基于生成对抗网络的医学诊断模型知识蒸馏对抗攻击方法

王小银[1]，吕硕[1]，孙家泽[1]，杨宜康[2]
（1.西安邮电大学计算机学院，710121，西安；2.西安交通大学自动化科学与工程学院，710121，西安）

摘要：为了提高医学诊断模型防御攻击的能力，提出了一种基于生成对抗网络的医学诊断模型知识蒸馏对抗攻击方法。首先创建医学对抗攻击端到端训练网络，并以残差网络作为对抗网络架构；其次在生成器特征块中融合扩张卷积块和通道注意力机制，采用马尔可夫判别器改进判别器网络结构；最后利用生成器和判别器组建生成对抗网络，使用对抗样本进行知识蒸馏对抗攻击，以训练医学诊断模型提高识别精度。采用对抗样本对所提对抗方法进行攻击验证，结果表明：本文方法对抗攻击的成功率为 92.6%，与所对比的主流方法相比，该方法的成功率提高了 20%，生成对抗样本的最大平均差异降低了 3.68%，峰值信噪比、结构相似性分别提升了 5.07%、20.29%。本文方法解决了医学诊断模型在对抗攻击中难以获取网络结构和参数信息的问题，生成的对抗样本更接近真实样本，网络效果更佳，为辅助医疗模型诊断及模型安全性提供了参考方案。

关键词：医学诊断模型；知识蒸馏；对抗攻击；生成对抗网络
中图分类号：TP391.41 **文献标志码**：A
DOI：10.7652/xjtuxb202207009 **文章编号**：0253-987X(2022)07-0076-10

OSID 码

Knowledge Distillation Adversarial Attack Method Based on Generative Adversarial Network for Medical Diagnosis Model

WANG Xiaoyin[1], LÜ Shuo[1], SUN Jiaze[1], YANG Yikang[2]
(1. School of Computer Science, Xi'an University of Posts and Telecommunications, Xi'an 710121, China;
2. School of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: To improve the ability of a medical diagnosis model to defend against attacks, a knowledge distillation adversarial attack method based on generative adversarial network for medical diagnosis model is proposed. Firstly, a medical adversarial attack end-to-end training network is created, and the residual network is used as the adversarial network architecture. Secondly, the dilated convolution block and the channel attention mechanism are integrated into the generator feature block, and the Markov discriminator is used to improve the discriminator network structure. Finally, the generator and the discriminator are used to form a generative adversarial network, and the adversarial samples are used to conduct knowledge distillation adversarial attacks to train the medical diagnosis model to improve the recognition accuracy. Adversarial samples are used to verify the attack of the proposed adversarial method. The results show that the success rate of this method is 92.6%, an increase by 20% compared with the mainstream method for reference. The maximum mean discrepancy decreased by 3.68%, and the

收稿日期：2021-11-18。 作者简介：王小银(1976—)，女，副教授，硕士生导师。 基金项目：陕西省重点研发计划资助项目(2020GY-010)。
网络出版时间：2022-04-07 网络出版地址：http://kns.cnki.net/kcms/detail/61.1069.T.20220402.1447.002.html

peak signal to noise ratio and structural similarity increased by 5.07% and 20.29%. The method solves the problem that the medical diagnosis model is difficult to obtain the network structure and parameter information in the adversarial attack. The generated adversarial samples are closer to the real samples, and the network effect is better. It provides a reference scheme for auxiliary medical model diagnosis and model security.

Keywords: medical diagnosis model; knowledge distillation; adversarial attack; generative adversarial network

医学影像对于医疗工作人员判别患者病情的发展具有重要意义，当前临床诊断中主要依赖医疗工作者的技术能力及医疗经验。随着全球新冠疫情的大流行和当前疾病的复杂程度，通过人工对医学影像进行判断面临着巨大的压力和挑战。伴随着医学影像技术和人工智能(artificial intelligence，AI)技术的发展，越来越多的研究人员把深度学习技术应用在大规模医学影像诊断任务中，AI 医学影像诊断模型已成为医学影像分析和病情研判的重要支撑工具。深度学习中用到的最主要技术是深度神经网络(deep neural network，DNN)，医疗工作人员通过借助 DNN 诊断模型，可以增强对疾病的识别精度和提高工作效率。

尽管 DNN 具有很高的性能，但其在疾病诊断中的实际应用存在诸多安全威胁。在训练饱和的医学诊断模型中添加对抗样本进行对抗攻击后，将会使 AI 系统检测异常。利用深度学习进行诊断、决策或医疗保险赔偿的系统都可以被恶意操纵，进而导致医疗模型产生致命性的错误[1]。由此构建的 AI 医疗诊断模型会存在难以快速筛查病情、难以保障诊断精准度、模型安全性差等问题[2]。

