(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112506797 B
(45)授权公告日 2022.05.24

(21)申请号 202011525218.0

(22)申请日 2020.12.22

(65)同一申请的已公布的文献号
申请公布号 CN 112506797 A

(43)申请公布日 2021.03.16

(73)专利权人 南京航空航天大学
地址 210016 江苏省南京市秦淮区御道街29号

(72)发明人 陈芳 成楚凡 张道强

(74)专利代理机构 南京瑞弘专利商标事务所(普通合伙) 32249
专利代理师 陈国强

(51)Int.Cl.
G06F 11/36(2006.01)

(56)对比文件
US 2016314064 A1,2016.10.27
WO 2008060022 A1,2008.05.22
CN 112052186 A,2020.12.08
CN 111782529 A,2020.10.16

审查员 伍小辉

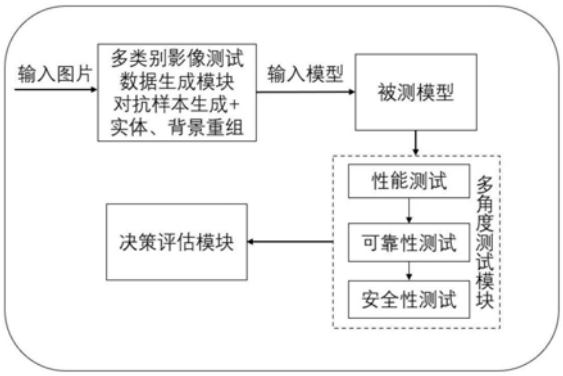
权利要求书3页 说明书8页 附图2页

(54)发明名称

一种针对医学影像识别系统的性能测试方法

(57)摘要

本发明公开了一种医学影像识别系统性能测试方法,包括1)多类别影像测试数据生成模块,它包含对抗样本生成网络和实体、背景重组方法;2)基于系统稳定性、可靠性及安全性性能的多角度测试模块;3)模型决策评估模块。本发明实现了多类别影像测试数据生成和多角度完备的系统测试,最终完成医学影像识别系统的决策评估,未来应用前景广泛。



CN 112506797 B

1.一种针对医学影像识别系统的性能测试方法,其特征在于:性能测试方法中包括了:多类别影像测试数据生成模块,所述多类别影像测试数据生成模块包括对抗样本生成网络和实体、背景重组方法;多角度测试模块,所述多角度测试模块包括性能测试、可靠性测试和安全性测试;决策评估模块,所述决策评估模块对所输入的测试结果进行分析,判断模型性能,并给出详细的测试报告;

网络输入一组待分类识别的图片,首次将图片输入到所述多类别影像测试数据生成模块中,经过图像增广后输入模型进行分类,并将分类结果输入至多角度测试模块,多角度测试模块对模型的学习结果进行测试,并将结果传递到决策评估模块,决策评估模块对所输入的测试结果进行分析,判断模型性能,并给出详细的测试报告;

所述对抗样本生成网络和实体、背景重组方法中包括对抗增广使用多损失混合对抗伪装,

多损失函数$\mathcal{L}$表示为:

$$\mathcal{L}=\left(\mathcal{L}_s+\mathcal{L}_c+\mathcal{L}_m\right)+\lambda\cdot\mathcal{L}_{adv} \tag{1}$$

其中:λ表示对抗强度、$\mathcal{L}_{adv}$表示对抗性损失、$\mathcal{L}_s$表示用于样式生成的样式损失、$\mathcal{L}_c$表示用于保存源图像内容的内容损失、$\mathcal{L}_m$表示用于确保扩展样本的平滑度的平滑度损失;

用户定义现有图像、目标攻击区域和预期目标样式,在所需的区域中生成所需的样式,在每个步骤中向生成的扩展样本中添加附加的物理适应训练;

两个图像之间的样式距离通过两个图像在样式表示中的差异来定义,

$$\mathcal{D}_s=\sum_{l\in S_l}\left\|\mathcal{G}\left(\widetilde{F}_l\left(x^s\right)\right)-\mathcal{G}\left(\widetilde{F}_l\left(x'\right)\right)\right\|_2^2 \tag{2}$$

其中:$\mathcal{D}_s$为特征距离,l是风格层特征,S_l是提取风格表示的风格层集,$\widetilde{F}_l$是风格样式的特征提取器,$\mathcal{G}$是从$\widetilde{F}_l$的一组样式层提取的深层特征的Gram矩阵,x^s是风格参考图像,x′是生成的对抗样本;

用于样式生成的样式损失$\mathcal{L}_s$所生成参考样式中的增强图像的内容与原始图像的内容非常不同;具体如下,

$$\mathcal{L}_c=\sum_{t\in c_t}\left\|\widetilde{F}_t(x)-\widetilde{F}_t\left(x'\right)\right\|_2^2 \tag{3}$$

其中:$\mathcal{L}_c$是内容损失,t是内容层特征,c_t是提取内容表示的内容层集,$\widetilde{F}_t$是内容层的特征提取器,x是原始图像,x′是生成的对抗样本;

通过减少相邻像素之间的变化,提高增强图像的平滑度;对于增强图像,平滑度损失定义如下,

$$\mathcal{L}_m=\sum\left(\left(x'_{i,j}-x_{i+1,j}\right)^2+\left(x'_{i,j}-x_{i,j+1}\right)^2\right)^{\frac{1}{2}} \tag{4}$$

