目录

[专题实验报告 2](#_Toc156427891)

[一、 可解释性方法的基本原理 2](#_Toc156427892)

[二、 LIME 4](#_Toc156427893)

[三、 Grad-CAM++ 6](#_Toc156427894)

[四、 ScoreCAM 7](#_Toc156427895)

[五、 Score-CAM的缺陷与改进 9](#_Toc156427896)

[六、 三种方法的可解释性分析结果比较 10](#_Toc156427897)

[七、 参考 15](#_Toc156427898)

# 专题实验报告

## 可解释性方法的基本原理

可解释性方法的基本原理是通过解释黑盒模型的决策过程，使其决策对人类更加可理解和透明。这些方法的核心思想是在不牺牲模型性能的前提下，提供关于模型预测的解释性信息。一般而言，可解释性方法可以分为几个主要类别：首先，局部解释性方法致力于解释模型对于个别样本的决策。这包括了诸如 LIME（Local Interpretable Model-agnostic Explanations）的方法，通过在输入空间中引入微小扰动，生成局部可解释的模型来逼近原始模型的决策函数。这使得模型的决策在特定样本附近变得更加透明，帮助理解为何模型会做出特定的预测。

其次，全局解释性方法旨在理解整个模型的行为，而非个别样本。例如，SHAP（SHapley Additive exPlanations）方法基于博弈论中的 Shapley 值，为每个特征分配一个重要性分数，从而提供了一个全局的特征重要性解释。这有助于理解模型整体上对于不同特征的依赖关系。

另外，基于注意力机制的方法，如Grad-CAM（Gradient-weighted Class Activation Mapping）等，通过利用模型中间层的梯度信息来生成图像中的注意力热图，标示模型在决策过程中关注的区域。这种方法使模型的决策可视化，增加了对模型预测的直观理解。

总体而言，可解释性方法在不同的层面上提供了对模型决策的解释，从而帮助用户理解模型的内在机制、提高模型的可信度，以及发现模型可能存在的问题。这些方法的综合使用有助于在实际应用中更好地理解和信任机器学习模型。

1. **LIME (Local Interpretable Model-agnostic Explanations)**

LIME是一种局部解释性模型，旨在解释复杂模型在特定实例附近的行为。它通过在输入空间中生成随机样本并观察模型对这些样本的反应，来逼近原始模型的行为。然后，LIME使用线性模型来拟合这些生成的样本，以提供更简单的解释性模型。

基本原理： LIME的基本思想是通过在输入空间中生成局部扰动来理解黑盒模型的预测。它尝试用一个简单可解释的模型（例如线性模型）来近似黑盒模型在输入附近的行为。

应用领域： 适用于文本分类、表格数据、图像分类等。

1. **RISE (Randomized Input Sampling for Explanation)**

RISE使用随机遮挡生成一系列遮挡后的图像，这些遮挡包括对图像的随机区域进行遮挡。这些遮挡可能是不同形状和大小的方块、矩形或其他形状。对于每个遮挡后的图像，使用训练好的图像分类模型进行预测，记录模型对每个遮挡后图像的预测结果。RISE通过对比有遮挡和无遮挡的情况，评估每个像素对模型输出的影响。如果一个像素对模型输出有很大影响，那么在遮挡时，模型的预测性能会受到较大影响。根据像素的重要性评估，生成一个重要性图，用于表示模型对图像的预测所关注的区域。在这个图中，亮度较高的区域表示对模型输出更为重要的区域。

1. **Grad-CAM (Gradient-weighted Class Activation Mapping)**

Grad-CAM的灵感来源于Class Activation Map（CAM），CAM是一种通过对全局平均池化层的输出进行权重加权来生成图像的类别激活热图的方法。Grad-CAM通过使用模型输出的类别分数对最后一个卷积层的梯度进行加权，产生了一个重要性图，该图指示了对于模型的最终预测，哪些区域在图像中是关键的。

