软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2022,33(9):3391−3406 [doi: 10.13328/j.cnki.jos.006640]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

目标检测模型的决策依据与可信度分析*

平昱恺[1], 黄鸿云[2], 江 贺[3], 丁佐华[1]

[1](浙江理工大学 信息学院, 浙江 杭州 310018)

[2](浙江理工大学 图书馆, 浙江 杭州 310018)

[3](大连理工大学 软件学院, 辽宁 大连 116081)

通信作者: 黄鸿云, E-mail: huanghongyun07@hotmail.com

摘 要: 目标检测模型已经在很多领域得到广泛应用, 但是, 作为一种机器学习模型, 对人类来说仍然是一个黑盒. 对模型进行解释有助于我们更好地理解模型, 并判断其可信度. 针对目标检测模型的可解释性问题, 提出将其输出改造为关注每一类物体存在性概率的具体回归问题, 进而提出分析目标检测模型决策依据与可信度的方法. 由于原有图像分割方法的泛用性较差, 解释目标检测模型时, LIME 所生成解释的忠诚度较低、有效特征数量较少. 提出使用 DeepLab 代替 LIME 的图像分割方法, 以对其进行改进. 改进后的方法可以适用于解释目标检测模型. 实验的对比结果证明了所提出改进方法在解释目标检测模型时的优越性.

关键词: 可解释性; 可信度分析; 目标检测; 机器学习; 深度学习

中图法分类号: TP18

中文引用格式: 平昱恺, 黄鸿云, 江贺, 丁佐华. 目标检测模型的决策依据与可信度分析. 软件学报, 2022, 33(9): 3391–3406. http://www.jos.org.cn/1000-9825/6640.htm

英文引用格式: Ping YK, Huang HY, Jiang H, Ding ZH. Decision Basis and Reliability Analysis of Object Detection Model. Ruan Jian Xue Bao/Journal of Software, 2022, 33(9): 3391–3406 (in Chinese). http://www.jos.org.cn/1000-9825/6640.htm

Decision Basis and Reliability Analysis of Object Detection Model

PING Yu-Kai[1], HUANG Hong-Yun[2], JIANG He[3], DING Zuo-Hua[1]

[1](School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

[2](Library, Zhejiang Sci-Tech University, Hangzhou 310018, China)

[3](School of Software Technology, Dalian University of Technology, Dalian 1160081, China)

Abstract: The object detection model has been widely applied in many fields; however, as a machine learning model, it remains a black box to humans. Interpreting the model is conducive to a better understanding of the model and can help judge whether the model is reliable. In view of the interpretability problem of the object detection model, this study proposes that the output of the model should be changed into a specific regression problem that focuses on the existence possibility of the objects of each class. On this basis, the methods to analyze the decision basis and reliability of the object detection model are put forward. Due to the poor versatility of the original image segmentation method, LIME generates unfaithful and ineffective interpretations when interpreting the object detection model. Therefore, the image segmentation method with LIME replaced by DeepLab is put forward and improved, and the improved method can interpret the object detection model. The experiment results prove the superiority of the improved method in interpreting the object detection model.

Key words: interpretability; reliability analysis; object detection; machine learning; deep learning

目标检测 (object detection) 是指在给定图片中, 标出目标物体的边界框与类别, 属于图像处理问题. 目标检测模型被广泛应用于医疗、监控、自动驾驶、安防等领域, 并且通常是以上领域软件系统中不可或缺的部分. 近年来, 基于深度学习的目标检测模型在检测能力上得到了巨大提升. 但是, 与传统检测算法不同, 由于深度学习模型

* 基金项目: 国家自然科学基金 (61751210)
收稿时间: 2021-06-23; 修改时间: 2021-08-08, 2021-10-07, 2021-11-10; 采用时间: 2021-12-28; jos 在线出版时间: 2022-05-24

的不可解释性, 使用者无法得知模型根据哪些信息做出决策, 因此, 其可靠性始终受到质疑. 如果目标检测模型始终保持其“黑盒”特性, 将永远无法被应用于需要高可靠性的场景中.

目前, 虽然存在许多对深度学习模型进行解释的方法, 但对目标检测模型可解释性的研究几乎是一片空白. 由于大多数目标检测模型基于 CNN 构建, 本文从现有解释思路出发, 对解释 CNN 分类模型的方法进行改造, 使其适用于目标检测模型.

LIME (local interpretable and model-agnostic explanations)[1]是一种可用于解释所有回归问题模型 (regression model) 的解释器, 其解释 CNN 分类模型的方法和结果对解释目标检测模型有重大启发意义, 但无法直接应用于目标检测模型. 导致这种情况的主要原因有两个, 一是目标检测模型的输出较多, 无法直接作为 LIME 的输入; 二是图像分割方法的泛用性较差所导致的解释忠诚度低与有效特征数量少. 为解决此问题, 本文首先对目标检测模型的输出进行改造, 将其转化为简单回归问题, 从而使模型的输出符合 LIME 的使用条件. 其次, 使用语义分割代替原始 LIME 的图像分割方法, 使 LIME 适用于目标检测模型. 最后, 使用 LIME+DeepLab 解释目标检测模型, 得到模型的决策依据, 并基于交并比 (IoU), 提出对模型预测可信度的评价方法.

对目标检测模型进行解释, 揭示其决策依据, 评估其可信度, 在模型部署、模型理解、模型优化等方面具有重大意义. 例如, 在医疗影像识别、辅助驾驶等任务场景中, 可对模型的决策可信度进行分析. 基于对解释方法实时性的考虑, 我们将其实际应用分为以下两类. 对于实时性要求较低的任务, 可在模型工作时将其决策依据与可信度提供给用户, 在提高系统可靠性的同时, 帮助用户做出更好的决策. 对于实时性要求较高的任务, 可在任务结束后, 对模型的可信度表现进行评估.

本文主要贡献包括以下几点.

(1) 提出利用语义分割模型 DeepLab 代替 LIME 所使用的图像分割方法, 从而使 LIME 适用于目标检测模型.

(2) 分析并揭示 LIME 解释目标检测模型时的问题: 局部线性回归模型的忠诚度太低、权重太小. 详见第 4 节.

(3) 在解释目标检测模型时, 将其输出改造为关注每一类物体存在性概率的具体回归问题. 详见第 3.1 节.

(4) 提出使用 IoU, 在得到决策依据后, 可以在有标签数据集中对模型每一次预测的可信度进行定量计算.

本文第 1 节对解释目标检测模型的相关工作进行介绍. 第 2 节交代必要的背景知识. 第 3 节对什么是目标检测模型的决策依据做出定义, 并给出可信度分析的方法. 第 4 节具体分析 LIME 在解释目标检测模型时的不足及原因, 并给出改进方法. 第 5 节对使用的模型和数据集进行设置. 第 6 节通过实验, 证明本文所作改进的有效性. 第 7 节对文章进行总结, 对文中方法的不足进行分析, 并给出对未来工作的展望.

