ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 5,Sep. 2023,pp. 1104－1111
DOI:10. 16337/j. 1004-9037. 2023. 05. 009

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于多损失混合对抗函数和启发式投影算法的逼真医学图像增强方法

王　见，成楚凡，陈　芳

（南京航空航天大学计算机科学与技术学院，南京 211106）

摘　要：早期发现新冠肺炎可以及时医疗干预提高患者的存活率，而利用深度神经网络（Deep neural networks，DNN）对新冠肺炎进行检测，可以提高胸部CT对其筛查的敏感性和判读速度。然而，DNN在医学领域的应用受到有限样本和不可察觉的噪声扰动的影响。本文提出了一种多损失混合对抗方法来搜索含有可能欺骗网络的有效对抗样本，将这些对抗样本添加到训练数据中，以提高网络对意外噪声扰动的稳健性和泛化能力。特别是，本文方法不仅包含了风格、原图和细节损失在内的多损失功能从而将医学对抗样本制作成逼真的样式，而且使用启发式投影算法产生具有强聚集性和干扰性的噪声。这些样本被证明具有较强的抗去噪能力和攻击迁移性。在新冠肺炎数据集上的测试结果表明，基于该算法的对抗攻击增强后的网络诊断正确率提高了4.75%。因此，基于多损失混合和启发式投影算法的对抗攻击的增强网络能够提高模型的建模能力，并具有抗噪声扰动的能力。

关键词：医学图像增强；对抗性攻击；多损失混合；启发式投影法；攻击迁移性

中图分类号：TP391　　**文献标志码：**A

Realistic Medical Image Augmentation by Using Multi-loss Hybrid Adversarial Function and Heuristic Projection Algorithm

WANG Jian，CHENG Chufan，CHEN Fang

（College of Computer Science and Technology，Nanjing University of Aeronautics & Astronautics，Nanjing 211106，China）

Abstract：Early detection of COVID-19 allows medical intervention to improve the survival rate of patients. The use of deep neural networks（DNN）to detect COVID-19 can improve the sensitivity and speed of interpretation of chest CT for COVID-19 screening. However，applying DNN for the medical field is known to be influenced by the limited samples and imperceptible noise perturbations. In this paper，we propose a multi-loss hybrid adversarial function（MLAdv）to search the effective adversarial attack samples containing potential spoofing networks. These adversarial attack samples are then added to the training data to improve the robustness and the generalization of the network for unanticipated noise perturbations. Especially，MLAdv not only implements the multiple-loss function including style，origin，and detail losses to craft medical adversarial samples into realistic-looking styles，but also uses the heuristic projection algorithm to produce the noise with strong aggregation and interference. These samples are proven to have stronger anti-noise ability and attack transferability. By evaluating on COVID-19 dataset，it is shown that the augmented networks by using adversarial attacks from the MLAdv algorithm can improve

收稿日期：2022-05-10；**修订日期：**2023-05-30

the diagnosis accuracy by 4.75%. Therefore, the augmented network based on MLAdv adversarial attacks can improve the ability of models and is resistant to noise perturbations.

Key words: medical image augmentation; adversarial attack; multi-loss hybrid; heuristic projection algorithm; attack transferability

引　　言

机器学习方法,特别是深度神经网络(Deep neural networks,DNN)被广泛应用于医学诊断任务中[1-2]。例如,深度卷积神经网络(Convolutional neural networks, CNN)被用来对胸部X光图像进行分类,并诊断气胸和新冠肺炎[3]。然而,深度神经网络模型依赖于极大的数据集来避免过度拟合问题。因此,除了在原始图像数据集上实现较高的准确率外,新冠肺炎检测系统还有望检测到在训练中很少看到的新冠肺炎样本,并抵抗现实世界图像中的意外噪声。

遗憾的是,在医学图像分析的许多应用领域,由于昂贵的图像采集设备和有限的患者病例,很难收集到大量的数据。如图1所示,新冠肺炎数据集通常具有分类结果的长尾分布,新冠肺炎样本的早期数据未被充分泛化学习,并且非常快速的数据变化使得线性整流函数(Rectified linear unit,ReLU)网络趋于形成不封闭的决策边界,这使得网络面临被任意噪声激活的风险[4]。本文提出了对抗性增强方法来有效地生成增广对抗样本,从而提高网络的健壮性,使激活边界能够收敛到样本的原始边界。