1. **Grad-CAM++ (Gradient-weighted Class Activation Mapping++)**

Grad-CAM++在Grad-CAM的基础上引入了额外的加权项，以更精细地定位关键区域。通过引入正向传播中的梯度信息，相对于Grad-CAM来说更准确地捕捉图像中的重要区域，尤其在细节方面更为敏感。

1. **ScoreCAM (Score-weighted Class Activation Mapping)**

ScoreCAM的基本思想源自CAM方法，但它进行了改进以更好地定位关键区域。与CAM一样，ScoreCAM使用模型输出的类别分数，但不同之处在于ScoreCAM通过对最后一个卷积层的激活值应用类别分数进行加权，得到每个通道的权重。将每个通道的激活值与对应的类别分数相乘，得到每个通道的加权激活值。最后，将所有通道的加权激活值相加，形成最终的热图，其中强调了对于模型最终预测的类别而言对预测的贡献最大的图像区域。

1. **LayerCAM**

LayerCAM是ScoreCAM的扩展，它使用了多层特征图，通过对多层特征图进行融合来提高解释性。这样可以更全面地捕捉模型对输入的理解。LayerCAM引入了对不同卷积层的权重，对于每个卷积层，计算其激活值，即层级激活值。这表示了每个卷积层在处理输入图像时的响应。与ScoreCAM类似，LayerCAM使用模型输出的类别分数，但是对每个卷积层的激活值应用类别分数进行加权，得到每个卷积层的权重，将每个卷积层的加权激活值相加，形成最终的热图。LayerCAM考虑了不同卷积层的响应，以更全面地捕捉深层特征对最终预测的影响。

1. **可解释性方法的分析与比较**

适用领域： LIME是模型无关的，适用于各种类型的模型；RISE、Grad-CAM系列、ScoreCAM系列主要用于图像分类。

解释方式： LIME通过生成局部近似模型解释预测，RISE通过随机采样图像进行解释，Grad-CAM系列和ScoreCAM系列通过对梯度或分数的加权得到激活热图，LayerCAM引入多层特征图融合。

可解释性： LIME提供局部解释，RISE提供整体解释，Grad-CAM系列、ScoreCAM系列提供局部解释，LayerCAM综合多层信息提供更全面的解释。

复杂性： LIME相对较简单，Grad-CAM系列和ScoreCAM系列较为直观，RISE和LayerCAM相对较复杂。

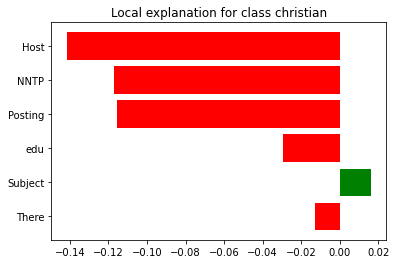
应用场景： 选择方法需根据应用场景，如需要全局解释、模型无关性等，可以选择LIME方法。

## LIME

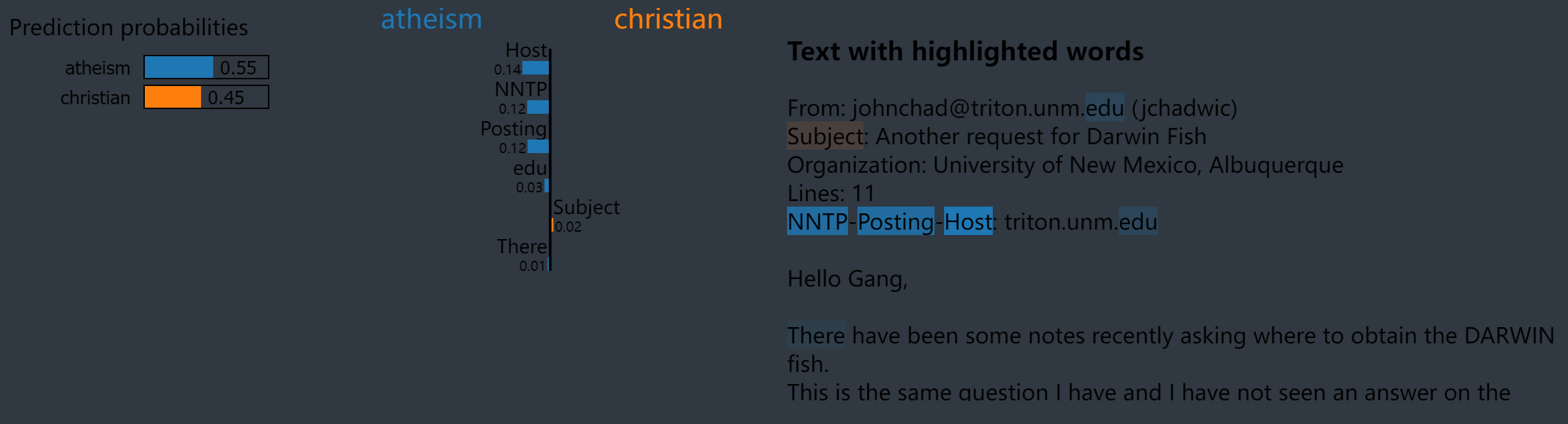
运行开源LIME代码https://github.com/marcotcr/lime/doc/notebooks/Lime - basic usage, two class case.ipynb