其中:$x'_{i,j}$为对抗样本(i,j)坐标处的像素值,$x_{i+1,j}$为原始图像(i+1,j)坐标处的像素

值，$x_{i,j+1}$为原始图像(i,j+1)坐标处的像素值：

对于对抗性损失$\mathcal{L}_{adv}$，使用以下交叉熵损失：

$$\begin{cases} log\left(p_y\left(x^{'}\right)\right), & \text{非目标类别，} \\ -log\left(p_{yadv}\left(x^{'}\right)\right)+log\left(p_y\left(x^{'}\right)\right), & \text{目标类别} \end{cases} \quad (5)$$

其中：$p_{yadv}()$和$p_y()$分别是目标模型F对标签yadv和y的概率输出，yadv为对抗样本的类别，y为原始图像的类别；

将现实条件引入到增强示例的生成过程中，如下所示：

$$\min_{x^{'}}\left(\left(\mathcal{L}_s+\mathcal{L}_c+\mathcal{L}_m\right)\right)+\max_{T\in\mathcal{T}}\lambda\cdot\mathcal{L}_{adv}\left(o+T\left(x^{'}\right)\right) \quad (6)$$

其中：o是在物理世界中采样的随机背景图像，T是旋转、调整大小和颜色偏移的随机变换，$\mathcal{T}$是变换的集合；

目标背景重组增广使用分割算法Mask R-CNN将目标从背景中分割出来，使用插值算法将背景中空白部分补充像素，最后随机组合目标和背景，实现图片增广。

2.根据权利要求1所述的针对医学影像识别系统的性能测试方法，其特征在于：

所述多角度测试模块中的性能测试，包含不同角度：识别准确率Accuracy判别判断，识别损失值Loss判别判断以及蜕变关系判别；准确率Accuracy和损失值Loss的判断，都是用增广前后的模型输出的准确率Accuracy和识别损失值Loss相减得到的扩充前后识别准确率差值百分比Δacc和扩充前后识别损失差值百分比Δloss；

蜕变测试定义为：C_i为原测试图像X_{test}^{i}的被图像识别系统分类标签，S_i为原测试图像X_{test}^{i}的置信分数；C_i'为结合蜕变关系利用X_{test}^{i}合成的新测试图像X_{test1}^{i}的分类标签，S_i'为结合蜕变关系利用X_{test}^{i}合成的新测试图像X_{test1}^{i}的置信分数，

蜕变关系表述为：

$C_i=C'_i$ and $\Delta S=|S_i-S'_i|<c$ (7)

其中：c为超参数，0<c<100，c设置为50，ΔS为扩充前后置信分数的差。

3.根据权利要求1所述的针对医学影像识别系统的性能测试方法，其特征在于：

所述多角度测试模块中的可靠性测试为鲁棒性certified robustness测试，在原始图像x满足置信度保证的条件下，在范数球半径R内能够免疫攻击：

$$g(x+\varepsilon)=g(x)\coloneqq argmax(z(x)),\forall\varepsilon\in B(x;R) \quad (8)$$

其中：z()为损失函数，g()是待优化的目标函数，∀指任意，ε为引入的噪声，B(x;R)为噪声集合，R为范数球半径，x是原始图像；

最后鲁棒性准确度robacc定义为：

$$robacc=\frac{\#\text{具有鲁棒性的样本数量}}{\#\text{总样本数量}} \quad (9)。$$

4.根据权利要求1所述的针对医学影像识别系统的性能测试方法，其特征在于：

所述多角度测试模块中的安全性测试为模型不变性测试，选择一个随机图像，使用下

面描述的四种方法之一选择一个像素的扰动，然后测量网络对该扰动的敏感度，第一种方法为裁剪Crop方法，在原始图像中随机选择一个正方形，并将正方形的大小调整为224x224px，然后，将正方形对角平移一个像素，以创建第二个图像，图像通过平移单个像素来与第一个图像不同；第二种方法为嵌入Embedding方法，先缩小图像，使图像最小尺寸为100px，同时保持纵横比，并嵌入到224x224px图像内的随机位置，同时用黑色像素填充图像的其余部分，然后将嵌入位置移位单个像素，再次创建两个相同的图像，直到移位单个像素；第三种方法中，先缩小图像，使图像最小尺寸为100px，同时保持纵横比，并将其嵌入到224x224px图像内的随机位置，然后使用简单的修复算法，即每个黑色像素被其邻域中的非黑色像素的加权平均所取代，第四个方法与第二个方法相同，先缩小图像，使图像最小尺寸为100px，但不移动嵌入位置，而是保持嵌入位置不变，并将嵌入图像的大小更改单个像素。

5.根据权利要求4所述的针对医学影像识别系统的性能测试方法，其特征在于：

在安全性测试中，用两种方法测量灵敏度作为模型的不变性测试，第一个是网络的TOP-1预测在单像素扰动后发生变化的概率，称之为P，即Top-1 Change，第二个是测量在顶层类的一个像素扰动之后，网络计算的概率的平均绝对值变化，称之为平均绝对变化，即MAC。

6.根据权利要求1所述的针对医学影像识别系统的性能测试方法，其特征在于：

所述决策评估模块对所输入的测试结果进行分析，判断模型性能，即Accuracy扩充后识别准确率，Loss扩充后识别损失，Δacc扩充前后识别准确率差，Δloss扩充前后识别损失差，CR模型鲁棒性用robacc来表征，ΔS扩充前后置信分数差，P网络的TOP-1预测在单像素扰动后发生变化的概率，MAC平均绝对变化，并给出测试报告，当对多个识别模型进行性能对比的时候，大量单独的性能指标往往对用户来说过于繁杂，使用户难以作出合理的判断，因此，定义了综合性能指标CM来反映不同识别系统的综合性能；公式如下：

$$CM_i = \sum_{j=1}^{N} \omega_j \left[\left(2 * max\left(M_j\right) - M_{ij} \right) / \left(2 * max\left(M_j\right) - min\left(M_j\right) \right) \right] \qquad (10)$$

