**计算机视觉结课报告**

**卷积神经网络在进行图像分类任务时的可解释性分析**

**U202115287-计卓2101-黄靖迪**

目录

[1. 绪论 1](#_Toc14378)

[1.1. 可解释性分析概述 1](#_Toc564)

[1.1.1. 诞生背景 1](#_Toc15206)

[1.1.2. 经典可解释性方法 1](#_Toc26580)

[1.2. 内容简介 1](#_Toc16557)

[2. 经典可解释性方法的理论 1](#_Toc10533)

[2.1. LIME（Local Interpretable Model-agnostic Explanations）方法 1](#_Toc11878)

[2.1.1. LIME方法的提出 1](#_Toc23447)

[2.1.2. LIME方法的特点 1](#_Toc29272)

[2.1.3. LIME方法的基本原理 2](#_Toc4997)

[2.1.4. LIME方法的优缺点 4](#_Toc32350)

[2.2. RISE（Randomized Input Sampling for Explanation）方法 5](#_Toc5796)

[2.2.1. RISE方法的提出 5](#_Toc14228)

[2.2.2. RISE方法的特点 5](#_Toc31469)

[2.2.3. RISE方法的基本原理 5](#_Toc21525)

[2.2.4. RISE方法的优缺点 7](#_Toc24051)

[2.3. Grad-CAM (Gradient-weighted Class Activation Mapping)方法 8](#_Toc17285)

[2.3.1. Grad-CAM方法的提出 8](#_Toc32703)

[2.3.2. Grad-CAM方法的特点 8](#_Toc24818)

[2.3.3. Grad-CAM方法的基本原理 8](#_Toc9636)

[2.3.4. Grad-CAM方法的优缺点 10](#_Toc24799)

[2.4. Grad-CAM++（Gradient-weighted Class Activation Mapping Plus Plus）方法 11](#_Toc3246)

[2.4.1. Grad-CAM++方法的提出 11](#_Toc17516)

[2.4.2. Grad-CAM++方法的特点 11](#_Toc12445)

[2.4.3. Grad-CAM++方法的基本原理 11](#_Toc8955)

[2.4.4. Grad-CAM++方法的优缺点 12](#_Toc4993)

[2.5. ScoreCAM（Score-Weighted Class Activation Mapping）方法 12](#_Toc26050)

[2.5.1. ScoreCAM方法的提出 12](#_Toc21365)

[2.5.2. ScoreCAM方法的特点 13](#_Toc11562)

[2.5.3. ScoreCAM方法的基本原理 13](#_Toc1811)

[2.5.4. ScoreCAM方法的优缺点 14](#_Toc5612)

[2.6. LayerCAM方法 15](#_Toc24209)

[2.6.1. LayerCAM方法的提出 15](#_Toc1097)

[2.6.2. LayerCAM方法的特点 15](#_Toc13983)

[2.6.3. LayerCAM方法的基本原理 15](#_Toc3537)

[2.6.4. LayerCAM方法的优缺点 16](#_Toc29580)

[2.7. 总结 16](#_Toc7829)

[3. 经典可解释性方法的复现 17](#_Toc28218)

[3.1. LIME方法复现 17](#_Toc7354)

[3.2. Grad-CAM++方法复现 19](#_Toc18908)

[3.3. ScoreCAM方法复现 20](#_Toc9665)

[3.4. 三种经典可解释性方法得出的结果和分析 21](#_Toc161)

[3.4.1. 图像1 21](#_Toc6919)

[3.4.2. 图像2 24](#_Toc9907)

[3.4.3. 图像3 25](#_Toc8006)

[3.5. 总结 27](#_Toc7733)

[4. 总结 27](#_Toc29985)

# 绪论

## 可解释性分析概述

### 诞生背景

卷积神经网络（Convolutional Neural Network，CNN）在图像分类任务中表现出色，但它们通常被认为是“黑盒”模型，难以理解其决策过程，这种不透明性可能会造成问题，特别是在需要对模型决策进行解释或解释性分析时。

### 经典可解释性方法

经典的可解释性方法有LIME, RISE, Grad-CAM, Grad-CAM++, ScoreCAM, LayerCAM等方法。

## 内容简介

本篇专题报告将简要阐述上述集中经典可解释性方法的原理，并对其进行比较和分析，还将会对其中的几个方法进行复现，并得到可解释性分析结果。

# 经典可解释性方法的理论

## LIME（Local Interpretable Model-agnostic Explanations）方法

### LIME方法的提出

该方法于2016年，在由Marco Tulio Ribeiro、Sameer Singh和Carlos Guestrin三位作者的论文《Why Should I Trust You? Explaining the Predictions of Any Classifier》中首次提出，这篇论文中详细介绍了LIME的原理和方法，并提出了通过生成局部可解释模型来解释任意分类器的预测。

### LIME方法的特点

LIME方法是一种局部可解释性方法，旨在提供对于模型在特定实例周围的行为的解释。它通过在输入空间中生成邻域样本，使用这些样本训练一个可解释的局部模型，然后解释这个局部模型的行为，以逼近原始模型的决策过程。

### LIME方法的基本原理

在本篇论文中，作者认为，模型有可解释性首先就需要为输入变量和输出提供定性的解释，而且同时也要考虑到使用者自身的局限性，面对不同专业水平的受众可解释性的概念不一样，另一个重要的标准是局部忠实度，直观理解就是局部的原模型输出和解释模型的输出要一致。当然局部的忠实度并不代表着全局的忠实度，对于全局解释进行这种代理分析还是比较困难的。

LIME的目的，就是给定一个复杂的模型，能够用一个简单的可解释模型，搭配可解释的特征进逼近这个复杂的模型，在局部上能够达到和复杂模型相似的效果。

在方法上，LIME主要通过扰动输入样本实例的特征来判断哪些特征对模型的预测输出影响较大。前提是这些扰动的特征，是人类可理解的。

作者在论文中介绍了局部探索抽样的方法（如图 2.1所示）。红蓝两个区域被非线性曲线分开，从全局来看，难以用一个简单的模型取拟合，但在局部可以用一个线性分类器（局部可解释模型）去划分。

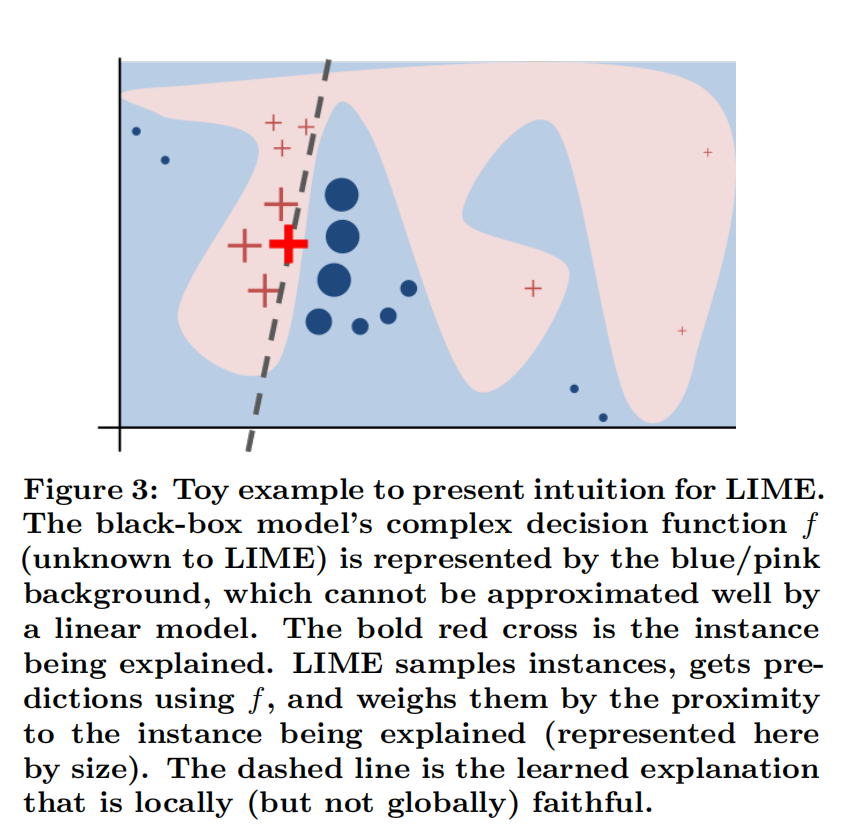


图 2.1 局部探索抽样

LIME的伪代码如图 2.2所示，该方法的工作流程大致如下：

* 1. 将原样本的特征向量转换成可解释的特征向量，使其更具解释性。
* 2. 随机扰动这些可解释的特征N次，得到N个样本。然后将这些样本带入局部的这个线性分类器得到预测结果。将原特征维度的向量带入原复杂分类模型得到预测结果。
* 3. 根据随机扰动得到的样本，我们可以计算它们与要解释的样本的距离，利用这个距离代入指数核计算得到原样本相对于扰动样本的权重，以定义要解释的样本的局部性。
* 4. 然后根据以上结果，通过公式 2.2去迭代学习出这个局部的可解释模型，例如线性回归或决策树。
* 5.分析训练得到的可解释模型，以理解原始模型在选择的实例上的预测原因。这可以通过查看模型的权重、特征重要性等来实现。

