数据集上进行实验。证明训练完成后，该模型可以直接对原始样本和对抗样本进行正确分类，对各类对抗攻击算法达到很好的防御效果，且对比已有方法防御效果好，增强了深度模型的鲁棒性。

关键词： 对抗样本；对抗生成网络；多特征融合；防御模型；注意力机制

DOI: 10.57237/j.se.2023.02.003

Network Defense Model Based on Multi scale Feature Fusion for Countering Cycle Generation

Huang Kaiming, Jia Shijie*, Chen Jie

School of Electric Engineering and Automation, Dalian Jiaotong University, Dalian 116028, China

Abstract: The deep neural network has been applied well in many fields at present, but the security problem of the deep model has become increasingly prominent. Recent research shows that the neural network model will output the wrong classification by confronting samples. Therefore, confrontation samples are a major obstacle that must be overcome for the further development of deep neural networks. At present, the main task is to design an efficient and powerful defense model with strong robustness that can defend against multiple attack algorithms. This paper proposes a defense model based on multi-scale feature fusion circular confrontation generation network by combining the generation of generic adversarial networks (GAN) with existing attack algorithms. First, use the confrontation samples generated by the attack algorithm as the training samples of GAN, and use the unique network structure of CycleGan to make the reconstructed image closer to the clean image and remove potential disturbances. On the generator side, this paper uses multi feature fusion TernausNet structure to ensure that the feature information of the image can be restored as much as possible during the process of

*通信作者：贾世杰, jsj@djtu.edu.cn

收稿日期: 2023-01-05; 接受日期: 2023-02-16; 在线出版日期: 2023-03-28

disturbance removal, and adds attention mechanism to the discriminator side to increase the receptive field, establish global dependency, and train a more robust discriminator to help GAN training through experiments on CIFAR-10 and ImageNet datasets. It is proved that after the training, the model can directly classify the original samples and confrontation samples correctly, and achieve good defense effect for all kinds of confrontation attack algorithms. Compared with the existing methods, the defense effect is good, and the robustness of the depth model is enhanced.

Keywords: Confrontation Sample; Confrontation Generation Network; Multi Feature Fusion; Defense Model; Attention Mechanis

1 引言

近几年深度神经网络的快速发展，在带来巨大的经济效益和社会效益的同时，也产生了人工智能的安全问题。有研究表明[1]深度神经网络易受到对抗样本的攻击，即在原始样本上添加精心设计的微弱扰动就能使深度神经网络判别错误。随着不断深入的研究深度学习模型，诸多自动驾驶图像识别等领域的安全问题也逐一暴露出来[2]。

在 2014 年 szegedy 等人[3]首先提出了“对抗样本”的概念，他们发现将精心设计的微小扰动注入模型的输入样本图像，极有可能使得模型输出错误的分类结果，这种扰动往往十分微小，几乎无法被人眼察觉，可以实现欺骗网络模型的目的。随后 Goodfellow 等人[4]解释了对抗样本的原理，提出神经网络易受对抗样本攻击的原因在于它们的高维线性特性，否定早期研究认为的神经网络高维非线性，并在此基础上提出了一种快速梯度法（Fast Gradient Sign Method, FGSM）来生成对抗样本。

与此同时还有不少针对攻击算法提出的防御方法，一种主要的对抗防御方法是增强网络内部结构的鲁棒性，以实现模型的对抗样本防御。例如，Guo [5]使用 JPEG 压缩，使用离散余弦变换来去除对人类视觉不那么重要的高频分量。He 等[6]提出参数噪声注入（Parametric Noise Injection, PNI），通过解决 Min-Max 优化问题，在每一层的激活或权重进行可训练的高斯噪声注入，实验结果表明 PNI 有效提升了对白盒和黑盒攻击的鲁棒性，Liao 等人[7]采用高级表示引导去噪器（HGD）框架作为预处理步骤来消除扰动。NeurIPS2017 国防竞争排名第 2 的 Xie 的方法[8]引入了两步准备处理图像，该思路是在图像进入分类网络之前加入了随机缩放（resize）层和随机填充（padding）层，然后通过结合这个随机层的方法抵御攻击。