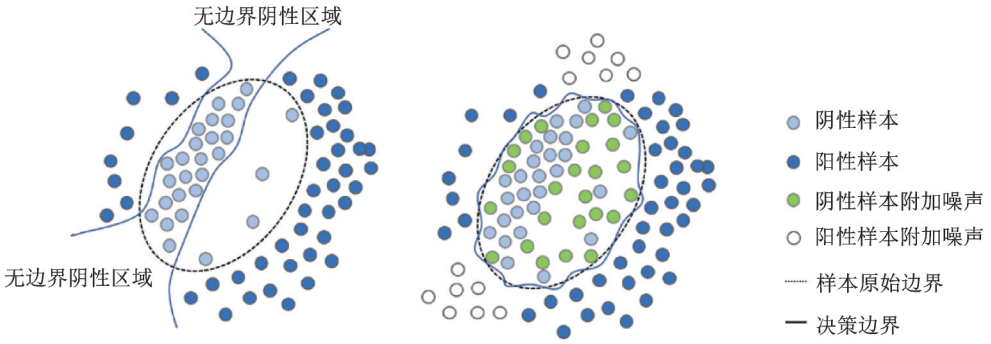

图1　预期的训练方案概念图

Fig.1　Concept map of expected training plan

常用的图像增强技术可分为传统的图像增强和基于机器学习的图像增强。传统的图像增强是通过几何变换或图像压缩等方法扩增原始数据集。文献[5]提出了在极坐标系中使用径向变换的图像增强,生成的图像包含以不同角度旋转的原始图像的多个副本[6]。基于机器学习的图像增强与传统的数据增强不同,基于学习的增强由数据驱动,可以从现实世界的底层数据分布中学习最优的图像变换。例如,开发一个两阶段生成性对抗网络(Generative adversarial network,GAN)来生成扩展的图像-掩码对的细胞核图像[7]。然而,在实践中,训练生成网络并不容易,因为它们对超参数调整很敏感[8],并且它可能会受到模式崩溃问题的影响[9]。

本文介绍了一种逼真的数据增强方法,它使用多损失混合对抗伪装,不求助于生成对抗网络。与现有方法相比,本文技术可以生成新的增强图像,这些图像对人类观察者来说似乎是合法的,并且不需要依赖大量数据。此外,还使用启发式投影算法对生成的噪声进行聚集,提高了噪声的不可去除性和攻击迁移性。

1 增强样本生成方法

1.1 多损失混合及投影方法

本文提出了一种多损失混合对抗伪装方法，以产生有效的对抗样本，其中包含了欺骗网络的噪声，最终对模型进行重新训练以提高模型泛化能力。

给定一个原始医学图像 $\boldsymbol{X}\in\mathbf{R}^{N\times H\times W}$（$N$、$H$ 和 W 分别为图像的通信数、高度和宽度，值的大小为0～255）的输出标签为 $\boldsymbol{Y}\in\mathbf{R}$，通过解决以下优化问题找到有效的对抗性样本 $\boldsymbol{X}_{\mathrm{a}}$。

$$\begin{cases}\text{minimize}\left\|\boldsymbol{X}-\boldsymbol{X}_{\mathrm{a}}\right\|_{p}+\mathcal{L}_{\text{hybrid}}(\boldsymbol{X},\boldsymbol{X}_{\mathrm{a}})\\ \text{s.t. } D(\boldsymbol{X},\boldsymbol{X}_{\mathrm{a}})\leqslant\varepsilon\end{cases}\tag{1}$$

式中：$\mathcal{L}_{\text{hybrid}}$ 为多损失函数，D 为原始样本与对抗样本之间的差异度量函数，本文使用 l_1 损失，ε 为阈值。利用这些对抗样本对模型进行再训练来提高模型的稳健性。

