基于 Grad-CAM 的视觉解释方法探索与实现

刘欣桐

摘要

在过去的十年中，卷积神经网络（CNN）模型在解决复杂的视觉问题方面取得了很大的成功。然而，考虑到缺乏对其内部功能的理解，这些深度模型被视为" 黑箱" 方法。近来，人们对开发可解释的深度学习模型产生了浓厚的兴趣，基于最近提出的称为 CAM 的方法，我们提出了一种称为 Grad-CAM 的广义方法，与现有技术相比，Grad-CAM 适用于各种 CNN 模型系列：（1）具有全连接层（例如 VGG）的 CNN，（2）用于结构化输出（例如字幕）的 CNN，（3）在具有多模态输入（例如视觉问题回答）或强化学习的任务中使用的 CNN，而无需架构改变或再训练。将 Grad-CAM 与现有的细粒度可视化相结合，创建一个高分辨率的类区分可视化 Guided Grad-CAM，并将其应用于图像分类、图像字幕和可视化问答（VQA）模型，包括基于 ResNet 的架构。在此基础上，通过 Grad-CAM++ 方法以及 Smooth Grad-CAM++ 方法对现有效果进行了改进，可以在更好的对象定位以及解释单个图像中多个对象实例的发生方面提供 CNN 模型预测的更好的视觉解释。

关键词：Interpretable ML, Convolutional Neural Networks, Computer Vision, 视觉解释

1 引言

近年来，基于卷积神经网络（Convolution Neural Network，CNN）的深度神经模型在各种计算机视觉任务实现了前所未有的突破。然而，现有的方法几乎都将 CNN 模型作为黑盒使用，缺乏“可解释性”这一痛点严重制约了其在医疗等涉及公众安全的关键领域应用。如果没有对预测背后的底层机制进行推理，深层模型就无法得到完全信任，这大大阻碍深度模型在与公平性、隐私性和安全性有关的关键应用程序中使用。为了构建安全、可信地部署深度模型，需要同时提供准确的预测和人类能领会的解释，特别是对于跨学科领域的用户，探究深度学习的可解释性并建立起“透明”的模型是十分必要的。从广义上讲，这种透明性在人工智能（AI）进化的三个不同阶段都是有用的。首先，当人工智能明显弱于人类，并且还不能可靠地“部署”时，透明度和解释的目标是识别故障模式，从而帮助研究人员将精力集中在最有成效的研究方向上。第二，当人工智能与人类不相上下，并且可靠地“可部署”（例如，图像分类是根据充分的数据训练的一组类别），目标是在用户中建立适当的信任和信心。第三，当人工智能明显强于人类时，解释的目标是机器教学。即：教人类如何做出更好决策的机器。我们试图揭示卷积神经网络分类模型中图像的空间特征与其类别权重之间的联系，利用热力图定位对分类结果起激活作用的图像区域。通过构建具有可解释性的模型，能够揭示模型做出该预测决策的具体理由，提高其在关键领域的应用范围。

2 相关工作

2.1 视觉可解释性

神经网络可解释性是指对于神经网络所做出的决策，进行合理的解释。这里的解释可以是从数学理论层面进行的先验解释，比如对于激活函数的差异分析、模型的泛化能力分析，也可以是对于网络

预测结果的后验解释，比如我们训练好的模型将一张图片分类为” 猫”，我们希望知道网络是通过什么因素或特征将它分类为” 猫” 这个类别的。本文关注的是后验解释，即解释已有模型的决策。而对于卷积神经网络，目前最常见的是通过可视化的方式来解释模型的决策。许多先前的工作通过突出显示“重要”像素（即这些像素的强度变化对预测的得分影响最大）来可视化 CNN 预测。具体而言，Simonyan 等人可视化了预测类别得分的偏导数 w.r.t. 像素强度，而引导反向传播和去卷积对原始梯度进行修改，从而导致定性改进。这些方法在后续工作中进行了比较。尽管生成了细粒度的可视化，但这些方法不是类区分的。

2.2 主流的视觉可解释性方法

2.2.1 基于梯度/特征的可解释性方法

采用梯度或特征来解释深度模型是最直接的解决方案，在图像和文本任务中被广泛使用。其关键思想是将梯度或隐藏的特征图值作为输入重要性的近似值。一般来说，在这类方法中，梯度或特征值越大，表示重要性越高。需要注意的是，梯度和隐藏特征都与模型参数高度相关，那么这样的解释可以反映出模型所包含的信息。经典方法例如：SA、Guided-BP 等。这些方法的关键区别在于梯度反向传播的过程以及如何将不同的隐藏特征图结合起来。Guided-Backpropagation 目前在可解释性方面存在较大争议，它是通过对于回传梯度进行一定过滤，从而得到更为干净和聚焦的可视化结果。然而在 Sanity Check 中，作者发现，该方法与模型参数无关，即初始化模型参数后，仍然能得到相似结果。

2.2.2 基于扰动的可解释性方法

基于扰动的方法采用一种全局的方式来定位图像中对于决策更重要的区域，被广泛用于解释深度图像模型。此类方法的做法符合人类直觉，通过将图像上部分的移除或者保留，来直接衡量该区域对于网络决策分数的影响。比如在一张” 猫” 的图片中，背景信息的移除通常不会降低” 猫” 类别上的置信度，而一些关键区域，如” 猫” 的耳朵，在移除后则可能造成置信度的下降。其根本动机是研究不同输入扰动下的输出变化。当重要的输入信息被保留（没有被扰动）时，预测结果应该与原始预测结果相似。基于扰动的方法采用不同的掩码生成算法来获得不同类型的掩码。需要注意的是，掩码可以对应节点、边或节点特征。经典基于扰动的方法，包括：GNN-Explainer、PG-Explainer、ZORRO 等。直观地讲，掩码捕捉到的重要输入特征应该传达关键的语义意义，从而得到与原图相似的预测结果。这些方法的区别主要在于三个方面：掩码生成算法、掩码类型和目标函数。然而，这类方法存在一个明显的劣势，即如何生成掩码。目前主要有两种思路，一种是通过采样的方式（比如随机采样或蒙特卡洛采样）生成多个掩码，然后来计算每一个掩码区域的重要性，这种方式往往需要生成大量掩码，计算量很大；第二种是通过优化的方式来生成掩码，初始化一个随机掩码，通过优化损失函数来不断更新掩码，这种方法的缺点是需要在损失函数中增加额外的正则化项，来使得生成的掩码面积尽可能的小，同时还能尽可能多的影响决策分数，由于存在优化过程，即使对于同一张图，方法每一次生成的解释都是不完全一致的。因此，这种方法不能直接应用于图模型，因为图数据是以节点和边来表示的，它们不能调整大小以共享相同的节点和边数，结构信息对图来说至关重要，可以决定图的功能。