另一种对抗防御方法是通过预处理的方式。生成对抗网络是 GoodFellow 等人[9]在 2014 年提出的一种无监督生成式学习算法，包含两部分网络，一是生成器网络（Generator），另一部分则是鉴别器网络（Discriminator）。从 2016 年 GAN 提出至今，广泛应用于多个领域。随着对抗样本的出现，一些研究工作已将 GAN 应用在对抗样本防御之中。Jin 等[10]提出了 APE-GAN 消除对抗样本的影响，APE-GAN 通过重建良性样本来消除对抗样本表面的噪声，在彩色图像和灰度图像都有良好的表现，但其训练过程极不稳定并且防御的范围有限。Kabkab 等[11]提出了 Defense-GAN 防御对抗样本；DefenseGAN 通过将靠近原始对抗样本的新的良性样本作为推理模型的输入来消除对抗样本的影响；可以连同任意分类器使用并且不会影响分类器结构，可以看作是一个分类任务之前的预处理。试验结果表明在 Minst 等灰度图像上表现良好，但在彩色图像上还有待提升。Mustafa 等人[12]提出利用小波滤波和超分辨对抗生成网络图像恢复对手，以最小化对抗效果，但是超分模型的参数较多推理速度较慢，无法更实际的推进到模型部署。但从 GAN 防御模型的角度出发，通常的手段是将非流形样本重新映射到自然图像流形上，但还存在特征提取不足、模型参数较多、推理速度慢和难以部署的情况。

针对这些问题，本文提出了基于多尺度特征融合的对抗循环生成网络的攻击防御模型。本文将图像上的多尺度特征进行融合在 GAN 进行重建图像，清除扰动，整个过程用 CycleGan 的特殊结构极大保证了图片在从对抗域迁移到干净域时不会造成信息的损失，从而使生成的图像更加接近于真实样本。本文用 SSP 攻击算法生成的对抗样本作为模型训练数据并用循环损失函数保证在训练过程的稳定性，另外把 TernausNet 作为对抗循环生成网络的生成器，来补全特征提取不足的问题，在判别器端加入注意力机制，提升感受野，使判别器更鲁棒性，从而提升神经网络对对抗样本的防御成功率。

本文的主要创新点如下主要分为两部分：

1. 提出了一种基于多特征融合循环对抗生成网络的防御方法，本文提出的融合注意力机制和 TernausNet 的对抗生成网络防御模型，采用循环一致性损失对图像重建的内容进行约束，促进攻防性能不断提高，在此基础上对生成器模块改进为 TernausNet 结构，充分提升图像深层特征的提取和多尺度特征的融合，在判别器端加入注意力机制扩大了感受野构成一个更鲁棒性的判别器。训练完成后可以模型可以防御多种对抗攻击。

2. 实验结果表明在公共数据集 ImageNet 和 Cifar10 上相比于主流的 NRP, adv.train 方法，都验证了其有效性，在经过防御模型后图像的潜在扰动更少并且在测试集上获得了更高的准确率。

2 相关工作

2.1 对抗攻击算法

2.1.1 DeepFool

2016 年 Moosavi-Dezfooli 等人[13]提出了一种基于超平面分类思想的对抗样本生成算法 DeepFool，提出鲁棒性指标来计算最优对抗扰动大小，该指标通过迭代的方式进行计算，每轮迭代的对抗扰动累加值都成为下一轮迭代的基础。

2.1.2 PGD

2018 年，Madry 等人[14]提出 FGSM 的改良版被称为是最强的一阶梯度攻击—投影梯度下降法（PGD），是一种多步迭代算法，是目前公认为最强的白盒攻击方法，也是用于评估模型鲁棒性的基准测试算法之一，同样具有的迁移攻击能力弱的问题

2.1.3 SSP

2020 年 Muzammal 等人[15]提出基于特征失真的自监督扰动算法，传统的强白盒攻击算法通常用于 AT 考虑已知的网络参数 θ，并干扰输入创建x'，不能很好地推广到其他网络，Muzammal 等人建议通过最大化特征损失来寻找对手的神经网络。这种方法不依赖于决策边界信息，因为其方法的“基于表示的”攻击通过解决以下优化问题直接扰动了特征空间。

2.2 自注意力机制

对于图像分类的任务中现有的神经网络仍有在分类图像上特征关注度不足的问题自注意力机制经过对凸显特征图中含有较多有效信息的特征表示，并抑制无用信息的影响，从而达到对网络模型的表征能力和图像分类性能上的优化。