研究人员已经提出了许多不同的方法来进行对抗攻击。攻击大致可以分为 3 类，基于梯度的迭代攻击、基于优化的迭代攻击和基于生成对抗网络(generative adversarial networks，GAN)的对抗攻击方法。基于梯度的迭代攻击按照深度学习模型的梯度变化最大的方向添加图像扰动。比如快速梯度符号法(fast gradient sign method，FGSM)[3]：通过在梯度上添加增量来使模型对样本做出错误分类。迭代梯度攻击法(project gradient descent，PGD)[4]：在 FGSM 算法基础上，在梯度迭代过程中进行多次迭代，控制扰动在规定的范围。基于优化的迭代攻击，主要思想是在迭代训练过程中将网络参数固定下来，把对抗样本当做唯一需要训练的参数，通过反向传递过程调整对抗样本。比如基于目标函数置信度攻击(Carlini&Wagner，C&W)[5]：使用约束范数进行优化对抗攻击过程。基于深度学习对抗扰动的攻击方法(Jacobian-based saliency map attack，JSMA)[6]：在对抗攻击时通过前向导数来调整干扰值进而生成对抗样本。GAN 主要使用生成器和判别器结构对样本输出进行优化。比如生成对抗样本网络(AdvGAN)[7]：核心思想是将干净样本通过 GAN 的生成器映射成对抗扰动，然后加在对应的干净样本中，判别器负责判别输入的样本是否为对抗样本。马尔可夫判别器(Patch-GAN)[8]：判别器最后的输出不是一个标量值，而是一个 $N\times N$ 的二维矩阵，矩阵中的元素表示输入图像的局部区域，这样训练使模型更加关注图像细节。通道注意力对抗网络(SE-GAN)[9]：一种利用多通道模型及语义辅助 GAN 生成同一场景在不同视角下的图像，效果超过了大部分的模型。上述攻击算法大多数是针对测试过程或测试数据集的，且需要一直对模型的体系结构和参数进行白盒访问。大部分黑盒攻击方法都是基于对抗样本的可迁移性，成功率都不高。传统生成医学对抗样本时无法真实地表征图像中特征信息的全局关系，且医学图像的 3 通道的像素点比较多，所以如基于梯度的迭代攻击和快速梯度符号法进行简单的编码解码结构无法提取更深层次的特征。在生成医学图像对抗样本时判别器的判别指标不够全面，不适用于精度要求高的医学图像。

综上所述，所提对抗攻击方法采用 GAN 网络构建医学诊断模型的知识蒸馏对抗攻击方法，对抗网络采用生成器和判别器结构，实现更加有效的高黑盒攻击成功率。与最新的基于 GAN 的对抗攻击方法进行攻击效果比较，实验结果表明所提方法的生成网络能够在较短的迭代中找到最优解，生成结果更精准，图像细节特征更优。

1　医学模型的对抗攻击

1.1　生成对抗网络

生成对抗网络 GAN 是一种无监督学习算法框架[10]，其算法思想受启发于二人零和博弈理论。

GAN 的网络结构主要由一个可以生成对抗样本的生成器 G 和一个可以判别对抗样本真实性的判别器 D 两部分组成，模型结构如图 1 所示。

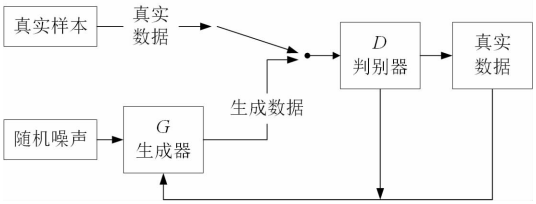

图 1 GAN 网络架构图

Fig. 1 The network architecture diagram of GAN

结合医学图像的特殊性，使用 GAN 生成的对抗样本进行对抗攻击，可以很好解决黑盒对抗攻击时的对抗样本可迁移问题，并且提高了对抗攻击的收敛速度，对抗攻击的有效性也优于使用传统对抗攻击的方法，随着大规模医疗诊断模型的落地，GAN 在 AI 医疗安全领域的使用会更加的广泛[11]。

1.2 知识蒸馏对抗攻击

模型蒸馏是模型压缩和加速的技术之一，Ba 等[12]提出知识蒸馏(knowledge distillation，KD)的概念，该模型能够模拟原始模型的输出，且具有较好的性能。Hinton 等[13]实现了模型蒸馏的流程，证实了蒸馏模型能够达到原始模型的泛化能力。Romero 等[14]提出了知识蒸馏网络模型(Hints Thin Nets，FitNet)方法，该方法得到的蒸馏网络性能足以媲美原始网络且参数量和计算开支显著降低。

模型蒸馏技术可以通过对目标模型进行迁移学习，利用模型压缩技术复制目标模型的网络结构，从而对蒸馏模型进行利用和迁移[15]。在对抗攻击时使用模型蒸馏技术替代黑盒攻击方式，能够对无法攻击的未知网络模型实现高可靠攻击，并增加对攻击对象的迁移性与通用性，可以对未知的医学诊断模型进行参数学习和模型复制，进而对高精度医学诊断模型实现有效对抗攻击。

本文方法使用知识蒸馏模型对医学诊断模型进行对抗攻击，运用模型蒸馏技术获取诊断模型父模型的本地复制作为子模型，模拟父模型的输出和决策分布。蒸馏模型构造方式如图 2 所示。

主要通过对子模型和父模型的输出结果计算交叉熵损失函数，最小化此函数。通过反向传播修改子模型的参数，使得子模型的输出逐渐逼近父模型，两个模型关键层的权值向量能够逐渐吻合，模型能够学习到模型的决策能力。使用对抗攻击方法针对子模型进行对抗训练，优化决策输出函数从而达到攻击效果最佳。将训练饱和的对抗攻击模型对父模型进行对抗攻击，根据输出判别对抗效果，能够提高对抗攻击的迁移性和通用性[16]。