2.2.3 基于类激活图的可解释性

这一类最为经典的为 Class Activation Mapping（CAM）方法。CAM 将最后一层的节点特征映射到输入空间，从而识别重要节点。它要求模型必须采用全局平均池化层和全连接层作为最终分类器。CAM 将最终的节点嵌入，通过加权求和的方式组合不同的特征图，从而获得输入节点的重要性分数。权重是从与目标预测连接的最终全连接层获得的。该方法非常简单高效，但仍有几大限制：1) CAM 对 CNN 结构有特殊要求，限制了它的应用和推广。2) 它假设最终的节点嵌入可以反映输入的重要性，这是启发式的，通常只激活有区别的对象区域，并且错误地包含许多与对象相关的背景。3) WSSS(弱监督语义分割) 模型只有固定的图像级别的标签，因此很难抑制激活目标对象会激活出的不同背景区域。

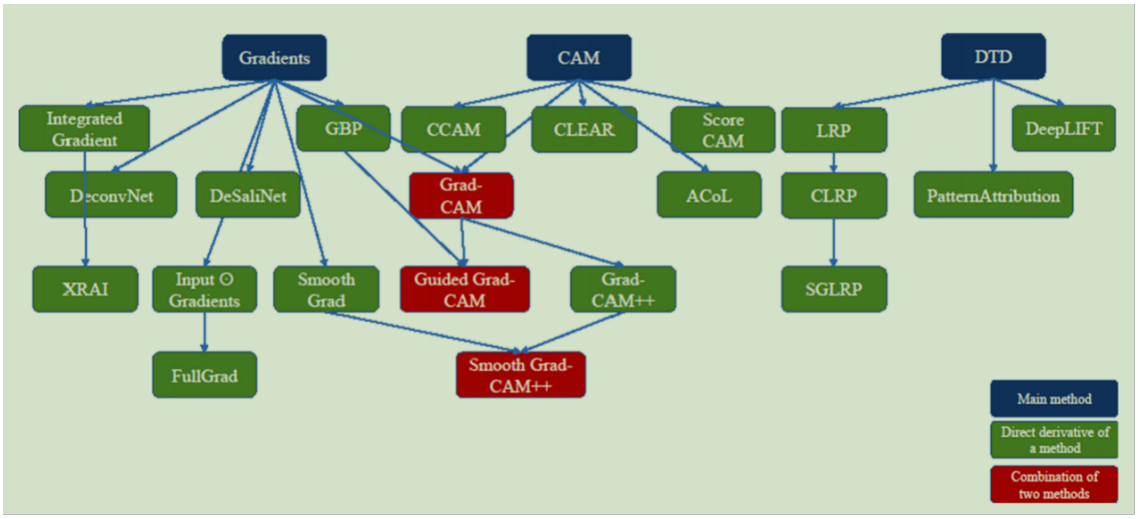

图 1: 主流的视觉可解释性方法

3 本文方法

3.1 Grad-CAM 方法概述

CAM 揭示了卷积神经网络分类模型中图像的空间特征与其类别权重之间的联系。然而，CAM 只适用于模型中有全局平均池化层并且只有一个全连接层的情形，如 ResNet，MobileNet 等。由于 CAM 算法中生成类激活图所需要的类别权重，即为全局平均池化层和全连接输出层之间的，对应着图片类别的权重。对于 VggNet，DenseNet 等有着多个全连接层的模型，CAM 则不再适用，因为无法获取到类别权重。为了解决这一问题，Grad-CAM 应运而生。

Grad-CAM 通过去除全局平均池化层的约束，基本思路是目标特征图的融合权重可以表达为梯度。同样，它也将最终的节点嵌入映射到输入空间来衡量节点重要性。但是，它没有使用全局平均池化输出和全连接层输出之间的权重，而是采用梯度作为权重来组合不同的特征图。Grad-CAM 总结下来就是下面这个公式：

$$L^{c}_{Grad-CAM} = ReLU\left(\sum_{k} \alpha^{c}_{k} A^{k}\right) \tag{1}$$

其中:

A 代表某个特征层，在论文中一般指的是最后一个卷积层输出的特征层

k 代表特征层 A 中第 k 个通道 (channel)

c 代表类别 c

A^k 代表特征层A 中通道k 的数据

α_k^c 代表针对 A^k 的权重

关于 α_k^c 的计算公式如下：

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{2}$$

其中：

y^c 代表网络针对类别 c 预测的分数 (score)，注意这里没有通过 softmax 激活

A_{ij} 代表特征层 A 在通道 k 中，坐标为 ij 位置处的数据

Z 等于特征图的宽度乘高度

通过计算公式 (2) 可知 α_k^c 就是通过预测类别 c 的预测分数 y^c 进行反向传播，然后利用反传到特征层 A 上的梯度信息计算特征层 A 每个通道 k 的重要程度。

接着通过对特征层 α_A 每个通道的数据进行加权求和，最后通过 RELU 激活函数得到 Grad-CAM（论文中说使用 ReLU 是为了过滤掉 Negative pixles，而 Negative pixles 很可能是归属于其他类别的 pixles）。Grad-CAM 方法示意图如下所示：