其中：CM_i代表第i个识别系统的综合性能值，ω_j代表云服务第j个性能指标值的权重，$max(M_j)$代表多个识别系统中第j个性能指标的最大值，$min(M_j)$代表多个识别系统中第j个性能指标的最小值，M_{ij}代表第i个识别系统的第j个性能指标值，N代表识别系统性能指标值的总数，通过利用公式$(2*max(M_j)-M_{ij})/(2*max(M_j)-min(M_j))$对$M_{ij}$的值进行标准化至[0，1]区间。

7.根据权利要求1所述的针对医学影像识别系统的性能测试方法，其特征在于：

所述决策评估模块最终会输出多角度测试模块测试的各项结果，并生成相应的测试报告表格。

一种针对医学影像识别系统的性能测试方法

技术领域

[0001] 本发明属于医学影像识别系统性能分析技术领域，特别涉及一种针对医学影像识别系统的性能测试方法。

背景技术

[0002] 医学影像识别系统在临床医学诊断中发挥着重要的作用，它大大改变了临床诊断模式，促进了临床医学的发展。智能医学影像识别是基于人工智能技术，对X线片、计算机断层扫描、磁共振成像等常用医学影像学技术扫描图像和手术视频进行分析处理的过程，其发展方向主要包括智能影像诊断、影像三维重建与配准、智能手术视频解析等。目前，对该领域的研究已取得一定的进展，正在逐步走向临床应用。因此，对医学影像识别系统性能的评估和测试对未来临床医学的发展尤为重要。FERET首次给识别算法设置了性能基准，定义了一系列的评测标准，极大地推进了识别技术的发展，其制定的评测标准和评价协议一直影响至今，为以后人脸识别技术的发展带来了深远的影响。虽然一般的图像识别系统有一些测试方案，但是针对医学影像识别系统的测试方案还没有被提出。而且由于当时的识别技术尚不成熟，参与FERET评测的识别系统也多是大学实验室里的原型系统，识别效果并不是非常令人满意。

[0003] 近年来，对识别模型进行分析，并自动进行性能分析的测试方法需求逐渐增加。随着深度学习技术快速发展，医学影像识别系统的性能指标也得到快速提高，大大提高了识别效率，所以如何测试这些模型的性能亟待解决。针对医学影像识别系统在测试中，需要考虑到生成的测试影像的视觉真实性，因此提出了对抗样本生成网络和实体、背景重组方法，充分保证生成样本的真实性；且医学识别系统对可靠性和安全性的测试要求更高，因此提出多角度测试方案，并将对抗样本应用到医学影像中，达到对医学识别模型更好分析的效果。

发明内容

[0004] 本发明提供了一种针对医学影像识别系统的性能测试方法，以解决现有技术中的问题。

[0005] 为了实现上述目的，本发明采用以下技术方案：

[0006] 一种针对医学影像识别系统的性能测试方法，性能测试方法中包括了：多类别影像测试数据生成模块，所述多类别影像测试数据生成模块包括对抗样本生成网络和实体、背景重组方法；多角度测试模块，所述多角度测试模块包括性能测试、可靠性测试和安全性测试；决策评估模块，所述决策评估模块对所输入的测试结果进行分析，判断模型性能，并给出详细的测试报告；

[0007] 网络输入一组待分类识别的图片，首次将图片输入到所述多类别影像测试数据生成模块中，经过图像增广后输入模型进行分类，并将分类结果输入至多角度测试模块，多角度测试模块对模型的学习结果进行测试，并将结果传递到决策评估模块，决策评估模块对

所输入的测试结果进行分析，判断模型性能，并给出详细的测试报告。

[0008] 进一步的，所述对抗样本生成网络和实体、背景重组方法中包括对抗增广使用多损失混合对抗伪装，

[0009] 多损失函数$\mathcal{L}$表示为：

[0010] $$\mathcal{L}=\left(\mathcal{L}_s+\mathcal{L}_c+\mathcal{L}_m\right)+\lambda\cdot\mathcal{L}_{adv} \quad (1)$$

[0011] 其中：λ表示对抗强度、$\mathcal{L}_{adv}$表示对抗性损失、$\mathcal{L}_s$表示用于样式生成的样式损失、$\mathcal{L}_c$表示用于保存源图像内容的内容损失、$\mathcal{L}_m$表示用于确保扩展样本的平滑度的平滑度损失；

[0012] 用户定义现有图像、目标攻击区域和预期目标样式，在所需的区域中生成所需的样式，在每个步骤中向生成的扩展样本中添加附加的物理适应训练；

[0013] 两个图像之间的样式距离通过两个图像在样式表示中的差异来定义，

[0014] $$\mathcal{D}_s=\sum_{l\in S_l}\left\|\mathcal{G}\left(\widetilde{F}_l\left(x^s\right)\right)-\mathcal{G}\left(\widetilde{F}_l\left(x'\right)\right)\right\|_2^2 \quad (2)$$

[0015] 其中：$\mathcal{D}_s$为特征距离，l是风格层特征，S_l是提取风格表示的风格层集，$\widetilde{F}_l$是风格样式的特征提取器，$\mathcal{G}$是从$\widetilde{F}_l$的一组样式层提取的深层特征的Gram矩阵，x^s是风格参考图像，x′是生成的对抗样本；

[0016] 用于样式生成的样式损失$\mathcal{L}_s$所生成参考样式中的增强图像的内容与原始图像的内容非常不同；具体如下，

[0017] $$\mathcal{L}_c=\sum_{t\in c_t}\left\|\widetilde{F}_t(x)-\widetilde{F}_t\left(x'\right)\right\|_2^2 \quad (3)$$