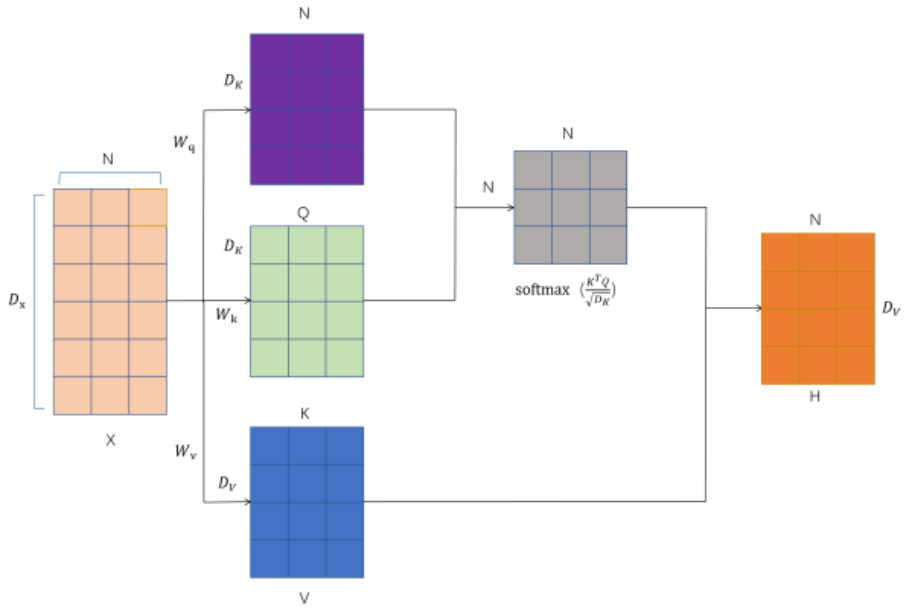

图 1 自注意力机制

假设此时输出为 X，则可以分为以下几个步骤：将输入的 X 线性投影到三个不同的空间，已分别得到查询向量q_i，键向量k_i，值向量v_i，得到三个不同的矩阵w_q,w_k,w_v。

2.3 对抗样本防御

对于对抗样本防御技术的研究思路主要分为两大类—数据层面的防御方法和模型层面的防御方法。数据层面的防御方法可以分为修改训练模型参数[16, 17]和在测试阶段修改的输入样本[5, 18]；模型层面的防御方法则可以分为修改网络结构[19, 20]和使用附加模型 [21, 22]，下面将介绍几个与本文实验相关的对抗防御方法。

2.3.1 对抗训练（Adv.Training）

对抗训练最早是由 Madry 等人[14]提出的一种对抗样本的防御方法。其主要思想是利用攻击算法生成的对抗样本作为原数据集的扩充，一并加入到模型的训练，被称为是最有效的对抗防御方式，该方法存在防御滞后和通用性较差无法防御多种攻击的问题。

2.3.2 Neural Representation Purifier

2020 年 Muzammal 等人[15]提出思想是用一种自监督的扰动来生成对抗样本。作者将深度网络特征空间中特征的损失作为 SSP 攻击算法，将其生成的对抗样本加入到 GAN 的训练之中去，该方法不仅促使两者更好地学习对抗样本和干净样本的分布，生成质量更高的图片，更重要的是使鉴别器具有更好的鲁棒性。但该算法模型推理速度较慢，模型计算量大的问题。