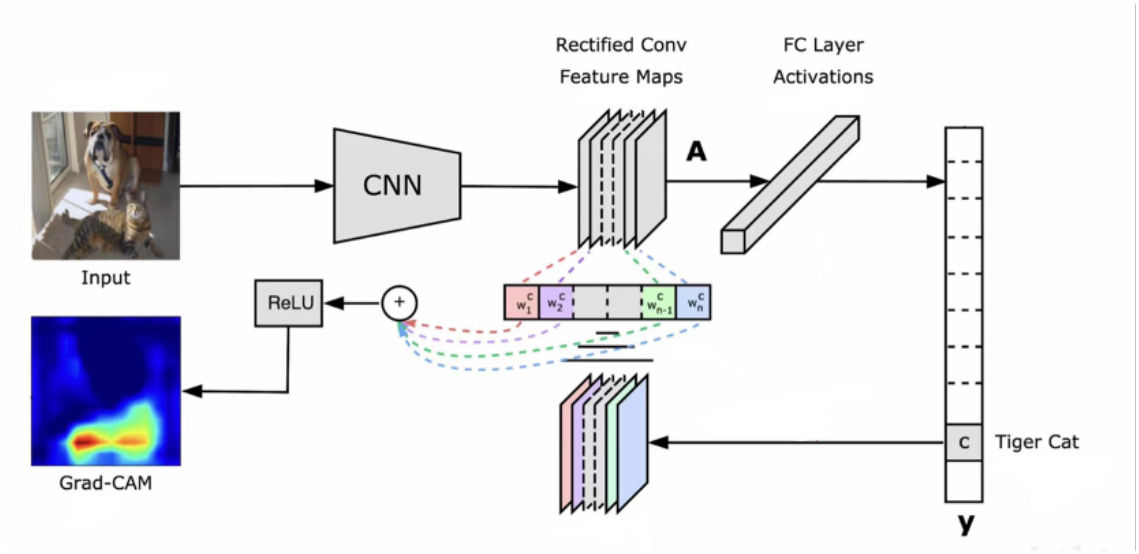

图 2: Grad-CAM 方法示意图

Grad-CAM 允许我们从基于 CNN 的模型中生成可视化的解释，其实并不一定要是分类问题，只要是可求导的激活函数，在其他问题也一样使用 Grad-CAM。这些模型将具有复杂得多的交互作用的卷积层级联起来。事实上，我们将 Grad-CAM 应用于“超越分类”的任务，包括利用 CNN 进行图像字幕和视觉问题回答（VQA）的模型。综上所述，Grad-CAM 可视化流程：

1) 输入：给定的一个图像和一个感兴趣的类别（例如：tiger cat）；

2) 通过模型的 CNN 部分进行向前传播，得到特定任务的各类别分数 ysoftmax 层之前）；

3) 将给定的类别（tiger cat）设置为 1，其他类别的梯度都设置为 0 ；

4) 将给定类别分数 yc 反向传播至卷积特征图，组合计算得到粗糙的梯度 CAM 定位（蓝色热力图）；

5) 将热力图与反向传播的结果进行点乘，得到高分辨率的特定 Grad-CAM 可视化图；

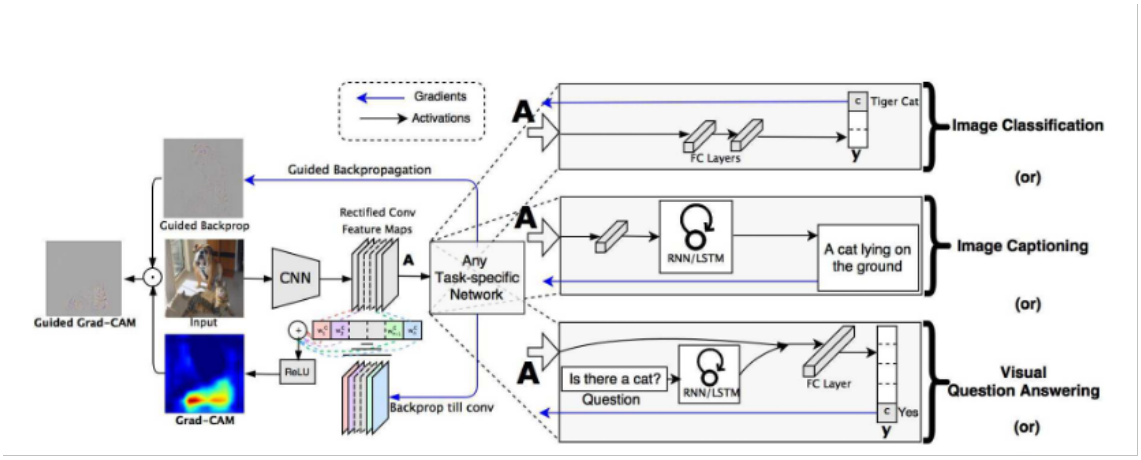

图 3: Grad-CAM 概述

3.2 Guided Grad-CAM 方法概述

尽管 Grad-CAM 可视化具有类区分性并且能够很好地定位相关图像区域，该方法缺乏像像素空间梯度可视化方法（引导反向传播和去卷积）显示细粒度重要性的能力。为了结合类判别性和高分辨率两个优点，通过逐点乘法融合了引导反向传播和 Grad-CAM 可视化这种可视化是高分辨率的（当感兴趣的类别是“老虎猫”时，它识别出重要的“老虎猫”特征，如条纹、尖耳朵和眼睛）和类别区分性的（它显示“老虎猫”，但不显示“拳击手（狗）”）。值得说明的是，使用过程中用反卷积替换引导反向传播能够得到类似的结果，但是发现反卷积有伪像（并且引导反向传播可视化通常噪音更小），所以这里选择引导反向传播而不是反卷积。

4 复现细节

4.1 与已有开源代码对比

复现过程参考了官方开源代码（https://github.com/ramprs/grad-cam/）官方在 torch 上进行实验，本次复现工作基于 pytorch 进行了重写与改进。

4.1.1 核心方法的重写：

方法核心思想仅需要对目标类别的 score 进行求导，然后追踪到目标特征图的梯度，对该梯队进行 element-wise 求平均 (GAP 操作) 即获得特征融合的权重。具体如下:

```
利用onehot的形式锁定目标类别
one_hot = np.zeros((1, output.size()[-1]), dtype=np.float32)
one_hot[0][index] = 1
one_hot = torch.from_numpy(one_hot).requires_grad_(True)
获取目标类别的输出,该值带有梯度链接关系,可进行求导操作
one_hot = torch.sum(one_hot * output)
self.model.zero_grad()
one_hot.backward(retain_graph=True) # backward 求导
获取对应特征层的梯度map
grads_val = self.extractor.get_gradients()[-1].cpu().data.numpy()
target = features[-1].cpu().data.numpy()[0, :] # 获取目标特征输出
weights = np.mean(grads_val, axis=(2, 3))[0, :] # 利用GAP操作, 获取特征权重
```

```
    cam = weights.dot(target.reshape((nc, h * w)))
    relu操作,去除负值, 并缩放到原图尺寸
    cam = np.maximum(cam, 0)
    cam = cv2.resize(cam, input.shape[2:])
    归一化操作
    batch_cams = self._normalize(batch_cams)
```