[0018] 其中：$\mathcal{L}_c$是内容损失，t是内容层特征，c_t是提取内容表示的内容层集，$\widetilde{F}_t$是内容层的特征提取器，x是原始图像，x′是生成的对抗样本；

[0019] 通过减少相邻像素之间的变化，提高增强图像的平滑度；对于增强图像，平滑度损失定义如下，

[0020] $$\mathcal{L}_m=\sum\left(\left(x'_{i,j}-x_{i+1,j}\right)^2+\left(x'_{i,j}-x_{i,j+1}\right)^2\right)^{\frac{1}{2}} \quad (4)$$

[0021] 其中：$x_{i,j}$为对抗样本(i,j)坐标处的像素值，$x_{i+1,j}$为原始图像(i+1,j)坐标处的像素值，$x_{i,j+1}$为原始图像(i,j+1)坐标处的像素值：

[0022] 对于对抗性损失$\mathcal{L}_{adv}$，使用以下交叉熵损失：

[0023] $$\begin{cases} log\left(p_y\left(x'\right)\right), & \text{非目标类别，} \\ -log\left(p_{yadv}\left(x'\right)\right)+log\left(p_y\left(x'\right)\right), & \text{目标类别} \end{cases} \quad (5)$$

[0024] 其中：$p_{yadv}()$和$p_y()$分别是目标模型F(F指代通用机器模型的目标函数，例如vgg的目标函数F为fc8，由此可以得到对应1000类的概率输出)对标签yadv(对抗样本的类别)

和y(原始图像的类别)的概率输出。

[0025] 将现实条件引入到增强示例的生成过程中,如下所示:

[0026] $$\min_{x'}\left(\left(\mathcal{L}_s+\mathcal{L}_c+\mathcal{L}_m\right)\right)+\max_{T\in\mathcal{T}}\lambda\cdot\mathcal{L}_{adv}\left(o+T\left(x'\right)\right) \quad (6)$$

[0027] 其中:o是在物理世界中采样的随机背景图像,T是旋转、调整大小和颜色偏移的随机变换,$\mathcal{T}$是变换的集合;通过根据原始图像x和背景图像o,生成的增强样本对于人类观察者来说基本是合法的;

[0028] 目标背景重组增广使用分割算法Mask R-CNN将目标从背景中分割出来,使用插值算法将背景中空白部分补充像素,最后随机组合目标和背景,实现图片增广。

[0029] 进一步的,多角度测试模块中的性能测试,包含不同角度:识别准确率Accuracy判别判断,识别损失值Loss判别判断以及蜕变关系判别;准确率Accuracy和损失值Loss的判断,都是用增广前后的模型输出的准确率Accuracy和识别损失值Loss相减得到的扩充前后识别准确率差值百分比Δacc和扩充前后识别损失差值百分比Δloss;

[0030] 蜕变测试定义为:C_i为原测试图像X_{test}^i的被图像识别系统分类标签,S_i为原测试图像X_{test}^i的置信分数;C_i'为结合蜕变关系利用X_{test}^i合成的新测试图像X_{test1}^i的分类标签,S_i'为结合蜕变关系利用X_{test}^i合成的新测试图像X_{test1}^i的置信分数,那么蜕变关系表述为:

[0031] $C_i=C'_i$ and $\Delta S=|S_i-S'_i|<c$ (7)

[0032] 其中:c为超参数,0<c<100,c设置为50,ΔS为扩充前后置信分数的差。

[0033] 进一步的,多角度测试模块中的可靠性测试为鲁棒性(certified robustness)测试,在原始图像x 满足置信度保证的条件下,在范数球半径R内能够免疫攻击:

[0034] $$g(x+\varepsilon)=g(x):=argmax(z(x)),\forall\varepsilon\in B(x;R) \quad (8)$$

[0035] 其中:z()为损失函数,g()是待优化的目标函数,∀指任意,ε为引入的噪声,B(x;R)为噪声集合,R为范数球半径,R为一个无线接近0的值,x是原始图像;

[0036] 最后鲁棒性准确度(robacc)定义为:

[0037] $$robacc=\frac{\#\text{具有鲁棒性的样本数量}}{\#\text{总样本数量}} \quad (9)$$

[0038] 多角度测试模块中的安全性测试为模型不变性测试,选择一个随机图像,使用下面描述的四种方法之一选择一个像素的扰动,然后测量网络对该扰动的敏感度,第一种方法为“裁剪(Crop)”方法,在原始图像中随机选择一个正方形,并将该正方形的大小调整为224x224px,然后,我们将该正方形对角平移一个像素,以创建第二个图像,该图像通过平移单个像素来与第一个图像不同;第二种方法为“嵌入(Embedding)”方法,先缩小图像,使其最小尺寸为100px,同时保持纵横比,并将其嵌入到224x224px图像内的随机位置,同时用黑色(Black)像素填充图像的其余部分,然后将嵌入位置移位单个像素,再次创建两个相同的图像,直到移位单个像素;第三种方法中,先缩小图像,使其最小尺寸为100px,同时保持纵横比,并将其嵌入到224x224px图像内的随机位置,然后使用简单的修复算法(每个黑色像素被其邻域中的非黑色像素的加权平均所取代),第四个方法与第二个协议相同,先缩小图像,使其最小尺寸为100px,但我们不移动嵌入位置,而是保持嵌入位置不变,并将嵌入图像的大小更改单个像素(例如,从大小100x100px更改为大小101x101px像素)。