1 相关工作

1.1 解释深度模型的方法

虽然目前没有针对目标检测模型的解释方法, 但是目标检测大多基于 CNN 实现, 因此在本节中对适用于 CNN 的解释器进行介绍.

对深度学习模型的解释主要可分为两类[2].

(1) 将模型看作黑盒的模型无关解释 (model-agnostic). 这类方法不关注模型内部的处理过程, 只关注模型的输入与输出. 通过观察输入与输出之间的关系, 对模型的行为进行分析. LIME[1]、SHAP[3]、Anchors[4], 以及 MUSE[5]都属于此类解释.

(2) 研究模型内部结构的解释. 这类方法对训练后模型的参数和输入数据的具体处理过程进行分析, 从而揭示模型内部的运行过程. 此类解释方法的主要关注点有两个: 一是模型内部参数代表的具体意义, 二是输入数据在模型处理过程中的不同阶段所代表的不同具体意义. Bach 等人[6]、Shrikuma 等人[7] 、Amini 等人[8]提出的方法都属于此类解释.

由于在使用第 2 类方法解释目标检测模型时存在较大难度, 我们仅对黑盒解释方法作简要描述.

从解释目标检测模型的角度来看, LIME、SHAP 和 Anchors 这 3 种方法都使用了不合理的输入图像分割算法, 因此均无法生成有效的解释. MUSE 虽然在理论上可用于解释目标检测模型, 但需要对解释器本身和目标检测

模型的输出进行复杂的改造, 因此在实现上存在一定难度. 本文提出的方法, 针对 LIME 的不足进行改进, 不仅保留了原始 LIME 的易懂性和简洁性, 而且能对目标检测模型进行解释.

1.2 评价方法

缺少被广泛认可的统一定量评价方法是解释深度学习模型领域中存在的另一个问题. 在深度学习模型的可解释性领域, 需要以下两类评价方法.

(1) 评价模型的质量, 指基于模型的解释结果, 对模型的可信度、质量等方面进行评价. 换言之, 就是对模型的定量解释. 在许多经典的解释方法中, 由于解释器的实现方式各不相同, 作者通常使用自定义的标准进行模型可信度评价, 目前并不存在统一的模型可信度评价方法.

(2) 评价解释器的质量, 指对用于解释深度学习模型的解释器进行评价, 判断该解释器是否能够正确地解释模型. Camburu[9]、Fan 等人[10]提出的方法可对解释器进行评价, 但其适用性不足以包括现存的大部分解释器. LIME 原文中提出的评价方法仅适用于自身可解释的机器学习模型, 比如决策树.

本文的重点在于对目标检测模型进行解释, 因此主要关注图像问题模型的可信度评价方法. 由于现存的解释方案大多对此类模型进行定性解释, 且缺乏评价的基准 (指模型的基准决策依据). 即使是泛用性强的解释器 (比如 LIME), 作者也大多选择避开对图像问题模型的两类评价方法进行研究, 转而使用 NLP 问题模型或普通的深度学习问题模型 (ANN) 进行定量评价. 因此, 提出评价目标检测模型可信度的方案具有一定意义.

1.3 目标检测模型

近年来, 学界对目标检测问题的研究越发深入, 提出了大量基于深度学习的目标检测模型, 并在各类数据集中表现出较好的性能. 一些基于传统 CNN 的目标检测模型, 如 Faster R-CNN[11]、YOLOv1[12]、YOLOv2[13]、YOLOv3[14]、YOLOv4[15]、YOLOX[16]等, 能在 PSACALVOC、COCO 等数据集中达到相当高的准确率. 除了基于 CNN 的模型, 随着自注意力机制[17]的广泛运用, 有大量基于 Transformer 的目标检测模型被提出, 如 DETR[18]、TPH-YOLOv5[19]、ViT-FRCNN[20]、Deformable DETR[21]等, 同样有不俗的性能. 此外, 也有许多基于某些特殊网络架构的模型, 比如基于脉冲神经网络的 Spiking-YOLO[22]和基于 Matrix Nets 的模型[23], 这些模型在某些特定领域 (比如小目标检测) 中有相当惊艳的表现.

本文的方法是由 LIME 改进而来, 继承了其强大的泛用能力, 只要模型把目标检测看作回归任务, 就能对其进行决策依据与可信度的分析. 基于深度学习的目标检测模型可被分为两类: One-Stage 和 Two-Stage[24]. 对于 One-Stage 模型, 本文方法可直接应用 (上述模型中只有 Faster R-CNN 属于 Two-Stage). 对于 Two-Stage 模型, 通过在候选框预测和物体分类两个阶段分别应用本文方法, 也可获得对应的解释, 但需要按本文第 3.1 节中对物体存在性概率的描述, 对模型两个阶段的输出进行类似处理.

2 背景知识

本文的核心工作是改进 LIME 方法, 以实现对目标检测模型进行决策依据与可信度分析. 因此, 目标检测模型与 LIME 的工作原理是本文的前置条件.

2.1 目标检测

目标检测是指在已经确定所有目标物体类别的情况下, 在给定图像中标定出所有物体的边界框与对应类别. 换言之, 回答了“图片中有哪些物体?分别在哪里?”这两个问题. 模型的输入是一张图片, 输出由以下 3 部分组成. 模型输出经过处理后的可视化检测结果如图 1 所示.

(1) 每个物体的边界框

边界框, 指限定物体位置范围的矩形框. 模型以输出矩形框在图片中二维坐标的形式, 确定每个物体对应的边界框. 若模型输出一个边界框, 则代表其认为该边界框中存在物体.

(2) 每个物体的类别概率

类别概率, 指边界框内物体从属于某一个类别的概率.

(3) 每个物体的置信度概率

置信度概率, 指边界框内物体存在的概率.

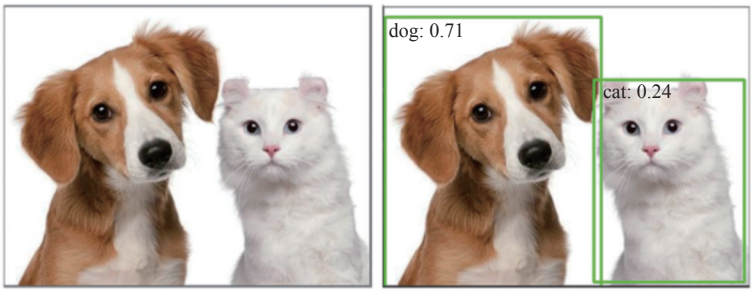

图 1　目标检测模型的处理结果

2.2 LIME

LIME (局部可解释的模型无关解释)[1]是一种将回归问题模型 (Regressor) 看作黑箱的局部[25]解释方法. 该方法针对模型的每一次预测 (Instance) 进行解释, 通过在局部使用线性回归模型对回归器的行为进行拟合, 给出每一个输入变量 (Feature) 对模型输出结果的影响. 理论上, LIME 的原理适用于所有回归器, 因此我们从 LIME 出发, 对目标检测模型进行解释.

