基于弱语义注意力的遥感图像可解释目标检测

周 勇[1,2],陈思霖[1,2],赵佳琦[1,2,3],张 迪[1,2],王瀚正[1,2]
(1. 中国矿业大学计算机科学与技术学院,江苏徐州 221116; 2. 矿山数字化教育部工程研究中心,江苏徐州 221116;
3. 灾害智能防控与应急救援创新研究中心,江苏徐州 221116)

摘 要: 近些年来随着遥感技术的快速发展,遥感图像目标检测成为了当前的研究热点. 针对遥感图像背景复杂以及现有目标检测模型缺乏可解释性等问题,本文提出了一种基于弱语义注意力的遥感图像可解释目标检测方法. 具体地,首先通过多层级特征金字塔来解决遥感图像中目标尺度变化范围大的问题. 其次,利用检测框的角度回归来解决遥感图像目标定向的问题. 然后,基于弱语义分割网络产生强化目标特征的注意力权重值,抑制背景噪声. 最终用网络剖析的分析方法,获取模型中卷积核对应的可解释性语义概念. 实验结果表明,本文提出的算法在遥感图像目标检测的准确性以及对背景噪声抑制上有较好的表现,并且通过可解释性算法在一定程度上使本文提出的模型易于理解.

关键词: 目标检测; 遥感图像; 注意力网络; 弱语义; 深度学习可解释性
中图分类号: TP753 **文献标识码:** A **文章编号:** 0372-2112 (2021)04-0679-11
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20200554

Weakly Semantic Based Attention Network for Interpretable Object Detection in Remote Sensing Imagery

ZHOU Yong[1,2], CHEN Si-lin[1,2], ZHAO Jia-qi[1,2,3], ZHANG Di[1,2], WANG Han-zheng[1,2]
(1. *School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu* 221116, *China*;
2. *Ministry of Education Engineering Research Center of Mine Digitization, Xuzhou, Jiangsu* 221116, *China*;
3. *Innovation Research Center of Disaster Intelligent Prevention and Emergency Rescue, Xuzhou, Jiangsu* 221116, *China*)

Abstract: In recent years, the object detection in remote sensing imagery has been a hot research spot with the development of remote sensing technique. To deal with the complex background in the imagery and the detection model's interpretability, we propose a weakly semantic based attention network for interpretable object detection model in remote sensing imagery. Firstly, a feature pyramid network is devised for the variation of object scales. Next, an angle is added to the regression to better locate the object. Thirdly, we add a weakly semantic segmentation network to enhance feature and filter the noisy information in the background. Finally, the model is dissected by the proposed method to get the interpretable semantic concepts of convolutional kernel. Experiments validated that the model has a good performance in the aspect of suppressing the background noise and make our model easy to comprehend.

Key words: object detection; remote sensing imagery; attention network; weakly semantic; deep learning interpretability

1 引言

近几年来,遥感技术的大力发展,使遥感数据的获取不再困难. 在数据充分的情况下,由于自然图像的目标检测方法存在预测尺度单一、水平框贴合目标效果差以及缺少对目标特征的强化等问题,因此无法较好

收稿日期:2020-06-10;修回日期:2020-12-11;责任编辑:李勇锋
基金项目:国家自然科学基金(No. 61806206);江苏省自然科学基金(No. BK20180639, No. BK20201346);江苏省六大高峰人才项目(No. 2015-DZXX-010)

地将这些方法应用于遥感图像目标检测. 遥感图像目标检测领域中存在的问题总结为如下四点.

(1)尺度变化问题:遥感图像具有较大的场景信息. 图像达到上百万个像素分辨率,因此目标相对图像尺度较小,导致无法获取目标精细的特征. 此外,遥感图像目标尺度变化范围广,不利于单一尺度的多类别目标检测.

(2)目标定向问题:遥感图像中的目标为定向密集排列. 目标排列的方向是无规律的,因此检测模型需要具有旋转不变性和更优质的检测框.

(3)背景复杂混乱:遥感图像背景复杂多样,包含大量背景冗余信息,比如山川、河流等. 这导致背景与目标之间的边界模糊,不利于模型对目标特征的提取.

(4)缺乏可解释性:深度学习是"黑盒"模型[1],缺乏对模型预测行为的解释信息,而遥感技术关乎国家安全问题,需要在一定程度上对模型进行可解释分析,增强模型预测结果的置信度.