3 基于多尺度特征融合的对抗循环生成网络的攻击防御模型

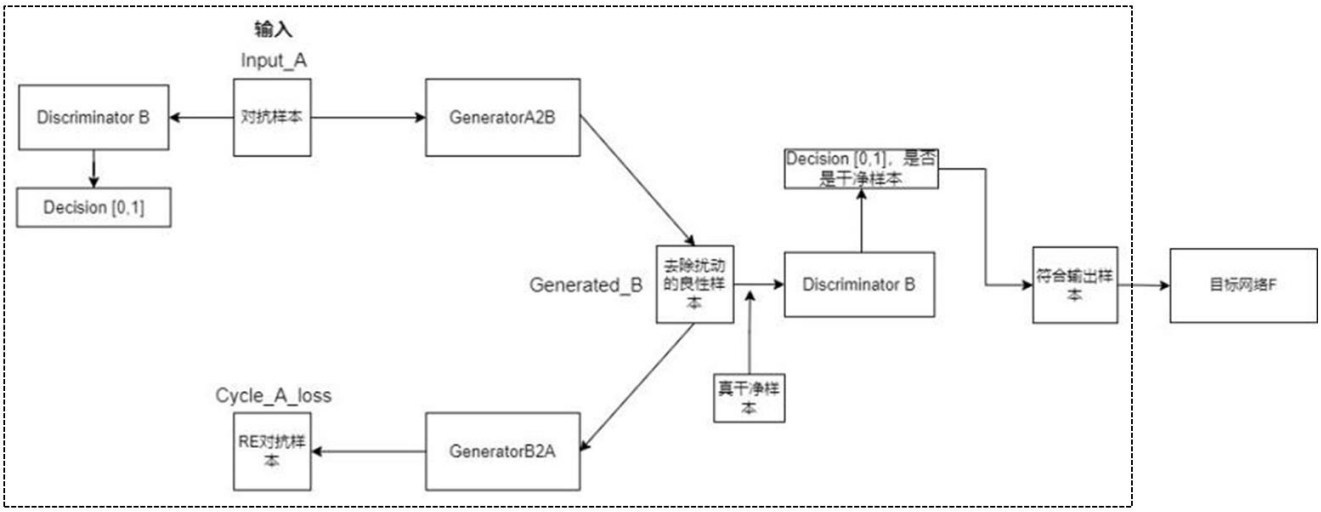

图 2 网络示意图

3.1 总体框架

本文将对抗训练和 GAN 防御对抗样本结合起来，提出了 CYCLEG-DefGAN，它的网络示意图如图 1 所示，当对抗样本输入到网络之中，经过训练好的生成器 GeneratorA2B 中，将对抗样本域里的图像尽可能的转换到干净样本域中然后交给判别器处理，GeneratorB2A 保证的是在样本从对抗域到干净域转换时不会过分学习干净域样本从而去过度改变对抗空间的样本，此时训练好的 DiscriminatorB 把生成器重建的图片和给定的干净图片进行对比，输出结果则代表的是预测值与真实标签的距离。当生成器输出的图片足够接近于真实样本时输出结果给目标网络 F，如此一来当有样本进入目标网络时，都会先经过本文的防御模型进行净化，从而起到良好的防御效果。

在 CYCLEG-DefGAN 中生成器和判别器不断博弈，生成器生成的样本越来越真实，判别器尽可能地提升分别样本的真假性。原始的干净样本通过 SSP 攻击算法形成对抗样本之后，有良好的攻击效果且具有较好的迁移性能，分类器不能正确分类。本文模型的思想是将 SSP 攻击算法生成的对抗样本和干净样本都加入到对抗生成网络训练中，判别器通过提供的干净样本不断促进生成器提升图片的质量从而将原本在图片上的扰动给破坏和清除掉，利用 GAN 的数据训练上的优势提升模型的鲁棒性，抵抗多种对抗样本，进一步提高对对抗样本防御的成功率。