2.2.1 LIME 解释 CNN 分类模型

目标检测模型大多基于 CNN 构建, 因此对 LIME 解释 CNN 分类模型的流程进行介绍, 以对 LIME 有基本了解.

将被解释的 CNN 分类模型称为 F, LIME 解释 F 的具体流程如下.

(1) 图像分割与图像扰动

图像分割, 指将 F 的输入图像分割为图像块, 并将图像块染色为该区域的颜色平均值. 具体样例如图 2 所示.

图 2　图像分割样例

经过分割, 原图可看作由 N 个图像块组成. 现规定图像向量, 值域为 {0, 1}, 大小为 N×1, 向量内某位置值若为 1, 代表该位置使用左图中图像块; 若值为 0, 代表使用右图中图像块. 所谓图像扰动, 即生成多个图像向量, 向其中随机填充 0、1, 从而得到某些图像块存在与不存在的多种情况.

(2) 在局部 (Local) 训练线性回归模型

假设图像扰动后, 得到 M 图像向量 (大小为 N×1), 将 M 个新图像输入给 F, 得到 M 个预测结果向量. LIME 的核心思想就在于: 观察某张图片中图像块变化对 F 预测结果的影响, 从而得到 F 的决策依据. 局部线性回归模型 (G) 的训练数据集是 M 个图像向量, 其 Label 是 M 个预测结果向量.

简而言之, 线性回归模型 G 就是对 F 的解释结果.

(3) 解释结果分析与可视化

定义 1. 特征图像块. 图片分割后得到的图像块.

G 中有 N 个权重 (weight), 对应原图分割出的 N 个特征图像块. 每个权重的值, 代表该特征图像块对 F 预测的影响 (支持它将图像分为某个类, 或反对).

若某个特征图像块对应的权重大于 0, 则它对模型预测该类起积极作用; 反之起消极作用. 绿色代表特征图像块权重大于 0, 红色代表小于 0. F 对某一图片类别进行预测的决策依据分析如图 3 所示, 展示的是 LIME 解释为什么 F 会将该图分类为“cat”.

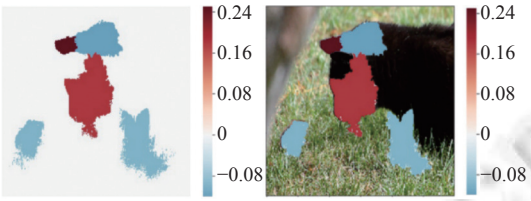

图 3 LIME 解释 CNN 分类模型样例

3 目标检测决策依据及可信度分析

从第 2.2 节中对 LIME 的描述可以看出, 使用 LIME 对 CNN 分类模型进行解释, 其实就是得到该模型的决策依据. 本文核心内容是对目标检测模型进行决策依据和可信度分析, 因此也属于对深度学习模型的可解释性研究.

为避免概念的混淆, 并对本文的两个核心内容做详细介绍, 在本节中对决策依据、可信度评价这两个名词做详细说明.

3.1 决策依据

跟从 LIME 方法, 将决策依据的形式定义为特征图像块及其对应的权重值[1].

将输入图片的图像块作为决策依据的意义在于, 找出模型基于哪些图像块做出决策. 以图 1 为例, 模型检测出“cat”与“dog”两个物体, 本文的目的是找出哪些图像块导致“cat”这一预测结果, 哪些又导致“dog”. 通过对其决策依据进行分析, 可以判断模型的决策是否合理、可靠.

由第 2.1 节可知, 目标检测模型的工作内容包括边界框定位与物体存在性预测两个方面, 因此, 对决策依据的分析理应从以上两方面出发.

(1) 物体存在性预测

对存在性预测作决策依据分析, 就是找出模型判断图片中存在某类物体的原因. 模型对物体存在性的预测包括两个内容: 存在性概率与类别概率. 若规定物体类别数量为 C, 模型预测存在 N 个边界框, 则有对应的 N 个物体存在性概率和 $N\times C$ 个对应的分类概率. 本文不探究每个具体预测出现的原因, 仅分析整体决策行为. 举例说明, 若模型预测图片中存在两个不同位置与置信度的“dog”类物体, 我们不具体分析每个“dog”被预测的原因, 仅分析其认为图片中存在“dog”类的原因.

定义 2. 类别存在性概率. 计算边界框内物体的存在性概率与类别概率之乘积, 可以得到该物体存在且属于某一类别的概率, 即边界框中的类别存在性概率.

因此, 对模型预测结果进行处理, 取每一类别中类别存在性概率的最大值, 得到每一类物体的最大存在性概率, 该结果应是大小为 C 的向量, 称之为物体存在性向量. 通俗来讲, 物体存在性向量的含义是“对于所有类别, 模型预测该类物体存在的概率是多少”. 至此, 我们将目标检测的第一类工作内容转化为具体的回归问题, 可以直接通过 LIME 方法进行解释.

所谓具体的回归问题, 指处理后目标检测模型输出的物体存在性向量的结构与 CNN 分类模型的分类概率向量类似. 事实上, 单纯从数值角度来看, 若把模型看作黑盒, 此时的目标检测模型与 CNN 模型没有任何区别, 因为其输入与输出格式完全一致; 区别仅在于输出中数值代表的意义不同.

(2) 边界框定位

对边界框定位作决策依据分析, 就是分析图像块会如何影响边界框的坐标值. 目标检测模型输出的边界框数

量较多, 且难以转化为具体的回归问题. 因此, 本文对如何解释边界框定位问题不作探究. 后文中提到的决策依据, 均指模型对物体存在性进行预测的决策依据.

3.2 对目标检测模型的预测进行可信度评价

本文中的可信度评价, 指第 1.2 节中的第一类评价方法, 对模型每一次预测的可信度进行计算, 其前置条件是获得模型的基准决策依据. 我们将模型可信度定义为模型实际决策依据与基准决策依据的重合度. 因此, 探究模型可靠性时, 我们不仅要知道“模型依据哪些图像块进行决策”, 还要知道“模型应该根据哪些图像块进行决策”.

定义 3. 基准决策依据. 在模型训练过程中, 人们希望模型去关注的部分输入.

在普通的机器学习任务中, 如 Monk’s Problem, 存在基本输入单位, 因此至少可以通过人工选择的方式得到基准决策依据. LIME 原文中通过人工制造类似数据集, 对使用其解释结果进行模型可信度评价的流程进行说明. 而在图像类任务中, 不存在基本输入单位, 更不可能人工制造数据集, 导致难以获取基准决策依据.

目标检测问题中数据集的独有性质使我们可以简便地获得基准决策依据. 数据集以边界框的形式标定所有物体的范围和对应类别, 直观看来, 框内的图片内容就是模型预测该物体时的基准决策依据. 如图 4 中左侧图所示, “car”类物体被其外接矩形所包裹, 代表在模型预测图片中存在“car”类物体时, 边界框内的图像即是基准决策依据.