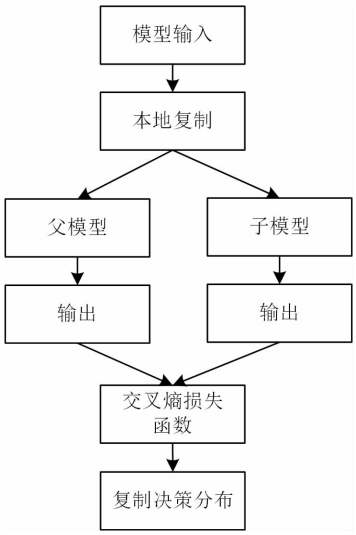

图 2 蒸馏模型构造方式

Fig. 2 Distillation model construction method

2 医学模型生成式网络对抗攻击

本文提出基于 GAN 的医学诊断模型知识蒸馏对抗攻击方法 AmdGAN(Adversarial Medical GAN)，使用知识蒸馏模型对目标模型进行复制，构建生成对抗网络。残差神经网络(ResNet)架构作为对抗攻击的生成器模型[17]，使用带有扩张卷积的残差块结构[18]，引入通道注意力机制(SENet)对生成器进行修改，使用马尔可夫判别器[19]网络作为判别器。整体对抗攻击流程图如图 3 所示。

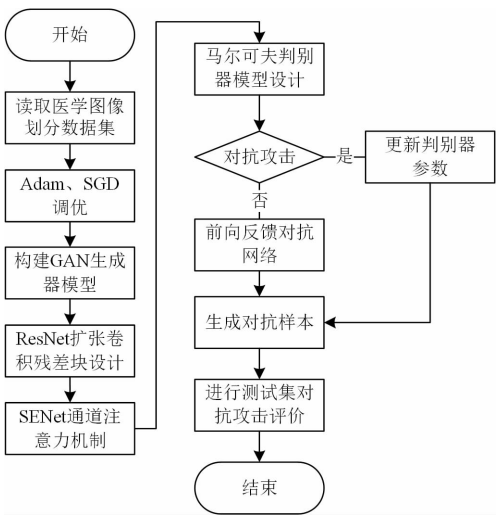

图 3 对抗攻击流程图

Fig. 3 The flow chart of adversarial attack

2.1　知识蒸馏对抗攻击架构

构建基于 GAN 的医学诊断知识蒸馏对抗攻击模型，包括生成器 G 和判别器 D，对抗攻击整体架构如图 4 所示。

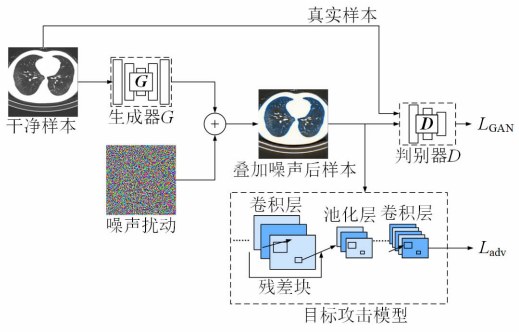

图 4　对抗攻击整体架构图

Fig. 4　The overall architecture diagram of adversarial attack

构造扰动生成器 G，有目标攻击下输入目标类别，输出为该样本所对应的扰动，可对目标模型进行半白盒和黑盒攻击。生成器 G 通过在原始医学影像实例 x 叠加干扰噪声后生成对抗样本 $x+G(x)$，判别器 D 主要通过优化损失函数进而引导 G 的训练过程，并且通过最小化损失值从而保证对抗样本的真实性，同时在扰动系数的限制下尽可能减小噪声。攻击目标模型 f，向 f 输入 $x+G(x)$，并输出损失，该损失在定向攻击时表示预测结果与目标结果间的距离，在非定向攻击时表示与真实类的距离，其中 GAN 的损失为

$$L_{GAN}=E_x\log D(x)+E_x\log\{1-D[x+G(x)]\} \tag{1}$$

式中：E_x 为生成器和判别器的期望值；判别器 D 的目的是将被扰动的数据 $x+G(x)$ 与原始数据 x 进行区分，以确保生成的对抗样本与真实图像的数据接近。

使用知识蒸馏模进行对抗攻击，主要是使子模型拟合父模型的权值向量，较少地依赖对抗样本的转移性。使用模型蒸馏技术获取目标模型的本地复制，使用与白盒攻击相同的攻击方式对蒸馏模型进行攻击。根据医学诊断场景构建目标模型父模型 b 的蒸馏子模型 f，对于模型 b 和模型 f 来说，要使得模型蒸馏过程中的目标函数最小化，即

$$V=\arg\min_x E_x H[f(x),b(x)] \tag{2}$$

式中：$f(x)$ 和 $b(x)$ 分别代表同一张图像在蒸馏模型和黑盒模型的输出结果；H 代表交叉熵损失。

通过最小化蒸馏过程目标函数，使得子模型 f 的输出结果逐渐逼近父模型 b，两个模型的某些关键权值向量能够逐渐吻合，子模型能够学习到父模型的泛化能力。得到本地复制子模型 f 之后，可以使用白盒攻击的方式对模型进行攻击。相较于传统利用对抗样本转移性的黑盒攻击方式，蒸馏模型能够显著提高黑盒攻击方式下的攻击成功率。

使用白盒攻击方法对知识蒸馏后的目标模型 f 进行对抗攻击，以生成具有最佳对抗攻击效果的对抗样本，最小化其损失值 L_{adv} 可以使对抗样本在攻击过程中的结果更接近于期望值。L_{adv} 的定义如下

$$L_{adv}=E_x l_f[x+G(x),t] \tag{3}$$

式中：l_f 表示目标模型 f 的损失函数；在有目标攻击中，t 为攻击目标。

2.2　生成器网络结构

生成器采用 SENet 通道注意力机制，使用 SENet 通道注意力机制对各通道的依赖性进行建模以提高网络的适应能力。注意力机制包括 3 个部分：挤压、激励和注意。