本文受自然图像目标检测网络 RetinaNet[2] 启发,提出一个更加有效的遥感图像目标检测模型. 该方法充分考虑遥感图像的特点以及深度学习模型缺乏可解释性的问题,在检测任务中引入角度回归,设计抑制背景噪声的弱语义注意力模块,并且对主干网络的卷积层进行可解释性分析.

其中弱语义注意力模块,不同于计算机视觉中的语义分割任务. 由于遥感图像没有语义分割中逐像素的标注,缺乏明确目标与背景的边界,因此定义弱语义分割是利用目标检测中检测框的标注信息生成以检测框为边界的弱语义掩膜(Weakly Semantic Mask),以此作为分割的监督信号,产生弱语义的目标背景分割图,进而获取强化目标特征、过滤背景噪声的注意力权重值. 在 DOTA[3] 数据集上对模型进行评估,与基准方法相比提升了 2.4% mAP 值. 本文具体贡献总结如下:

(1)提出了一个弱语义注意力模块来提高遥感图像中目标与背景的差异性,提升模型在复杂背景中的检测效果.

(2)设计了一个新的遥感图像目标检测框架 WS-R^{2}etinaNet(Weakly Semantic Rotational RetinaNet),有效地解决了目标定向等问题,得到更加精确的检测框.

(3)对模型进行可解释性分析. 通过改进网络剖析[4]方法,解剖模型主干网络的隐藏层结构,得到与卷积核关联的可解释性语义量化值,对该可解释性量化值加以分析.

2 相关工作

2.1 自然图像目标检测算法

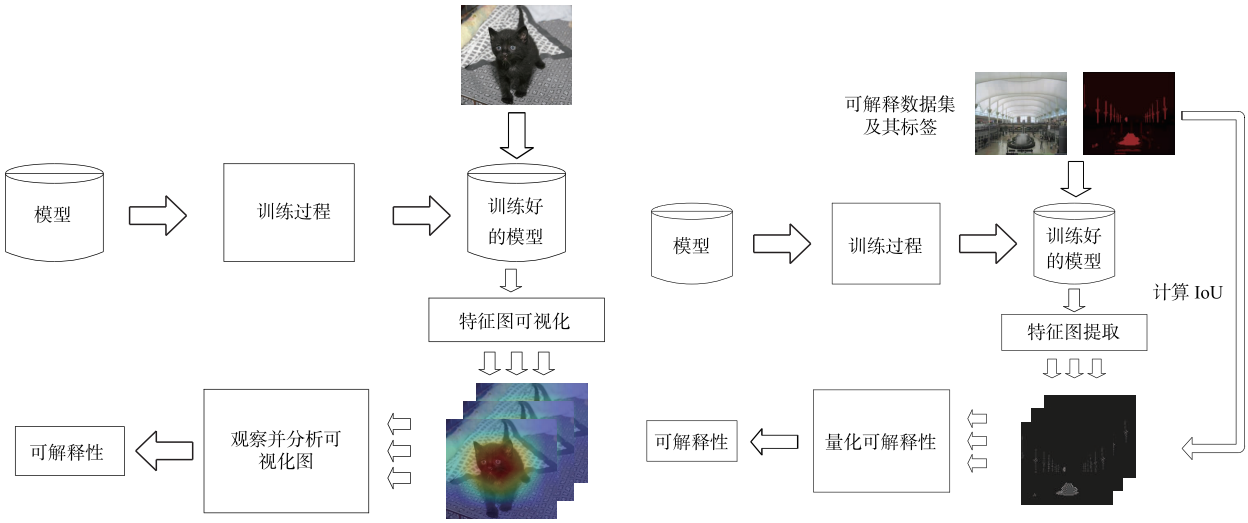
目标检测是计算机视觉领域的一项基本任务,主要是定位图像中特定物体出现的区域并判定目标类别[5]. 基于卷积神经网络的目标检测方法主要分为单阶段目标检测(One-Stage)和双阶段目标检测(Two-Stage). 双阶段目标检测方法以 R-CNN[6](Region-Based Convolutional Neural Networks)为开端,首先生成包含目标的区域提案(Region Proposal),而后对这些提案框坐标进行微调和分类,预测出目标位置和类别. 后续工作为了提高检测的精确度和速度,发展出多个基于 R-CNN 的框架,Fast R-CNN[7] 利用共享卷积层来简化计算,并且提出了感兴趣区域池化层(RoI Pooling)解决区域提案尺度不一致问题,Faster R-CNN[8] 摒弃选择性搜索算法(Selected Search,SS),提出了基于锚点框(Anchor Box)的区域生成网络(Region Proposal Network,RPN),提高了运行的速度. 为了进一步提高检测精准性,Mask R-CNN[9]、Cascade R-CNN[10] 等检测方法在 Faster R-CNN 上做出相应改进. 实验表明双阶段方法更加准确,但是运行速度较慢.

