RISE: Randomized Input Sampling for Explanation of Black-box Models

Vitali Petsiuk,Abir Das,Kate Saenko

摘要

摘要: 深度神经网络越来越多地用于自动化数据分析和决策，但它们的决策过程很大程度上不清楚，很难向最终用户解释。在本文中，我们解决了以图像作为输入和输出类概率的深度神经网络的可解释人工智能问题。我们提出了一种称为 RISE 的方法，它生成一个重要性图，指示每个像素对模型预测的突出程度。与使用梯度或其他内部网络状态来估计像素重要性的白盒方法相比，RISE 可以在黑盒模型上工作。它通过对输入图像的随机掩蔽版本来探测模型并获得相应的输出，通过经验估计重要性。我们比较了我们的方法和最先进的重要性提取方法，同时使用了自动删除/插入度量和基于人工注释的对象段的指向度量。在几个基准数据集上的广泛实验表明，我们的方法匹配或超过了其他方法的性能，包括白盒方法。

关键词：黑盒模型；随机掩蔽版本；解释性

1 引言

对于图像识别和分类，深度神经网络已经发展得较为成熟，但是它们的识别方式和过程对用户来说几乎是透明的，为了解决解释性问题，作者提出了 RISE 的方法，该方法以图像作为输入，使用图像的随机遮掩版本来探测模型，并以概率图为输出，该概率是根据深度神经网络得到的，用于解释该深度神经网络是如何做出图像分类的，指示出了每个像素对模型预测的影响程度。这种方法适用于黑盒模型，因为它不需要考虑模型的参数，结构等等。此外，对于解释性的评估，使用了删除/插入来度量效率，当评估删除时，会从原图像开始，不断模糊一部分的区域像素，当然，会优先从影响程度较高的像素区域开始模糊，直到模糊整个图像，可以得到下降曲线，并定义 AUC 即曲线下面积作为衡量指标，曲线下降的越快，表示解释得越好，插入的方法类似于此。综上所述，该方法很好地解释了黑盒模型，也可视化了模型是如何对图像进行预测的过程，且与最先进的重要性提取方法相比，该方法匹配超过了其它方法的性能，也包括白盒方法。

2 Motivation

最近，深度神经网络的成功导致了人工智能（AI）研究的显著增长。尽管取得了成功，但目前仍然不清楚一个特定的神经网络是如何做出决定的，它对这个决定有多确定，是否以及何时可以信任，或者何时需要纠正。在决策可能产生严重后果的领域，尤其重要的是，决策模式必须是透明的。在认知心理学，哲学和机器学习研究中，有大量的证据表明解释对理解和建立信任的重要性。在本文中，我们解决了可解释人工智能的问题，即为人工智能模型的决策提供了解释。具体来说，我们感兴趣的是解释深度神经网络对自然图像的分类决策。现有的方法计算一个给定的基本模型（要解释的模型）和一个输出类别的重要性。然而，它们需要访问基本模型的内部内容，例如输出相对于输入、中间特

征图或网络权值的梯度。许多方法也局限于某些网络架构和/或层类型。在本文中，我们提倡一种更通用的方法，它可以为任意网络生成一个显著性映射，而不需要访问其内部结构，也不需要对每个网络体系结构重新实现。LIME 提供了这样一种黑盒方法，通过在要解释的实例周围抽取随机样本，并拟合一个近似的线性决策模型。然而，它的显著性是基于超像素，这可能不能捕获正确的区域。作者提出了一种新的黑盒估计像素显著性的黑盒方法，称为随机输入采样解释（RISE）。我们的方法是通用的，只需要把它作为一个完整的黑盒，而不假设访问其参数、特征或梯度。其关键思想是通过随机掩模对输入图像进行子采样，并记录其对每个掩模图像的响应，来探测基本模型。最终的重要性图是随机二进制掩模的线性组合，其中组合权值来自于掩模图像上的基础模型预测的输出概率。

3 相关工作

3.1 线性决策模型 (LIME)

产生解释的重要性已经在机器学习内外的多个领域得到了广泛的研究。在历史上，使用规则或决策树来表示知识的已经被发现是人类可以解释的。另一种研究方向集中于近似可解释性较差的模型（例如，神经网络、非线性支持向量机等）。使用简单的、可解释的模型，如决策规则或稀疏线性模型。在最新进展中, 提出了拟合一个更可解释的近似线性决策模型（LIME）。LIME 是一种可解释的近似线性决策模型，虽然局部性近似效果较好，但是对于一个足够复杂的模型，线性近似可能不会导致非线性模型的贴合表现。LIME 模型可以应用于黑盒模型，但是它对超像素的依赖导致了低重要性图，