2.2.1　挤压操作流程

将全局空间信息压缩到通道中，利用图像特征的局部接受域，使用全局平均池生成通道特征信息。各通道的全局空间特征作为该通道的公式为

$$z=F(u)=\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}u(i,j) \tag{4}$$

式中：u 为局部接受域的信道描述符；i 和 j 表示通道位；z 由 u 通过 HW 收敛而产生，这些局部描述符的统计信息可以表达整个图像的特征信息。

2.2.2　激励操作流程

对各通道的聚合信息及依赖关系进行捕获，计算各通道的关联程度，激励函数为

$$s=F(z,W)=\sigma[g(z,W)]=\sigma[W_2\delta(W_1 z)] \tag{5}$$

式中：δ 指 ReLU 函数；W_1、W_2 属于通道描述符。

2.2.3　注意操作流程

将体系结构中的残差模块更改为 SENet 注意力初始网络后，与构造好的 ResNet 残差网络进行结合，从而使网络能够提高其对信息特征的敏感性，通过挤压和激励重新校准滤波器响应。生成器 SE 注意力机制网络结构如图 5 所示。

通过 SENet 对通道间的依赖关系进行建模，可以自适应的调整各通道的特征响应值[20]，将通道块添加到构造好的对抗攻击网络中，仅需要增加很小的计算消耗，就可以极大地提升网络性能。

2.2.4　扩张卷积网络结构

对生成器的特征通道块采用扩张卷积网络结

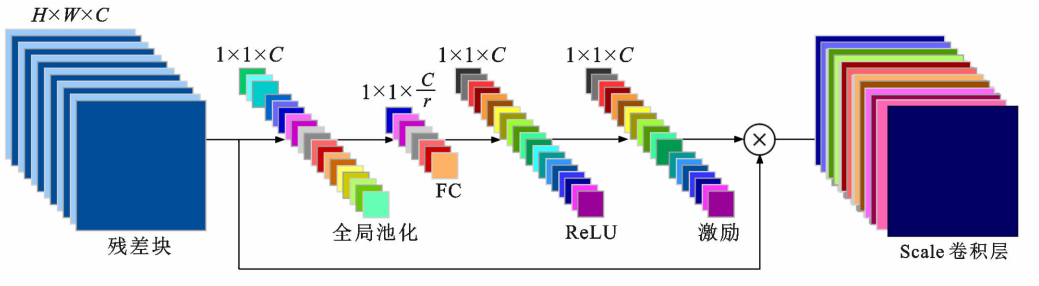

图 5 生成器 SE 注意力机制网络结构

Fig. 5 The SE attention mechanism network structure of generator

构，每层的第一组输入进行降采样，卷积滤波器对偶数行和偶数列进行评估。每一层的卷积层记为 g，其中每一层通过特征映射扩展到多个特征，每一层理想化的输出公式是

$$(g_i^l * f_i^l)(p) = \sum_{a+b=p} g_i^l(a) f_i^l(b) \tag{6}$$

式中：p 对应每一层的特征映射；g_i 表示组中的第 i 层；f_i 是对 g_i 进行特征过滤；a 是输入特征图；b 是输出特征图。

通过对图像进行上采样操作来提高图像的分辨率，改变 g^4、g^5 卷积层的卷积算子为两个空洞卷积

$$(g_1^4 * 2f_1^4)(p) = \sum_{a+2b=p} g_1^4(a) f_1^4(b) \tag{7}$$

原始的 ResNet 对输入图像每维下采样 32 倍，使用扩张卷积后对输入图像下采样 8 倍。感受野和原 ResNet 对应层一样，可以帮助目标模型对全局图像特征进行有效的识别，从而提高目标识别的准确率。

2.3 判别器网络架构

通过使用马尔可夫判别器增强对抗样本的图像纹理细节[21]，对抗网络判别器的网络结构图如图 6 所示。将医学影像分割为 $P_1, P_2, \cdots, P_n$ 切片，使用判别器对每个切片进行差别化判定作为判别结果，将 n 个判别结果进行加权平均输出更为精确的全局判别结果。

使用马尔可夫判别器修改原始 GAN 对抗网络的判别器，以单标量输出判别结果决定图像真实性。通过使用 L_1 损耗，生成器不仅可以欺骗判别器，还可以减小 L_1 与真实输入的距离，从而捕获图像纹理的局部连续性和图像中的普遍全局特征。对抗攻击损失的函数为

$$L_{\mathrm{GAN}}(D) = E_x \log D(x) + E_y \log[1 - D(G(\bar{x}))] \tag{8}$$

式中：L_{GAN} 主要由训练阶段引入的马尔可夫判别器网络 D 的损失函数计算；x 为输入真实图像的像素点；$\bar{x}$ 为损失像素点；y 为预测像素点。生成器和判别器在训练时要使 L_{GAN} 最小。