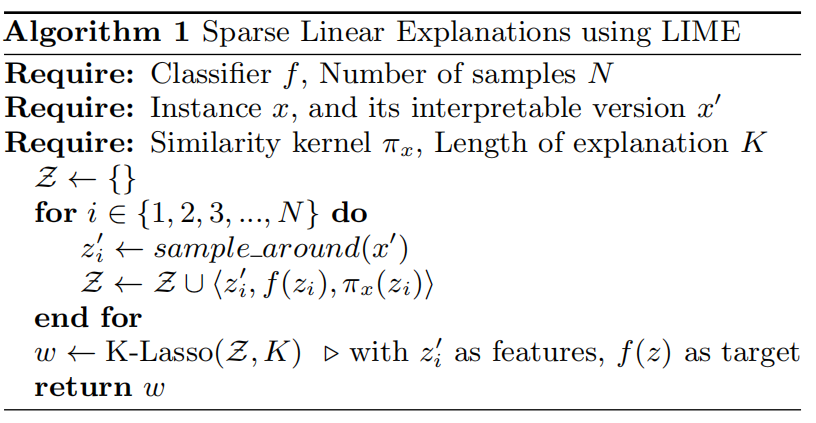


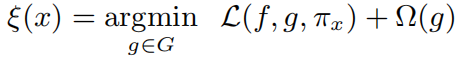
图 2.2 LIME伪代码

在将特征向量转换为可解释的特征向量的过程中，对不同的特征向量有不同的处理方法。例如对文本分类任务，可通过设置可解释性模型为词袋并设置词语数上限K来确保可稀释的可能性，该可解释模型的复杂度如公式 2.1所示。对于图像分类，也可利用“超像素”得到具有相同复杂度（公式 2.1）的可解释模型，其中可解释的图像是一个二进制向量，其中1表示原始超像素，0表示灰化的超像素。



公式 2.1

对于公式 2.2，该公式的目标是最小化局部模型与原模型的差异，同时权衡、局部模型的复杂度。



公式 2.2

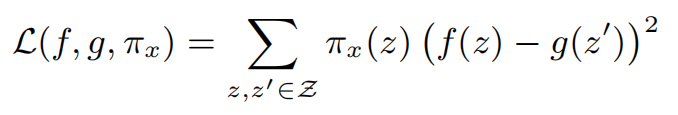
在该公式中各个符号对应含义如下：

* f：原模型，即复杂的分类器
* g：可解释模型，即局部的简单的可解释模型。论文用的是稀疏线性回归模型，即g(z’)=w·z’
* πx(z)：即样本x和样本z的距离相近程度的衡量，在作者采用的稀疏线性回归回归作为解释模型中如公式 2.3所示



公式 2.3

在本论文用的稀疏线性回归模型中，三者构成的函数关系可以由公式 2.4表示。



公式 2.4

* Ω(g)：模型g的复杂度。（如Dtree中可以用树的深度来表示，线性模型则可以用非零权重的数量表示）
* G：局部的可解释模型的假设空间
* x：原特征维度的样本
* x’：可解释特征维度的样本
* z’：可解释特征维度的随机扰动得到的样本
* z：z’ 恢复到原特征维度对应的样本

对于选取的特征个数，论文设为常数K，经过扰动后，得到N个样本，利用N个样本，然后利用这N个样本做lasso回归，最后得到的线性函数即局部解释模型的权重即为各特征对应的模型预测该样本的贡献度。

在论文中，作者认为LIME对单个样本的预测提供了解释。但是单个样本的解释还不足以让使用者信服整个模型，因此提出了SP-LIME算法，目的是为了能够挑出尽可能少的样本但是这些样本能够尽可能多的覆盖一些更重要的特征，这样通过这些样本就能够较全面的了解模型做出的预测更依赖于哪些特征，以便用户能较快的判断该模型是否能够信任，在此处不做过多介绍。

### LIME方法的优缺点

优点：

* 1. 模型无关性：LIME是一种模型无关的解释方法，适用于各种类型的机器学习模型，包括深度学习模型和其他黑盒模型。
* 2. 导出可解释模型： LIME 使用可解释的模型（例如线性模型或决策树）来近似原始模型的行为。这样的模型更容易理解，提供了关于模型决策的更直观解释。
* 3. 普适性：由于 LIME 的模型无关性，它可以用于各种领域和任务，包括图像分类、自然语言处理等。
* 4. 提供置信度： LIME 生成的可解释模型不仅提供解释，还提供了每个特征的权重，可以用作解释的置信度度量。

缺点：

* 1. 采样不足可能导致偏差： LIME 的解释依赖于对输入空间的采样，如果采样不足或者不够代表性，可能导致对模型行为的不准确解释。
* 2. 局限于局部解释： LIME 主要提供局部解释，即对于给定的实例，只提供在该实例附近的解释。这可能无法捕捉全局模型行为。
* 3. 可解释模型的选择： LIME 需要选择一个可解释的模型来近似原始模型的行为。这个选择可能会影响解释的准确性，因为不同的可解释模型对于不同的数据分布和模型行为可能产生不同的效果。
* 4. 依赖超参数：LIME方法中的一些超参数，例如采样次数和扰动的范围，需要进行调整以达到最佳效果，需要经验进行调试。
* 5. 需要大量计算开销： LIME通过在输入空间中生成大量的局部样本，然后对这些样本进行目标模型的预测，从而构建一个可解释的局部模型。LIME需要训练局部模型来近似目标模型在输入空间的行为。

## RISE（Randomized Input Sampling for Explanation）方法

### RISE方法的提出

在2018年在作者为A. Petsiuk, H. Xiao, A. Kozlovskiy和M. Hein的论文《RISE: Randomized Input Sampling for Explanation of Black-box Models》中提出

### RISE方法的特点

RISE方法方法是基于随机输入采样的图像解释方法，通过观察模型对扰动样本的反应来推断模型对输入的敏感性和特征的重要性。

### RISE方法的基本原理

该方法较为简单，大概处理流程可如图 2.3所示：

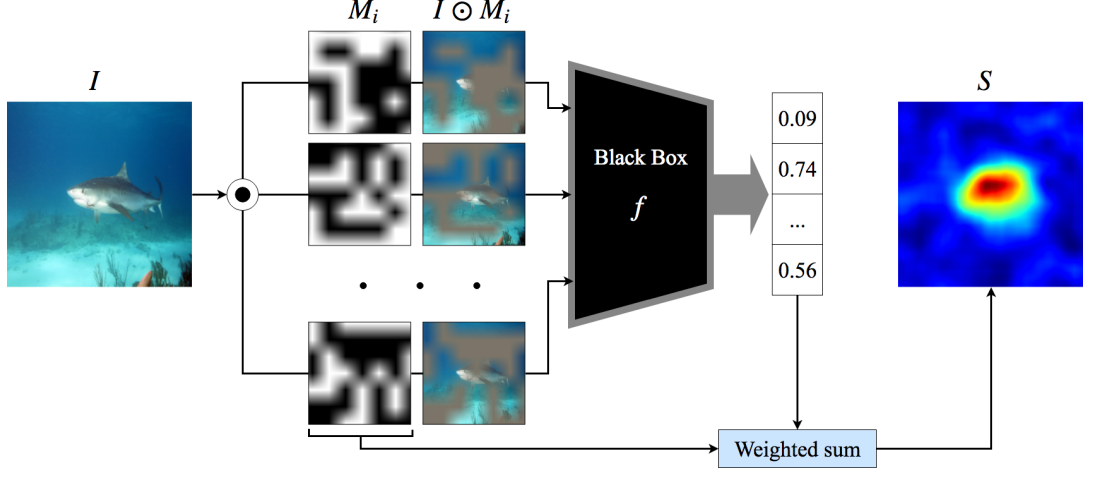


图 2.3 RISE方法的大致工作流程

RISE是一种基于蒙特卡罗采样的解释方法，大致工作流程如下所示：

* 1. 随机采样：RISE的第一步是对输入图像进行蒙特卡罗采样，通过在图像的像素上引入随机噪声或者随机遮挡的方式来生成多个扰动样本。
* 2. 模型预测：对于每个生成的扰动样本，使用深度学习模型进行预测。记录模型对每个扰动样本的输出结果，即模型对每个变体图像的预测概率或分数。
* 3. 统计分析：对模型在所有扰动样本上的输出进行统计分析。一般来说，RISE会考虑目标类别的概率分布或者模型的置信度，以确定哪些像素对于最终预测起到了关键作用。这可以通过计算每个像素在所有扰动样本中的出现频率或者平均预测概率来实现。
* 4. 生成特征重要性图：根据统计分析的结果，RISE方法生成一个特征重要性图，也可以被视为热力图。这个图用来显示在输入图像的不同区域上哪些像素对于模型的决策具有更大的影响，即哪些区域对于模型的预测起到了关键作用。
* 5.可视化解释：最终，RISE提供了一个可视化的解释，让用户能够直观地理解模型对于输入图像的决策过程。用户可以通过观察热力图来识别模型关注的区域，从而了解哪些特征对于模型的决策具有重要性，如图 2.4所示