4.1.2 生成 cam 方法重写：

```
    def gen_cam(image, mask):
    """
    生成CAM图
    :param image: [H,W,C],原始图像
    :param mask: [H,W],范围0~1
    :return: tuple(cam,heatmap)
    """
    # mask转为heatmap
    heatmap = cv2.applyColorMap(np.uint8(255 * mask), cv2.COLORMAP_JET)
    heatmap = np.float32(heatmap) / 255
    heatmap = heatmap[..., ::-1] # gbr to rgb

    # 合并heatmap到原始图像
    cam = heatmap + np.float32(image)
    return norm_image(cam), heatmap
```

4.1.3 改进策略一：Grad-CAM++

为了优化 Grad-CAM 的结果，定位会更精准，也更适用于目标类别物体在图像中不止一个的情况。Grad-CAM 是利用目标特征图的梯度求平均 (GAP) 获取特征图权重，可以看做梯度 map 上每一个元素的贡献是一样。而本文认为梯度 map 上的每一个元素的贡献不同，因此增加了一个额外的权重对梯度 map 上的元素进行加权。将权重参数 $\alpha_{i,j}^{kc}$ 进行了如下改进：

$$\alpha_{i,j}^{kc} = \frac{\frac{\partial^2 Y^c}{\left(\partial A_{i,j}^k\right)^2}}{2\frac{\partial^2 Y^c}{\left(\partial A_{i,j}^k\right)^2} + \sum_a \sum_b A_{a,b}^k \frac{\partial^3 Y^c}{\left(\partial A_{i,j}^k\right)^3}} \tag{3}$$

核心代码：

```
# 获取特征权重的过程
# 反向传播
self._backprop(scores, class_idx)
# 注意,这里用乘法,因为论文中将2次偏导和3次偏导进行了2次方和3次方的转化
grad_2 = self.hook_g.pow(2)
grad_3 = self.hook_g.pow(3)
# 获取alpha权重map
alpha = grad_2 / (2 * grad_2 + (grad_3 * self.hook_a).sum(axis=(2, 3), keepdims
    =True))
# 利用alpha 权重map去获取特征权重
return alpha.squeeze_(0).mul_(torch.relu(self.hook_g.squeeze(0))).sum(axis=(1,
    2))
```

4.1.4 改进策略二：Smooth Grad-CAM++

这一项改进是结合 smoothGrad 来优化已有 CAM 方法的效果。Smooth Grad-CAM++ 的做法为多次输入加入随机噪声的图片，对结果并求平均，用以消除输出 saliency maps 的”噪声”，达到“引入噪声”来“消除噪声”的效果。将权重参数 $\alpha_{i,j}^{kc}$ 进行了如下改进：

$$W_k^c = \sum_i \sum_j \alpha_{i,j}^{kc} \text{ ReLU}\left(\frac{1}{n}\sum_1^n D_1^k\right) \tag{4}$$

Smooth Grad-CAM++ 的伪代码可描述如下：

Procedure 1 Algorithm 1: Smooth Grad-CAM++ algorithm

Input: Image X_0, Model , $f(X)$, class c, layer l

Output: $L_{Grad-CAM++}^{c}l$

// get activation of layer l;

$A_l \leftarrow f_l(X)$

$M_c(x) = \frac{1}{n}\sum_1^n M_c\left(x + \mathcal{N}\left(0, \sigma^2\right)\right)$

$C \leftarrow$ the number of channels in A_l

for *i* ***in*** *range(1,3)* **do**

$D_i^k \leftarrow$ gradient $\left(A_l^k\right)$

end

$$\alpha_{i,j}^{kc} = \frac{\frac{1}{n}\sum_1^n D_1^k}{2\frac{1}{n}\sum_1^n D_2^k + \sum_a \sum_b A_{a,b}^k \frac{1}{n}\sum_1^n D_3^k}$$

$$W_k^c = \sum_i \sum_j \alpha_{i,j}^{kc} \text{ ReLU}\left(\frac{1}{n}\sum_1^n D_1^k\right)$$

$$L_{SmoothGrad-CAM++}^{c} = \text{ReLU}\left(\sum_k W_k^c A^k\right)$$

核心代码：

```
for i in range(self.n_samples): # 进行n_samples次加噪声操作
self.model.zero_grad()
# 输入图片增加高斯噪声
x_with_noise = torch.normal(mean=x, std=std_tensor).requires_grad_()
score = self.model(x_with_noise)
score[0, idx].backward(retain_graph=True) # 求梯度
activations = self.values.activations
gradients = self.values.gradients
n, c, _, _ = gradients.shape
# 获取alpha, 和grad-cam++一致
numerator = gradients.pow(2)
denominator = 2 * gradients.pow(2)
ag = activations * gradients.pow(3)
denominator += \
ag.view(n, c, -1).sum(-1, keepdim=True).view(n, c, 1, 1)
denominator = torch.where(
   denominator != 0.0, denominator, torch.ones_like(denominator))
alpha = numerator / (denominator + 1e-7)
relu_grad = F.relu(score[0, idx].exp() * gradients)
# 获取weights
weights = (alpha * relu_grad).view(n, c, -1).sum(-1).view(n, c, 1, 1)
# 对特征层加权融合, 并进行relu+归一化操作
cam = (weights * activations).sum(1, keepdim=True)
```

```
    cam = F.relu(cam)
    cam -= torch.min(cam)
    cam /= torch.max(cam)
    total_cams += cam
total_cams /= self.n_samples # 求平均操作
return total_cams.data
```

4.2 实验环境搭建

4.2.1 依赖

python 3.6.x

pytoch 1.0.1+

torchvision 0.2.2

opencv-python

matplotlib

scikit-image

numpy

4.2.2 使用说明

python main.py

–image-path examples/pic1.jpg

–network densenet121

–weight-path /opt/pretrainedmodel/densenet121-a639ec97.pth

4.2.3 参数说明

image-path：需要可视化的图像路径 (可选, 默认./examples/pic1.jpg)

network: 网络名称 (可选, 默认 resnet50)

weight-path: 网络对应的与训练参数权重路径 (可选, 默认从 pytorch 官网下载对应的预训练权重)

layer-name: Grad-CAM 使用的层名 (可选, 默认最后一个卷积层)

class-id：Grad-CAM 和 Guided Back Propagation 反向传播使用的类别 id（可选, 默认网络预测的类别)

output-dir：可视化结果图像保存目录 (可选，默认 results 目录)