这段代码演示了如何使用LIME来解释一个基于随机森林的文本分类器在20个新闻组数据集上的预测结果。

获得可解释性分析结果如下：



**图：基督教社会类别文档分类的解释**



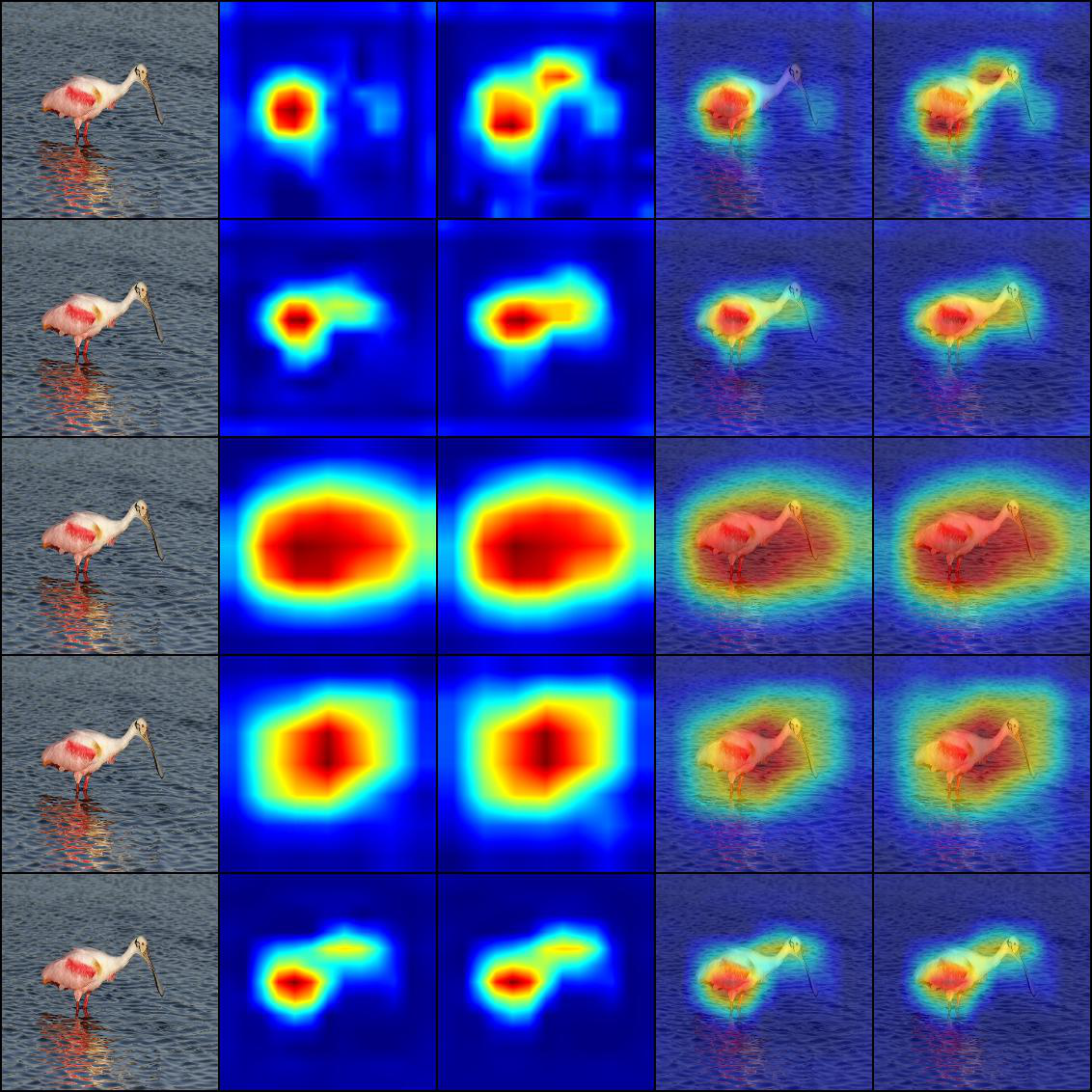
**图：使用D3.JS在html和notebook中渲染的可视化结果图表**

LIME生成的解释结果中包含了对于文本中每个词语的权重，反映了模型在做出预测时对于各个词语的重要性。例如，在将一篇文档判断为“无神论”类别时，“Host”、“NNTP”、“Posting”等词语在模型做出判断的过程中具有很高的重要性（权重）。通过在原始文本中去除一些特定特征（词语），LIME展示了模型对于这些特征的敏感性。这种分析有助于用户理解模型预测的鲁棒性，以及对于不同词语的依赖关系。

## Grad-CAM++



**图：输入图片water-bird.jpg**



**图：五种不同模型对同一张图片的Grad-CAM++可解释性分析结果对比**

开源代码分别演示了使用Grad-CAM++方法对alexnet、vgg、resnet101、densenet161、squeezenet1.1五种预训练的模型对给定图像的分类的可视化解释。通过使用不同模型的Grad-CAM++解释器生成可解释性热图，直观地显示出了不同模型对相同输入图像关注区域的差异。由输出结果可以看出resnet101和densenet161表现出相对较大的感兴趣区域（覆盖包含水鸟在内的一片区域），表示对图像的全局结构和上下文信息有较高的关注度；在感兴趣区域中，可能存在对整体图像特征的强烈响应。而**AlexNet、VGG16和SqueezeNet1.1**显示较小的感兴趣区域，更强调图像的局部结构和细节（关注区域更贴合水鸟的轮廓，特别是水鸟的身体）。这些模型更专注于捕捉水鸟的局部特征，特别是水鸟的身体部分。

## ScoreCAM

<https://colab.research.google.com/drive/1m1VAhKaO7Jns5qt5igfd7lSVZudoKmID?usp=sharing>