单阶段目标检测方法是把检测和分类任务统一看作为回归问题. 经典的单阶段目标检测方法是 SSD(Single Shot Multi-box Detector)[11] 和 YOLO(You can Only Look Twice)[12~14] 系列. 这些方法利用锚点框直接回归出检测框的位置以及相应类别,虽然精度不及双阶段检测,但速度得到明显提升. 为解决单阶段方法在预测准确性上存在的问题,Lin T[2] 提出了 RetinaNet,不仅速度上达到 5fps,并且准确性高于双阶段方法,因此本文的网络基于 RetinaNet 进行改进.

2.2 深度学习可解释性方法

目前深度学习可解释性的研究分为多个分支,其中可视化的方法是重要的研究方向之一. Samek W[15] 提出敏感度分析,量化模型对输入变量的敏感程度,并可视化敏感程度高的区域,说明该区域主要影响模型决策. 另一些可视化方法采样卷积核激活值最大的图像块[16,17],而后可视化这些激活图像块,分析网络如何获取信息. Goyal[18] 使用两种可视化技术:遮挡和引导的反向传播,在图像中找到相对重要的区域.

如图 1 所示,上述的可解释性可视化算法对网络特征图或激活图进行可视化,缺乏对这些可视化特征的进一步分析,且这些方法利用人类视觉观察分析获取网络模型的可解释性,容易出现人类主观上的判断失误. 本文在网络剖析算法上进行改进,对可视化后的特征图进行进一步计算,量化特征分布并利用相关数据集将特征关联到可解释性语义上,相较于其他可解释性技术,本文算法不需要人类的主观分析,算法通过自适应选择阈值的方式,分析图像生成特征,获取模型内部的语义,使模型更加易于理解.



(*a*)传统可视化可解释性方法　　(*b*)本文可解释性方法

图1　传统可视化可解释性方法和本文可解释性方法对比

3　基于弱语义注意力的遥感图像可解释目标检测

3.1　概述

本文基于传统的单阶段目标检测方法 RetinaNet,提出一种弱语义注意力的遥感图像可解释目标检测算法,解决遥感图像目标检测中尺度变化大、背景复杂、目标定向以及缺乏可解释性的问题. 首先,利用特征金字塔网络可以获取丰富语义特征的优势,降低遥感图像中目标尺度变化大对检测性能的影响. 其次,由于弱语义注意力网络可以强化目标特征、弱化背景噪声,因此采用弱语义注意力网络缓解遥感图像中复杂背景噪声对目标识别与检测的影响. 然后,在遥感图像目标检测中引入角度值回归使检测框与目标完全匹配,以此解决图像中目标定向密集排列问题. 最终,网络剖析算法可以通过计算主干网络隐藏层表征信息的可解释性量化值,以此理解模型主干网络的工作原理,使算法更加可靠. 本文提出的网络 WS-R^{2}etinaNet 具体结构如图 2 所示.