数据集使用：ILSVRC2012 数据集（ImageNet-1k）

4.3 创新点

Grad-CAM 弥补了开山之作 CAM 方法的以下三方面缺陷：1) GAP 层必须替换整个连接层 2) 只能分析卷积层最后一层的输出，但不能分析中间层 3) 仅适用于图像分类任务。实现了任何中间层都可以进行分析，与此同时带来了更广泛的应用场合。但是在实验过程中仍然发现以下两点存在改进空

间：1）定位同一类别多个对象时性能下降 2）对于单个对象图像，Grad-CAM 热图通常不能完整地捕捉整个对象。对此提出了两种改进策略进行完善和效果提升。

Grad-CAM++ 创新思想在于，在求每个特征图的权重参数时，是将该图片上所有的像素点做一个全局平局。容易理解的是，通常一张图片当中往往是图片靠近中间的位置有用信息量更多，而四角或者边缘位置有用信息数量往往是比较少的，而直接将所有像素点一并处理效果必然有待提升。此方法可简单的理解为通过对特征图求一节、二阶甚至三节导数的方式，经过较为复杂的公式对像素点的权重进行了刻画。其主要的改进效果是定位更准确，更适合同类多目标的情况。

Smooth Grad-CAM++ 创新思想在于结合梯度平滑策略。在显示与图像分类结果相关区域过程中，往往存在许多噪声，且很难探究噪声的组成。smoothGrad 的做法为多次输入加入随机噪声的图片，对结果并求平均，用以消除输出 saliency maps 的”噪声”，达到“引入噪声”来“消除噪声”的效果。策略的结合使得可视化效果大大提升。

5 实验结果分析

5.1 图像分类

本次实验在 ImageNet-1K 数据集上使用 resnet50 网络模型，网络与对应的训练参数权重路径采取 pytorch 官网预训练权重参数，可视化类别为 231 的图像。Grad-CAM 在图像分类任务可视化效果如图四所示。

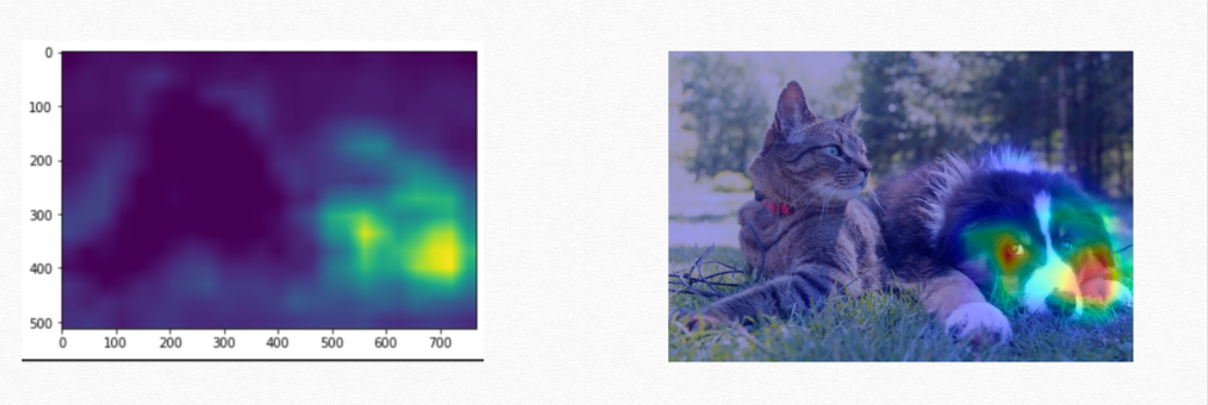

图 4: Grad-CAM 视觉解释

为进一步提高可视化效果，在实现具有类判别性的 Grad-CAM 方法基础上，将矩阵与反卷积做点乘，生成既具有类判别性又具有高分辨率的 Guided Grad-CAM。效果如图五所示。

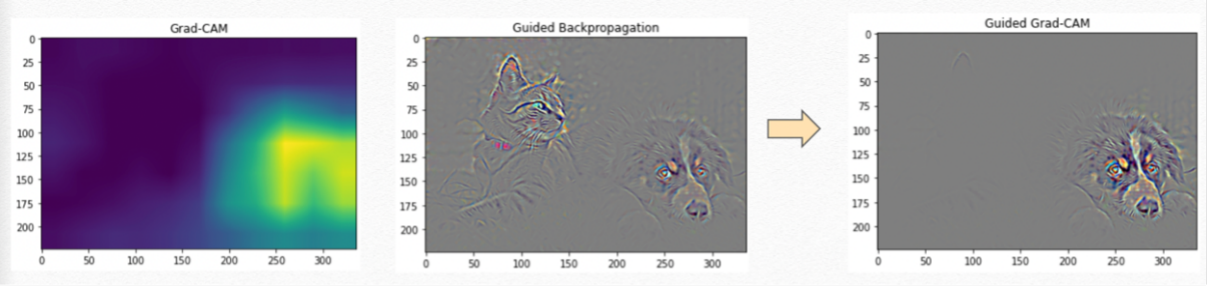

图 5: Guided Grad-CAM 视觉解释

在完成论文复现主要工作的基础上，对现有方法进行了两点改进，即：Grad-CAM++ 和 Smooth Grad-CAM++ 方法。为保证比较的全面以及详细，此处我们共比较了七种网络模型（vgg16, vgg19, resnet50, resnet101, densenet121, inceptionv3、mobilenetv2 Visual interpretation），四种基于梯度的视觉

解释方法（Grad-CAM、Grad-CAM++、Smooth Grad-CAM++、Guided Grad-CAM）。分别在单个对象、多个对象两种情况下展开比较。效果如图六、七所示：