3 实验结果与评价

将本文方法 AmdGAN 与主流的基于 GAN 的对抗攻击方法在对医学影像诊断模型进行对抗攻击时的攻击效果进行比较，将 AmdGAN 与 Adv-

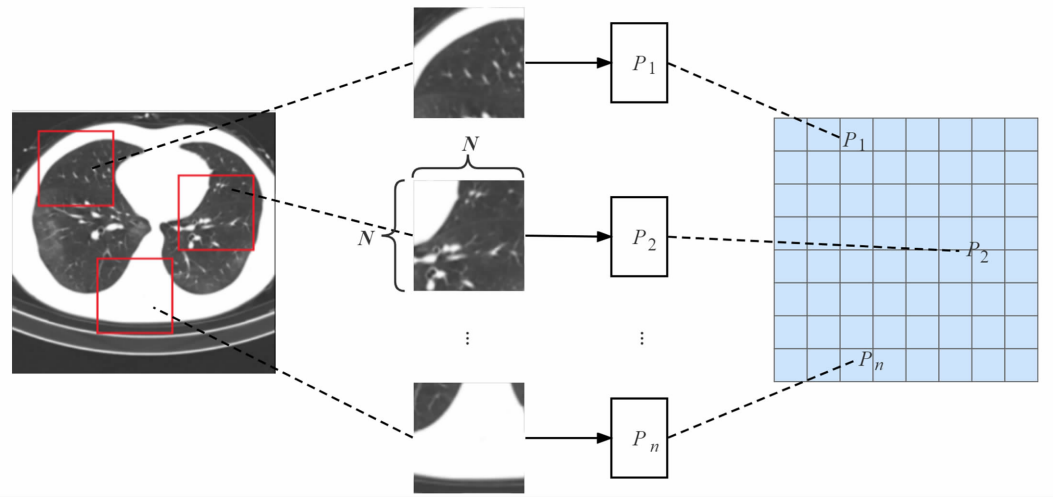

图 6 马尔可夫判别器的网络结构图

Fig. 6 The network structure diagram of Markov discriminator

GAN、Patch-GAN、SE-GAN 方法进行比较，通过对抗攻击成功率、对抗网络整体评价指标来对该方法的攻击效果进行评价。

3.1　数据集

使用新冠肺炎数据集（COVID-19 Radiography Database，CD）[22]、糖尿病视网膜病变图像数据集（DRIMDB，DB）[23]、皮肤癌数据集（Dermatoscopic，DC）[24]、医学领域数据集（MedMNIST，MT）[25]分别构建医学诊断模型，这些数据集公开可用，并且在医学深度学习领域具有代表性。

3.2　构建医学诊断目标模型

构建面向医学影像诊断的残差神经网络目标模型，使用 AmdGAN 在数据集上根据病人的医学影像构建医学影像数据集，医学影像数据集的训练集和测试集划分比例为 8∶2。

分别使用 VGG-16、ResNet-50、ResNet-101、ResNet-152 搭建迁移学习目标模型，构建残差单元，调节模型训练参数。在训练过程中选用 Adam＋SGD 梯度下降策略，Adam 用作快速下降算法，SGD 用作调优。模型训练时均使用同样的数据增强方法，训练目标模型直至达到收敛状态后保存目标模型。医学影像目标模型诊断正确率结果如表 1 所示。

根据实验结果，ResNet-152 在医学影像识别正确率和平均正确率较高，并且对于数据集的适应能力和识别稳定性高于 VGG-16、ResNet-50 和 ResNet-101。在下一阶段选择 ResNet-152 目标模型和适应性较好的 CD 数据集进行对抗攻击。

表 1　医学影像目标模型诊断正确率

Table 1　The diagnostic accuracy of medical imaging target model

目标模型	正确率/%				平均正确率/%			
	CD	DB	DC	MT	CD	DB	DC	MT
VGG-16	95.35	91.67	85.33	87.98	96.63	94.90	86.93	91.85
ResNet-50	96.66	89.72	81.62	79.58	97.80	95.12	89.71	86.03
ResNet-101	95.15	90.77	82.61	90.36	98.25	90.12	87.30	94.28
ResNet-152	96.10	87.37	84.75	91.18	98.43	91.02	88.64	95.48

3.3　医学模型对抗攻击

为验证知识蒸馏方法在黑盒对抗攻击场景下的效果，实验组使用知识蒸馏方法获取本地模型，对照组利用直接查询的方式对黑盒模型进行攻击。训练深度神经网络、卷积神经网络和一个对照的医学诊断目标模型网络，知识蒸馏黑盒对抗攻击方法采用前后的对抗攻击成功率如表 2 所示。使用知识蒸馏方法的对抗成功率比直接查询方式高 38.1%。

表 2　知识蒸馏对抗攻击对比

Table 2　Knowledge distillation adversarial attack comparison

组别	对抗成功率/%
对照组	43.17
实验组	81.23

通过攻击目标模型，并进行对抗攻击测试，迭代次数 Epoch 设置为 300，测试集数量为 600 张医学影像。通过控制数据集相同输入变量，将 AmdGAN 和单独使用 GAN 网络的 AdvGAN、使用 SENet 作为生成器模型的 SE-GAN、使用马尔可夫判别器作为判别器的 Patch-GAN 对抗攻击方法进行对抗攻击实验。图 7 为进行对抗攻击时的效果对比图。AmdGAN 在对抗攻击实验中，对抗攻击的成功率为 92.6%，而其他 3 种模型在 72%～78%之间，对抗攻击成功率上优于其他主流的基于 GAN 的对抗攻击策略，对抗攻击效果显著。

调整测试集从 400 张到 2 000 张数量之间分别对 AmdGAN 和另外 3 种攻击方法进行测试，表 3 为调整对抗攻击测试样本数量后的 4 种方法进行对抗后的诊断模型对抗攻击成功率。