开源代码演示的是使用预训练的VGG16模型生成图像解释。可视化解释如图所示：

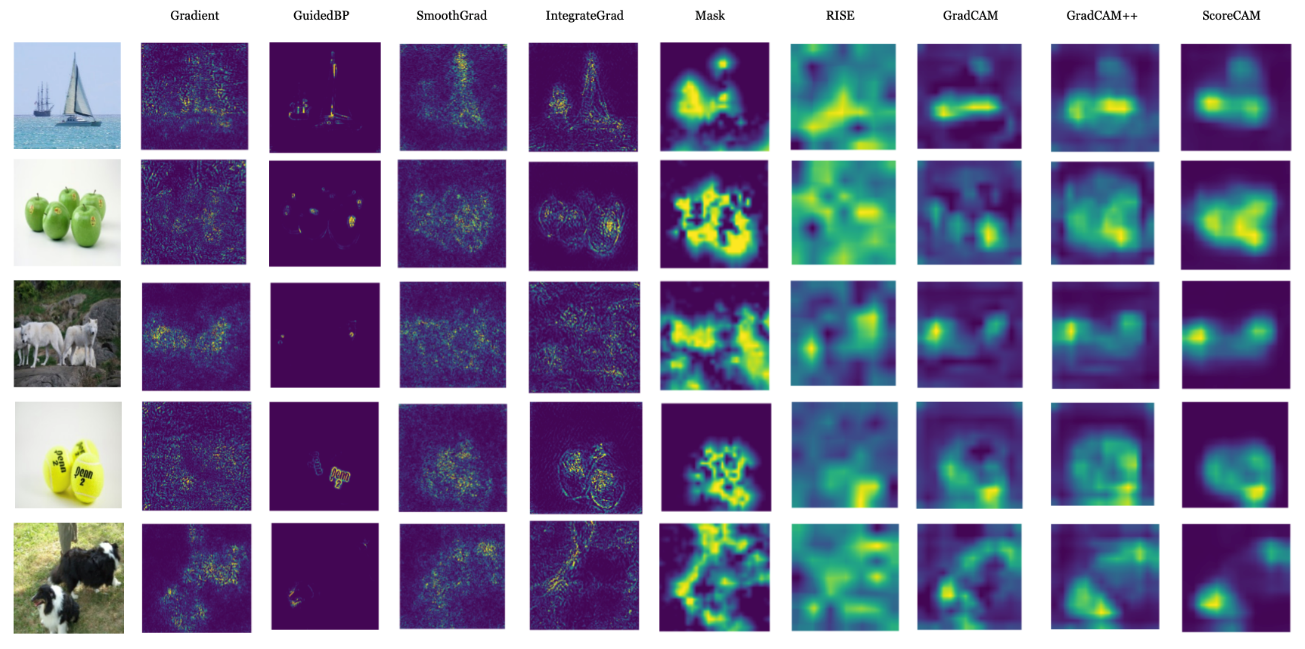


**图：输入**



**图：输出**

由输出结果可以看出Score-CAM更倾向于关注全局特征，而忽视局部解释。



**图：Score-CAM与其它方法的对比**

从以上对比图可以看出ScoreCAM 生成的热图更加清晰、直观，强调模型对图像的关键区域。以第二张输入图片为例，可以看到针对苹果的特定部分有更强的颜色强调，表明这些区域对于模型的决策起到关键作用；同时热图具有比较明显的边界，让用户更容易理解模型的关注点。而RISE的热图则都呈现出一种整体的、模糊的效果，表明RISE强调模型对整个图像的感知。Grad-CAM 以及 Grad-CAM++ 生成的热图同样突出显示了模型对于帆船/苹果/羊等的局部区域，但相对于 ScoreCAM分辨率明显更低一些。

## Score-CAM的缺陷与改进

空间信息不足：Score-CAM 主要通过使用模型中最后一个卷积层的通道权重来生成热图。这意味着生成的热图主要关注在通道级别的特征，而不是在像素级别上进行细致的关注。同时，由于使用了全局平均池化，Score-CAM 在计算通道权重时对整体图像进行了平均处理。这导致了对于图像中各个局部区域的感知可能不足，尤其是对于模型预测的决策区域。在生成热图时，Score-CAM 并未直接考虑像素之间的空间关系。这使得生成的热图可能不够平滑，对于图像中局部变化的响应不够灵敏。

改进方法：

1. Spatial Regularization：引入空间正则化项，以考虑像素级别的空间关系。可以通过在热图生成过程中添加平滑性惩罚项或采用空间滤波技术，确保生成的热图在局部区域上更加一致和平滑。

引入平滑性正则化项，可以通过在生成热图的过程中添加平滑性惩罚项来实现：

*def* spatial\_regularization\_loss(heatmap):

    smoothness\_loss = torch.sum(torch.abs(heatmap[:, :, 1:] - heatmap[:, :, :-1])) + \

                       torch.sum(torch.abs(heatmap[:, 1:, :] - heatmap[:, :-1, :]))

    return smoothness\_loss

# 在生成热图的过程中添加平滑性正则化

raw\_heatmap = generate\_raw\_heatmap(model, image)

smoothness\_loss = spatial\_regularization\_loss(raw\_heatmap)