图 4 目标检测模型基准决策依据的具体样例

在 CV 领域中, 交并比 (IoU) 常被用于计算图片中的区域重合度, 且是计算目标检测模型性能的重要指标之一. 因此, 我们在计算模型可信度时直接套用这一方法, 通过计算实际决策依据与基准决策依据之间的 IoU, 可以得到简单的模型可信度. 上文所述的可信度计算过程, 完全符合本文对可信度的定义.

假设得出的解释结果在图片中所占部分为 A, 基准决策依据所占部分为 B, 其 IoU 的计算公式如公式 (1) 所示. 由于在目标检测的训练数据集中, 边界框是物体的外接矩形, 因此本文认为若 IoU 达到 0.5 以上, 则可说明模型的该次预测较为可信.

$$IoU = \frac{A \cap B}{A \cup B} \tag{1}$$

具体案例见图 4 中右侧图, 绿色部分为 LIME+DeepLab 得出的解释, 可视化结果中标出了目标检测模型检测出图片中存在“car”物体时所用到的决策依据图像块, 计算 IoU 指标, 得到其值为 0.76. 通过在数据集中进行大量实验, 可以得出某一目标检测模型的总体可信度.

4 基于 LIME 的目标检测模型解释

4.1 传统 LIME 在目标检测模型上应用的局限性

理论上, LIME 可以用于解释所有回归器, 因此也可用于目标检测模型. 但在不对模型输出进行处理的情况下, LIME 的解释结果是没有意义的. 通过遵循第 3.1 节中对模型输出的处理, 使得 LIME 可以真正用于解释目标检测模型.

在此基础上, 通过观察解释结果, 我们发现, 尽管修改后的目标检测模型符合 LIME 的使用条件, 但 LIME 在目标检测模型上并不适用, 无法得到有效的决策依据 (解释结果).

在对目标检测模型进行决策依据分析时, 原始的 LIME 方法主要存在以下两个问题.

(1) 局部线性回归模型的 R^2 过小 (忠诚度过低)

LIME 的核心是通过训练局部的线性回归模型 (G), 揭示每个输入变量对模型预测结果的影响程度. 因此, G 对被解释模型 (F) 是否具有足够的忠诚度 (Faithfulness), 是评判解释器本身可信度的重要因素. 如果 G 的忠诚度高, 说明其能在局部可靠地拟合 F, 即能得到可信的解释器. 所谓忠诚度, 也就是线性回归模型在局部的行为与目标检测模型有多接近. G 训练的数据来自于 F 的输入和输出, 即训练集是对 F 行为的描述, 因此, 训练后 G 在训练集上的性能就可以体现它对原始模型的忠诚度.

R^2 是广泛被使用的评判线性回归模型的标准之一, 也被称为相关指数或拟合优度. 其定义如下所示. 公式中 y_i 代表数据集中的标签值 (label), $\hat{y}_i$ 代表线性回归模型的预测值, $\bar{y}_i$ 代表数据集中标签值的平均值.

$$R^2=\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2} \tag{2}$$

在线性回归模型中, R^2 值代表的是因变量的变化能被自变量用线性关系解释的比例. 因此, 我们可以使用 R^2 来判断, 简单的线性回归模型能否在局部拟合一个复杂的目标检测模型. 换言之, 判断目标检测模型的行为在局部能否用图像块的线性关系来解释.

线性回归模型的训练集是对目标检测模型行为的描述, 因此, R^2 值越高, 证明线性回归拟合的正确度越高, 且因变量能更好地被自变量解释. 同时, 较高的 R^2 从侧面说明用于拟合的数据之间存在线性关系, 这也符合 LIME 使用的前提条件: 被解释模型的预测行为应在局部是线性的, 否则无法使用线性回归模型进行拟合. 由于 LIME 原文中没有具体提到如何对生成的局部模型的忠诚度进行计算, 我们暂时只使用 R^2 作为忠诚度的评价标准. 在使用原始的 LIME 解释目标检测模型时, 发现有一部分线性回归模型的 R^2 值非常小 (在 0.3 以下). 观察发现, 此类存在问题的线性回归模型在第 5.2 节中定义的大物体图片中出现十分频繁. 在大物体数据集中进行实验, 发现得到的线性回归模型 R^2 平均值只有 0.46. 由此可以初步得出结论, 原始的 LIME 方法在解释输入图片为大物体图片的模型预测时, 无法得到忠诚度高的线性回归模型.

原因猜测: 目标检测模型与 CNN 图像分类模型对图片中内容关注的角度不同, 目标检测模型更加关注图片中物体的整体性, 因此若使用原始的分割方法, 将图像分成小块, 在扰动过后, 由于物体的整体性没有受到较大的破坏, 目标检测模型的预测结果基本不变, 从而导致无法获得一个忠诚度高的局部线性回归模型. 简而言之, 对于大物体图片, 扰动后的模型预测结果几乎不变, 导致在无效数据集上训练线性回归模型.

原因证明 (1): 增加 QuickShift[26]分割算法的核值, 使其对图像块边界更不敏感, 即在分割后, 得到的每个图像块面积更大, 图像块总数更少, 更改核值前后的分割结果对比如图 5 所示. 修改分割算法后, 在同一数据集进行测试, 得到线性回归模型的 R^2 平均值可达到 0.71, 具体实验结果如表 1 所示. 初步可以证明小块的图像分割算法不适用于解释目标检测模型.

图 5 使用不同核值的 quickshift 图像分割结果

原因证明 (2): 上述实验过程中, 通过观察改变分割块大小后 R^2 平均值的变化程度, 来侧面证明小块的分割不适用于目标检测模型. 有更直接的证明方法, 观察改变核值前后目标检测模型预测的变化程度, 即可用数据直接证

明, 目标检测模型是否对图像中物体的微小变化不敏感.

通过实验, 同样在大物体数据集上进行测试, 对每种分割方法的图片产生 800 个扰动图片, 计算目标检测模型对其预测的变化, 结果如表 2 所示.

表 1 不同核值下的局部线性回归模型 R^2 平均值

目标检测模型	QuickShift 核值	R^2平均值
YOLOv1	4	0.46
	10	0.71
YOLOv2	4	0.32
	10	0.50
YOLOv3	4	0.27
	10	0.52

表 2 不同核值下的目标检测模型预测值变化比例

目标检测模型	QuickShift 核值	预测值平均变化比例	预测值最小变化比例	预测值变化比例小于0.2占比
YOLOv1	4	0.37	0.08	0.25
	10	0.46	0.17	0.02
YOLOv2	4	0.14	0.02	0.74
	10	0.29	0.05	0.34
YOLOv3	4	0.08	0.02	0.85
	10	0.27	0.04	0.39

将扰动前的模型预测值定义为 $P1$, 扰动后的模型预测值定义为 $P2$, 表中的变化比例定义为:

$$\frac{|P1-P2|}{P1} \tag{3}$$

实验结果说明, 使分割图像块增大后, 目标检测模型对图像的扰动更加敏感, 尤其显著的指标是 YOLOv1 模型的实验中, 在改变分割算法后, 预测结果变化率小于 0.2 的样本数仅占 0.02. 在一定程度上可以证明, 因为目标检测模型对物体局部的变化不敏感, 所以图像块较小的分割方法不适用于解释目标检测模型, 从而导致无法在局部得到忠诚度较高的线性回归模型.