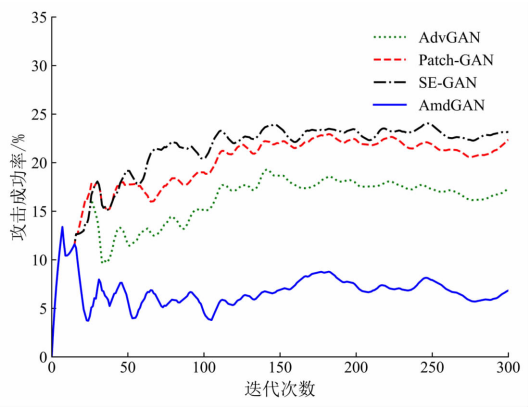

图 7　对抗攻击效果对比图

Fig. 7　The comparison diagram of adversarial attack

AmdGAN 在小样本和大样本数据集时对抗攻击成功率在 92.6%左右波动，而其他 3 种对抗攻击方式的对抗攻击成功率在 80%左右，且成功率依赖于样本数量。随着样本数量的增加，AmdGAN 的对抗攻击成功率比较稳定，攻击模型收敛速度更快，在小样本和大样本时对抗攻击成功率都很高。

表 3 在 9 种不同规模数据集对诊断模型的对抗攻击成功率

Table 3 Success rate of adversarial attacks against diagnostic models on 9 datasets of different scales

对抗攻击方法	对抗成功率/%								
	400 张	600 张	800 张	1 000 张	1 200 张	1 400 张	1 600 张	1 800 张	2 000 张
AdvGAN	77.71	76.96	80.06	78.39	80.51	82.68	83.95	84.62	84.51
Patch-GAN	81.65	78.63	79.76	80.91	81.11	80.63	80.31	80.44	80.57
SE-GAN	77.70	79.62	75.37	80.16	79.02	77.34	79.06	78.31	78.80
AmdGAN	92.67	91.97	93.71	92.95	92.26	93.02	92.33	92.48	92.54

3.4 对抗攻击网络模型评价

在对抗攻击领域，对抗生成的图像缺乏良好的定量评估指标，GAN 的评价指标要能够量化生成的样本和真实样本之间的差异性，并且可以评估生成的对抗样本图像的质量，保证生成样本的准确性和多样性。使用均方误差(mean square error，MSE)、最大平均差异(kernel maximum mean discrepancy，kernel MMD)、峰值信噪比(peak signal to noise ratio，PSNR)、结构相似性(structural similarity，SSIM)[26]来评估该方法对抗攻击的效果。

综合评价 AmdGAN 和其他 3 种方法的均方误差指数，图 8 为 4 种对抗方法的均方误差评价，由图可知该方法的均方误差较小，说明使用 AmdGAN 生成的对抗样本在对抗攻击时的攻击成功率更加可信，预测精确度也最高。

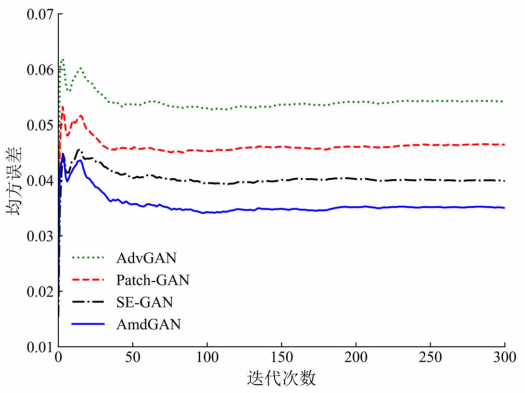

图 8 4 种对抗方法的均方误差评价

Fig. 8 MSE evaluation of four adversarial methods

4 种对抗方法的最大平均差异评价如图 9 所示。图像中的 4 种方法最大平均差异在 4.1～4.7 之间，AdvGAN 和 Patch-GAN 的变化趋势相似，最大平均差异过高即说明 GAN 网络结构不佳。AmdGAN 和 SE-GAN 的变化趋势相似并且最大平均差异差距很小，表明通过改变生成器的结构为 SE 通道注意力机制可以很好的提高 GAN 网络的性能，生成的对抗样本和原始图像的分布会更加接近，生成的图片的质量越好。

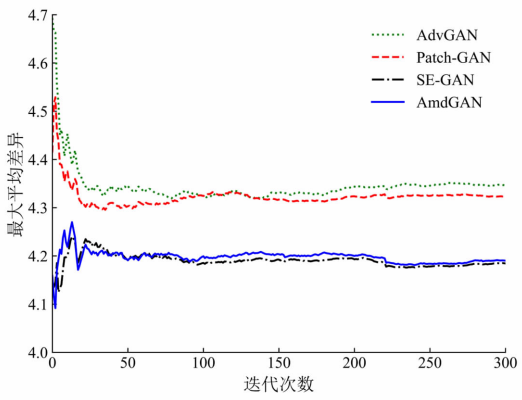

图 9 4 种对抗方法的最大平均差异评价

Fig. 9 Kernel MMD evaluation of four adversarial methods

综合评价 AmdGAN 和其他 3 种方法的峰值信噪比指数，4 种对抗方法的峰值信噪比评价如图 10 所示，其中 SE-GAN 和 AmdGAN 的峰值信噪比优于其他两种方法，说明通过修改生成器结构在图像保留真实度方面优于原始 GAN 网络和仅修改判别器结构的 Patch-GAN，AmdGAN 采用了扩张卷积和通道注意力，该方法的峰值信噪比收敛后在 71 dB 左右，优于其他 3 种方法，图像失真也最小，说明生成的对抗样本更加接近真实图像。