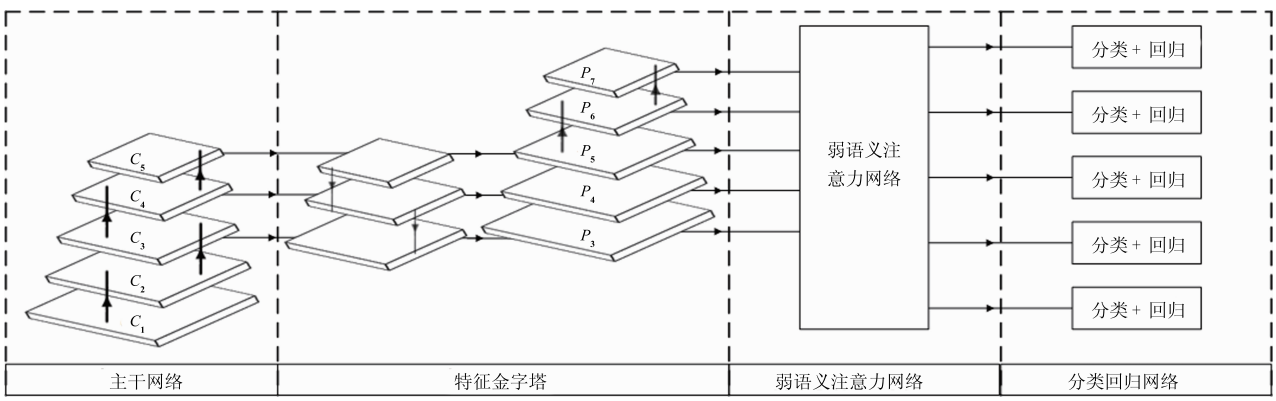


图2　WS-R^{2}etinaNet的整体结构

3.2　特征金字塔网络

在深度学习卷积神经网络中,特征图的语义信息随着卷积层数的加深而愈加丰富,在底层提取的特征较原始,层次越高,提取的特征越抽象,在高层已经是一种语义组合[19]. 因此,大多数网络在顶层特征图上进行检测. 然而,这种方法忽略了目标尺度问题,例如尺度小的目标(仅仅只有几个像素)在特征提取阶段会出现特征消失的现象,导致检测器无法获取该目标的特征信息. 并且在遥感图像数据集中,存在目标尺度不一致的缺陷(比如车辆和港口相差数十倍的尺度). 因此,本文采用特征金字塔网络模型(Feature Pyramid Network,FPN)[20]进行多尺度检测. 如图 2 所示,该模型利用侧连接方式建立一个自顶向下的多尺度特征金字塔结构,每层特征包含相邻层级上目标的特征信息,提高对不同尺度目标预测的准确性. 本文的方法与传统 FPN 不同,只利用主干网络的三层特征,通过额外的卷

积层获取步长更大、语义更丰富的特征图,以达到在遥感图像上更加准确的预测效果.

输入的图像首先经过一个自下而上的特征提取阶段. 该阶段本文使用 ResNet[21] 作为主干网络,五个卷积阶段保留五个不同深度的特征图$\{C_1,C_2,C_3,C_4,C_5\}$. 然后根据这些特征图构建特征金字塔$\{P_3,P_4,P_5,P_6,P_7\}$. 具体地,首先将 C_5 特征图通过一个 $1\times1\times256$ 的卷积操作得到特征图 P_5,然后将 C_4 特征图通过一个$1\times1\times256$ 的卷积操作降维,使其深度与 P_5 深度一致. 利用双线性插值法对 P_5 特征进行上采样,输出尺寸与 C_4 保持一致. 最终 P_5 特征图与 C_4 特征图进行融合,特征融合的计算公式如式(1)所示:

$$P_4=\frac{1}{2}\sum_{i=1}^{H\times W}(C_4^i+P_5^i) \tag{1}$$

其中 H 与 W 表示 C_4 特征图的长、宽,i 表示特征图上第 i 个像素点的像素值. P_3 的获取方法类似 P_4. 这些输出的特征图通过 $3\times3\times256$ 卷积操作得到最终的前三层金字塔特征$\{P_3,P_4,P_5\}$. 为了获取语义更加丰富的特征信息,P_5 通过两次 $3\times3\times256$ 的卷积操作得到 P_6、P_7,最终自顶而下得构建出五层金字塔特征图$\{P_3,P_4,P_5,P_6,P_7\}$. 该组特征包含各种尺度目标,且保留低层特征的位置细节信息以及高层特征的语义信息,极大的提高了检测各个尺度目标的准确性.

3.3 弱语义注意力网络

由于鸟瞰视角的特殊性,遥感图像中存在大量真实世界中复杂的场景(比如山川、河流等),这些噪声信息严重影响网络对目标的识别与检测.

在语义分割任务中,网络通过图像的语义标签来学习目标与背景的特征,从而达到像素级分割预测[22]. 然而,在目标检测数据集中缺乏这种逐像素的语义标签,因此,本文利用目标检测数据集的真实标注框来生成弱语义掩膜,并设计了一个弱语义注意力模块来过滤背景噪声,强化目标特征. 具体地,获取图像中像素点的位置信息,并与标注框对比. 若像素点位于框内或框的边缘,则设置该像素点值为 1;反之则设置为 0. 最终生成的弱语义掩膜有助于对目标与背景特征的分割.