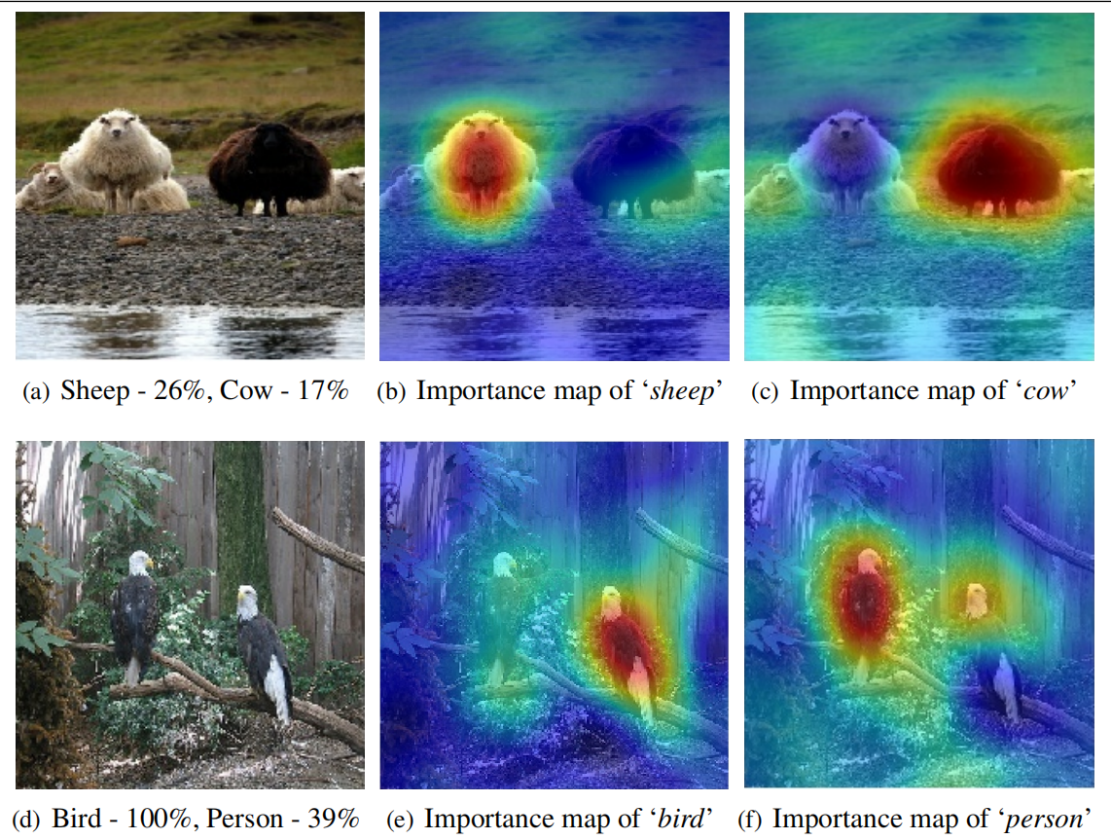


图 2.4 RISE方法解释的不同图像特征对分类的贡献

### RISE方法的优缺点

优点：

* 1. 模型无关性：RISE是一种模型无关的解释方法，适用于各种类型的深度学习模型，包括黑盒模型。
* 2. 全局解释：通过引入随机扰动并统计分析多个扰动样本，RISE 可以提供相对全局的模型解释。这使得用户可以更全面地理解模型对输入图像的整体决策过程。
* 3. 直观性：RISE提供了一个直观的特征重要性图，通过热力图的形式展示模型关注的区域。这样的可视化结果使得用户能够直观地理解模型对于输入图像的关注点。
* 4. 普适性： 由于RISE 的模型无关性，它可以用于各种领域和任务，包括图像分类、自然语言处理等。

缺点：

* 1. 计算开销较高：因为需要生成大量的随机扰动样本，并在模型上进行预测，RISE的计算开销相对较高。
* 2. 局限性：RISE提供的解释主要是基于输入图像的全局特性，而在某些情况下可能无法提供足够细致的局部解释。对于某些局部决策或者对单个像素的重要性的解释可能相对有限。
* 3. 依赖超参数：RISE方法中的一些超参数，例如采样次数和扰动的范围，需要进行调整以达到最佳效果，需要经验进行调试

## Grad-CAM (Gradient-weighted Class Activation Mapping)方法

### Grad-CAM方法的提出

该方法是在2017年在作者为Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh和Dhruv Batra的论文《Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization》中提出。

### Grad-CAM方法的特点

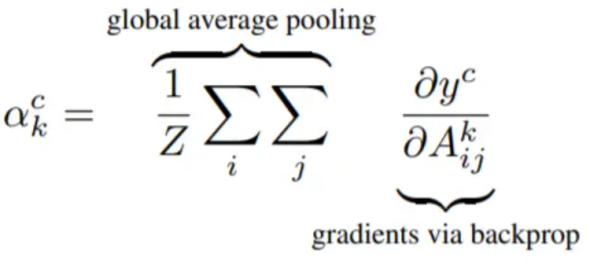
Grad-CAM是用梯度信息反映的CAM（Class Activation Map）方法。使用目标类别的梯度信息来加权卷积层的特征图，生成一个热力图，以显示模型关注的区域。

### Grad-CAM方法的基本原理

卷积神经网络的卷积操作可以看做是滤波器对图片进行特征提取，通过滑动窗口的方式实现，因此特征图和输入图片存在空间上的对应关系。特征图的权重可以认为是被层层卷积核过滤后而保留的有效信息，其值越大，表明特征越有效，对网络预测结果越重要。一个深层的卷积神经网络，通过层层卷积操作，提取空间和语义信息。一般存在其他更难理解的层，例如分类的全连接层、softmax层等，很难以利用可视化的方式展示出来。所以，CAM的提取一般发生在卷积层，尤其是最后一层卷积。

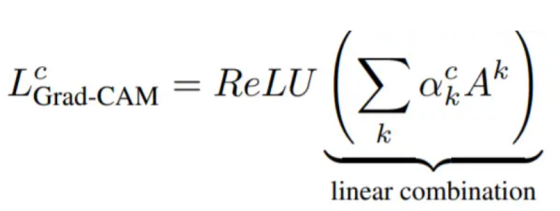
Grad-CAM方法的步骤：

* 1. 前向传播：首先，通过对输入图像进行前向传播，获取模型的输出。
* 2. 计算目标类别的分数：选择模型预测的目标类别，计算该类别的分数（概率）。这个分数用于表示模型对于这个类别的置信度。
* 3. 反向传播：将目标类别的分数对于模型最后一个卷积层的输出进行反向传播。通过反向传播，计算得到目标类别分数对于每个卷积层输出的梯度。
* 4. 全局平均池化与权重计算： 对每个卷积层的梯度进行全局平均池化，得到每个通道的权重，计算算法如公式 2.5所示。这一步的目的是计算每个通道对于目标类别分数的贡献。



公式 2.5 为特征图Ak分配重要性权重ak

* 5. 加权和：将每个通道的权重与对应的卷积层输出相乘并相加，得到最终的类激活映射（Class Activation Map）。这个映射表示了每个位置对于目标类别分数的重要性，即高亮的区域是网络在该图像上关注的区域。
* 6. ReLU激活：将生成的类激活映射经过ReLU激活函数，去除负值，强调网络关注的区域，计算过程如公式 2.6所示。



公式 2.6 由

Grad-CAM方法的基本过程如图 2.5所示。

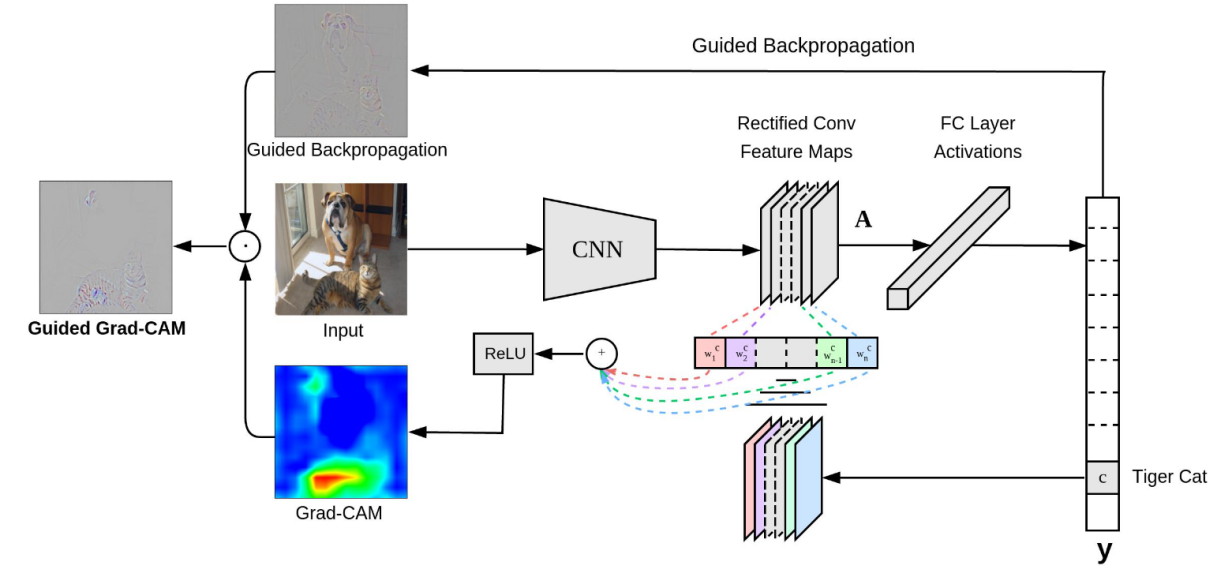


图 2.5 Grad-CAM方法的过程图

最终生成的Class Activation Map（CAM）是一幅与输入图像相同尺寸的热力图，如图 2.6所示。



图 2.6 Class Activation Map

### Grad-CAM方法的优缺点

优点：

* 1. 直观性：Grad-CAM生成的可视化结果是激活热图，直观地显示了模型对于输入图像的哪些区域有更强的响应。

缺点：

* 1. 过度平均化：Grad-CAM在生成激活热图时使用了全局平均池化，这可能导致对于较大物体的定位效果不佳。因为全局平均池化会丢失物体的空间信息，使得热图中的高亮区域可能覆盖了整个物体而不是局部区域。
* 2. 对背景的敏感性：Grad-CAM对输入图像中的背景区域也可能有一定的敏感性，这可能导致生成的激活热图中包含一些与目标物体无关的信息。
* 3. 不能解释全局信息： Grad-CAM主要关注于局部信息，无法提供关于整个模型决策的全局解释。
* 4. 模型支持性不佳：对CNN模型效果最佳。