图2显示了本文的模型健壮性改进方法流程。用户定义干净的源图像，预期的目标样式。然后，本文提出的方法在所需的区域中生成具有所需样式的增强样本。首先通过特征提取对高维特征进行提取，并采用多损失方法对 $\mathcal{L}_{\text{hybrid}}$ 进行优化，生成扩展样本 $\boldsymbol{X}_{\mathrm{a}}$；然后引入基于启发式投影的扰动迭代算法来提高对抗样本的攻击迁移性。本文计算目标模型的梯度 $\nabla J(\boldsymbol{X}_{\mathrm{a}},Y)$，并制作梯度图，其中 $J(\boldsymbol{X}_{\mathrm{a}},Y)$ 为交叉熵损失。进一步在每次迭代的图像中加入一个小扰动 $\varepsilon\cdot\text{sign}(\nabla J(\boldsymbol{X}_{\mathrm{a}},Y))$，以实现扰动迭代。

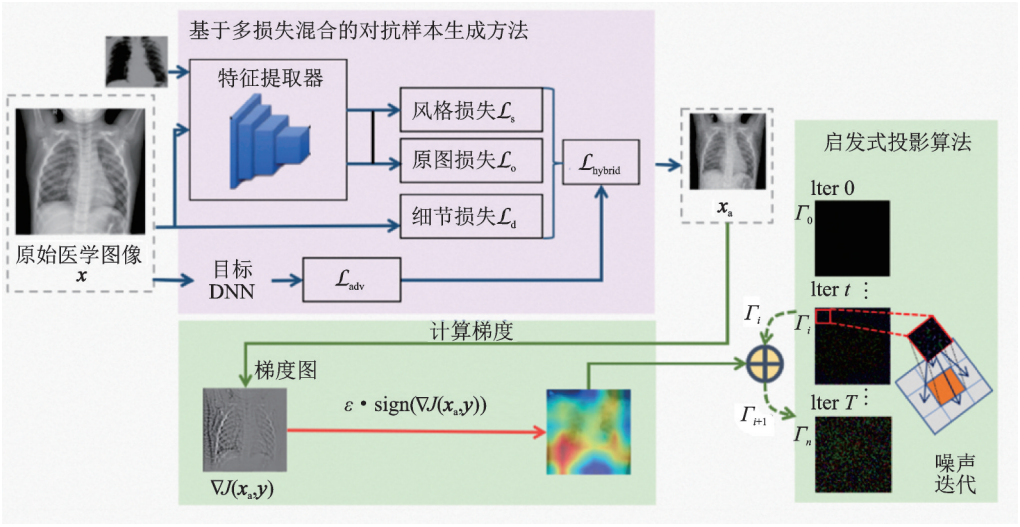

图2 本文方法流程

Fig.2 Flow of the proposed method

首先设计了一个关于风格、原图和细节的多损失混合函数 $\mathcal{L}_{\text{hybrid}}$，用于将医学增强样本制作成和逼近于原图的外观样式。最终的多损失函数是对抗性损失 $\mathcal{L}_{\text{adv}}$、用于样式生成的风格损失 $\mathcal{L}_{\mathrm{s}}$、用于保持源图像的原始特征的原图损失 $\mathcal{L}_{\mathrm{o}}$ 和用于确保增强样本在细节上平滑的细节损失 $\mathcal{L}_{\mathrm{d}}$ 的组合，定义为

$$\begin{cases}\mathcal{L}_{\text{hybrid}}=(\mathcal{L}_{\mathrm{s}}+\mathcal{L}_{\mathrm{o}}+\mathcal{L}_{\mathrm{d}})+\tau\cdot\mathcal{L}_{\text{adv}}\\ \text{s.t. } D(\boldsymbol{X},\boldsymbol{X}_{\mathrm{a}})\leqslant\varepsilon\end{cases}\tag{2}$$

式中：$\mathcal{L}$ 为超参数，本文设置为0.5。其次，进一步使用基于投影的启发式扰动迭代算法，将增强后样本中的噪声投影到周围区域，以保证强聚集和干扰。本文最大化目标模型的交叉熵损失 $J(\boldsymbol{X}_{\mathrm{a}}^{\mathrm{t}},Y)$，$\boldsymbol{X}_{\mathrm{a}}^{\mathrm{t}}$ 为