[0039] 进一步的,在安全性测试中,用两种方法测量灵敏度作为模型的不变性测试,第一个称之为 P(Top-1 Change),是网络的TOP-1预测在单像素扰动后发生变化的概率;第二称之为“平均绝对变化”(MAC),测量在顶层类(即在两个帧的第一帧中具有最高概率的类)的一个像素扰动之后,网络计算的概率的平均绝对值变化(即在两个帧的第一个帧中具有最高概率的类)。

[0040] 进一步的,决策评估模块对所输入的测试结果进行分析,判断模型性能【Accuracy扩充后识别准确率,Loss扩充后识别损失,Δacc扩充前后识别准确率差,Δloss扩充前后识别损失差,CR 模型鲁棒性(用robacc来表征),ΔS扩充前后置信分数差,P(Top-1 Change)网络的TOP-1预测在单像素扰动后发生变化的概率,MAC平均绝对变化)】,并给出详细的测试报告,当我们对多个识别模型进行性能对比的时候,大量单独的性能指标往往对用户来说过于繁杂,使用户难以作出合理的判断,因此,我们将不同指标对于识别系统的综合影响考虑到性能指标设计中,然后定义了一个综合性能指标CM(Composite Value)来反映不同识别系统的综合性能;公式如下:

[0041] $$CM_i = \sum_{j=1}^{N} \omega_j \left[\left(2 * max\left(M_j\right) - M_{ij} \right) / \left(2 * max\left(M_j\right) - min\left(M_j\right) \right) \right] \quad (10)$$

[0042] 其中:CM_i代表第i个识别系统的综合性能值,ω_j代表云服务第j个性能指标值的权重,$max(M_j)$代表多个识别系统中第j个性能指标的最大值,$min(M_j)$代表多个识别系统中第j个性能指标的最小值, M_{ij}代表第i个识别系统的第j个性能指标值,N代表识别系统性能指标值的总数,通过利用公式 $(2*max(M_j)-M_{ij})/(2*max(M_j)-min(M_j))$对$M_{ij}$的值进行标准化至[0,1]区间。可以看出,CM 的值越大,识别系统的综合性能越好。

[0043] 对于一些识别系统性能指标来说,例如Loss,P(Top-1 Change),MAC等,其值M_{ij}越小, CM综合性能指标值越大,这些性能指标的值可以直接代入上述公式中,而对于识别准确率等指标,其值越大,说明识别性能更好,但是直接代入公式会造成CM值的降低,这并不符合预期,所以我们需要对这些性能指标的值进行处理,使用$(1-M_{ij})$来替代公式中的M_{ij}值。

[0044] 进一步的,决策评估模块最终会输出多角度测试模块测试的各项结果,并生成相应的测试报告表格,如表1-3所示。由于不同任务场景的要求不同,测试系统也会给出相应的建议。

[0045] 与现有技术相比,本发明具有以下有益效果:

[0046] 本发明实现了多类别影像测试数据生成和多角度完备的系统测试,最终完成医学影像识别系统的决策评估,未来应用前景广泛。

附图说明

[0047] 图1是本发明的框架流程图;

[0048] 图2是本发明中对抗增广流程图;

[0049] 图3是本发明中目标背景重组增广流程图;

[0050] 图4是本发明中决策评估模块。

具体实施方式

[0051] 下面结合实施例对本发明作更进一步的说明。

[0052] 一种针对医学影像识别系统的性能测试方法，如图1所示，性能测试方法中包括了：多类别影像测试数据生成模块、多角度测试模块和决策评估模块，所述多类别影像测试数据生成模块包括对抗样本生成网络和实体、背景重组方法；所述多角度测试模块包括性能测试、可靠性测试和安全性测试；决策评估模块，所述决策评估模块对所输入的测试结果进行分析，判断模型性能，并给出详细的测试报告；

[0053] 网络输入一组待分类识别的图片，首次将图片输入到所述多类别影像测试数据生成模块中，经过图像增广后输入模型进行分类，并将分类结果输入至多角度测试模块，多角度测试模块对模型的学习结果进行测试，并将结果传递到决策评估模块，决策评估模块对所输入的测试结果进行分析，判断模型性能，并给出详细的测试报告。

[0054] 对抗样本生成网络和实体、背景重组方法，考虑到医学影像的特性和生成的测试影像的视觉真实性，使用了对抗样本生成联合实体、背景重组方案。对抗增广使用多损失混合对抗伪装，该技术可以生成人类观察者看起来合法的新的增强图像，而不需要依赖大量的数据来训练生成网络。我们的目标是开发一种机制来生成具有自定义样式的扩展样本，利用样式变换技术实现图像增强，利用对抗攻击技术实现图像的隐蔽性。最终的多损失函数$\mathcal{L}$是对抗强度λ与对抗性损失$\mathcal{L}_{adv}$的乘积、用于样式生成的样式损失$\mathcal{L}_s$、用于保存源图像内容的内容损失$\mathcal{L}_c$和用于确保扩展样本的平滑度的平滑度损失$\mathcal{L}_m$的组合。

[0055] 多损失函数$\mathcal{L}$表示为：

[0056] $$\mathcal{L}=\left(\mathcal{L}_s+\mathcal{L}_c+\mathcal{L}_m\right)+\lambda\cdot\mathcal{L}_{adv} \qquad (1)$$

[0057] 其中：λ表示对抗强度、$\mathcal{L}_{adv}$表示对抗性损失、$\mathcal{L}_s$表示用于样式生成的样式损失、$\mathcal{L}_c$表示用于保存源图像内容的内容损失、$\mathcal{L}_m$表示用于确保扩展样本的平滑度的平滑度损失；