3.2 类激活映射方法 (CAM)

为了解释图像中的分类决策，之前的工作要么是视觉上对决策，有强烈支持的地面图像区域，要么是生成为什么决策的文本描述。视觉基础通常表示为一个显著性或重要性图，它显示了每个像素对模型决策的重要性。现有的深度神经网络解释方法要么设计“可解释的”网络架构，要么试图解释或“证明”现有模型所做的决策。类激活映射方法通过计算所有通道在该位置的特征激活值的加权和来实现图像的每个位置的对于类的重要性。然而，这种方法只能应用于一种特定的 CNN 架构，在这种架构中，在分类层之前的卷积特征映射通道上执行全局平均池。Grad-CAM 通过用类分数的平均梯度来权衡每个位置的特征激活值。

4 本文方法

4.1 本文方法概述

衡量图像区域重要性的一种方法是模糊或者“干扰”它，并观察这种变化对黑盒模型的决策有多大的影响。例如，可以将像素强度设置为零，模糊区域或者通过添加噪声来实现。RISE 方法中，通过生成很多个遮掩的版本，通过将像素强度降低为零来实现遮掩，然后不断通过这种方法生成很多个掩膜版本然后放入到模型中，来获取每个像素的重要性。具体地，如题 1，我们通过调整像素的随机组合，将它们的强度降低到零来估计像素的重要性。通过将一副图像与一个 [0,1] 值的矩阵来生成掩膜。

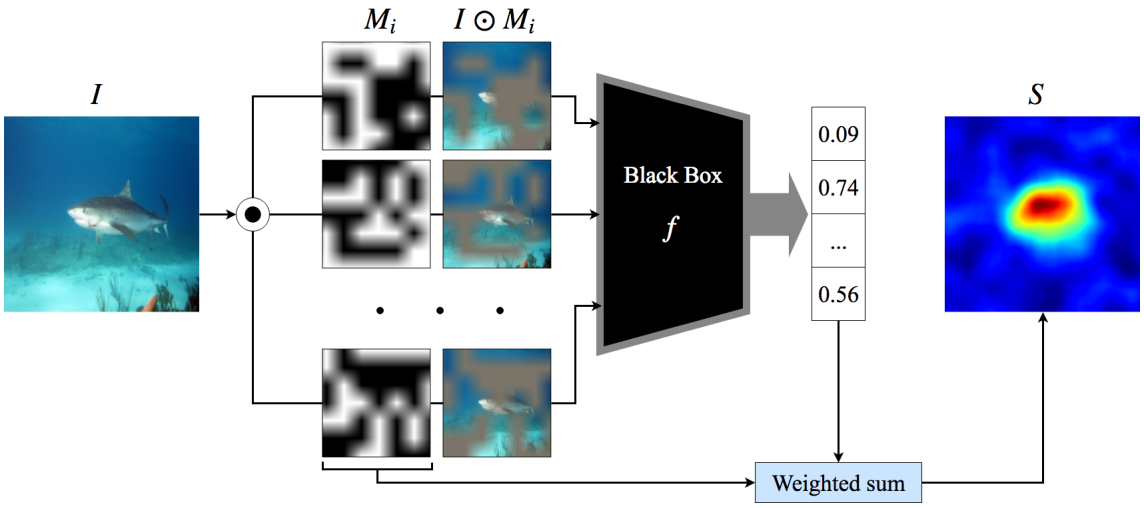

图 1: RISE 方法

4.2 掩膜生成

独立掩蔽像素可能会导致对抗性效应：像素值的轻微变化可能会导致模型的置信度分数的显著变化。此外，通过将其元素独立设置为 0 和 1 来生成掩模，将产生大小为 2H×W 的掩模空间。首先对更小的二进制掩模进行采样，然后使用双线性插值将上采样到更大的分辨率。插值后，掩模不再是二进制的，而是在 [0,1] 中有值。最后，为了允许更灵活的掩蔽，我们在两个空间方向上通过随机数量的像素来移动所有掩蔽。