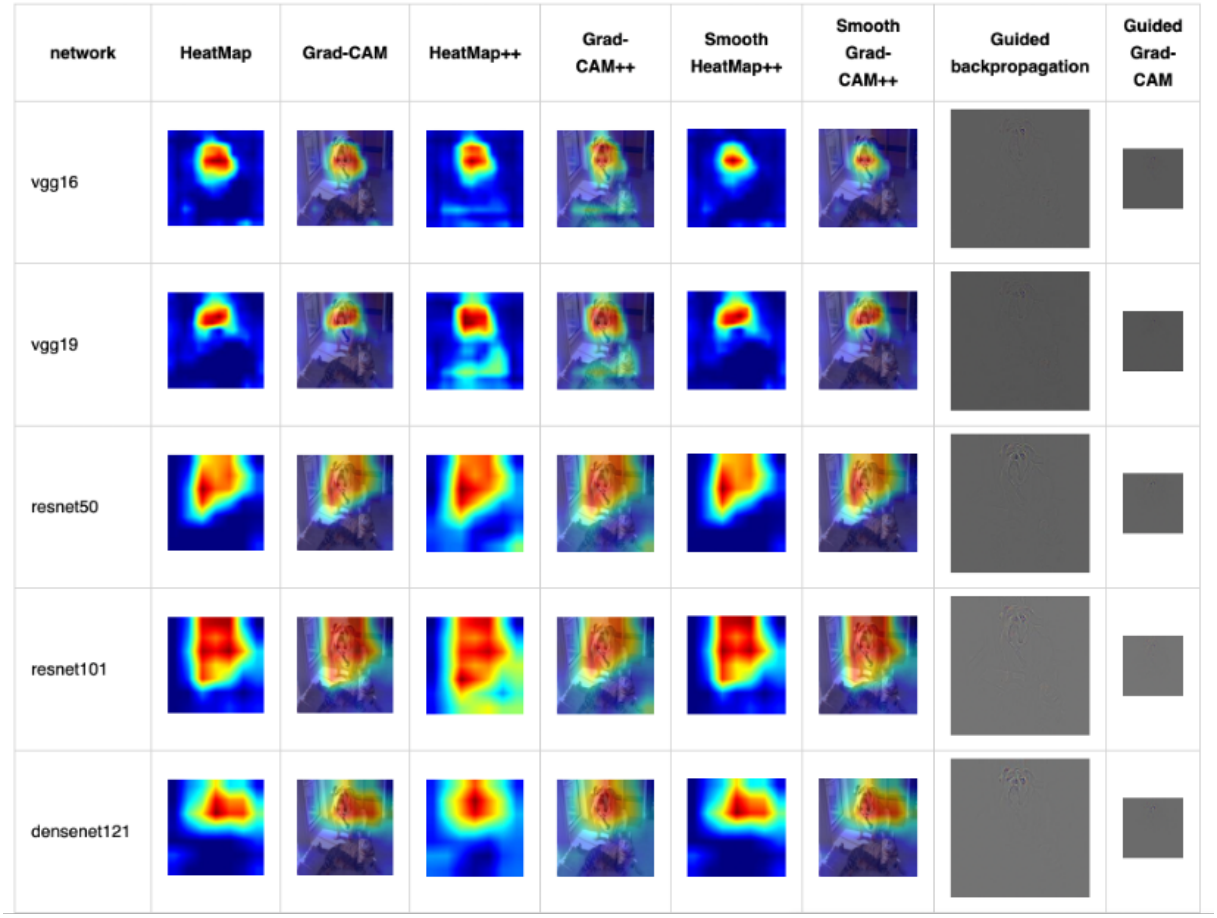

图 6: 不同模型单个对象比较效果

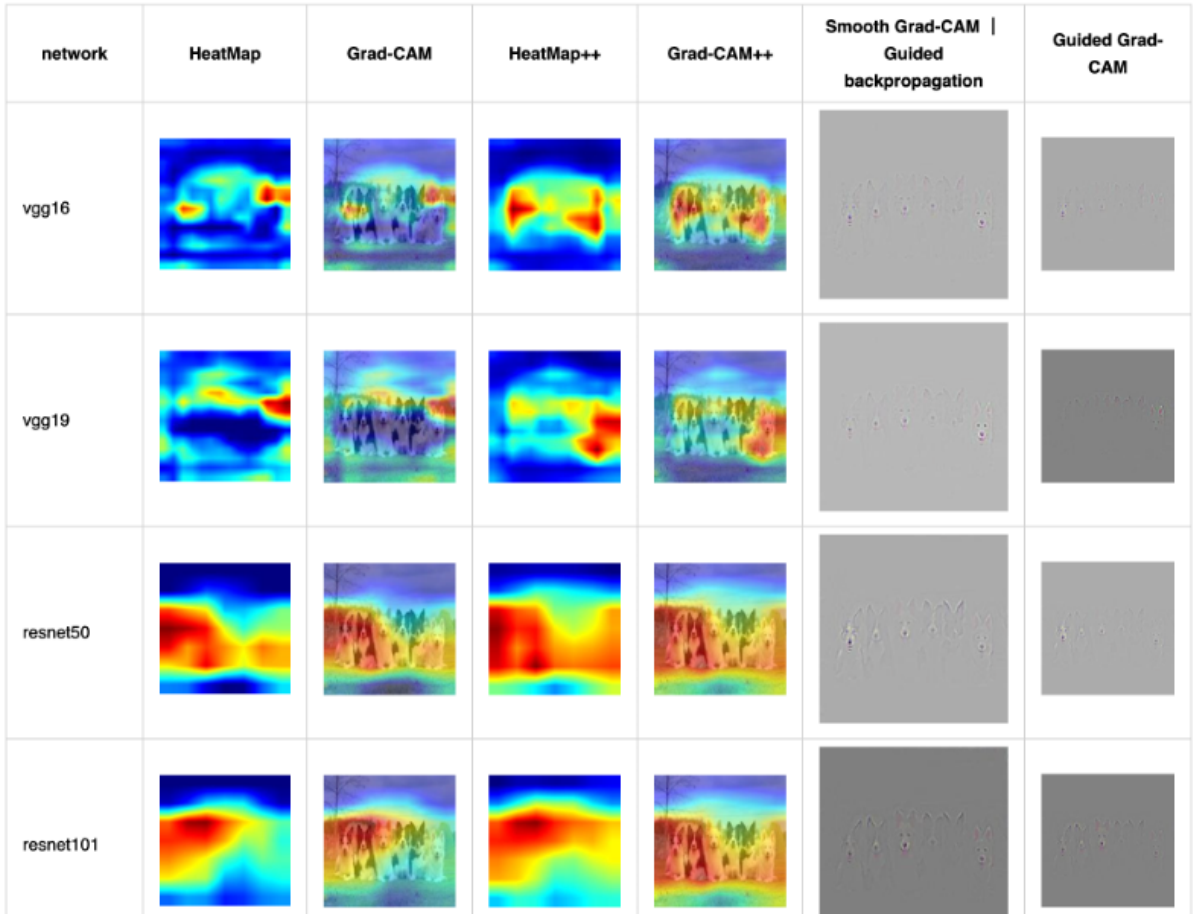

图 7: 不同模型多个对象比较效果

通过横向对于方法模型的比较和纵向对于不同网络的比较，可以得到以下几点结论：

1. vgg 模型的 Grad-CAM 并没有覆盖整个对象, 相对来说 resnet 和 denset 覆盖更全, 特别是 densenet; 从侧面说明就模型的泛化和鲁棒性而言 densenet>resnet>vgg