目标对抗样本,并用$f(x)$来表示DNN的预测标签,以确保$f(\boldsymbol{X}_{\mathrm{a}}^{\mathrm{t}})\neq Y$。具体操作为

$$\begin{cases} x'_{\mathrm{a}}=\mathrm{Clip}\left(x_{\mathrm{a}}+\varepsilon\cdot\mathrm{sign}\left(\nabla J\left(x_{\mathrm{a}},\ \mathrm{y}\right)\right)+\gamma\cdot\mathrm{sign}\left(W_{\mathrm{o}}*P\right)\right) \\ \text{s.t. } x_{\mathrm{a}},x'_{\mathrm{a}}\in[0,255] \end{cases} \tag{3}$$

式中:W_{o}定义为投影核,P定义为所产生的扰动。本文使用启发式投影算法将超过阈值ε的噪声投影到周围区域。因为超过ε阈值的像素具有更高的概率处于类激活图可视化区域的高亮区域[10],所以本文通过使用该投影方法自然地扩展了这些区域中的扰动聚集。

1.2　多损失混合方法的详细设计

1.2.1　风格损失

图像隐蔽性定义为$\|\boldsymbol{x}-\boldsymbol{x}_{\mathrm{a}}\|_p$,其中$\|\cdot\|_p$通常使用$L_p$范数,也通常使用$L_2$和$L_\infty$。对于本文提出的图像增强,风格相似度是由增强样本和样式参考图像$\boldsymbol{X}_{\mathrm{s}}$之间的样式度量来定义的。两个图像之间的样式距离可以通过它们在样式表示中的差异来定义,即

$$\mathcal{L}_{\mathrm{s}}=\sum_{l\in P_l}\left\|\mathcal{G}\left(\widetilde{D}_l(\boldsymbol{X}_{\mathrm{s}})\right)-\mathcal{G}\left(\widetilde{D}_l(\boldsymbol{X}_{\mathrm{a}})\right)\right\|_2^2 \tag{4}$$

式中:$\widetilde{D}_l$为深度神经网络的特征抽取器,$\mathcal{G}$为在$\widetilde{D}_l$的一组样式层上提取的深度特征的格拉姆矩阵[11]。由于不同的层可以学习不同的风格,所以本文使用网络的所有卷积层作为风格层。

1.2.2　原图损失

上述风格损失可用于生成参考样式的增强图像,但增强图像的内容可能与原始图像的内容非常不同。原始图像的内容可以通过原图损失来保存,具体为

$$\mathcal{L}_{\mathrm{o}}=\sum_{l\in O_l}\left\|\widetilde{D}_1(\boldsymbol{X})-\widetilde{D}_1(\boldsymbol{X}_{\mathrm{a}})\right\|_2^2 \tag{5}$$

通过这种方式可以确保增强后的图像在深度表示空间中具有与原始图像非常相似的内容。本文使用特征提取网络的更深一层作为内容层。

1.2.3　细节损失

通过减少相邻像素之间的变化确保画面细节平滑。对于增强图像x_{a},细节损失定义为

$$\mathcal{L}_{\mathrm{d}}=\sum\sqrt{\left(X_{\mathrm{a}\,i,j}-X_{i+1,j}\right)^2+\left(X_{\mathrm{a}\,i,j}-X_{i,j+1}\right)^2} \tag{6}$$

式中:$X_{\mathrm{a}\,i,j}$为图像$\boldsymbol{X}_{\mathrm{a}}$的坐标$(i,j)$处的像素。直观地说,这将鼓励图像具有低方差的局部补丁。Sharif等[12]指出,光滑项有助于提高物理环境中增强样本的稳健性。