3.1.1 生成器结构

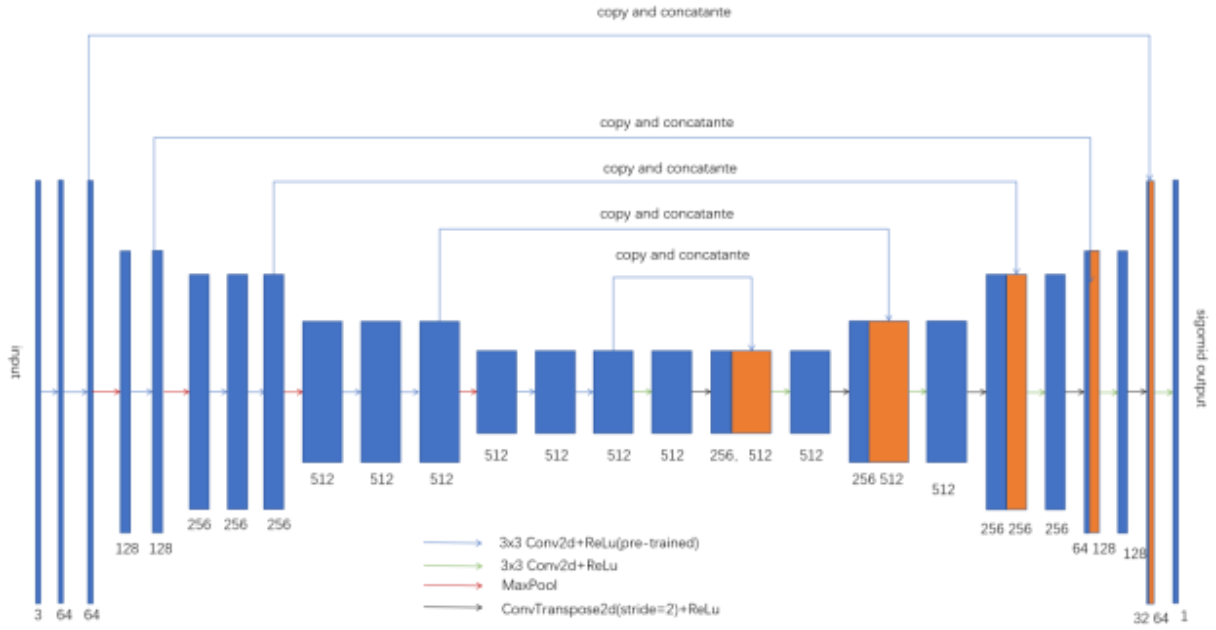

图 3 生成器结构图

生成器的网络结构如图 3，采用的是 TernausNet 卷积编码器-解码器架构，U-Net 该网络被认为是最先进的图像分割架构之一，本文认为它也适用于像素级图像转换任务[23]，在基础上本文使用的 TernausNet 是一种在 ImageNet 上预先训练有 VGG16 编码器的 U-Net 结构，通过加载预训练权重的 U-Net 来轻松提升生成器的性能，通过预训练方式，保留几层与训练数据，实现多特征融合。加快了网络的收敛速度。如上图 3 作为这个结构中的编码器，参考传统的 VGG 系列，所有卷积层都有 3x3 的卷积核，随着网络的加深，在每次最大池化操作结束之后，信道数会翻倍直到 512。在解码器端使用转置卷积层，将特征映射的大小增加一倍，同时将通道数减少一半。然后将转置卷积的输出与解码器的相应部分的输出级联。生成器负责的是去除对抗样本的扰动，生成接近原始数据集真实分布的样本。

3.1.2 鉴别器结构

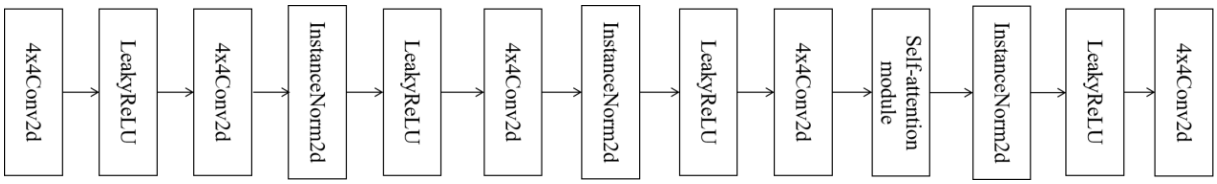

图 4 鉴别器结构图

判别器体系结构也是基于 VGG 网络[24]的，它由五个卷积块组成，其中包含卷积层然后是 batch-norm 和 leaky-relu，最后再连接一个全连接层组成。在此基础上加入自注意力机制来增加判别器的感受野，目的是通过训练得到一个更鲁棒性的判别器，自注意力机制则利用特征图自主学习权重分布，再将其与特征进行融合，加强训练过程对分类任务贡献大的特征，降低贡献度小的特征的影响，提高判别器分类准确率，因此可以获得一个更好鲁棒性的判别器使他的分辩干净样本与对抗样本。

3.2 损失函数

关于整体的损失函数，本文一共定义了四个损失

函数，如下所示

$$L(G,F,D_X,D_Y)=L_{GAN}(G,D_Y,X,Y)+L_{GAN}(G,D_x,Y,Y)+\lambda L_{cyc}(G,F)+\lambda L_{identity}(G,F) \quad (1)$$

$$L_{GAN}(G,D_Y,X,Y)=E_{y\sim p_{data(y)}}[\ln D_Y(y)]+E_{x\sim p_{data(x)}}[\ln(1-D_Y(G(x)))] \quad (2)$$

$$L_{cyc}(G,F)=E_{x\sim p_{data}(x)}[||F(G(x))-x||_1]+E_{y\sim p_{data}(y)}[||G(F(y))-y||_1] \quad (3)$$

$$L_{identity}(G,F)=E_{y\sim p_{data}(y)}[||G(y)-y||_1]+E_{x\sim p_{data}(x)}[||F(x)-x||_1] \quad (4)$$

3.2.1 生成对抗损失函数

对于映射关系从对抗域迁移到干净域 X→Y，生成器 G 的目标是将 X 空间中的样本转化成 Y 空间中的样本，可以构造下面损失函数：

$$L_{GAN}(G,D_Y,X,Y)=E_{y\sim p_{data(y)}}[\ln D_Y(y)]+E_{x\sim p_{data(x)}}[\ln(1-D_Y(G(x)))] \quad (5)$$

D 为判别器，输出值[0,1]，$D_y=1$代表输出来自 Y 空间。x表示输入生成器 G 从 X 空间取出的样本，G(x)表示生成器 G 生成的图像，$D_Y(y)$表示判别器D_Y判断 y 是否 Y 空间中取出的样本的概率，这个值越接近 1 越好。G 试图生成与来自域 Y 的图像相似的图像G(x)，而判别器D_Y的目的是区分翻译后的样本G(x)和真实的样本 y。G 的目标是最小化这个目标，以对抗一个试图最大化它的对手 D，即$min_G max_{D_Y} L_{GAN}(G,D_Y,X,Y)$。同理得 Y→X$min_F max_{D_X} L_{GAN}(G,D_X,Y,X)$。