2. Grad-CAM++ 相对于 Grad-CAM 也是覆盖对象更全面，特别是对于同一个类别有多个实例的情况下,Grad-CAM 可能只覆盖部分对象，Grad-CAM++ 基本覆盖所有对象; 但是这仅仅对于 vgg 而言, 想 densenet 直接使用 Grad-CAM 也基本能够覆盖所有对象 MobileNet V2 的 Grad-CAM 覆盖也很全面

3. Inception V3 和 MobileNet V2 的 Guided backpropagation 图轮廓很模糊，但是 ShuffleNet V2 的轮廓则比较清晰

建立“透明”模型，定位显示出对图像分类结果做出依据的部分在实际应用中具有重要意义。本次实验以识别数据集中的偏见为例，原始数据集为一个男性医生数量远远大于女性医生数量的偏见数据集，分类任务为实现数据集对医生和护士两种职业的分类。通过中间热力图标记的范围可以看出，未调整的偏见数据集作出分类的主要依据在于性别，显著标记更多的为头发等区分男性女性的图像区域。而了解到此点，在调整过后的数据集上进行实验可以看出右侧图像中，正确解释应该为通过识别工作人员手中的工具等作为主要依据进行两种职业的区分，而并非性别。

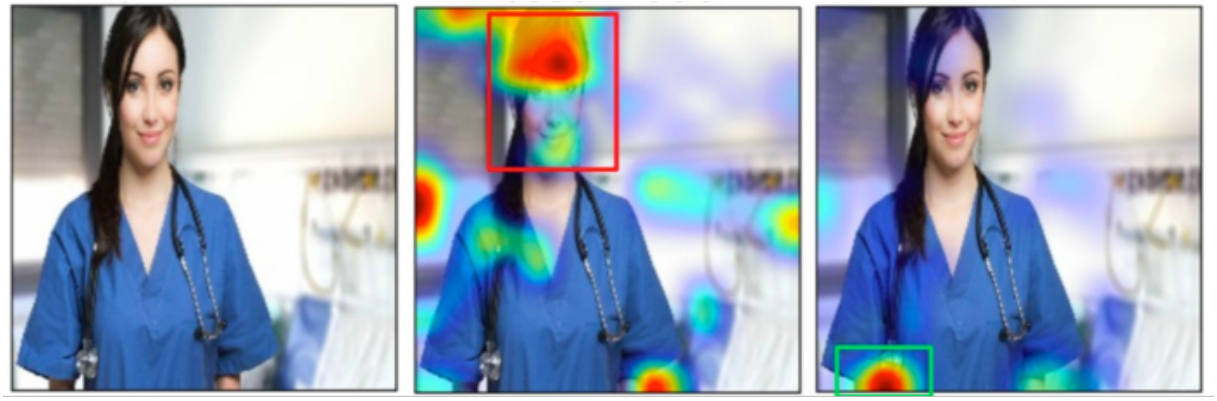

图 8: 识别偏见数据集

5.2 VQA

Grad-CAM 具有广泛的应用场景，在 VQA 场景上也有所体现。如下图所示，我们建立一个视觉问答，想要询问消防栓的颜色，可得到不同回答，但每种回答都可以给出做出此项决策的依据区域。效果如图九、十所示：