将 3.2 节中多层级特征金字塔$\{P_3,P_4,P_5,P_6,P_7\}$均输入弱语义注意力模块,从而获取各层级特征图的注意力权重. 如图 3 所示,假设当前操作的特征图为 $P_x(x=3,4,5,6,7)$,首先将 P_x 输入一个全卷积语义分割网络[23](Fully Convolutional Networks for Semantic Segmentation, FCN)得到分割结果 A_x. 然后对 A_x 进行上采样操作,使其输出尺寸与 P_x 保持一致. 最终上采样后的 A_x 通过 *Softmax* 函数,得到对应特征图 P_x 的注意力权重值.

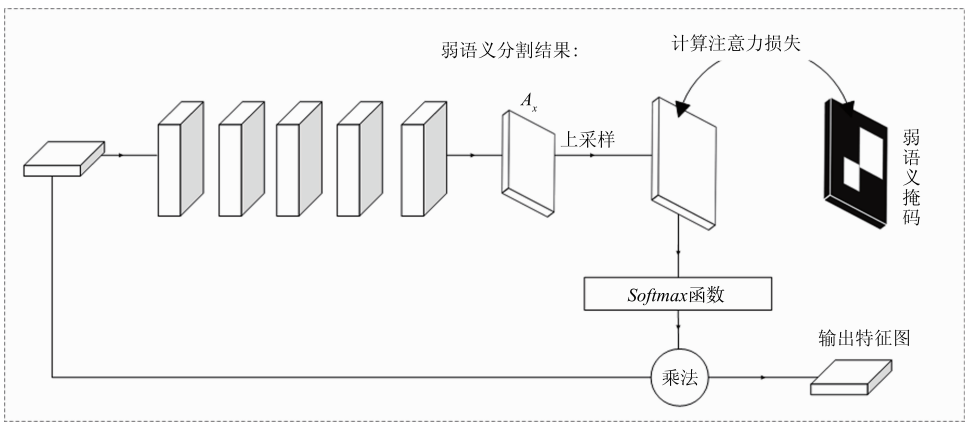


图3 弱语义注意力网络结构

注意力网络在训练过程中,利用弱语义掩膜学习图像中目标与背景的差异信息,得到的分割结果表示目标和背景的置信度,该置信度即为强化目标的注意力权重值. 将注意力权重值与输入特征进行逐元素乘法计算,强化目标特征的同时弱化了背景信息,因此达到了过滤背景噪声的作用. 值得注意的是,注意力网络的损失采用交叉熵损失的形式,计算损失时,需要重置分割结果的尺寸,使其与弱语义掩膜尺寸保持一致.

3.4 分类回归网络

分类回归网络在 RPN 网络的基础上加入角度值回归. 首先特征图上的每一个像素点根据预先设置好的锚点框数量 k(实验设定 $k=21$)生成不同尺寸、纵横比的锚点框,以满足多尺度目标检测的条件. 然后将锚点框特征输入网络,最终通过网络中的两个分支,分别完成分类任务以及检测框位置的回归.

回归网络的输出表示锚点框与其真实标签的位置偏移量. 回归目标的计算方法如式(2)和式(3)所示:

$$t_x=(x-x_a)/w_a,\ ty=(y-y_a)/h_a \tag{2}$$
$$t_w=\log(w/w_a),\ t_h=\log(h/h_a),\ t_\theta=\theta-\theta_a$$
$$t'_x=(x'-x_a)/w_a,\ t'_y=(y'-y_a)/h_a \tag{3}$$
$$t'_w=\log(w'/w_a),\ t'_h=\log(h'/h_a),\ t'_\theta=\theta'-\theta_a$$

其中(x,y,w,h,θ)表示矩形中心点的坐标和该矩形的

长度、宽度、角度值. x 表示的是真实框,x_a 表示的是锚点框,x' 表示的是预测框,其他代数形式及角标与 x 类似. 定义 θ 表示框长边与水平轴的夹角,固定角度范围为[-90°,0]. 由于在低层特征图中需要保留更多尺寸较大的目标语义信息,而高层中则需要保留更多小目标的语义信息. 因此,在各层构建锚点框时选择不同的锚点框尺寸.