final\_heatmap = raw\_heatmap + *lambda* \* smoothness\_loss

1. Guided Backpropagation Integration：结合 Guided Backpropagation 等梯度反传的改进方法，以获得更准确的像素级别权重。这有助于更好地捕捉模型对图像局部特征的贡献。

guided\_backprop = compute\_guided\_backpropagation(model, image)

raw\_heatmap = generate\_raw\_heatmap(model, image)

final\_heatmap = guided\_backprop \* raw\_heatmap

1. Spatial Attention Mechanisms: 引入空间注意力机制，以依据图像中的局部区域给予不同的权重。这可以通过引入空间注意力网络或结合现有的注意力机制来实现，从而更好地关注对模型决策有重要影响的区域。

*class* SpatialAttention(*nn*.*Module*):

*def* *\_\_init\_\_*(self, in\_channels):

        super(SpatialAttention, *self*).*\_\_init\_\_*()

*self*.conv = nn.Conv2d(in\_channels, 1, kernel\_size=1)

*def* forward(self, x):

        attention = torch.sigmoid(*self*.conv(x))

        return attention \* x

spatial\_attention = SpatialAttention(in\_channels=model.last\_conv.out\_channels)

raw\_heatmap = generate\_raw\_heatmap(model, image)

final\_heatmap = spatial\_attention(raw\_heatmap)

1. Feature Importance Estimation: 在生成热图时，考虑每个特征的重要性。这可以通过引入对特征的权重调整或在生成过程中考虑每个特征的贡献度来实现，以更全面地反映模型对于图像各个特征的关注。

权重调整和特征贡献度计算实现：

raw\_heatmap = generate\_raw\_heatmap(model, image)

feature\_importance = compute\_feature\_importance(model, image)