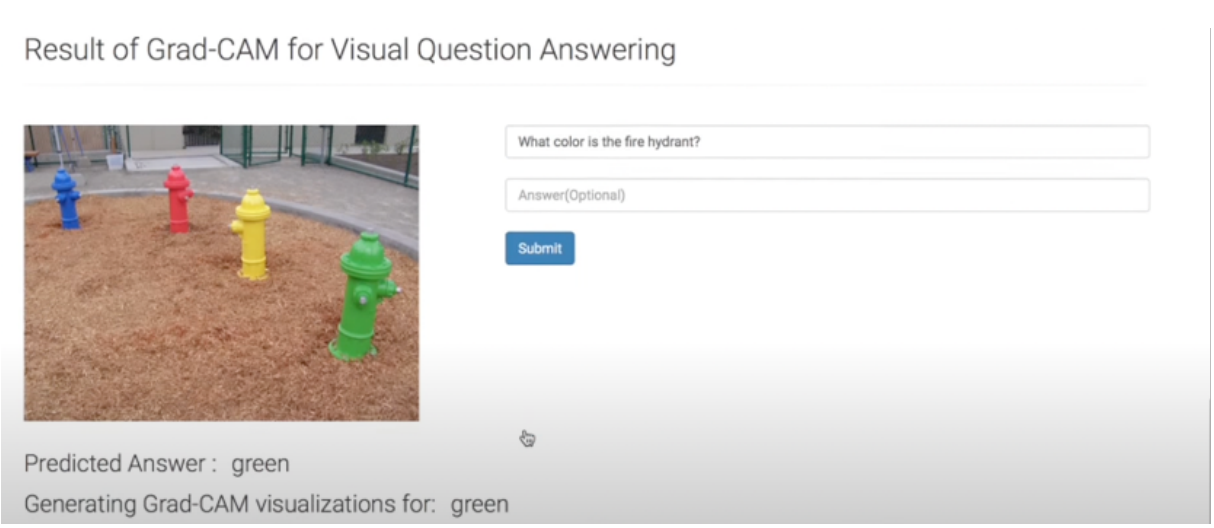

图 9: VQA 视觉问答

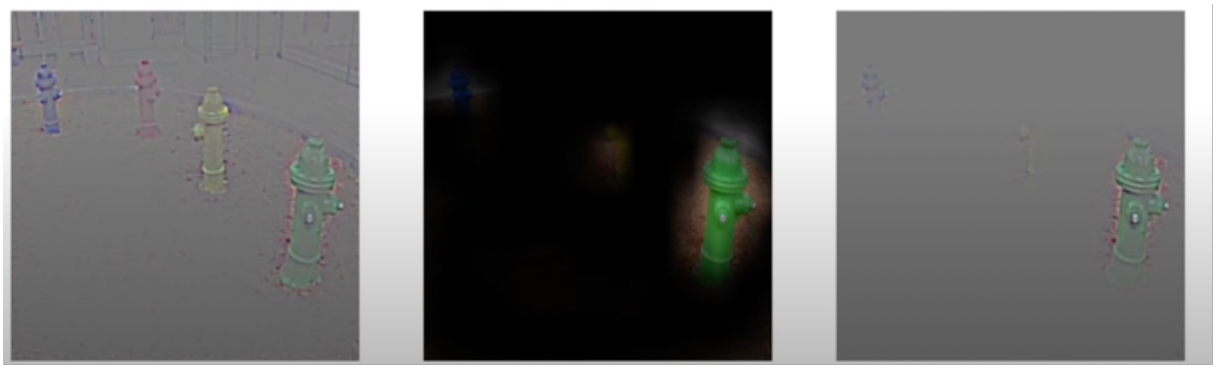

图 10: VQA 回答具有可解释性

6 总结与展望

本次工作中，复现了一种新的类区分定位技术-梯度加权类激活映射（Grad-CAM）-通过产生视觉解释使任何基于神经网络的模型更加透明。在此基础上，将 Grad-CAM 定位功能与现有的高分辨率可视化相结合，以获得既有高分辨率又具备类别区分的 Guided Grad-CAM 方法。Grad-CAM 在以下两个方面优于所有现有方法：一方面彰显可解释性，另一方面体现在对原始模型的忠实性。广泛的人类研究表明，此种方法可以更准确地区分类别，更好地揭示分类器的可信度，并帮助识别数据集中的偏见。除此之外，尝试了 Grad-CAM 对各种现成架构的广泛适用性，包括图像分类和 VQA，为可能的模型决策提供忠实的视觉解释。最后，我用本次复现的内容在自己的研究课题上进行了实验和测试，在先前对血细胞分类模型训练好的模型上进行了视觉解释，从而定位出病灶的位置，大大推进了后续的研究工作。

我认为，一个真正的人工智能系统不仅应该是智能的，还应该能够推理其信念和行为，让人类信任它。本次课程尝试的模型按照方法分类来讲，均属于基于梯度/特征的可解释性方法，然而基于梯度的方法往往会遭受梯度饱和的问题，这导致反向传播梯度减小，并不利地影响可视化的质量。未来的工作可以关注 grad-free 相关方法的思想与质量，例如：Score CAM、SS CAM、Ablation-CAM 等方法。除此之外，未来的工作希望尽可能推进解释算法扩展到解释由其他神经网络架构（如递归神经网络、长短时记忆网络和生成对抗网络）所做的决策的可能性。

参考文献

[1] Selvaraju R R, Das A, Vedantam R, et al. Grad-CAM: Why did you say that?[J]. arXiv preprint arXiv:1611.07450, 2016.

[2] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE international conference on computer vision. 2017: 618-626.

[3] Chen L, Chen J, Hajimirsadeghi H, et al. Adapting grad-cam for embedding networks[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020: 2794-2803.

[4] Fu R, Hu Q, Dong X, et al. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns[J]. arXiv preprint arXiv:2008.02312, 2020.

[5] Panwar H, Gupta P K, Siddiqui M K, et al. A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images[J]. Chaos, Solitons Fractals, 2020, 140: 110190.

[6] Lee S, Lee J, Lee J, et al. Robust tumor localization with pyramid grad-cam[J]. arXiv preprint arXiv:1805.11393, 2018.

[7] Zhang Y, Hong D, McClement D, et al. Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging[J]. Journal of Neuroscience Methods, 2021, 353: 109098.

[8] Joo H T, Kim K J. Visualization of deep reinforcement learning using grad-CAM: how AI plays atari games?[C]//2019 IEEE Conference on Games (CoG). IEEE, 2019: 1-2.

[9] Kim J Y, Kim J M. Bearing fault diagnosis using grad-CAM and acoustic emission signals[J]. Applied Sciences, 2020, 10(6): 2050.

[10] He T, Guo J, Chen N, et al. MediMLP: using Grad-CAM to extract crucial variables for lung cancer postoperative complication prediction[J]. IEEE journal of biomedical and health informatics, 2019, 24(6): 1762-1771.

[11] Choi J, Choi J, Rhee W. Interpreting neural ranking models using grad-cam[J]. arXiv preprint arXiv:2005.05768, 2020.

[12] Moujahid H, Cherradi B, Al-Sarem M, et al. Combining CNN and Grad-Cam for COVID-19 Disease Prediction and Visual Explanation[J]. Intelligent Automation Soft Computing, 2022, 32(2).

[13] Jahmunah V, Ng E Y K, Tan R S, et al. Explainable detection of myocardial infarction using deep learning models with Grad-CAM technique on ECG signals[J]. Computers in Biology and Medicine, 2022, 146: 105550.

[14] Li Y, Yang H, Li J, et al. EEG-based intention recognition with deep recurrent-convolution neural network: Performance and channel selection by Grad-CAM[J]. Neurocomputing, 2020, 415: 225-233.

[15] Jiang H, Xu J, Shi R, et al. A multi-label deep learning model with interpretable grad-CAM for diabetic retinopathy classification[C]//2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC). IEEE, 2020: 1560-1563.

[16] Morbidelli P, Carrera D, Rossi B, et al. Augmented Grad-CAM: Heat-maps super resolution through augmentation[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 4067-4071.

[17] Kim J K, Jung S, Park J, et al. Arrhythmia detection model using modified DenseNet for comprehensible Grad-CAM visualization[J]. Biomedical Signal Processing and Control, 2022, 73: 103408.

[18] Umair M, Khan M S, Ahmed F, et al. Detection of COVID-19 using transfer learning and Grad-CAM visualization on indigenously collected X-ray dataset[J]. Sensors, 2021, 21(17): 5813.

[19] Chakraborty T, Trehan U, Mallat K, et al. Generalizing Adversarial Explanations with Grad-CAM[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 187-193.

[20] Sattarzadeh S, Sudhakar M, Plataniotis K N, et al. Integrated grad-CAM: Sensitivity-aware visual explanation of deep convolutional networks via integrated gradient-based scoring[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 1775-1779.

[21] Papandrianos N I, Feleki A, Moustakidis S, et al. An explainable classification method of SPECT myocardial perfusion images in nuclear cardiology using deep learning and grad-CAM[J]. Applied Sciences, 2022, 12(15): 7592.