3.2.2 循环一致损失函数

在与判别器进行二元极大极小博弈后，达到纳什均衡点，生成器能够生成更接近 Y 空间的样本，然而，因为没有匹配的数据集，去扰动的图像可能不会保留原始对抗样本的内容信息。基于风格迁移中的 CycleGAN，本文提出循环一致损失，防止 G 过分学习 Y 空间的样本而过度改变 X 空间的样本。

$$L_{cyc}(G,F)=E_{x\sim p_{data}(x)}[||F(G(x))-x||_1]+E_{y\sim p_{data}(y)}[||G(F(y))-y||_1] \quad (6)$$

3.2.3 身份损失函数

Loss_identity 的作用是为了防止输入与输出之间的色彩构成过多。通过映射以保留输入和输出之间的颜色组成是有帮助的。

$$L_{identity}(G,F)=E_{y\sim p_{data}(y)}[||G(y)-y||_1]+E_{x\sim p_{data}(x)}[||F(x)-x||_1] \quad (7)$$

4 实验与分析

4.1 实验环境

主要用到的实验硬件和软件设置如下：处理器为 Intel Xeon（R）W-2245；显卡为 NVIDIA GeForce RTX 3090；操作系统为 Windows 10；软件平台为 PyCharm、Python 3.7、MATLAB；深度学习框架为 Pytorch。

4.2 数据集

(1) CIFAR10 数据集的由 10 个类的共计 60000 张彩色图像组成，图像大小均为 32×32 像素，其中训练样本有 50000 个，测试样本有 10000 个。(2) ImageNet-NeurIPS [25]是 NeurIPS Challenge 数据集中专门用来测试防御模型鲁棒性的对抗攻击数据，是随机挑选 ImageNet 各类别样本通过攻击算法生成，ImageNet 数据集是目前世界上图像识别最大的数据库，提供超过 1000000 个样本。

4.3 实验参数设置

超参数设置：batchSize=16，lr=0.0002，decay_epoch=3, Adam 优化器一阶矩估计的指数衰减率为 0.5，二阶矩估计的指数衰减率为 0.999。

根据各个损失函数对训练结果的影响，为保证训练生成器完美的学习到训练，以使得训练得到的判别

器对于原始样本和对抗样本具有较强的分类。将损失函数的权重设置为 1:1:10:10。

4.4 实验结果分析

本文防御模型为保证通用性，随机选取 25k 张 COCO 数据集进行训练，这样经过训练的防御模型可以抵御多种类型对抗样本的攻击，对于不同的网络不需要再额外的训练可以直接切换到别的目标网络进行防御，有较高的通用性。我们对抗训练的预算扰动给定为$\epsilon = 16/255$，SSP 攻击算法会在预定的范围内寻找最优解来生成样本，此前提下对抗样本作为我们防御模型的训练集。

在训练 GAN 防御模型时为了确保模型训练充分，本文将鉴别器的训练损失函数可视化进行分析如图 5 所示，随着迭代次数的增加损失函数在后半段开始快速的收敛，知道稳定表明判别器训练完成，意味着得到一个判别器对对样本的真假判定及防御对抗样本的防御模型。