事实上, 仅改变图像分割时分割图像块的大小并不能解决这一问题. 本文认为 LIME 的作者也考虑到保持物体完整性的问题, 因此没有使用简单的均匀分块方法对图像进行分割, 而是使用 QuickShift 算法, 保证局部的物体完整. 但在目标检测模型中, QuickShift 算法已经不能提供解释所需的整体的物体完整性, 因此本文提出使用语义分割模型对图片进行分割.

(2) 局部线性回归模型的权重过小

使用原始 LIME 方法对目标检测模型进行解释时的另一个问题是: 局部线性回归模型的权重值非常小, 基本处于 10^{-2} 级别, 偶尔有值会大于 0.1. 此问题说明原始 LIME 方法无法从输入图像中找出对目标检测模型决策影响力较大的图像块, 即无法找出有效决策依据.

在 VOC2007 数据集上, 设置核值为 4 进行实验时, 将 0.1 作为特征图像块对应权重的阈值 (权重绝对值大于 0.1 则视为有效特征, 反之视为无效特征), 大部分情况下, 每张图片中仅能得到权重值较低的 1~2 个有效特征, 权重平均值为 0.04 且最大权重值为 0.31. 根据原文观点, 可以认为通过 LIME 解释后, 发现所有特征图像块对预测结果的影响程度均较低. 若设置核值为 10 进行实验, 可以观察到权重高的特征数明显增加. 换言之, 当使用的特征图像块较小时, 原始的 LIME 无法从图像中找出决定性决策依据.

使用具体的例子对上述问题进行说明, 使用核值为 4 的 QuickShift 分割方法进行解释的具体样例见图 5 与图 6, 沿用上文的可视化设置. 由上述两图可见, 使用核值为 4 时, 解释所得的有效特征图像块的数量与面积占比均极低. 若仅增大核值, 则会导致物体与背景融合, 得到不合理的决策依据.

图 6 核值为 4 的 QuickShift 图像分割与模型解释结果

上述问题可总结为 3 点.

(1) 特征图像块权重整体偏小.

(2) 有效特征数量低.

(3) 原始分割方法泛用性差

初步猜测其原因是目标检测模型更加关注物体的整体, 导致图像分割后得到的小图像块无法对模型的预测结果造成显著影响. 局部线性回归模型中的较小权重, 正是对这一特点的体现. 此问题同样在大物体图片中出现地更为频繁. 由此可初步得出结论, 为解决该问题, 需要一种更合理的图像分割算法, 以找出图片中的决定性决策依据.

注意, 在实际使用 LIME 解释时, 有效特征图像块的权重大小应随它对目标检测模型的影响而有所不同, 上文实验测算权重值大小的目的是通过几乎所有的权重值均较小, 以证明原始的分割方法无法找出对目标检测模型影响大的图像块. 因为在实验数据集中, 必然存在能被模型所检测出的物体, 此时若无法找出对应的决策依据, 说明解释模型时的图像分割方法不合适.

4.2 改进 LIME

本节使用 DeepLab 代替原本的图像分割方法, 得到了可用于解释目标检测模型的解释器 (LIME+DeepLab).

4.2.1 问题分析

通过对第 4.1 节中两个问题的研究发现, 图像块的分割大小与 R^2、权重值、变化率之间可能存在某种正相关的联系, 可能与目标检测模型在局部行为的线性程度有关. 因此, 使用原分割算法, 并增大其核值可能是一种方法.

由 QuickShift 分割方法的原理可知, 在增加核值后, 虽然可以观察到以上几个指标明显增加, 但是会产生新的问题: 物体与背景的边界被模糊. 增大核值本质是使分割算法对分割块的边界更加不敏感, 会导致部分背景与物体无法区分, 从而无法解释目标检测模型的决策依据是物体还是背景. 具体案例见图 5, 当核值为 10 时, 可以观察到分割后相当一部分背景与主要物体混合在一起.

基于对决策依据简洁性的要求, 我们认为在分割图像时应做到物体与物体区分, 物体与背景区分. 显然, QuickShift 等方法并不合适, 因此需要一种新的图像分割方法, 并满足以下条件.

(1) 能够区分物体与背景;

(2) 能够区分不同的物体;

(3) 尽可能保留每个物体的完整性.

4.2.2 新的图像分割方法

使用语义分割 (semantic segmentation) 模型: DeepLab[27], 可以满足以上条件. DeepLab 是一种基于深度学习的语义分割模型, 其图像分割结果如图 7 所示.

图 7 DeepLab 的图像分割结果

第 6.2 节的实验结果证明, DeepLab 确实在目标检测问题上解决了 LIME 的局限性. 在此, 我们对 DeepLab 起作用的原因做出分析. 在第 4.1 节中已经说明, 由于目标检测模型更加关注图片中物体的整体性, 所以 LIME 存在局限性的根本原因是原始图像分割方法不能分割出对模型决策有较大影响力的图像块. 然而, DeepLab 作为语义

分割模型, 在分割出大影响力图像块这方面有着先天优势. 具体来说, 目标检测的任务是找出图片中的目标, 语义分割的目标是标记出图片中含有语义信息的部分, 这两个任务存在很大程度上的重叠. 在绝大多数情况下, 图片中的目标肯定包含广义上的语义信息, 因此, 只要使用一个性能足够好的语义分割模型, 必然能从图片中找出对目标检测模型的决策具有较大影响力的图像块, 也就解决了原始图像分割方法的不足.

一个不容忽视的问题是, 作为一种深度学习模型, DeepLab 本身具有一定的不可解释性, 将它作为解释目标检测模型过程中的一种工具, 是否可行? 我们从 DeepLab 自身的性能和不可解释性是否会传递到解释结果中这两个角度来论证使用 DeepLab 作为解释过程中特征选取方法的可行性.

首先, DeepLab 模型的性能较强, 在 VOC2012 数据集中测试的 mIOU 指标达到 87.8%. 在保留物体完整性的同时, 对不同物体、背景的分辨能力很强. 从图像语义分割质量的角度来看, DeepLab 的性能已经足够.

其次, DeepLab 的不可解释性不会传递到解释结果中. DeepLab 在进行语义分割的过程中, 存在深度学习模型普遍的不可解释性, 换言之, DeepLab 决策背后的原因是未知的. 但在对目标检测模型进行解释时, 研究对象是目标检测模型, 且仅使用 DeepLab 的分割结果; 单从分割结果来看, 在完成分割后, 形状确定的图像块不具有不可解释性.