## Grad-CAM++（Gradient-weighted Class Activation Mapping Plus Plus）方法

### Grad-CAM++方法的提出

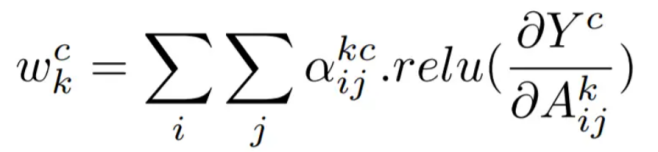
在2018年在作者为Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader和Vineeth N Balasubramanian的论文《Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks》中提出。

### Grad-CAM++方法的特点

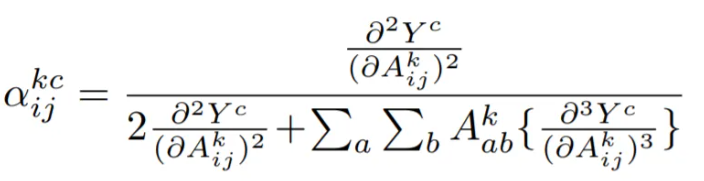
Grad-CAM++是对 Grad-CAM 方法的改进，旨在提高可解释性图生成的准确性和清晰度，也是一种用梯度信息反映的热力图方法。

### Grad-CAM++方法的基本原理

Grad-CAM是利用目标特征图的梯度求平均(GAP)获取特征图权重，可以看做梯度图上每一个元素的贡献是一样。 而Grad-CAM++ 认为梯度图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图上的元素进行加权，更新的算法如公式 2.7、公式 2.8所示



公式 2.7



公式 2.8

Grad-CAM与Grad-CAM++方法的区别如图 2.7所示。

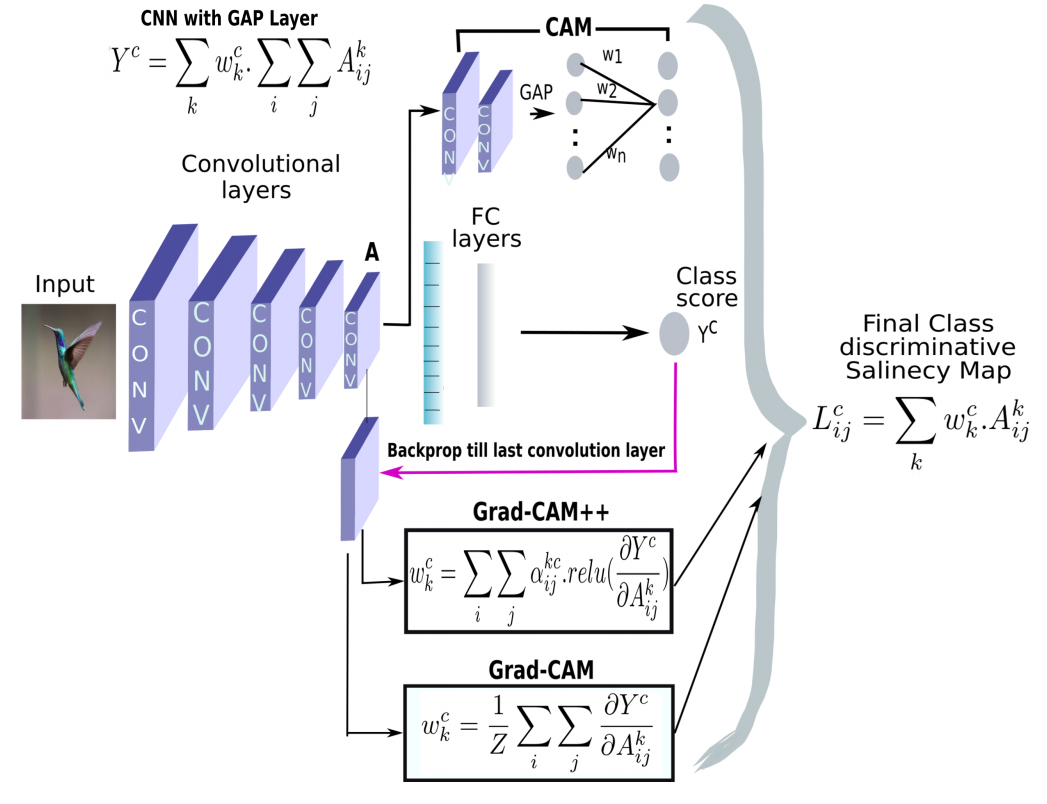


图 2.7 Grad-CAM算法与Grad-CAM+算法的区别

### Grad-CAM++方法的优缺点

因Grad-CAM++方法与Grad-CAM方法相似，故仅讨论二者之间相比较的优缺点

优点：

* 1. 更精准的定位：Grad-CAM++引入了正则化项和对权重的修正，以减小激活热图中的噪声，增强对目标物体的定位精度，能提供更准确的目标区域定位。
* 2. Grad-CAM++对权重的计算进行了修正，提高了梯度的准确性，使得生成的激活热图更加准确。这可以降低热图中不相关区域的影响，使可解释性结果更清晰。

缺点：

* 1. Grad-CAM++的性能可能受到超参数的影响，因此需要进行调参来获得最佳结果，依赖于测试和经验。

## ScoreCAM（Score-Weighted Class Activation Mapping）方法

### ScoreCAM方法的提出

在2020年在作者为Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel和Xia Hu的论文《Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks》中提出。

### ScoreCAM方法的特点

通过加权类别激活映射（CAM）来生成可视化解释，使用模型对特定类别的分数来加权类激活映射。这样，模型在分类时给予高置信度的区域更大的权重。使得解释更加关注对模型决策贡献较大的区域。