final\_heatmap = raw\_heatmap \* feature\_importance

## 三种方法的可解释性分析结果比较

**LIME（Local Interpretable Model-agnostic Explanations）**

**优点：**

模型无关性：适用于各种类型的模型，包括复杂的深度学习模型。

局部解释：提供对特定实例的局部解释，有助于理解模型在个别样本上的决策过程。

可解释性结果可视化：LIME 提供了可视化的方式展示解释结果，例如生成的权重和对应的可视化图，使解释结果更容易理解。

适用性广泛：LIME 不仅适用于文本数据，还可用于表格数据、图像数据等多种数据类型，使其在各种任务和领域中都具有广泛的适用性。

用户友好：LIME 的使用相对简单，通过生成局部样本并使用可解释的模型，用户可以较容易地获得对模型预测的解释结果。

**缺点：**

样本生成开销：LIME 需要在附近生成大量样本以构建局部模型，这可能会导致在大规模数据集或高维数据上的计算开销较大。

模型误差：LIME 构建的局部模型是一个简化的可解释模型，与原始模型相比，可能引入一些误差。尤其是在复杂的模型或非线性关系下，误差可能会比较大。

超参数选择：LIME 中一些超参数的选择可能对解释结果产生影响，例如样本生成时的权重选择等，需要根据具体情况进行调整。

不考虑全局结构：LIME 主要关注局部解释，对于全局结构的理解相对较弱，无法很好地捕捉整体模型的行为。

依赖可解释模型：LIME 在解释时需要选择一个可解释的模型作为局部模型，但并非所有任务都有合适的可解释模型可用。

**Grad-CAM++（Gradient-weighted Class Activation Mapping++）**

**优点：**

高分辨率可视化：Grad-CAM++ 生成的热图相对较高分辨率，能够更清晰地显示模型对图像的关注区域，使解释结果更直观。

梯度权重考虑：Grad-CAM++ 在计算梯度权重时引入了额外的正则化项，更全面地考虑了梯度信息，有助于提高解释的准确性。

对多通道特征的支持：Grad-CAM++ 能够对多通道的特征图进行解释，对于一些复杂模型，特别是在含有多通道特征的情况下具有更好的适用性。

**缺点：**

复杂度高：Grad-CAM++ 的计算相对较复杂，需要额外的正则化项和复杂的梯度计算，计算开销较大。

不稳定性：在某些情况下，Grad-CAM++ 可能表现出对输入图像微小变化较为敏感，这可能导致解释结果的不稳定性。

全局结构理解：类似于传统的 Grad-CAM，Grad-CAM++ 主要关注局部解释，对于全局结构的理解相对有限。

不适用于所有模型：Grad-CAM++ 的效果依赖于模型的梯度信息，对于某些结构较为特殊的模型，可能无法得到良好的解释结果。

**ScoreCAM（Score-weighted Class Activation Mapping）**

**优点：**

直观可解释性：ScoreCAM 直观地将预测分数与每个像素的权重相关联，易于理解模型对图像每个位置的关注程度。

对多通道特征的支持：ScoreCAM 能够对多通道的特征图进行解释，对于一些复杂模型，特别是含有多通道特征的情况，具有更好的适用性。

相对简单：相较于一些复杂的解释方法，ScoreCAM 的实现较为简单，易于理解和使用。

**缺点：**

不考虑空间信息：ScoreCAM在计算权重时不考虑空间信息，可能导致对图像结构的理解不够全面。