此外, 本文所描述的解释, 是针对模型的每一次预测进行的 (instance-wise). 对使用者来说, 首先见到被分割后的模型输入图片, 然后见到每个图像块的对应权重, 整个解释过程对使用者来说完全可理解, 且与 DeepLab 的不可解释性无关.

综上, 可以使用 DeepLab 作为解释目标检测模型的工具.

4.2.3 替换原始 LIME 的图像分割方法

LIME 解释模型预测的过程如下.

(1) 对输入进行分割;

(2) 对输入进行扰动, 得到数据集;

(3) 训练局部线性回归模型, 得到解释.

在第 (1) 步中, 原始 LIME 使用 QuickShift 等传统图像分割方法对输入图片进行分割, 为了处理不同特性的图片, 还需要人为设定分割算法的参数. 本文对原始分割算法的替换非常直接: 在第 (1) 步中, 舍弃原有的分割方法, 使用 DeepLab V3 模型对输入图片进行分割. 这一替换不仅使改进后的 LIME 能用于解释目标检测模型, 而且使 LIME 的自适应性有所提升. 在解释一些特定的模型时, 比如小目标检测模型, 不需要对图像分割算法的参数做出调整.

5 实验设置

本节对实验过程中使用的目标检测模型和数据集做出设置.

5.1 实验所用模型设置

首先, 对实验中被解释模型做出设置. 当前性能较好的目标检测模型可分为两类.

(1) Two-Stage 模型

此类模型将边界框预测与类别预测分为两个阶段进行, 首先使用定位模型预测出物体边界框的位置, 再将边界框中的图片内容输入 CNN 分类器, 得到类别预测结果. Faster R-CNN[10]、R-CNN[28]等方法均属于此类.

(2) One-Stage 模型

此类模型将目标检测问题视为回归问题, 通过网络结构的设计, 将边界框预测与类别预测同时进行.YOLO 系列模型[12−14]、SSD[29]等方法均属于此类.

由于第 2 类模型性能普遍较强, 且是目标检测领域的主流研究趋势, 本文实验过程中将使用端到端模型中的经典范例: YOLOv1、YOLOv2、YOLOv3. 这 3 种目标检测模型将作为实验中被解释的目标, 其具体训练设置及性能测试结果如表 3 所示.

表 3 所用模型的训练设置

名称	训练集	测试集	mAP (%)
YOLOv1	VOC2007 (train)	VOC2007 (val)	63.4
YOLOv2	VOC2007 (train)	VOC2007 (val)	76.8
YOLOv3	VOC2007 (train)	VOC2007 (val)	82.3

第二, 对实验中用到的图像分割方法做出设置. 原始 LIME 使用的图像分割方法是 QuickShift[26]. 对于 DeepLab 模型, 我们使用其作者在 Github 上开源的 DeepLab V3, 具体设置如表 4 所示.

表 4 DeepLab V3 的设置

名称	训练集	测试集	mIOU (%)
DeepLab V3	VOC2012 (train)	VOC2012 (val)	87.8

5.2 实验所用数据集设置

VOC2007 数据集是目标检测领域的经典数据集之一, 物体类别数量为 20; 包含 5011 个训练图片、4952 个测试图片. 将该数据集用于目标检测模型的训练, 选择其作为本文实验数据集的原因如下.

(1) 图片资源非富, 模型训练后性能较强.

(2) 物体类别数量较少, 相比于 COCO 数据集的 80 类与 ImageNet 数据集的 21 841 类, 更适合解释目标检测模型的初期工作.

(3) YOLOv1 和 YOLOv2 模型在 COCO 和 ImageNet 数据集上的测试性能较差, 而本文的主要目的是对收敛性和性能较好的模型进行解释, 因此不在这两个数据集上进行实验.

考虑到解释器性能的限制, 并使实验结果更加直观, 对 VOC2007 数据集进行筛选, 仅使用图片中出现的物体类别数小于 3, 且每类物体个数为 1 的图片, 制作成精简数据集. 此数据集作为解释目标检测模型时的输入图片.

在使用未经修改的 LIME 解释目标检测模型时发现, 解释器忠实度低等问题在某一类图片中尤为明显, 因此将 VOC2007 数据集中的 100 张此类图片挑选出, 作为测试问题是否解决的小型数据集. 我们将此类图片定义为大物体图片, 此小型数据集定义为大物体数据集, 此类图片定义如下, 具体样例如图 8 所示.

图 8 大物体图片的具体样例

定义 4. 大物体图片. 至少有一个物体的边界框占据整张图片面积的 50% 以上的图片.

6 实验结果

6.1 决策依据分析结果直观对比

沿用上文的解释结果可视化设置, 并额外添加影响力权重阈值为 0.1, 对目标检测模型一次预测进行决策依据分析的结果如图 9 所示. 添加此阈值的效果是: 忽略线性回归模型中权重绝对值小于 0.1 的特征所对应的图像块, 使用此阈值的理由如下.

图 9 修改图像分割算法前后解释结果对比图

我们在解释模型时, 遵从 LIME 原文, 使用 K-LASSO[1]进行线性回归, 且设置 K 值为 10, 从而最多得到 10 个影响力权重最大的图像块. 对于特征数为 10 的多元线性回归问题, 若某一输入特征对应的权重值小于 0.1, 显然说明该特征在线性回归模型中影响力较低. 而线性回归模型代表目标检测模型在这一输入图片上的预测行为, 进而说明该图像块对目标检测模型的影响力较低, 不具有显著意义. 超过阈值的图像块, 称之为有效图像块.

图 9 中, 模型预测边界框内存在“car”类, 且置信度为 0.65, 图片下方为使用不同图像分割方法的决策依据分析结果. 左 1 与左 2 使用 LIME 原文中的 QuickShift[26]分割方法, 右 1 使用 DeepLab V3[27](一种语义分割模型) 分割, 具体实验设置与部分指标对比如表 5 所示.

表 5 修改图像分割算法前后解释结果对比表

名称	局部线性回归模型R^2值	最大影响力权重值
QuickShift (kernel=4)	0.39	0.05
QuickShift (kernel=10)	0.67	0.31
DeepLab V3	**0.96**	**0.51**

分析表 5 中的数据与可视化结果可知, 修改图像分割算法为 DeepLab V3 后, LIME 在解释目标检测模型时的效果显著提升, 不仅局部线性回归模型的忠实度极大提高, 能从输入图像中找出对目标检测模型决策影响力较大的图像块, 且解释结果十分直观. 对使用 QuickShift 分割方法时的几个直观问题简要分析如下: 左 1 中, 所有图像块的影响力权重均低于阈值, 因此没有图像被标出; 左 2 中, 可以明显看出部分车辆已与背景图片混合.

以上分析, 直观地说明, 我们所做的改进使 LIME 可以应用于解释目标检测模型.