4 种对抗方法的结构相似性评价如图 11 所示，由图可知 AmdGAN 的结构相似性中位数较大、下四分位点和上四分位点也较大，其次是 SE-GAN，然后是 Patch-GAN，AdvGAN 的中位数最小。由基础数据的核密度分布可知该方法生成的对抗样本与真实图像的纹理和全局内容相匹配，生成的对抗样

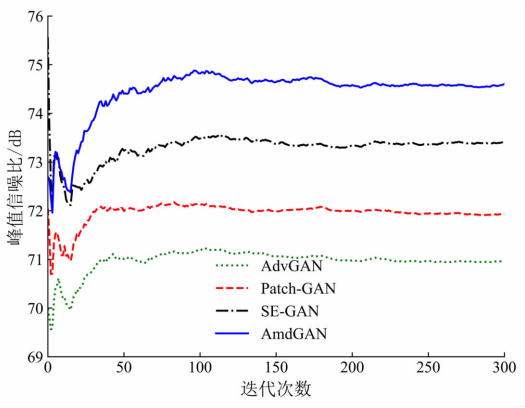

图 10　4 种对抗方法的峰值信噪比评价

Fig. 10　PSNR evaluation of four adversarial methods

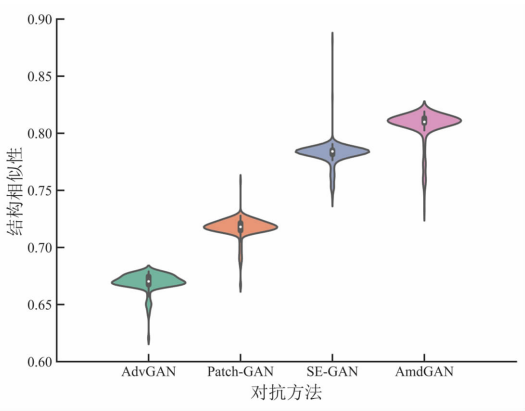

图 11　4 种对抗方法的结构相似性评价

Fig. 11　SSIM evaluation of four adversarial methods

本具有更清晰的细节，并且图像失真最小。

3.5　实验结果

实验表明，对于一个训练完美的医学影像诊断模型，通过添加微小的扰动，训练蒸馏网络模型，再采用 GAN 就可以生成对抗样本来误导检测结果，从而轻易攻击成功医学诊断模型。采用本文方法 AmdGAN 进行医学诊断模型的对抗攻击的对抗攻击成功率优于使用 AdvGAN、Patch-GAN、SE-GAN 进行对抗攻击，在生成的对抗样本真实度上也优于其他 3 种方法。4 种基于 GAN 的对抗攻击方式整体效果上 AmdGAN 最优，Patch-GAN 和 SE-GAN 的效果次之，使用基础 GAN 网络结构的 AdvGAN 最差，说明通过修改判别器网络为 Patch 架构可以很好地增强对抗攻击效果，通过使用通道注意力机制和扩张卷积的残差块生成器结构在提升对抗攻击成功率的同时也可以让生成的对抗样本真实度显著增加。

其他基于 GAN 的对抗攻击不能有效的攻击的原因可能是没有关注到高分辨率医学图像的细节特征，传统的基于 GAN 的判别器不能够对医学影像捕获核心特征信息，并且在模型迭代训练时迭代下降的比较慢。本文方法主要是从上述方面对 GAN 进行优化和改进，从而达到了较优的对抗性能。

4　结论与展望

本文提出一个基于 GAN 的医学诊断模型对抗攻击方法，经过良好训练的生成器可以基于有限的数据集训练后用于特定的医疗对抗攻击场景，对白盒模型和半白盒模型对抗攻击的同时，也可以对无法访问的黑盒模型进行知识蒸馏模型攻击。当应用 AmdGAN 来验证医学模型的安全性问题时，无需了解诊断模型的网络架构，并且在不知道防御是否存在的情况下生成对抗样本时，所生成的对抗样本能够保持较高的感知质量，并且在攻击成功率方面优于同类对抗攻击方法。

通过使用 AmdGAN 对医学诊断模型进行攻击找到模型的脆弱性，可以指导模型针对性的防御。AmdGAN 可以改善医疗诊断模型的识别结果，并且降低模型训练的成本，对于改善医疗诊断模型的安全问题以及促进医疗行业发展具有重要的指导意义。在下一步的研究中，医疗从业者以及机器学习研究人员应该共同致力于把对抗攻击算法融入到医疗诊断系统中，从而能够提高现阶段医疗诊断的安全性和防止医疗欺诈。

参考文献：

[1]　于振华，黄山阁，杨波，等. SLEIR 新冠肺炎传播动力学模型构建与预测 [J]. 西安交通大学学报，2022，56(5)：1-11.
YU Zhenhua，HUANG Shange，YANG Bo，et al. Construction and prediction of SLEIR COVID-19 transmission dynamics model [J]. Journal of Xi'an Jiaotong University，2022，56(5)：1-11.

[2]　HIRANO H，MINAGI A，TAKEMOTO K. Universal adversarial attacks on deep neural networks for medical image classification [J]. BMC Medical Imaging，2021，21(1)：9.

[3]　GOODFELLOW I J，SHLENS J，SZEGEDY C. Explaining and harnessing adversarial examples [C]// 3rd International Conference on Learning Representations (ICLR). San Diego，CA，USA：ICLR，2015：20193407343614.