1.2.4　对抗性损失

对于对抗性损失$\mathcal{L}_{\mathrm{adv}}$,本文使用以下交叉熵损失

$$\mathcal{L}_{\mathrm{adv}}=\begin{cases} \lg\left(p_y(\boldsymbol{X})\right) & \text{无目标攻击} \\ -\lg\left(p_{\mathrm{yadv}}\left(\boldsymbol{X}^{\mathrm{t}}\right)\right)+\lg\left(p_y\left(\boldsymbol{X}^{\mathrm{t}}\right)\right) & \text{有目标攻击} \end{cases} \tag{7}$$

式中:$\boldsymbol{X}^{\mathrm{t}}$为目标样本,$p_{\mathrm{yadv}}(\ )$为目标模型F的概率输出(logits上的SoftMax)。

1.3　启发式投影算法

图3给出了由FGSM、GAMA和本文方法为ResNet50模型生成的增强样本,其中第1行为生成的增强样本,第2行展示了与原始图像之间的噪声差异,第3行展示了使用BM3D进行去噪后的效果,最后一行为类激活图。Li等[13]证明了区域均匀的扰动在攻击模型中有很强攻击性。因此,本文认为在这些区分区域(图3最后一行类激活图中圈出的部分)具有聚集特征的噪声更有可能攻击成功,因为它们

扰乱了更重要的信息。

从类激活图可视化图像(CAM)可以看出,区分不同类别的区域往往聚集在几个特定的部分,与快速梯度下降攻击方法(Fast gradient sign method,FGSM)和基于生成对抗网络的攻击方法(Generative adversarial model-based attacks,GAMA)相比,本文的噪声也具有很强的聚集性,更容易成功攻击网络。同时,很多医学分类任务在训练前都需要去噪,本文使用了目前比较先进的块匹配三维协同滤波(Block-matching and 3D filtering,BM3D)方法来比较去噪前后的效果。从图3的第3行可以看出,启发式投影法产生的噪声很难去除,因此本文方法具有很强的抗噪声能力。

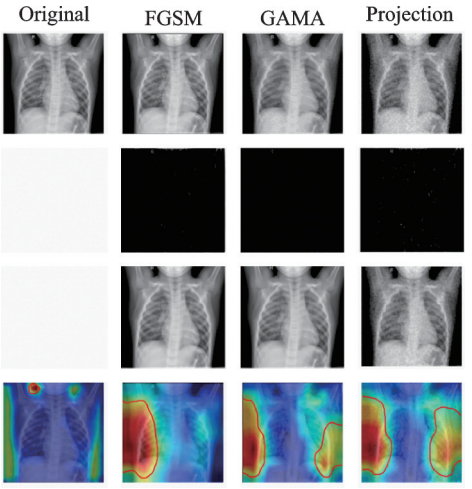

图3 由FGSM、GAMA和本文方法为ResNet50模型生成的增强样本

Fig.3 Enhanced samples generated by FGSM, GAMA and the proposed method for ResNet50 model

因此,本文结合前面的观测结果,提出了一种启发式投影算法,其灵感来自Rosen投影梯度法[14]:当迭代点位于可行域的边缘时,投影梯度方向。为了减少计算开销,将超过阈值的噪声投影到周围区域,从而导致噪声具有更强的聚集性。本文认为噪声向量中更容易打破ε限制的部分有更高的概率出现在区分区域的高亮区域。

如果增强图像$\boldsymbol{X}_{\mathrm{a}}$的$L_{\infty}$范数超过阈值,通过以下方式消除扰动

$$P=\mathrm{clip}\left(|\boldsymbol{X}_{\mathrm{a}}|-\varepsilon,0,\infty\right)\cdot\mathrm{sign}(\boldsymbol{X}_{\mathrm{a}}) \tag{8}$$

然后,最终的增强图像被定义为

$$\boldsymbol{X}_{\mathrm{a}}'=\mathrm{clip}\left(\boldsymbol{X}_{\mathrm{a}}+\varepsilon\cdot\mathrm{sign}\left(\nabla J(\boldsymbol{X}_{\mathrm{a}},Y)\right)+\gamma\cdot\mathrm{sign}(\boldsymbol{W}_{\mathrm{o}}*P)\right) \tag{9}$$