6.2 DeepLab 对决策依据分析效果的提升

将图片扰动算法改变为 DeepLab 后, 在大物体数据集中进行实验, 得到的局部线性回归模型的 R^2 对比结果如表 6 所示. 从表中数据可以发现, 更换分割算法为 DeepLab 后, R^2 的平均值达到 0.95, 说明几乎在每一次解释中, 产生的线性回归模型都能在局部非常忠实地拟合目标检测模型的行为.

将图片扰动算法改变为 DeepLab 后, 在大物体数据集中进行实验, 得到的目标检测模型对扰动后图片的预测变化如表 7 所示. 从表中数据可以发现, 更换分割算法为 DeepLab 后, 扰动图片能对目标检测模型产生更大的影响. DeepLab 另一个重要的优点是: 物体的完整性得到了保留 (对比图 5 与图 7 可以得出).

表 6 不同分割方法下的局部线性回归模型 R^2 平均值

目标检测模型	分割方法	R^2平均值
YOLOv1	QuickShift (kernel=4)	0.46
	QuickShift (kernel=10)	0.71
	DeepLab	**0.95**
YOLOv2	QuickShift (kernel=4)	0.32
	QuickShift (kernel=10)	0.50
	DeepLab	**0.95**
YOLOv3	QuickShift (kernel=4)	0.27
	QuickShift (kernel=10)	0.52
	DeepLab	**0.94**

表 7 不同分割方法下的目标检测模型预测值变化比例

目标检测模型	分割方法	预测值平均变化比例	预测值最小变化比例	预测值变化比例小于0.2占比
YOLOv1	QuickShift (kernel=4)	0.37	0.08	0.25
	QuickShift (kernel=10)	0.46	0.17	0.02
	DeepLab	**0.61**	**0.14**	**0.01**
YOLOv2	QuickShift (kernel=4)	0.14	0.02	0.74
	QuickShift (kernel=10)	0.29	0.05	0.34
	DeepLab	**0.51**	**0.06**	**0.02**
YOLOv3	QuickShift (kernel=4)	0.08	0.02	0.85
	QuickShift (kernel=10)	0.27	0.04	0.39
	DeepLab	**0.57**	**0.02**	**0.01**

从以上实验可以发现, 局部模型忠诚度过低的问题已被解决, 保留物体完整性的问题也被解决. 如果能解决局部模型权重较小的问题, 则 DeepLab 就是一个合适的图像分割算法.

在 VOC2007 数据集上实验, 使用 DeepLab 时, 有效特征图像块的权重平均值为 0.68, 最大值可达 0.81. 所有实验结果说明, 替换图像分割方法为 DeepLab 后, 能找到决定性的特征图像块, 使 LIME 对目标检测模型进行解释.

对于其原因, 我们分析如下. 目标检测模型经过训练后, 对于物体存在性预测, 其决策依据就是整个物体, 关注的是物体的整体. 所以在将背景与每个物体都独立分割后, 目标检测模型的预测值改变得非常明显, 更容易在局部用线性回归去拟合. 换言之, 在新的图片分割算法下, 目标检测模型的决策行为变得更加线性.

7 结论与展望

本节中, 将对本文的工作进行总结, 并对本文存在的不足以及未来的工作方向进行说明.

7.1 结 论

在用 LIME 直接对目标检测模型进行解释时, 发现局部线性回归模型的忠诚度与权重值过小这两个问题. 我们通过实验分析, 揭示了目标检测模型在预测时关注物体整体这一性质, 并将问题的原因定位在图像分割方法不合理. 本文通过将图像分割方法替换为语义分割模型 DeepLab, 并对解释内容作出定义, 成功解决 LIME 存在的问题, 并将其应用于解释目标检测模型. 通过实验证明, 采用 DeepLab+LIME, 可以得到可信度较高且直观的决策依据分析结果.

另一方面, 基于 IoU、模型解释结果、基准决策依据, 本文提出了一种在有标签数据集中评价目标检测模型可信度的方法, 一定程度上填补了目标检测领域模型中, 对模型可信度评价的空白.

7.2 展 望

7.2.1 模型预测可信度评价方法的不足

分析第 3.2 节中的评价方法, 可以发现其不足: 即使是解释结果完美标出了物体, IoU 指标也无法达到 1, 因为

目标检测数据集中的基准决策依据不只包括物体, 也包括一部分背景. 因此, 该评价方法仅可作为一定程度上的参考. 如何对模型预测的可信度进行评价, 仍然是一个值得探索的方向.

7.2.2 缺少评价解释器的方法

从第 6 节中的实验数据可知, 本文对解释器 (LIME) 的一些指标分析, 是把线性回归模型作为解释核心的基础上进行的. 因此, 本文中的分析仅适用于 LIME 类的解释, 不具有较强的泛用性. 结合第 1.2 节中对解释器评价方法的介绍可以发现, 如何提出一个好的解释器评价方法, 是机器学习模型可解释性领域中另一值得研究的方向.

7.2.3 如何对边界框定位进行决策依据分析

本文中仅对目标检测模型的物体存在性进行了决策依据分析, 没有考虑边界框定位这一问题. 边界框存在性本质上和物体存在性相同, 因此剩下的是如何解释边界框定位的问题, 也就是边界框中 4 个用于定位的坐标值如何变化.

同样基于 LIME, 我们已经对大物体图片进行过简单的实验, 举例如下.

首先, 我们猜测边界框位置会随物体某一部分的缺失而产生变化. 实验时, 我们将物体在水平 (垂直) 方向上进行切片, 不断增加物体的缺失面积 (将缺失部分变为黑色), 观察对应的边界框位置变化情况. 然而, 一般情况下, 除非物体的缺失面积达到 65% 以上, 边界框的位置与其对应物体存在性概率都不会产生较大的变化. 我们认为其原因仍然是目标检测模型更加关注物体的整体, 甚至存在一定的预测惯性. 所谓预测惯性, 指的是目标检测模型会在一定程度上盲目地扩大边界框的面积, 从而导致对图片的部分变化不敏感.

对大物体图片进行垂直 (水平) 方向上的切片样例如图 10 所示.