4.3 模型解释

作者使用了 VOC 和 MSCOCO 数据集的特定版本，以与相同数据集和相同基础模型上的最新报告进行公平的比较。使用了由 ResNet50 和 VGG16 网络作为基础模型。对于 ImageNet，从 pytorch 模型动物园下载了相同的基础模型。而对于模型解释，掩膜生成后，使用模型进行预测，判断掩膜的像素点是否对预测结果具有影响，从而生成像素对应的影响度。周而复始，可以得到整张图片中每个像素的影响度。最后按照该影响度展示图片，以来解释模型是如何进行预测的。如图 2 所示。

图 2: 模型解释

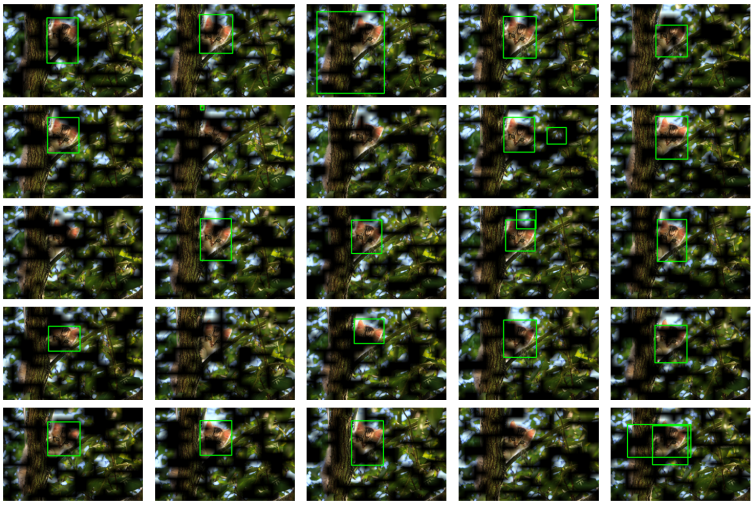

图 3: 掩膜对模型预测的影响

5 复现细节

5.1 与已有开源代码对比

复现了掩膜生成和模型解释的代码，引用了源代码中的评估代码，以来测试掩膜生成和模型解释是否正确。具体地，生成指定数量的只有 0 和 1 的等长等宽的矩阵，其中的 0 和 1 也随机，然后与图像的像素矩阵相乘作为掩膜后的图像，再放入 ResNet50 模型中进行预测，将生成的矩阵与处理前的矩阵进行对比，使用 IoU 的方法来计算每个像素点的得分。最后，将得分作为像素的权重输出图像。在掩膜生成之后，简单展示了一下掩膜对模型预测的影响情况。根据模型对掩膜后的图片的预测对每个像素进行权重打分，显然模型预测的物体的像素分数会相对较高。

5.2 解释结果的评估

尽管越来越多的研究集中于可解释的机器学习，但对于如何衡量机器学习模型的可解释性仍然没有达成共识。因此，人类评估一直是评估模型解释的主要方法，通过从透明度、用户信任或人类对模型决策的理解的角度来衡量它。而对于结果的评估，作者提出了两个自动评估指标：删除和插入。删除度量的判断是，删除“原因”将迫使基础模型改变其决策。具体来说，这个度量测量的是随着越来越重要的像素被删除，预测类的概率的较少，其中的重要性是从重要性图中获得的。在概率曲线下的急剧下降，从而导致低面积（作为移除像素的分数的函数）意味着一个很好的解释。另一方面，插入度量采用了一种互补的方法。它测量了随着越来越多的引入概率的增加，更高的 AUC 表明更好的解释。

5.3 评估模型的伪代码

Procedure 1 Deletion

Input: black box *f,* image *I,* importance map *S* number of pixels *N* removed per step
Output: deletion score *d*

$n \leftarrow 0$
$h_n \leftarrow f(I)$
while I has non-zero pixels **do do**
    $According\ to\ S,\ set\ next\ N\ pixels\ in\ I\ to\ 0$
    $n \leftarrow n+1$
    $h_n \leftarrow f(I)$
end
$b \leftarrow$ AreaUnderCurve ($h_i$vs.i/n, $\forall$i = 0, ...n)
return b

Procedure 2 Insertion

Input: black box *f,* image *I,* importance map *S* number of pixels *N* removed per step
Output: insertion score *d*

$n \leftarrow 0$
$I' \leftarrow Blur(I)$
while $I \neq I'$ **do**
    $According\ to\ S,\ set\ next\ N\ pixels\ in\ I'\ to\ corresponding\ pixels\ in\ I$
    $n \leftarrow n+1$
    $h_n \leftarrow f(I)$
end
$d \leftarrow$ AreaUnderCurve ($h_i$vs.i/n, $\forall$i = 0, ...n)
return d