对抗性攻击敏感：ScoreCAM 在某些情况下可能对对抗性攻击敏感，即对输入图像进行微小修改可能导致解释结果的变化。

全局结构理解：ScoreCAM 主要关注局部解释，对于全局结构的理解相对有限。

不适用于所有模型：ScoreCAM 的效果依赖于模型的激活值分布，对于某些结构较为特殊的模型，可能无法得到良好的解释结果。

存在定位误差：在某些情况下，ScoreCAM 可能产生定位误差，即在生成的热图中的最大响应位置并不总是与实际关注区域完全一致。

**综合比较**

**适用场景：**

LIME：适用于各种机器学习模型，尤其擅长对黑盒模型的局部解释，可以解释文本、表格和图像等不同类型的数据。

Grad-CAM++：主要应用于卷积神经网络，特别是图像分类任务，能够高分辨率地可视化模型对图像的关注区域。

ScoreCAM：适用于卷积神经网络，能够生成清晰的热图，强调模型对图像的关键区域。

**解释方式：**

LIME：通过生成局部可解释的线性模型，提供对单个样本的局部解释，关注于样本附近的局部特征。

Grad-CAM++：利用梯度权重生成高分辨率的热图，强调模型决策时的重要区域，更关注全局结构的影响。

ScoreCAM：基于类激活映射和分数权重，生成对模型预测决策的可视化解释，注重模型对图像的权重分布。

**计算开销：**

LIME：由于需要生成多个解释样本，计算开销较大，尤其是在样本较复杂的情况下。

Grad-CAM++：基于梯度信息生成热图，计算相对较轻量，适用于大规模图像数据集。

ScoreCAM：相对简单的计算过程，生成热图的计算开销相对较小。

**可解释性：**

LIME：提供局部可解释性，对单个样本进行解释，生成简单的线性模型，易于理解。

Grad-CAM++：通过高分辨率的热图展示全局结构信息，可解释性较好，直观呈现模型对图像的关注点。

ScoreCAM：提供对模型预测决策的清晰可视化，突出关键区域，具有直观的可解释性。

**稳定性与鲁棒性：**

LIME：在采样点数量和解释模型复杂度选择上对稳定性较为敏感，可能因参数选择不当而产生较大波动。

Grad-CAM++：相对稳定，梯度信息提供了相对鲁棒的解释结果，对输入微小扰动的鲁棒性较好。

ScoreCAM：在一定程度上对输入扰动敏感，可能产生定位误差，但相对较稳定。

## 七、参考

[1]. <https://github.com/marcotcr/lime>

[2]. https://github.com/adityac94/Grad\_CAM\_plus\_plus

[3]. <https://github.com/haofanwang/Score-CAM>

[4]. "Why Should I Trust You?": Explaining the Predictions of Any Classifier

[5]. RISE: Randomized Input Sampling for Explanation of Black-box Models

[6]. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

[7]. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks

[8]. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks

[9]. LayerCAM: Exploring Hierarchical Class Activation Maps for Localization