式中:$\boldsymbol{W}_{\mathrm{o}}$为大小为$w\times w$的特殊均匀投影核,$\mathrm{sign}(\boldsymbol{W}_{\mathrm{o}}*P)$是切割扰动的可行方向,$\boldsymbol{W}_{\mathrm{o}}$可简单定义为

$$W_{\mathrm{o}}[i,j]=\begin{cases}0 & i=\lfloor w/2\rfloor,j=\lfloor w/2\rfloor\\ 1/\left(w^2-1\right) & \text{其他}\end{cases} \tag{10}$$

综上所述,本文提供了一种多损失方法来生成决策边界附近的对抗性样本,并且这种对抗样本是难以检测的;并利用启发式投影算法对生成的噪声进行聚合,提高了模型的噪声不可去除性和攻击迁移性,然后将这些扩充后的样本送入目标模型进行训练,以增强模型的泛化能力。

2 实验和结果

通过新冠肺炎上的一个图像分类任务来评估该图像增强方法的有效性。该数据集中共有3类图像,其中正常图像1 341张,COVID-19阳性图像1 200张,病毒性肺炎图像1 345张,图像尺寸均为224像素×224像素。

2.1 样本设置

为了评估本文图像增强方法如何帮助深度学习模型提高泛化性能,在两类数据集(原始新冠肺炎图像和增强图像数据集)上对标准医学图像分类模型ResNet50进行了训练和测试。在实验中,原始数据集x被分为x_{train}和x_{test}两部分,比例为0.85∶0.15。此外,用本文方法和其他方法扩充的数据集x'也被

分为x'_{train}和x'_{test},比例相同。使用的原始图像训练数据集是x_{train},而增强的图像训练数据集是x_{train}和x'_{train}的组合。

2.2　多损失函数的消融影响

图4展示了具有或不具有两个可选损失项(原图损失$\mathcal{L}_o$和细节损失$\mathcal{L}_d$)的3组增广样本。当包含一个损失项时,通过式(2)将其直接添加到最终对象中。可以观察到,原图损失项可以帮助保留原始内容,而细节损失可以帮助产生平滑的对象表面。因此,多损失函数对于生成人类观察者看来合法的增广样本是有效的,而投影算法确保所生成的噪声是高度聚集且难以去除的。将本文多损失结合启发式投影方法与FGSM、GAMA和RayS进行比较,以验证其攻击可转移性。在表1中,最顶行标记了攻击类型:白盒或者黑盒。使用ResNet50生成了FGSM[15]、GAMA[16]、RayS[17]和本文方法的对抗样本;然后利用生成的攻击样本对ResNet101、Vgg16和ResNe152进行攻击,验证了不同攻击方法的攻击可迁移性。如表1所示,与其他攻击方法相比,本文方法的多次损失可以使攻击成功率平均提高28.9%,在攻击ResNet101时,可以提高35.1%的攻击成功率。这是因为本文方法的扰动具有聚集性,攻击具有更强的可迁移性。将本文数据扩增方法与传统的几何数据扩增方法(旋转与中心对称)及mixup扩增方法进行了对比,结果如表2所示。通过几何变化的数据扩增方法并不能体现出对抗样本的特点,从而对模型的分类精度提升非常小,mixup方法[18]通过对样本进行混类操作扩增数据集,一定程度上提升了模型的分类性能。而本文所提出的多损失函数能产生更有效的对抗性样本,并且投影算法方法在性能方面优于所有多损失混合方法。

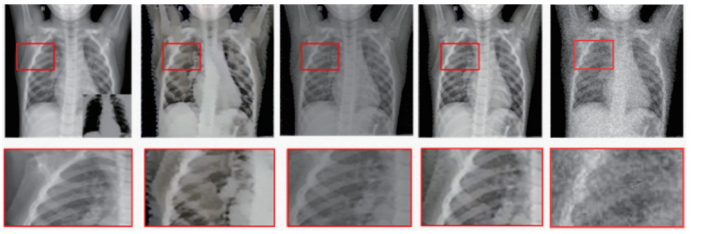

图4　不同损失项生成的增强样本

Fig.4　Enhanced samples generated using different loss terms

表1　加入投影算法的攻击成功率对比

Table 1　Comparison of attack success rates by incorporating projection algorithm　%