### ScoreCAM方法的基本原理

ScoreCAM方法的伪代码如图 2.8所示，结构如图 2.9所示。

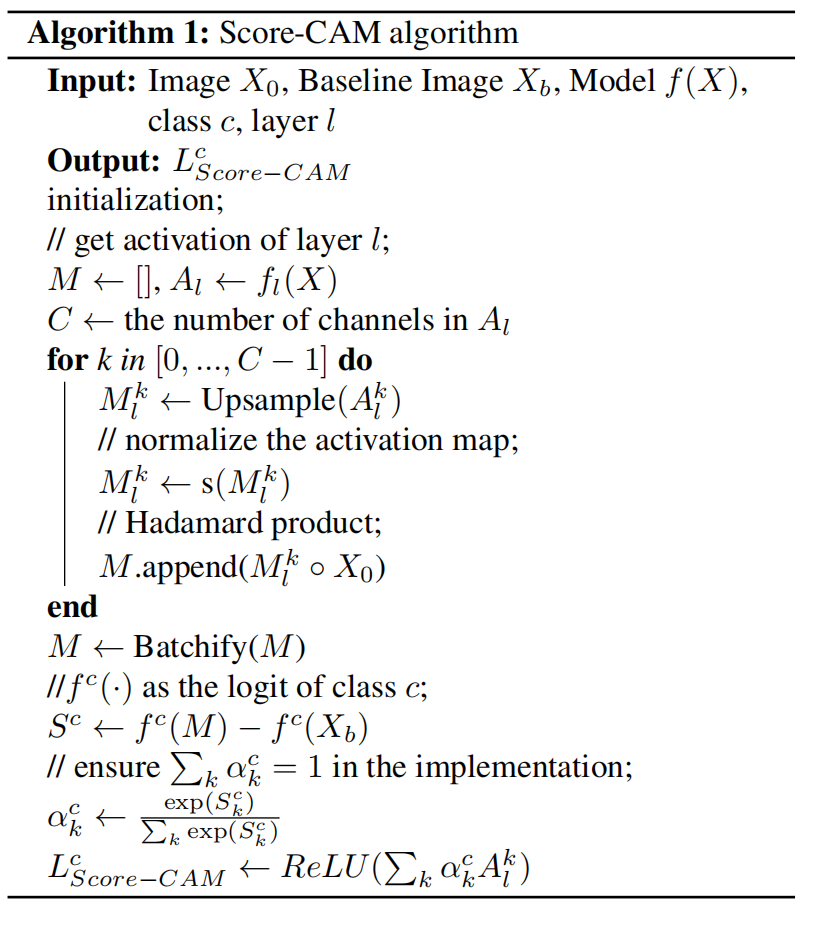


图 2.8 ScoreCAM方法伪代码

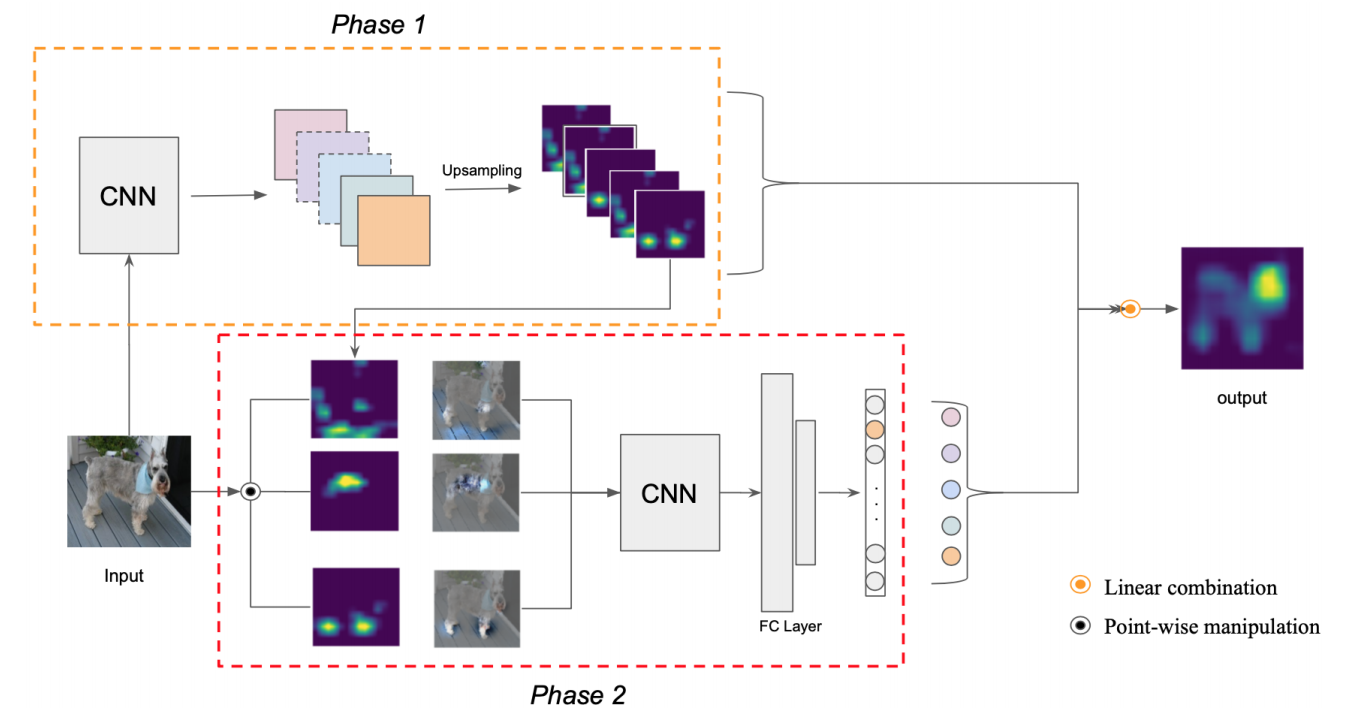
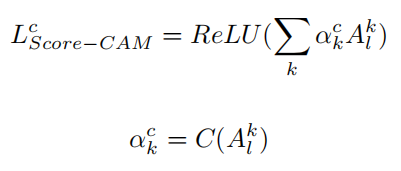


图 2.9 ScoreCAM方法

ScoreCAM方法中像素的贡献度计算如公式 2.9所示，不同于之前的Grad-CAM与Grad-CAM++，ScoreCAM摆脱了梯度信息的依赖，基于类别激活分数进行计算，通过对目标类别的类别激活分数进行反向传播，计算每个像素对于该类别的重要性。然后，通过将每个通道的权重应用于原始图像的相应通道，生成一个加权的图像表示。



公式 2.9

### ScoreCAM方法的优缺点

优点：

* 1. 直观性： ScoreCAM生成的热图提供了直观的可视化，强调了对于网络分类决策而言最重要的图像区域。这有助于理解模型是如何做出决策的。

缺点：

* 1. 单一解释： ScoreCAM生成的解释是基于单个目标类别的。对于多类别任务，你需要分别运行ScoreCAM来解释每个类别。
* 2. 忽略上下文信息： ScoreCAM主要关注目标类别的激活分数，可能会忽略全局上下文信息。在某些情况下，这可能导致解释不够全面。
* 3. 对抗性攻击敏感： 与一些基于梯度的方法一样，ScoreCAM可能对对抗性攻击敏感，即对输入进行微小的扰动可能导致解释的改变。
* 4. 固定的激活分数： ScoreCAM使用前向传播中的类别激活分数来生成解释。这些分数在每次前向传播中都是固定的，可能无法捕捉到模型对于不同输入的动态变化。

## LayerCAM方法

### LayerCAM方法的提出

在2020年在作者为Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng和 Yunchao Wei 的论文《LayerCAM: 探索用于定位的分层类激活图》中提出。

### LayerCAM方法的特点

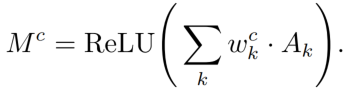
可以为CNN的不同层生成可靠的类激活图，这个属性能够收集到对象定位信息从粗略（粗略的空间定位）到精细（精确的细粒度细节）级别。在Grad-CAM的基础上，对map每个位置的激活值乘一个梯度权重。

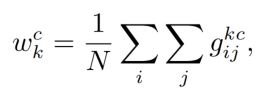
### LayerCAM方法的基本原理

在Grad-CAM的基础上，不再用全局权重，而是用像素级权重，也就是对应于特征图上的一个位置，使用ReLU函数机型计算，如果梯度为正，则用这个正梯度作为权重，如果梯度为负，则权重为零。

在之前的**错误！未定义书签。**节中，以及简述了Grad-CAM的基本原理，即为每个特征图 Ak 分配一个权重 ak。然后对特征 A 中的所有特征图进行线性加权求和。最后，应用 ReLU 函数操作从类激活图中去除负面影响。

在本篇论文中，作者对Grad-CAM方法每个特征图Ak设权重为wk，如公式 2.10所示。





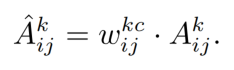
公式 2.10

作者在论文中提出：与特征图中某个位置对应的正梯度表明增加该位置的强度将对目标类别的预测分数产生积极影响。对于具有正梯度的位置，我们使用它们的梯度作为权重。那些具有负梯度的位置分配为零。

第 k 个特征图中空间位置 (i, j) 的权重可以写为如公式 2.11所示，为了获得某一层的类激活图，LayerCAM 首先将特征图中每个位置的激活值乘以一个权重，如公式 2.12所示。最后将结果 Aˆk 沿通道维度线性组合，得到类激活图，公式如公式 2.13所示。



公式 2.11



公式 2.12



公式 2.13

至此就得到了LayerCAM的方法的可解释性结果。

### LayerCAM方法的优缺点

优点：

* 1. 更精确的类激活图： LayerCAM能生成较Grad-CAM与Grad-CAM++更为精确、对象更细粒度的类激活图，以帮助分析图形对神经网络分类的贡献。
* 2. 网络适配性：LayerCAM 易于用于任何现成的基于 CNN 的图像分类器，无需修改网络架构和改变反向传播方式。

缺点：

* 1. 增加计算量

## 总结

在本篇中，粗略地介绍了LIME、RISE、Grad-CAM、Grad-CAM++、ScoreCAM和LayerCAM方法的原理和特点，按创造可解释模型的依赖划分，大致可以分为三类：

1. 局部样本拟合：LIME
2. 随机遮挡采样：RISE
3. 依赖目标类别的梯度：Grad-CAM、Grad-CAM++与LayerCAM
4. 基于CNN类别激活分数：ScoreCAM

这些方法采用了不同的方式得到可解释性模型，每种方法都有各自的特点。

1. LIME方法特别适用于黑盒模型，能够提供局部可解释性，且可适用于文本模型。
2. RISE方法通用性强，不依赖于具体的模型结构，能够捕捉模型对不同图像区域的敏感度。
3. Grad-CAM、Grad-CAM++、ScoreCAM和LayerCAM特别适用于CNN网络下的图像分类任务，Grad-CAM++更提供较Grad-CAM更精确的结果，依赖于梯度计算，LayerCAM在Grad-CAM++的基础上进一步提高了可解释模型的精细度。
4. ScoreCAM则不依赖于图像梯度进行计算，而是基于类别激活分数生成解释，直接与 CNN 模型相关。

# 经典可解释性方法的复现

在本报告的实现中，采用了Inception\_v3预训练模型作为实验中的模型，1因该模型是一个有1000多种分类的模型，故报告中会选取该模型预测概率前5的标签进行可视化解释。

## LIME方法复现

根据LIME方法，采用线性回归模型进行训练将LIME方法的代码实现分为4步：

1. 创建扰动图像
2. 利用模型预测、获取模型对当前类别的预测结果，即扰动样本对原始样本的预测概率。
3. 利用线性回归模型进行局部性解释
4. 输出预测结果，在原图上叠加，得到可解释性图像

首先是创建扰动图像，随机选择每个分割区域是否是“活动”的，其中 activate为True 表示活动，为False 表示非活动。若当前区域是非活动的 (active 为 False)，则将该区域的像素值设为0；否则，保持不变，以达到随机扰动图像的的效果，代码如图 3.1所示。



图 3.1 扰动图像的生成

完成上述扰动图像的创建后，接着便是使用模型预测扰动图像对应原样本的概率，得到预测标签Preds，如图 3.2所示

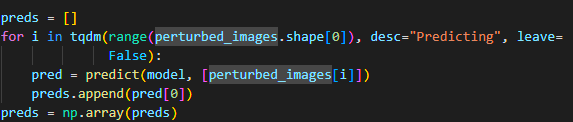


图 3.2 使用模型预测扰动图像的概率

接下来是通过对扰动图像预测得到的结果训练线性回归模型，使用线性回归模型拟合扰动样本的权重（weights）和模型对扰动样本的预测结果（class\_preds）之间的关系，然后获取线性回归模型的系数，即对每个扰动方式的权重，然后创建一个与图像分割相同形状的零数组，用于存储每个分割区域的权重，再创建一个与图像分割相同形状的零数组，用于存储二值掩码，将heatmap与原始图像结合起来对于每个分割区域及其对应的权重，进行循环遍历。

若权重为正，则表示该分割区域对模型的预测有正面影响，将权重赋值给对应的分割区域将对应的二值掩码位置置为1，用于后续结合heatmap和原始图像。若权重为负，表示该分割区域对模型的预测有负面影响。将对应的二值掩码位置置为0。代码如图 3.3所示。