[0058] 如图2所示，显示了对抗增广方法的概述。用户定义现有图像、目标攻击区域和预期目标样式，在所需的区域中生成所需的样式，如图2右侧所示。为了使扩展样本对各种环境条件(包括照明、旋转等)具有鲁棒性，在每个步骤中向生成的扩展样本中添加附加的物理适应训练；

[0059] 两个图像之间的样式距离通过两个图像在样式表示中的差异来定义，

[0060] $$\mathcal{D}_s=\sum_{l\in S_l}\left\|\mathcal{G}\left(\widetilde{F}_l\left(x^s\right)\right)-\mathcal{G}\left(\widetilde{F}_l\left(x'\right)\right)\right\|_2^2 \qquad (2)$$

[0061] 其中：$\mathcal{D}_s$为特征距离，l是风格层特征，S_l是提取风格表示的风格层集，$\widetilde{F}_l$是风格样式的特征提取器，$\mathcal{G}$是从$\widetilde{F}_l$的一组样式层提取的深层特征的Gram矩阵，x^s是风格参考图像，x′是生成的对抗样本；

[0062] 用于样式生成的样式损失$\mathcal{L}_s$所生成参考样式中的增强图像的内容与原始图像的内容非常不同；具体如下，

[0063] $$\mathcal{L}_c=\sum_{t\in c_t}\left\|\widetilde{F}_t(x)-\widetilde{F}_t\left(x'\right)\right\|_2^2 \qquad (3)$$

[0064] 其中：$\mathcal{L}_c$是内容损失，t是内容层特征，c_t是提取内容表示的内容层集，$\widetilde{F}_t$是内容层的特征提取器，x是原始图像，x′是生成的对抗样本；

[0065] 通过减少相邻像素之间的变化，提高增强图像的平滑度；对于增强图像，平滑度损失定义如下，

[0066] $$\mathcal{L}_m=\sum\left(\left(x'_{i,j}-x_{i+1,j}\right)^2+\left(x'_{i,j}-x_{i,j+1}\right)^2\right)^{\frac{1}{2}} \quad (4)$$

[0067] 其中：$x'_{i,j}$为对抗样本(i,j)坐标处的像素值，$x_{i+1,j}$为原始图像(i+1,j)坐标处的像素值，$x_{i,j+1}$为原始图像(i,j+1)坐标处的像素值：

[0068] 对于对抗性损失$\mathcal{L}_{adv}$，使用以下交叉熵损失：

[0069] $$\begin{cases} log\left(p_y\left(x'\right)\right), & \text{非目标类别，} \\ -log\left(p_{yadv}\left(x'\right)\right)+log\left(p_y\left(x'\right)\right), & \text{目标类别} \end{cases} \quad (5)$$

[0070] 其中：$p_{yadv}()$和$p_y()$分别是目标模型F(F指代通用机器模型的目标函数，例如vgg的目标函数F为fc8，由此可以得到对应1000类的概率输出)对标签yadv(对抗样本的类别)和y(原始图像的类别)的概率输出。

[0071] 为了使对抗图像样本在现实世界中可实现，我们在生成增强样本的过程中对现实条件进行了建模。由于现实世界环境通常涉及条件波动，如视点移动、图像噪声和其他自然转换，因此我们使用一系列调整来适应这些不同的条件。特别地，我们使用了一种类似于期望过转换(EOT)的技术。我们的目标是提高扩充样本对不同物理条件的适应性。因此，我们考虑了用于模拟物理世界条件波动的变换，包括旋转、缩放、颜色偏移(以模拟光照变化)和随机背景。在这里，将现实条件引入到增强示例的生成过程中，如下所示：

[0072] $$\min_{x'}\left(\left(\mathcal{L}_s+\mathcal{L}_c+\mathcal{L}_m\right)\right)+\max_{T\in\mathcal{T}}\lambda\cdot\mathcal{L}_{adv}\left(o+T\left(x'\right)\right) \quad (6)$$

[0073] 其中：o是在物理世界中采样的随机背景图像，T是旋转、调整大小和颜色偏移的随机变换，$\mathcal{T}$是变换的集合；通过根据原始图像x和背景图像o，生成的增强样本对于人类观察者来说基本是合法的；

[0074] 目标背景重组增广使用分割算法Mask R-CNN将目标从背景中分割出来，使用插值算法将背景中空白部分补充像素，最后随机组合目标和背景，实现图片增广，总体方法框架如图3所示。

[0075] 多角度测试模块中的性能测试，包含不同角度：识别准确率Accuracy判别判断，识别损失值 Loss判别判断以及蜕变关系判别；准确率Accuracy和损失值Loss的判断，都是用增广前后的模型输出的准确率Accuracy和识别损失值Loss相减得到的扩充前后识别准确率差值百分比Δacc和扩充前后识别损失差值百分比Δloss；

[0076] 蜕变测试定义为：C_i为原测试图像X^i_{test}的被图像识别系统分类标签，S_i为原测试图像X^i_{test}的置

[0077] 信分数；C_i'为结合蜕变关系利用X^i_{test}合成的新测试图像X^i_{test1}的分类标签，S_i'为结

合蜕变关系利用X_{test}^{i}合成的新测试图像X_{test1}^{i}的置信分数，那么蜕变关系表述为：

[0078] $C_i = C'_i$ and $\Delta S = |S_i - S'_i| < c$ (7)

[0079] 其中：c为超参数，0<c<100，c设置为50，ΔS为扩充前后置信分数的差。

[0080] 多角度测试模块中的可靠性测试为鲁棒性(certified robustness)测试，在原始图像x满足置信度保证的条件下，在范数球半径R内能够免疫攻击：