方法	使用ResNet50生成对抗样本				
	白盒	黑盒			
	ResNet50	ResNet101	Vgg16	ResNet152	Average
FGSM	80.9	29.6	19.4	20.3	23.1
GAMA	100	38.0	33.1	33.9	35.0
RayS	99.8	54.1	43.5	50.9	49.5
本文方法	100	73.1	51.2	67.4	63.9

表2　不同数据扩增方法分类精度的比较

Table 2　Comparison of classification accuracy between different data augmentation methods

方法	Accuracy	Recall	Precision	F_1_score
源数据集	0.916 6	0.916 6	0.926 7	0.915 7
旋转对称	0.916 8	0.916 8	0.923 0	0.916 3
mixup	0.945 7	0.945 7	0.946 2	0.934 6
只使用$\mathcal{L}_s$	0.944 0	0.944 0	0.945 5	0.933 2
使用$\mathcal{L}_s+\mathcal{L}_o$	0.946 5	0.946 5	0.947 7	0.946 4
使用Multi-loss	0.961 5	0.961 5	0.962 0	0.961 5
加入投影算法	0.964 1	0.964 1	0.964 3	0.964 1

2.3　结果分析

图5展示了在ROC曲线和AUC值方面,ResNet50在原始集合和组合集合(使用本文的方法和原始数据集)上训练的性能。此外,还使用ROC曲线及其AUC(曲线下面积)来评估分类性能,并且在组合数据集上训练的ResNet50显著优于在原始数据集上训练的模型。从图5可以看出,2类(COVID类)的

曲线下面积明显增大,这是因为本文的组合数据集增强了分类模型的泛化性能,使得在组合集合上训练的模型性能优于原始数据集。

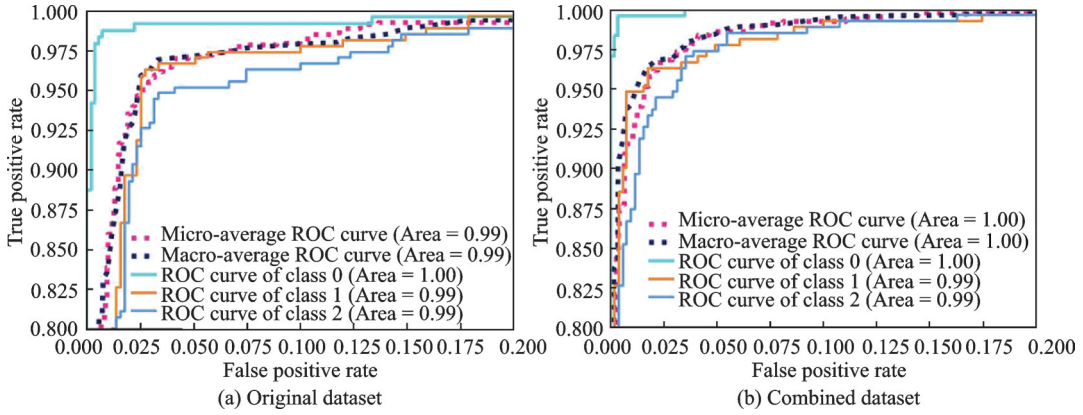

图5 ResNet50在原始数据集和组合数据集上训练的ROC曲线和AUC值方面的表现

Fig.5 Performance of ResNet50 in terms of ROC curves and AUC values for training on the original and combined datasets

3 结束语

本文提出了一种新的数据增强方法,它结合了对抗性攻击技术和包括风格、原图和细节损失在内的多损失函数来生成增强的医学样本。在新冠肺炎分类任务上的实验结果表明,该方法增强的图像样本可以提高4.75%的疾病诊断正确率。因此,基于多损失混合和启发式投影算法的对抗样本生成医学图像增强方法可以提高模型的通用性和可移植性,并在低数据场景下提供显著的改进。

参考文献:

[1] XU Mengting, ZHANG Tao, ZHANG Daoqiang. MedRDf: A robust and retrain-less diagnostic framework for medical pretrained models against adversarial attack[J]. IEEE Transactions on Medical Imaging, 2022, 41(8): 2130-2143.