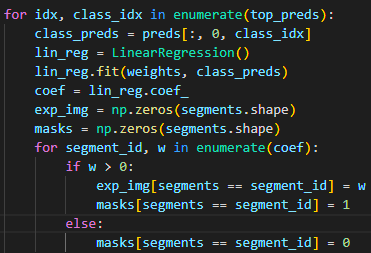


图 3.3 线性回归模型的训练和二值掩码的创建

在得到二值掩码后，将权重进行归一化，将其映射到0到1的范围。通过颜色映射得到heatmap，然后与原图进行叠加，即可得到使用LIME方法的图像可解释化，代码如图 3.4所示。

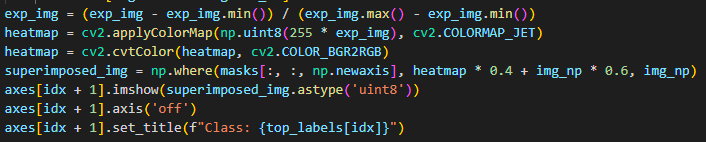


图 3.4 得到HeatMap并与原图进行叠加，得到可解释化模型

## Grad-CAM++方法复现

在实现Grad-CAM++方法复现的过程中，选定了预训练模型Inception\_v3的Mixed\_7c.branch\_pool.conv卷积层作为对象，因为该层是一个比较深的卷积层，它位于模型的后半部分，处在后半部分的该层会提取更加复杂和抽象的特征，这些特征对理解图像的内容非常重要，故选择了该层作为Grad-CAM++的目标层。

首先，使用一个前向钩子函数和一个反向钩子函数实现。前向钩子函数forward\_hook在目标图层进行前向传播时被调用，用于捕获目标图层的输出特征图。backward\_hook 在反向传播时被调用，用于捕获目标类别的梯度。对模型的每个模块进行迭代，当找到目标图层是，注册前向和反向钩子函数，代码如图 3.5所示。

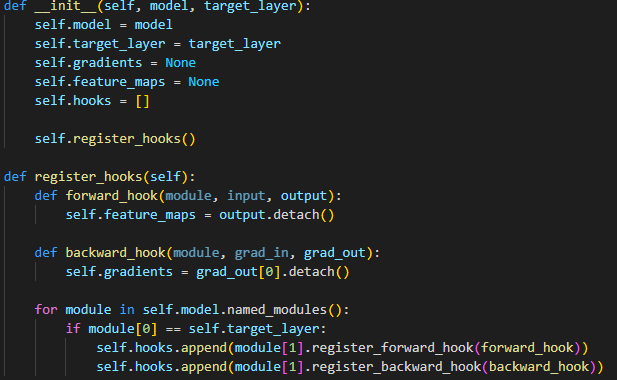


图 3.5 Grad-CAM++方法钩子函数

然后便是利用前向钩子捕获到的目标图层的输出特征图和反向钩子捕获到的目标类别的梯度，计算Grad-CAM++的热力图。使用自适应平均池化 （F.adaptive\_avg\_pool2d） 计算梯度的权重 （weights），然后利用得到的权重对目标图册的输出特征图进行加权，得到加权特征图，再对该加权特征图进行ReLU操作以去除负值，最后进行归一化，映射为热力图，即可得到Grad-CAM++方法产生的可解释化图像，代码如图 3.6所示。

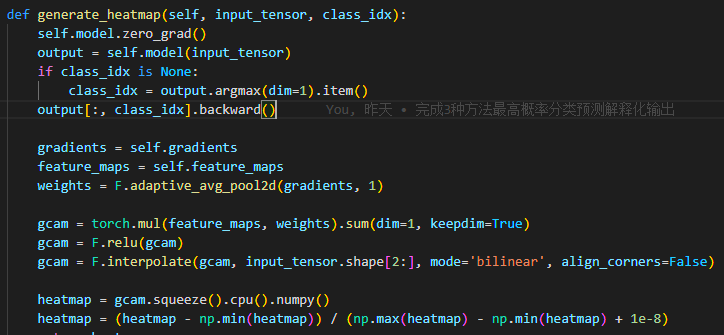


图 3.6 Grad-CAM++方法 热力图生成

## ScoreCAM方法复现

在ScoreCAM方法中，也使用了与在Grad-CAM++方法中同样的Mixed\_7c.branch\_pool.conv卷积层作为目标层，原因类似，并且选取一样的目标层能更好地比较两种方法的差异。

先将输入图像通过预处理为模型可接受的张量格式，对模型进行前向传播，通过前向传播的钩子函数以捕获目标层的特征图。然后对特征图进行插值操作并将其调整为与输入图像相同的尺寸。

接着初始化热图，对特征图进行处理，在每个通道上，通过插值操作将特征图调整为与输入图像相同的尺寸，然后将调整后的特征图与输入图像相乘，并计算模型对于该部分图像的预测分数（score），然后将得到的分数与调整后的特征图相乘，并累积到全局热力图中，最后将热力图归一化，生成热力图，代码如图 3.7所示。

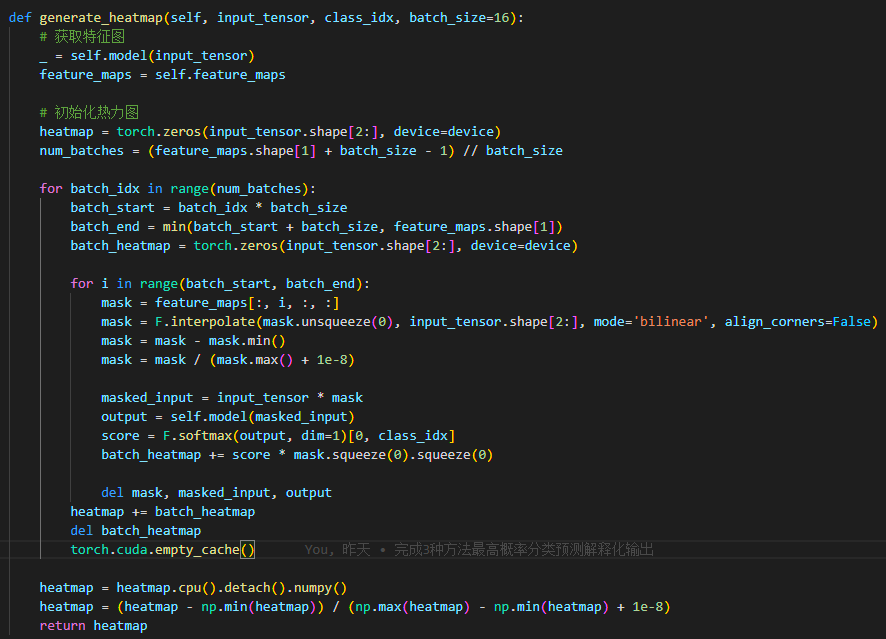


图 3.7 ScoreCAM方法

## 三种经典可解释性方法得出的结果和分析

对图像用Inception\_v3预训练模型进行预测后得出的概率前5的标签，分别使用LIME、Grad-CAM++、ScoreCAM可解释性方法得到的结果如下所示

注1：下列结果中使用了LIME方法得出的结果都是生成了10000张扰动图像进行预测出的结果。

注2：最后得到的可解释性图像的图表一共有6张子图，从左到右依次为：

1. 原图
2. 概率第1
3. 概率第2
4. 概率第3
5. 概率第4
6. 概率第5

此排版规则在后面不再阐释

### 图像1

对下图（图 3.8）使用Inception\_v3预训练模型进行预测，概率前五的结果依次为：

1. tabby（虎斑猫）
2. tiger\_cat（虎猫）
3. chow（松狮犬）
4. Egyptian\_cat（埃及猫）
5. Australian\_terrier（澳大利亚牧羊犬）



图 3.8 图像1

使用LIME可解释性方法得到的结果如下图（图 3.9）所示，在LIME可解释方法中，由于是对图像进行随机扰动并进行预测得到的线性模型，生成的热力图呈多边形形状，在最后生成的热力图在我试图通过设置阈值去掉对预测贡献无关的图像部分，但是效果不太好，就呈现出了这种破碎的预测结果，后面的图片也相同。

注：原图不计算在结果子图中，结果子图从图表中的第2张子图开始，序号从左往右依次为1、2、3、4、5，下同

可以看出，在使用LIME方法对图片进行解释时，LIME对第3、第5张结果子图预测结果为狗的可解释结果中，较为准确地将预测贡献解释较大的部分放在狗所在的位置，而对猫的部分较少，在第1张结果子图预测为猫的结果中，较准确的将较大贡献定位在了猫的位置，在第2张结果子图预测结果为猫的结果中，猫和狗的权重相似。