[0081] $$g(x+\varepsilon) = g(x) \coloneqq argmax(z(x)), \forall \varepsilon \in B(x;R) \quad (8)$$

[0082] 其中：z()为损失函数，g()是待优化的目标函数，∀指任意，ε为引入的噪声，B(x;R)为噪声集合，R为范数球半径，R为一个无线接近0的值，x是原始图像；

[0083] 最后鲁棒性准确度(robacc)定义为：

[0084] $$robacc = \frac{\#\text{具有鲁棒性的样本数量}}{\#\text{总样本数量}} \quad (9)$$

[0085] 多角度测试模块中的安全性测试为模型不变性测试，选择一个随机图像，使用下面描述的四种方法之一选择一个像素的扰动，然后测量网络对该扰动的敏感度，第一种方法为“裁剪(Crop)”方法，在原始图像中随机选择一个正方形，并将该正方形的大小调整为224x224px，然后，我们将该正方形对角平移一个像素，以创建第二个图像，该图像通过平移单个像素来与第一个图像不同；第二种方法为“嵌入(Embedding)”方法，先缩小图像，使其最小尺寸为100px，同时保持纵横比，并将其嵌入到224x224px图像内的随机位置，同时用黑色(Black)像素填充图像的其余部分，然后将嵌入位置移位单个像素，再次创建两个相同的图像，直到移位单个像素；第三种方法中，先缩小图像，使其最小尺寸为100px，同时保持纵横比，并将其嵌入到224x224px图像内的随机位置，然后使用简单的修复算法(每个黑色像素被其邻域中的非黑色像素的加权平均所取代)，第四个方法与第二个协议相同，先缩小图像，使其最小尺寸为100px，但我们不移动嵌入位置，而是保持嵌入位置不变，并将嵌入图像的大小更改单个像素(例如，从大小100x100px更改为大小101x101px像素)。

[0086] 在安全性测试中，用两种方法测量灵敏度作为模型的不变性测试，第一个称之为P(Top-1 Change)，是网络的TOP-1预测在单像素扰动后发生变化的概率；第二称之为“平均绝对变化” (MAC)，测量在顶层类(即在两个帧的第一帧中具有最高概率的类)的一个像素扰动之后，网络计算的概率的平均绝对值变化(即在两个帧的第一个帧中具有最高概率的类)。

[0087] 如图4所示，决策评估模块对所输入的测试结果进行分析，判断模型性能【Accuracy扩充后识别准确率，Loss扩充后识别损失，Δacc扩充前后识别准确率差，Δloss扩充前后识别损失差，CR 模型鲁棒性(用robacc来表征)，ΔS扩充前后置信分数差，P(Top-1 Change)网络的TOP-1预测在单像素扰动后发生变化的概率，MAC平均绝对变化)】，并给出详细的测试报告，当对多个识别模型进行性能对比的时候，大量单独的性能指标往往对用户来说过于繁杂，使用户难以作出合理的判断，因此，将不同指标对于识别系统的综合影响考虑到性能指标设计中，然后定义了一个综合性能指标CM(Composite Value)来反映不同识别系统的综合性能；公式如下：

[0088] $$CM_i = \sum_{j=1}^{N} \omega_j \left[\left(2 * max\left(M_j\right) - M_{ij}\right) / \left(2 * max\left(M_j\right) - min\left(M_j\right)\right) \right] \quad (10)$$

[0089] 其中:CM_i代表第i个识别系统的综合性能值,ω_j代表云服务第j个性能指标值的权重,max(M_j)代表多个识别系统中第j个性能指标的最大值,min(M_j)代表多个识别系统中第j个性能指标的最小值,M_{ij}代表第i个识别系统的第j个性能指标值,N代表识别系统性能指标值的总数,通过利用公式 (2*max(M_j)-M_{ij})/(2*max(M_j)-min(M_j))对M_{ij}的值进行标准化至[0,1]区间。可以看出,CM 的值越大,识别系统的综合性能越好。

[0090] 对于一些识别系统性能指标来说,例如Loss,P(Top-1 Change),MAC等,其值M_{ij}越小,CM综合性能指标值越大,这些性能指标的值可以直接代入上述公式中,而对于识别准确率等指标,其值越大,说明识别性能更好,但是直接代入公式会造成CM值的降低,这并不符合预期,所以我们需要对这些性能指标的值进行处理,使用(1-M_{ij})来替代公式中的M_{ij}值。

[0091] 决策评估模块最终会输出多角度测试模块测试的各项结果,并生成相应的测试报告表格,如表 1-3所示。由于不同任务场景的要求不同,测试系统也会给出相应的建议。

[0092] 表1.性能指标报告

[0093]

性能指标 / 模型名称	Accuracy(%)	Loss(%)	Δacc(%)	$\Delta loss$(%)	ΔS(%)	满足蜕变关系
模型名称	扩充后识别准确率	扩充后识别损失	扩充前后识别准确率差值百分比	扩充前后识别损失差值百分比	扩充前后置信分数差	(True/False)

[0094] 表2.安全性指标报告

[0095]

模型名称 / 安全性指标	模型名称
P（top-1 change）	
Crop	使用裁剪方式下的单像素扰动后发生变化的概率
Black	填充黑色方式下的单像素扰动后发生变化的概率
MAC	
Crop	使用裁剪方式下网络计算的概率的平均绝对值变化
Black	填充黑色方式下网络计算的概率的平均绝对值变化

[0096] 表3.模型稳定性指标报告以及模型综合性能报告

[0097]