[4]　MADRY A，MAKELOV A，SCHMIDT L，et al. Towards deep learning models resistant to adversarial attacks [C]// 6th International Conference on Learning

Representations (ICLR). Vancouver, BC, Canada: ICLR, 2018: 20193507364303.

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

[6] PAPERNOT N, MCDANIEL P, JHA S, et al. The limitations of deep learning in adversarial settings [C]// 2016 IEEE European Symposium on Security and Privacy (EuroS&P). Piscataway, NJ, USA: IEEE, 2016: 372-387.

[7] XIAO Chaowei, LI Bo, ZHU Junyan, et al. Generating adversarial examples with adversarial networks [C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence. New York, NY, USA: ACM, 2018: 3905-3911.

[8] HOU Le, SAMARAS D, KURC T M, et al. Patch-based convolutional neural network for whole slide tissue image classification [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2016: 2424-2433.

[9] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2018: 7132-7141.

[10] GOODFELLOW I J, POUGET A J, MIRZA M, et al. Generative adversarial networks [J]. Advances in Neural Information Processing Systems, 2014, 3: 2672-2680.

[11] 李策, 李兰, 宣树星, 等. 采用超复数小波生成对抗网络的活体人脸检测算法 [J]. 西安交通大学学报, 2021, 55(5): 113-122.
LI Ce, LI Lan, XUAN Shuxing, et al. Face anti-spoofing algorithm using generative adversarial networks with hypercomplex wavelet [J]. Journal of Xi'an Jiaotong University, 2021, 55(5): 113-122.

[12] BA J, CARUANA R. Do deep Nets really need to be deep? [J]. Advances in Neural Information Processing Systems, 2014, 27(8): 2654-2662.

[13] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network [J]. Computer Science, 2015, 14(7): 38-39.

[14] ROMERO A, BALLAS N, KAHOU S E, et al. Fitnets: hints for thin deep nets [C]// 3rd International Conference on Learning Representations (ICLR). San Diego, CA, USA: ICLR, 2015: 2019340734 3611.

[15] 侯静怡, 齐雅昀, 吴心筱, 等. 跨语言知识蒸馏的视频中文字幕生成 [J]. 计算机学报, 2021, 44(9): 1907-1921.
HOU Jingyi, QI Yayun, WU Xinxiao, et al. Cross-lingual knowledge distillation for Chinese video captioning [J]. Chinese Journal of Computers, 2021, 44(9): 1907-1921.

[16] 王曙燕, 金航, 孙家泽. GAN 图像对抗样本生成方法 [J]. 计算机科学与探索, 2021, 15(4): 702-711.
WANG Shuyan, JIN Hang, SUN Jiaze. Method for image adversarial samples generating based on GAN [J]. Journal of Frontiers of Computer Science & Technology, 2021, 15(4): 702-711.

[17] YU F, KOLTUN V, FUNKHOUSER T. Dilated residual networks [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2017: 636-644.

[18] 黄文涵, 殷国栋, 耿可可, 等. 基于扩张卷积特征自适应融合的复杂驾驶场景目标检测 [J]. 东南大学学报(自然科学版), 2021, 51(6): 1076-1083.
HUANG Wenhan, YIN Guodong, GENG Keke, et al. Object detection in complex driving scene based on dilated convolution feature adaptive fusion [J]. Journal of Southeast University (Natural Science Edition), 2021, 51(6): 1076-1083.

[19] ISOLA P, ZHU Junyan, ZHOU Tinghui, et al. Image-to-Image translation with conditional adversarial networks [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2017: 5967-5976.

[20] 陈法法, 成孟腾, 杨蕴鹏, 等. 融合双注意力机制和 U-Net 网络的锈蚀图像分割 [J]. 西安交通大学学报, 2021, 55(12): 119-128.
CHEN Fafa, CHENG Mengteng, YANG Yunpeng, et al. A segmentation method based on dual attention mechanism and U-Net for corrosion images [J]. Journal of Xi'an Jiaotong University, 2021, 55(12): 119-128.

[21] 张西宁, 雷威, 杨雨薇, 等. 采用自适应基因粒子群算法优化隐马尔科夫模型的方法及应用 [J]. 西安交通大学学报, 2018, 52(8): 1-8.
ZHANG Xining, LEI Wei, YANG Yuwei, et al. Adaptive genetic particle swarm algorithm for optimization hidden Markov models with applications [J]. Journal of Xi'an Jiaotong University, 2018, 52(8): 1-8.

[22] RAHMAN T, KHANDAKAR A, QIBLAWEY Y, et al. Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray ima-

ges [J]. Computers in Biology and Medicine, 2021, 132: 104319.

[23] GULSHAN V, PENG L, CORAM M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs [J]. JAMA, 2016, 316(22): 2402-2410.

[24] ESTEVA A, KUPREL B, NOVOA R A, et al. Dermatologist-level classification of skin cancer with deep neural networks [J]. Nature, 2017, 542(7639): 115-118.

[25] YANG Jiancheng, SHI Rui, NI Bingbing. MedMNIST classification decathlon: a lightweight AutoML benchmark for medical image analysis [C] // 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). Piscataway, NJ, USA: IEEE, 2021: 191-195.

[26] ARMANIOUS K, JIANG Chenming, FISCHER M, et al. MedGAN: medical image translation using GANs [J]. Computerized Medical Imaging and Graphics, 2020, 79: 101684.

（编辑　荆树蓉　刘杨）