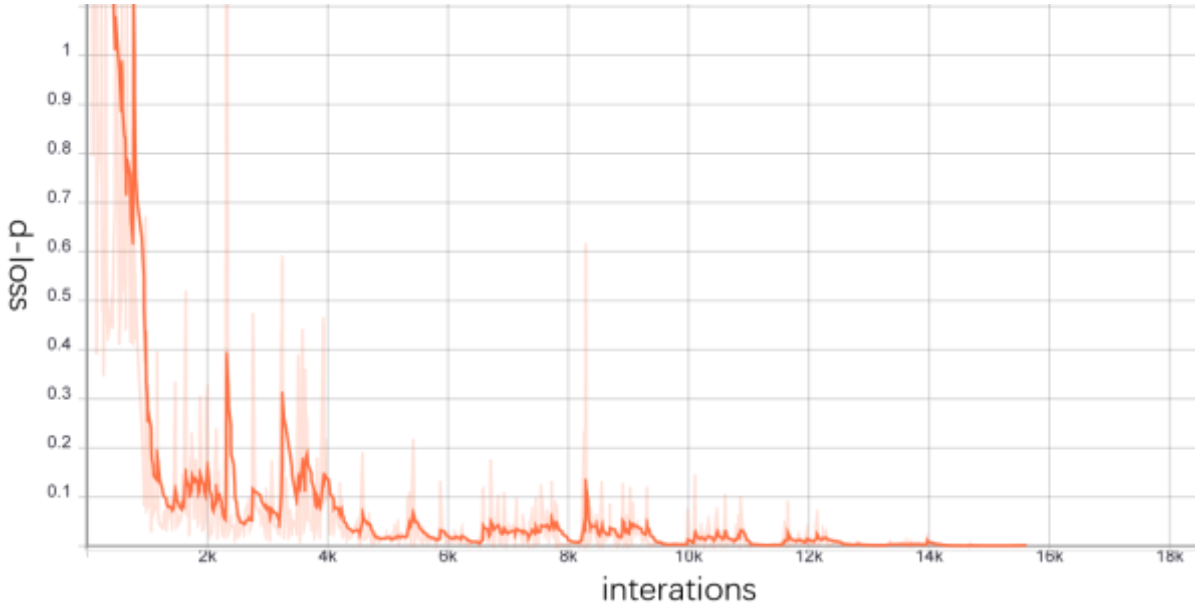

图 5 判别器损失函数随迭代次数变化

为分析不同情况下 CYCLEG-DefGAN 对于 DeepFool、PGD、SM、SSP 等各类白盒黑盒攻击方法产生的对抗样本的防御准确率，在测试方面本实验分别使用了 CIFAR10 和 ImageNet-NeurIPS 公共数据集进行测试验证，并与其他主流方法 NRP, Adv.train 进行对比分析。

表 1 ImageNet-NeurIPS 数据集中 CYCLEG-DefGAN 与各防御模型对各类对抗样本的防御准确率

目标网络 F	**Clean**		**DF**		**SM**		**PGD**		**SSP**	
准确率	**Rank-1**	**Rank-5**	**Rank-1**	**Rank-5**	**Rank-1**	**Rank-5**	**Rank-1**	**Rank-5**	**Rank-1**	**Rank-5**
无防御模型	75.4	92.45	10.45	88.55	12.25	76.6	14.00	85.75	25.30	42.15
NRP [15]	72.00	90.25	70.25	93.10	54.80	74.70	71.40	92.45	76.80	93.50
Our	70.89	90.10	72.65	93.55	56.05	75.50	73.15	93.95	78.20	93.65

表 2 CIFAR-10 数据集中 CYCLEG-DefGAN 与各防御模型对各类对抗样本防御准确率

目标网络 V	**Clean**		**DF**		**SM**		**PGD**		**SSP**	
准确率	**Rank-1**	**Rank-5**	**Rank-1**	**Rank-5**	**Rank-1**	**Rank-5**	**Rank-1**	**Rank-5**	**Rank-1**	**Rank-5**
无防御模型	92.58	99.72	0.18	89.44	0.00	99.72	0.08	99.72	49.33	89.43
NRP [15]	90.08	99.51	54.08	97.83	71.98	99.39	82.59	99.45	82.49	98.78
Adv.train [14]	90.09	99.57	77.86	99.22	85.18	99.50	86.14	99.51	83.70	99.06
Our	83.84	99.12	78.55	98.89	85.39	99.02	84.54	99.09	84.63	99.11

4.5 实验结果分析

表 1 描述了本文方法与其他防御方法分类结果的对比，在同一 ImageNet 测试集中对同一目标网络 FResenet152，攻击算法的扰动阈值一致的情况下进行攻击，在没有防御网络时目标模型基本没有办法正确分类数据。实验结果可得

(1) 在对于用白盒 DeepFool 和 PGD 及黑盒 SSP 产生的对抗样本进行测试时 CYCLEG-DefGAN 相比 NRP 防御模型有更好分类准确率，准确率提升了百分之 1.65 到百分之 2.45。

(2) 对于 SM 产生的对抗样本防御能力上本文的模型虽然比 NRP 效果要好，可能是因为特征攻击算法随着网络深度的增加不断提取特征而受到影响，因此防御模型准确率还需要进一步提升。综上，在 ImageNet 数据集上，在这几种防御机制比较下，其中 CYCLEG-DefGAN 对各对抗攻击的防御效果表现更好。

从表 2 可得，在 CIFAR10 数据集上，

(1) 在面对 DF 和 SM 以及 SSP 攻击算法产生的攻击防御模型时准确率有提升，CYCLEG-DefGAN 都优于其他主流算法。

(2) 但面对由 PGD 生成的对抗样本时准确率略低于 Adv.Train 百分之 1.6，这可能是因为本文模型基于生成对抗网络进行防御，生成对抗网络的优势在生成图像的方面，相比之下还稍显不足。