性能指标 / 模型名称	CR	CM
模型名称	模型鲁棒性	模型综合性能

[0098] 以上所述仅是本发明的优选实施方式,应当指出:对于本技术领域的普通技术人员来说,在不脱离本发明原理的前提下,还可以做出若干改进和润饰,这些改进和润饰也应视为本发明的保护范围。

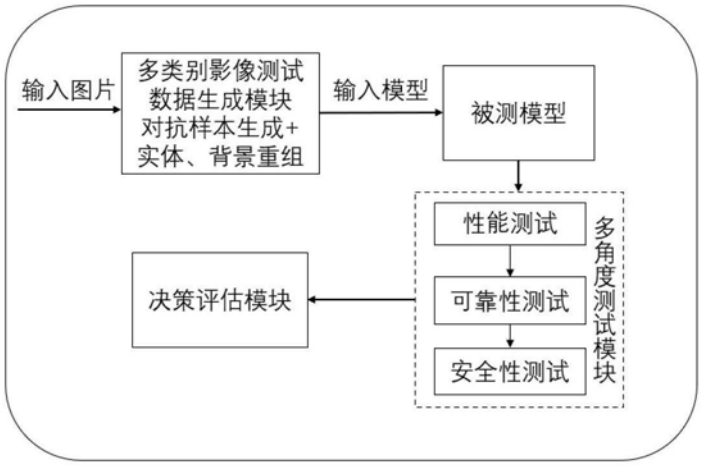

输入图片
多类别影像测试数据生成模块
对抗样本生成+实体、背景重组
输入模型
被测模型
性能测试
可靠性测试
安全性测试
多角度测试模块
决策评估模块

图1

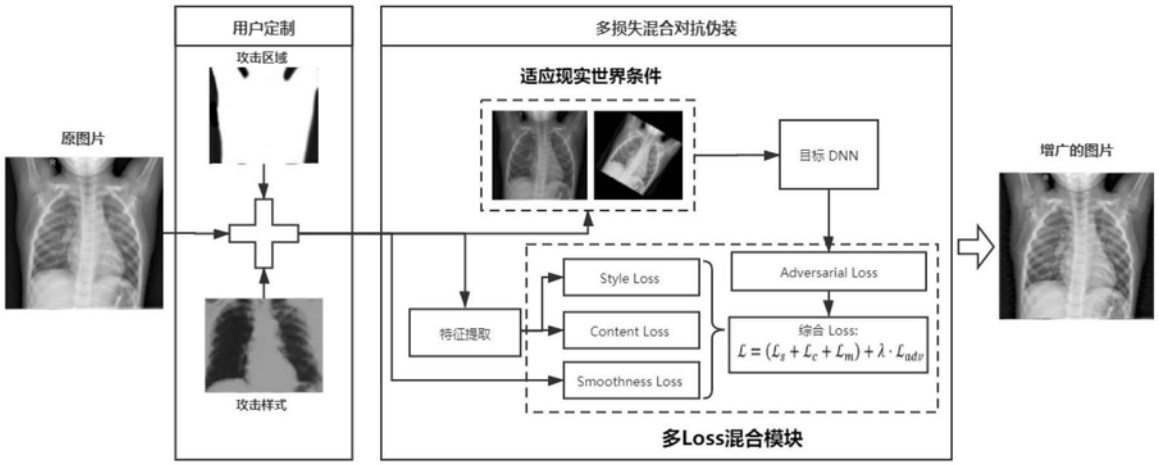

用户定制
攻击区域
原图片
攻击样式
多损失混合对抗伪装
适应现实世界条件
目标 DNN
增广的图片
Style Loss
Content Loss
Smoothness Loss
特征提取
Adversarial Loss
综合 Loss:
$\mathcal{L} = (L_s + L_c + L_m) + \lambda \cdot L_{adv}$
多Loss混合模块

图2

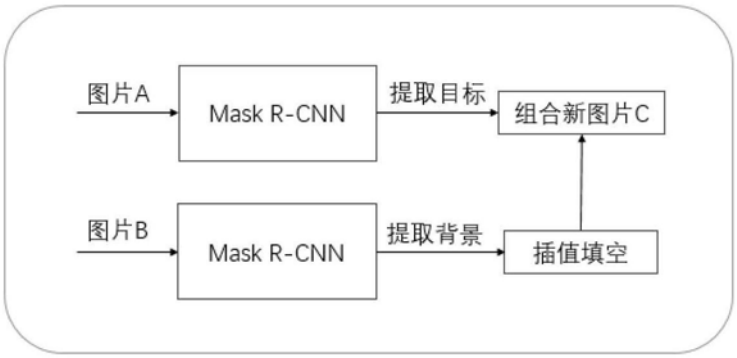

图片A
Mask R-CNN
提取目标
组合新图片C
图片B
Mask R-CNN
提取背景
插值填空

图3

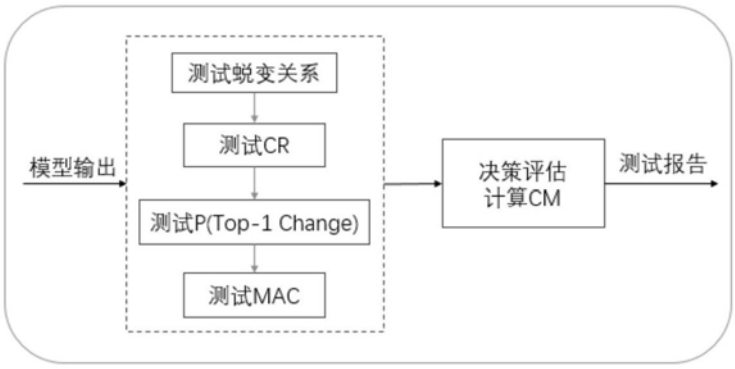


图4