该LIME方法受限于扰动图像的选择优化和扰动图像生成数量的限制，效果不是很理想。

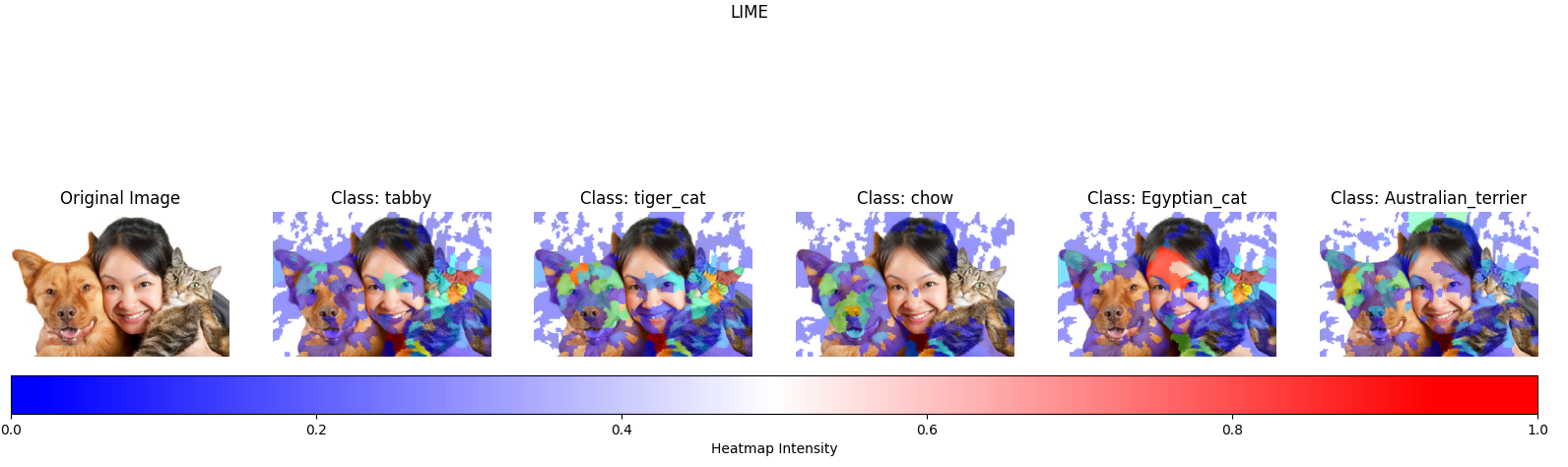


图 3.9 LIME方法 图像1

使用Grad-CAM+可解释性方法得到的结果如下图（图 3.10）所示。可以看到在第3、5张结果子图预测为狗的结果中，将较大贡献值定在狗的位置，而在第1、2、4张结果子图预测为猫的结果中，将较大贡献值定在猫的位置，效果很不错，较LIME方法有了很大提升。

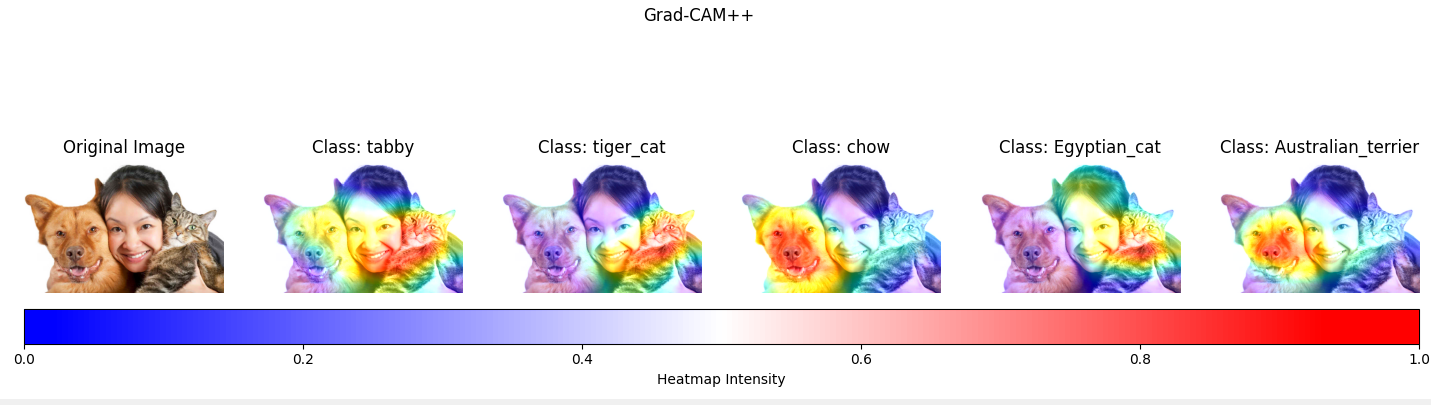


图 3.10 Grad-CAM++方法 图像1

使用ScoreCAM可解释性方法得到的结果如下图（图 3.11）所示。可以看到在第3、5张结果子图预测为狗的结果中，将较大贡献值定在狗的位置，而在第1张结果子图预测为猫的结果中，将较大贡献值定在猫的位置，但在第2、4张结果子图预测为猫的结果中，将较大贡献值放在了猫与狗的中间的人脸部分上，出现了偏差，效果不如Grad-CAM++好。

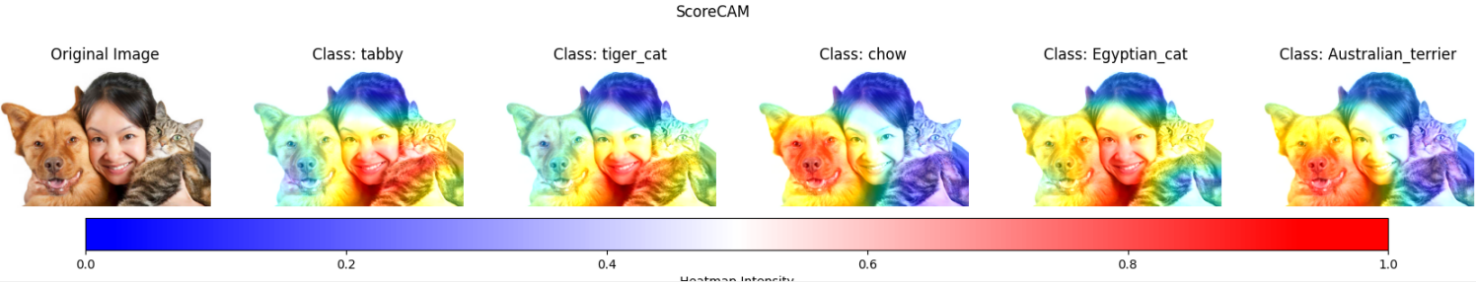


图 3.11 ScoreCAM方法 图像1

### 图像2

对下图（图 3.12）使用Inception\_v3预训练模型进行预测，概率前五的结果依次为：

1. Labrador\_retriever（拉布拉多犬）
2. golden\_retriever（金毛犬）
3. clumber（克伦伯犬）
4. Persian\_cat（波斯猫）
5. Brittany\_spaniel（布列塔尼猎犬）



图 3.12 图像2

使用LIME可解释性方法得到的结果如下图（图 3.13）所示。可以看出，在第1、2、3、5张结果子图中预测为狗的结果中，LIME方法将大部分贡献都定位在了狗身上，而在第4张结果子图预测为猫的结果中，LIME还是将一部分贡献定位在了狗头上，效果一般。

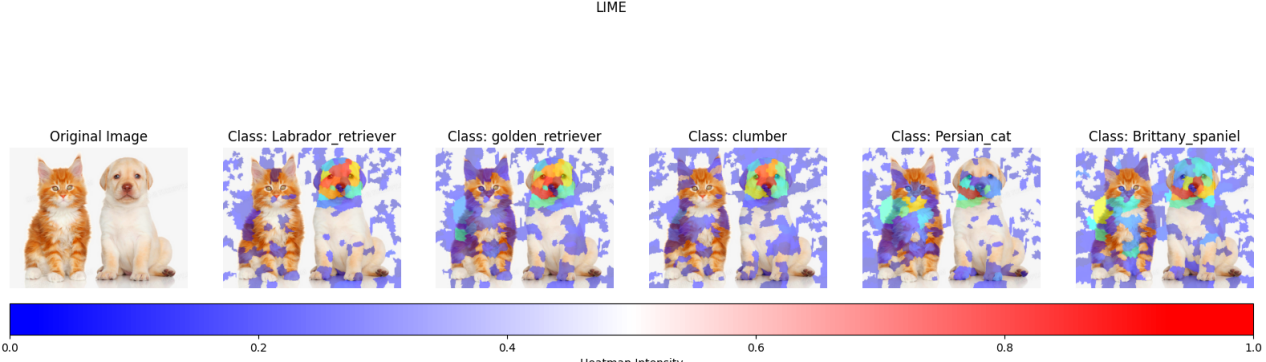


图 3.13 LIME方法 图像2

使用Grad-CAM+可解释性方法得到的结果如下图（图 3.14）所示。在第1、2、3、5张结果子图中预测为狗的结果中，Grad-CAM++方法将贡献都定位在了狗身上。在第4张结果子图预测为猫的结果中，则定位在了猫身上，很好地完成了任务。

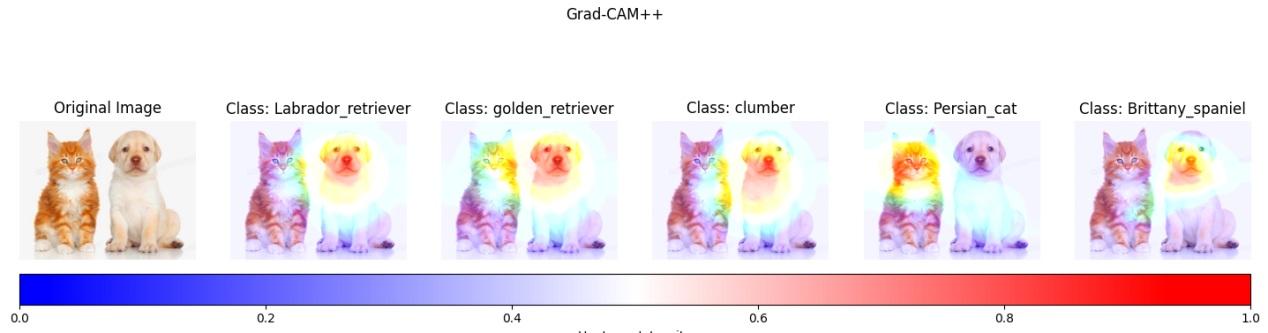


图 3.14 Grad-CAM++方法 图像2

使用ScoreCAM可解释性方法得到的结果如下图（图 3.15）所示。在第1、2、5张结果子图中预测为狗的结果中，ScoreCAM方法将贡献的绝大部分都定位在了狗身上，但在第3张结果子图中，定位到了猫和狗的中间。在第4张结果子图预测为猫的结果中，则定位在了猫身上，但仍有一些权重定位在了狗的脸上，较好地完成了任务。