如图 6 的图片所示，将本文的对抗样本防御已训练好的模型其输出结果和主流方法结果输出如下，本文防御模型重建的图片过滤噪声的效果在同等参数下比 NRP 的效果要好，生成的图像质量要明显优于 NRP 生成的图像，相比之下 NRP 生成图像模糊，具有较多的噪声残留，而本文生成的图像甚至比原图更为明亮清晰，残留的噪声更少。

图 6 防御模型效果图

5 结论

针对深度学习模型受到的对抗攻击问题，本文将对抗生成网络和 SSP 攻击算法结合，从修改网络结构的角度思考出了一种防御模型 CYCLEG-DefGAN。本文所提出模型使用 TernausNet 网络作为生成器，在判别器端加入自注意力机制的循环对抗生成网络，充分提取图像特征，增强图像关键特征在分类任务中的贡献度，模型可以对对抗样本正确分类，提升了对抗样本的防御成功率。

CYCLEG-DefGAN 通过在 CIFAR10 以及 ImageNet 数据集上多种攻击算法测试验证了其有效性，取得了更好的实验结果。然而，该模型仍有不足之处，例如当扰动阈值=0 时模型会影响深度神经网络的分类能力。在接下来的工作中，进一步探究如何减少模型对原始样本准确率的影响。

参考文献

[1] CHAKRABORTY A, ALAM M, DEY V, et al. Adversarial attacks and defences: A survey [J]. 2018.

[2] 张思思, 左信, 计算机学报 刘 J. 深度学习中的对抗样本问题 [J]. 2019, 42 (8): 1886-904.

[3] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks [J]. 2013.

[4] GOODFELLOW I J, SHLENS J, SZEGEDY C J A P A. Explaining and harnessing adversarial examples [J]. 2014.

[5] GUO C, RANA M, CISSE M, et al. Countering adversarial images using input transformations [J]. 2017.

[6] HE Z, RAKIN A S, FAN D. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack; proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2019 [C].

[7] LIAO F, LIANG M, DONG Y, et al. Defense against adversarial attacks using high-level representation guided denoiser; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2018 [C].

[8] XIE C, WANG J, ZHANG Z, et al. Mitigating adversarial effects through randomization [J]. 2017.

[9] MIRZA M, OSINDERO S J A P A. Conditional generative adversarial nets [J]. 2014.

[10] SHEN S, JIN G, GAO K, et al. Ape-gan: Adversarial perturbation elimination with gan [J]. 2017.

[11] SAMANGOUEI P, KABKAB M, CHELLAPPA R J A P A. Defense-gan: Protecting classifiers against adversarial attacks using generative models [J]. 2018.

[12] MUSTAFA A, KHAN S H, HAYAT M, et al. Image super-resolution as a defense against adversarial attacks [J]. 2019, 29: 1711-24.

[13] MOOSAVI-DEZFOOLI S-M, FAWZI A, FROSSARD P. Deepfool: a simple and accurate method to fool deep neural networks; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2016 [C].

[14] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks [J]. 2017.

[15] NASEER M, KHAN S, HAYAT M, et al. A self-supervised approach for adversarial robustness; proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2020 [C].

[16] TRAMèR F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: Attacks and defenses [J]. 2017.

[17] KABILAN V M, MORRIS B, NGUYEN H-P, et al. Vectordefense: Vectorization as a defense to adversarial examples [M]. Soft Computing for Biomedical Applications and Related Topics. Springer. 2021: 19-35.

[18] DAS N, SHANBHOGUE M, CHEN S-T, et al. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression [J]. 2017.

[19] HINTON G, VINYALS O, DEAN J J A P A. Distilling the knowledge in a neural network [J]. 2015, 2 (7).

[20] LIN Y K, WANG C F, CHANG C-Y, et al. An efficient framework for counting pedestrians crossing a line using low-cost devices: the benefits of distilling the knowledge in a neural network [J]. 2021, 80 (3): 4037-51.

[21] LEE H, HAN S, LEE J J A P A. Generative adversarial trainer: Defense to adversarial perturbations with gan [J]. 2017.

[22] ZHOU J, LIANG C, CHEN J. Manifold projection for adversarial defense on face recognition; proceedings of the European Conference on Computer Vision, F, 2020 [C]. Springer.

[23] ZHU J, SHI L, YAN J, et al. Automix: Mixup networks for sample interpolation via cooperative barycenter learning; proceedings of the European Conference on Computer Vision, F, 2020 [C]. Springer.

[24] SIMONYAN K, ZISSERMAN A J A P A. Very deep convolutional networks for large-scale image recognition [J]. 2014.

[25] NeurIPS Challenge [Z]. Kaggle. 2017.