[2] KIM H E, COSA-LINAN A, SANTHANAM N, et al. Transfer learning for medical image classification: A literature review [J]. BMC Medical Imaging, 2022, 22(1): 69.

[3] SOOMRO T A, ZHENG L, AFIFI A J, et al. Artificial intelligence (AI) for medical imaging to combat coronavirus disease (COVID-19): A detailed review with direction for future research[J]. Artificial Intelligence Review, 2022.55(2): 1409-1439.

[4] HEIN M, ANDRIUSHCHENKO M, BITTERWOLFET J, et al. Why RELU networks yield high-confidence predictions far away from the training data and how to mitigate the problem[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 41-50.

[5] SALEHINEJAD H, VALAEE S, DOWDELL T, et al. Image augmentation using radial transform for training deep neural networks[C]//Proceedings of 2018 IEEE ICASSP. [S.l.]:IEEE, 2018: 3016-3020.

[6] OKAFOR E, SMIT R, SCHOMAKER L, et al. Operational data augmentation in classifying single aerial images of animals [C]//Proceedings of 2017 IEEE International Conference on Inovations in Intelligent Systems and Applications. [S.l.]: IEEE, 2017: 354-360.

[7] PANDEY S, SINGH P R, TIAN J, et al. An image augmentation approach using two-stage generative adversarial network for nuclei image segmentation[J]. Biomedical Signal Processing and Control, 2020, 57: 101782.

[8] LEI Na, AN Dongsheng, GUO Yang, et al. A geometric understanding of deep learning[J]. Engineering, 2020, 6(3): 361-374.

[9] CHEN C, QIN C, QIU H, et al. Realistic adversarial data augmentation for MR image segmentation[C]//Proceedings of Medical Image Computing and Computer Assisted Invention. Lima, Peru: Springer, 2020.

[10] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of IEEE International Conference on Computer Vision. [S.l]: IEEE, 2017: 618-626.

[11] GATYS L A, ECKER A S, BETHGE M, et al. Image style transfer using convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l]: IEEE, 2016: 2414-2423.

[12] SHARIF M, BHAGAVATULA S, BAUER L, et al. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition[C]//Proceedings of 2016 ACM SIGSAC Conference on Computer and Communications Security. [S.l.]: ACM, 2016.

[13] LI Yingwei, BAI Song, XIE Cihang, et al. Regional homogeneity: Towards learning transferable universal adversarial perturbations against defenses[C]//Proceedings of Computer Vision—ECCV 2020: 16th European Conference. Glasgow, UK: [s.n], 2020.

[14] ROSEN J B. The gradient projection method for nonlinear programming. Part I. Linear constraints[J]. Journal of the Society for Industrial and Applied Mathematics, 1960, 8(1): 181-217.

[15] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. (2014-02-09). https://doi.org/10.48550/arXiv.1312.6199.

[16] AICH A, TA C K, GUPTA A, et al. GAMA: Generative adversarial multi-object scene attacks[J]. Advances in Neural Information Processing Systems, 2022, 35: 36914-36930.

[17] CHEN J, GU Q. Rays: A ray searching method for hard-label adversarial attack[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.[S.l.]: ACM, 2020: 1739-1747.

[18] ZHANG Hongyi, CISSE M, DAUPHIN Y N, et al. Mixup: Beyond empirical risk minimization[J]. (2018-04-27). https://doi.org/10.48550/arXiv.1710.09412.

作者简介:

王见(1999-),男,硕士研究生,研究方向:医学图像处理、深度神经网络对抗攻击与防御,E-mail: wang_jian@nuaa.edu.cn。

成楚凡(1998-),男,硕士研究生,研究方向:医学图像处理、深度神经网络对抗攻击与防御。

陈芳(1991-),**通信作者**,女,博士,副教授,研究方向:医学图像分析、计算机辅助手术导航,E-mail: chenfang@nuaa.edu.cn。

(编辑:王静)