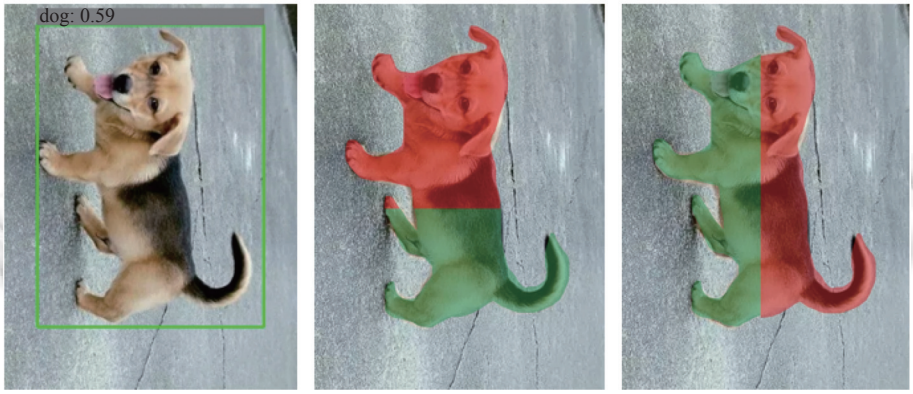

图 10 大物体图片上的简单实验

7.2.4 DeepLab 分割导致的细节信息缺失

由第 3.2 节可知, 在目标检测模型的训练数据集中, 基准决策依据不仅包括物体, 还包括物体外接矩形中的一部分背景, 因此, 本文中仅使用 DeepLab 对输入图片进行分割, 会使得一些细节的缺失. 对于这个问题, 我们考虑对背景进行再次分割, 即在使用 DeepLab 分割出背景后, 再对背景图像块使用 QuickShift 方法进行分割, 从而保留之前被忽略的细节.

此外, 如果只关注图片中的语义信息, 意义有限, 且可能会丢失在模型决策时起到作用的其他特征. 因此, 在未来的工作中需要结合其他的特征信息, 对模型决策进行解释.

7.2.5 一个有趣的模型预测结果

在对模型预测结果进行决策依据分析时, 我们在此过程中发现了一张有趣的模型预测图片, 如图 11 所示. 在此图片中, 目标检测模型在某一扰动图片中预测出了两个“bird”类物体, 且其中一个的存在性概率高达 0.57. 此样例也许这只是偶然, 但通过这一预测结果我们可以发现两个问题.

图 11 一个有趣的预测结果

第一, 通过本文实验, 虽然发现模型的决策依据与图片中的语义信息具有较高程度的一致性, 但在图 11 中所示的情况下, 如果我们把这张扰动图片作为输入, 由于 3 个带有形状的色块并不具有直观上的明显语义信息, 因此将无法得到人类可直接理解的解释. 无法解释某些特定的图片, 这是本文方法的缺陷.

第二, 部分类型样本的预测可能过度依赖图像的外轮廓, 或是其他的某些特征, 从而导致模型输出错误的结果.

References:

[1] Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. San Francisco: ACM, 2016. 1135–1144. [doi: 10.1145/2939672.2939778]

[2] Pedreschi D, Giannotti F, Guidotti R, Monreale A, Ruggieri S, Turini F. Meaningful explanations of black box AI decision systems. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence and the 31st Innovative Applications of Artificial Intelligence Conf. and 9th AAAI Symp. on Educational Advances in Artificial Intelligence. Honolulu: AAAI, 2019. 1213. [doi: 10.1609/aaai.v33i01.33019780]

[3] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 4768–4774. [doi: 10.5555/3295222.3295230]

[4] Ribeiro MT, Singh S, Guestrin C. Anchors: High-precision model-agnostic explanations. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence and the 13th Innovative Applications of Artificial Intelligence Conf. and 8th AAAI Symp. on Educational Advances in Artificial Intelligence. New Orleans: AAAI, 2018. 187. [doi: 10.5555/3504035.3504222]

[5] Lakkaraju H, Kamar E, Caruana R, Leskovec J. Faithful and customizable explanations of black box models. In: Proc. of the 2019 AAAI/ACM Conf. on AI, Ethics, and Society. Honolulu: ACM, 2019. 131–138. [doi: 10.1145/3306618.3314229]

[6] Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One, 2015, 10(7): e0130140. [doi: 10.1371/journal.pone.0130140]

[7] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: JMLR.org, 2017. 3145–3153. [doi: 10.5555/3305890.3306006]

[8] Amini A, Schwarting W, Soleimany A, Rus D. Deep evidential regression. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1251. [doi: 10.5555/3495724.3496975]

[9] Camburu OM. Explaining deep neural networks. arXiv: 2010.01496, 2020.

[10] Fan M, Wei WY, Xie XF, Liu Y, Guan XH, Liu T. Can we trust your explanations? Sanity checks for interpreters in android malware analysis. IEEE Trans. on Information Forensics and Security, 2020, 16: 838–853. [doi: 10.1109/TIFS.2020.3021924]

[11] Girshick R, Donahue J, Darrell T, Malik J. Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2016, 38(1): 142–158. [doi: 10.1109/TPAMI.2015.2437384]

[12] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 779–788. [doi: 10.1109/CVPR.2016.91]

[13] Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6517–6525. [doi: 10.1109/CVPR.2017.690]

[14] Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv: 1804.02767, 2018.

[15] Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: Optimal speed and accuracy of object detection. arXiv: 2004.10934, 2020.

[16] Ge Z, Liu ST, Wang F, Li ZM, Sun J. YOLOX: Exceeding YOLO series in 2021. arXiv: 2107.08430, 2021.

[17] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the

31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010. [doi: 10.5555/3295222.3295349]

[18] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 213–229. [doi: 10.1007/978-3-030-58452-8_13]

[19] Zhu XK, Lyu S, Wang X, Zhao Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision Workshops. Montreal: IEEE, 2021. 2778–2788. [doi: 10.1109/ICCVW54120.2021.00312]

[20] Beal J, Kim E, Tzeng E, Park DH, Zhai A, Kislyuk D. Toward transformer-based object detection. arXiv: 2012.09958, 2020.

[21] Zhu XZ, Su WJ, Lu LW, Li B, Wang XG, Dai JF. Deformable DETR: Deformable transformers for end-to-end object detection. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2020.

[22] Kim S, Park S, Na B, Yoon S. Spiking-YOLO: Spiking neural network for energy-efficient object detection. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 11270–11277. [doi: 10.1609/aaai.v34i07.6787]

[23] Rashwan A, Kalra A, Poupart P. Matrix Nets: A new deep architecture for object detection. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision Workshop. Seoul: IEEE, 2019. 2025–2028. [doi: 10.1109/ICCVW.2019.00252]

[24] Zou ZX, Shi ZW, Guo YH, Ye JP. Object detection in 20 years: A survey. arXiv: 1905.05055, 2019.

[25] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv: 1702.08608, 2017.

[26] Vedaldi A, Soatto S. Quick shift and kernel methods for mode seeking. In: Proc. of the 10th European Conf. on Computer Vision. Marseille: Springer, 2008. 705–718. [doi: 10.1007/978-3-540-88693-8_52]

[27] Chen LC, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. arXiv: 1706.05587, 2017.

[28] Ren SQ, He KM, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]

[29] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC. SSD: Single shot multibox detector. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 21–37. [doi: 10.1007/978-3-319-46448-0_2]

平昱恺(1998－), 男, 硕士生, 主要研究领域为软件工程, 深度学习.

江贺(1980－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为软件可靠性, 软件测试, 编译系统.

黄鸿云(1977－), 女, 硕士, 讲师, 主要研究领域为智能软件测试, 多媒体数据分析.

丁佐华(1964－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为智能系统软件建模、分析与测试.