6 实验结果分析

在生成掩膜和像素权重分数后，基本就完成了对模型预测的解释，然后利用源代码中的评估技术对得到解释模型进行评估。评估技术主要分为两种，删除和插入。删除是指从原图像开始，逐渐删除解释模型中分数较高的像素点，然后再放入网络中得到预测结果。显然，从分数较高的像素点开始删除，模型预测的结果会迅速下降，如果将预测结果画成一个曲线，那么可以使用曲线下面积作为评估标准 (AUC)，对于删除来说，AUC 应该越小越好，表示解释模型中分数较高的像素对模型预测具有明显的影响，简单来说，AUC 越小，解释的模型更好。对于插入来说，是从一张完全掩膜的图片开始，不断加入分数较高的像素，具体过程与删除相反，不再赘述。对于解释模型，可以得到以下的评估。4 5 6

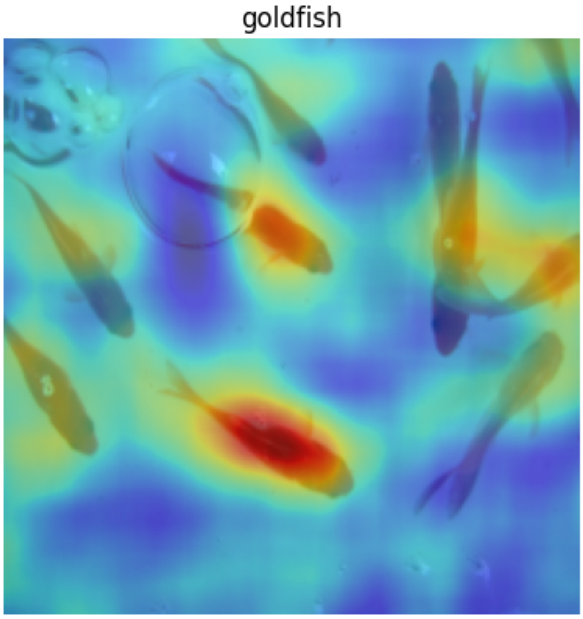

图 4: 解释模型样例

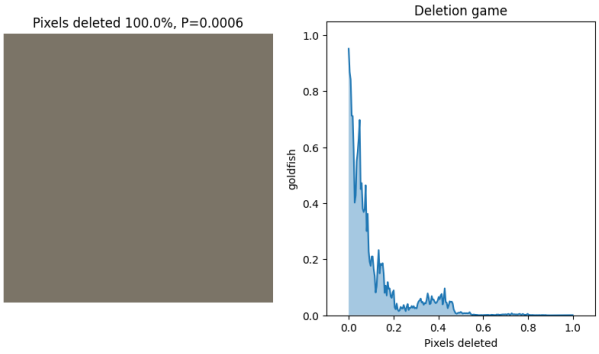

图 5: deletion

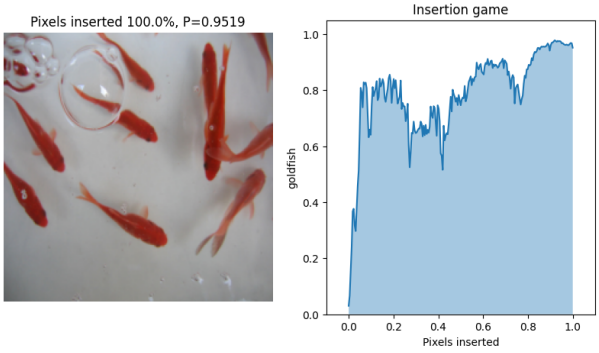

图 6: insertion

7 总结与展望

本文提出了一种通过估计输入图像区域对模型预测的重要性来解释黑盒模型的方法。尽管它的简单些和通用性，该方法在自动因果度量方面优于现有的解释方法。未来的工作有望利用该方法的普遍性来解释和其他领域的复杂网络所做的决策。但是 RISE 同样具有一定的缺点，RISE 无法摆脱背景噪声的影响，虽然 RISE 确实获得了更重要的特征，但是受噪声的影响，解释出来的模型不那么明显，同时，评估出来的效果同样不好，因为噪音容易与实体对象具有相似性，造成对模型预测的干扰。