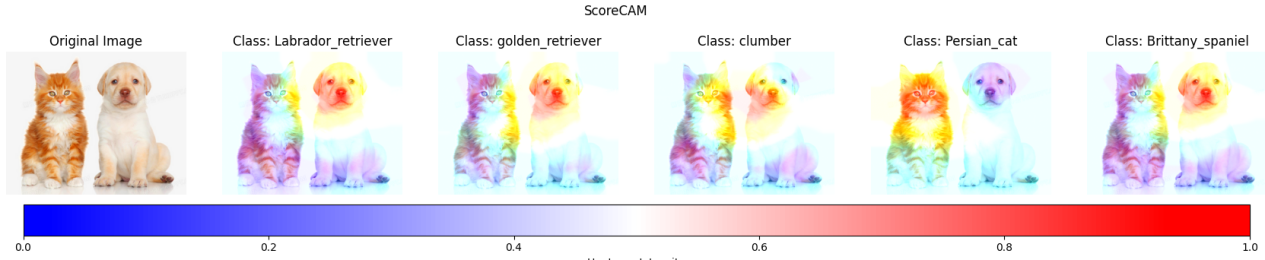


图 3.15 ScoreCAM方法 图像2

### 图像3

对下图（图 3.16）使用Inception\_v3预训练模型进行预测，概率前五的结果依次为：



图 3.16 图像3

1. Border\_collie（边境牧羊犬）
2. tabby（虎斑猫）
3. collie（柯利犬）
4. Bernese\_mountain\_dog（伯恩山犬）
5. tiger\_cat（虎猫）

使用LIME可解释性方法得到的结果如下图（图 3.17）所示。在第1、3、4张结果子图中预测为狗的结果中，LIME方法将大部分贡献都定位在了狗身上，效果不错。在2、5张结果子图预测为猫的结果中，LIME还是将一部分贡献定位在了狗头上，效果较好。

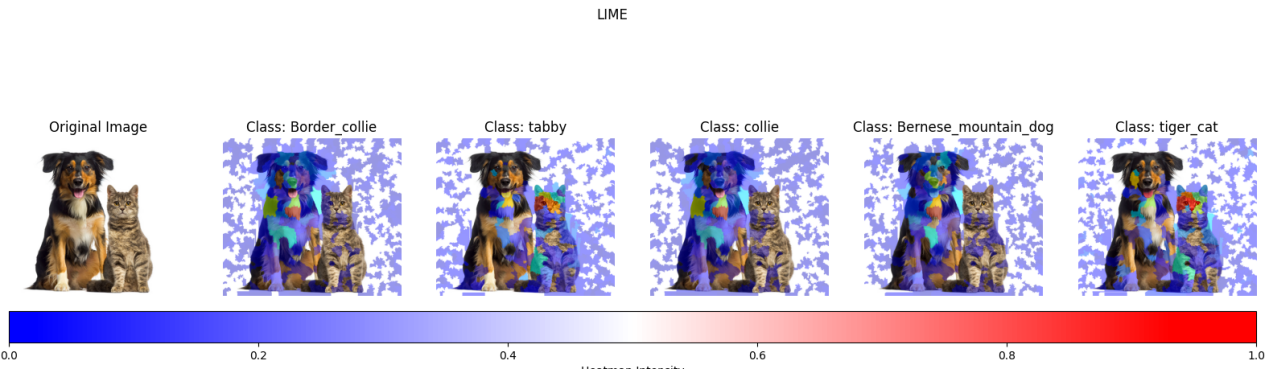


图 3.17 LIME方法 图像3

使用Grad-CAM+可解释性方法得到的结果如下图（图 3.18）所示。在第1、4张结果子图中预测为狗的结果中，Grad-CAM++方法将贡献都定位在了狗身上，但在第3张结果子图中有部分权重落在2了狗和猫地中间。在第2、5张结果子图预测为猫的结果中，则大部分定位在了猫身上，很好地完成了任务。

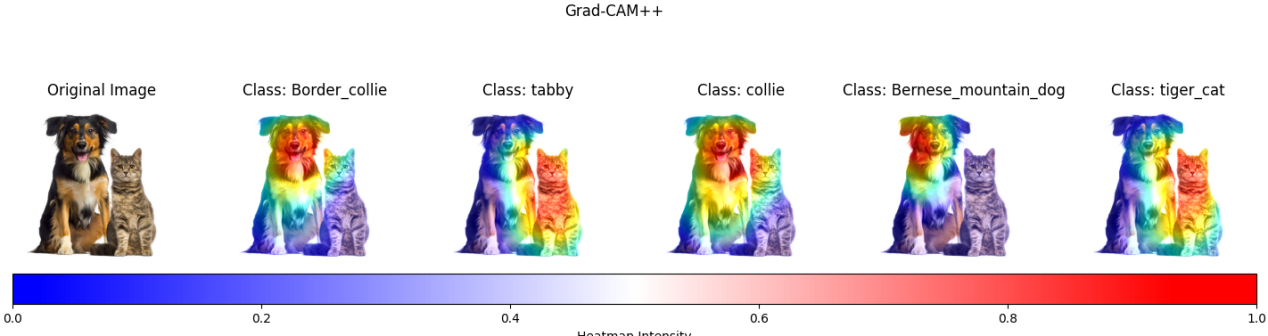


图 3.18 ScoreCAM方法 图像3

使用ScoreCAM可解释性方法得到的结果如下图（图 3.19）所示。在第1、3、4张结果子图中预测为狗的结果中，ScoreCAM方法将贡献大部分，但仍有较大一部分落在狗和猫中间。在第2、5张结果子图预测为猫的结果中，则大部分定位在了猫身上，较好的完成了任务。

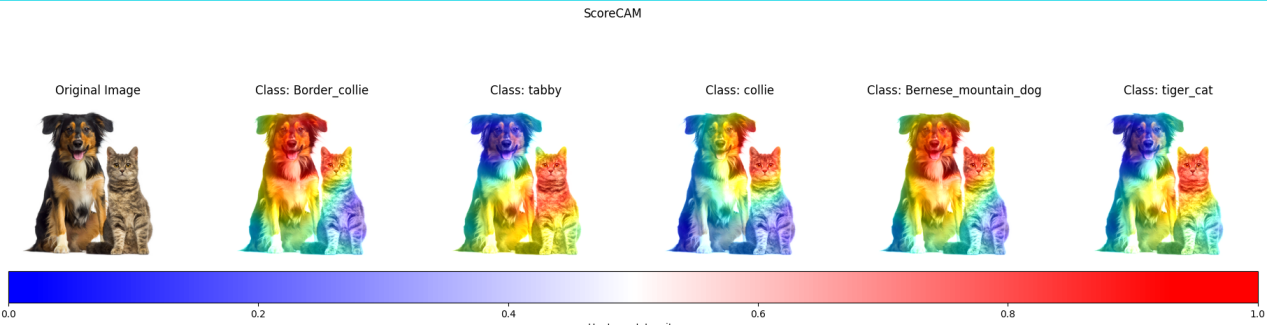


图 3.19 ScoreCAM方法 图像3

## 总结

在对该图像的前5个预测结果的图像可解释化采用不同可解释性方法得到的结果中，使用Grad-CAM++方法得到的结果最为准确，能准确地找出对预测结果贡献度最大地位置，以确定是图像上的哪个部位让预测模型输出这个预测结果。

使用ScoreCAM方法得到的结果较Grad-CAM+方法得到的结果高贡献度覆盖的范围更大，没有那么精确，并且位置也出现了偏差。

而LIME可解释模型受限于扰动图像的数量和扰动图像创建方法的限制，效果最差，仅能粗略的展示图像中对预测结果的贡献，也可能是LIME方法对该种多分类模型效果不好，因为在该报告中使用的预训练模型中猫和狗类的预测概率可能很相似，扰动图像稍微扰动一下就产生了误差。

在性能上，因为Grad-CAM++方法仅需读取反向传播计算目标类别对于最后一个卷积层的梯度，故计算量较小。而ScoreCAM因为需要将然后将特征图与输入图像相乘，并计算模型对于该部分图像的预测分数，然后将得到的分数与调整后的特征图相乘，故在性能上较Grad-CAM++方法较差，在计算的过程中也占用了更多资源。LIME方法因为需要产生大量的扰动图像进行预测，故性能最差，占用了很多计算资源，效率较低。

# 总结

在这门课中老师通过理论与实践相结合的方法，显示在理论课上为我们讲述了计算机视觉相关的理论知识，让我打下了坚实的理论基础。随后通过4个上机实验和结课实验，使我深入了解了如何使用神经网络进行数据二分类任务、MNIST手写数据集识别、双通道的MNIST手写数据集比对训练和这几种经典的图像可解释化方法的使用，特别是最后的结课报告，让我能利用Python这门语言复现了论文中的方法，让我深入学习了计算机视觉这门课程，也让我熟练掌握了Pytorch库的使用和使用Matplotlib库、tqdm库去进行训练过程和结果的可视化展示，让我对机器学习这个分支有了更深的了解以及产生了更浓厚的兴趣，感谢老师为我们一直以来的付出和努力。