不同于传统回归框,本文采用倾斜矩形交并比(Intersection over Union,IoU)算法[24]. 倾斜框 IoU 的计算需要考虑相交部分是一个任意多边形的问题. 因此,利用上述算法修改非极大值抑制[24],更好地过滤检测结果中的冗余框.

3.5　损失函数

如3.3 节中所述,弱语义注意力网络主要利用弱语义分割的方法生成注意力权重,并且采用了监督学习的方式,利用弱语义掩膜引导注意力权重的学习. 根据监督信号优化弱语义注意力网络的损失函数为交叉熵损失,具体形式如式(4)所示:

$$L_{\text{att}}(u_{ij},u'_{ij}) = -\frac{1}{H\times W}\sum_{i}^{H}\sum_{j}^{W}u_{ij}\log u'_{ij} \tag{4}$$

其中,H 与 W 表示弱语义掩膜的长度与宽度. u_{ij} 和 u'_{ij} 分别表示注意力网络输出(i,j)点的权重值和在弱语义掩膜上(i,j)点的像素值.

如3.4 节所述,回归分类网络包含两个分支,因此需要分别计算分类网络的损失以及回归网络的损失. 其中分类损失采用 Focal loss[2],如式(5)所示.

$$p_t=\begin{cases}p_n, & \text{if } t_n=1\\ 1-p_n, & \text{otherwise}\end{cases}$$

$$L_{\text{cls}}(p_n,t_n) = -\frac{1}{N}\sum_{n=1}^{N}\alpha\,(1-p_t)^{\gamma}\log(p_t) \tag{5}$$

其中,N 表示预测框的总数,p_n 表示多个类别的概率分布,t_n 表示目标的类别标签. Focal loss 中 α 与 γ 为超参数,分别设置为 0.25 与 2.

除了分类损失,在分类回归网络中还利用 smooth $L1$ 损失作为回归任务的损失函数,如式(6)所示:

$$L_{\text{reg}}(v'_{nj},v_{nj}) = \begin{cases}\dfrac{1}{N}\displaystyle\sum_{n=1}^{N}t'_n\sum_{j\in\{x,y,w,h,\theta\}}0.5\,(v'_{nj}-v_{nj})^2, & \text{if } |v'_{nj}-v_{nj}|<1\\ \dfrac{1}{N}\displaystyle\sum_{n=1}^{N}t'_n\sum_{j\in\{x,y,w,h,\theta\}}|v'_{nj}-v_{nj}|-0.5, & \text{otherwise}\end{cases} \tag{6}$$

其中 N 表示预测框的总数,t'_n 表示置信度(若值为 1 表示前景,值为 0 表示背景),v'_{*j} 表示预测的坐标向量而 v_{*j} 表示真实标签坐标向量.

因此,本文中模型训练过程多任务损失函数形式如式(7)所示:

$$L_{\text{total}} = \lambda_1 L_{\text{reg}} + \lambda_2 L_{\text{att}} + \lambda_3 L_{\text{cls}} \tag{7}$$

其中 λ_1、λ_2、λ_3 是多任务损失的平衡参数,L_{reg} 为上述回归损失,L_{att} 为注意力损失,L_{cls} 是分类损失.

3.6　网络解剖

神经网络以其黑盒表示的低可解释性为代价取得了优越的性能[25]. 但在医疗、驾驶、遥感等关乎人类或社会安全的领域,深度学习模型既需要卓越的效果,也需要对决策给出一定依据. 近些年来,一些可解释性可视化算法对网络特征图或激活图进行可视化,再对模型决策进行可解释性分析. 这些方法利用人类视觉的主观分析易出现判断失误. 本文对 Bai D[4] 提出的网络剖析的方法做出改进. 网络剖析算法的基本原理是根据一组预定义的人类可解释性语义及一个包含这些可解释性语义标注的数据集,利用图像在卷积网络中前向传播留下的特征图,探究特征图激活值的分布,进而通过计算该分布与数据集中可解释性语义标注的相似程度获取到网络中卷积核的可解释性语义信息. 算法具体描述如下:

首先将文献[4]中定义的人类可解释语义概念{场景、目标、部件、材料、纹理、颜色}按照符合人类理解的方式进行划分,将{场景、目标、部件}视为高级语义的概念而{材料、纹理、颜色}视为低级语义概念. 其次,利用网络剖析的算法来计算语义概念的评分值. 最终,量化可解释性并用其对卷积核进行编码. 计算过程如图 4 所示. 以主干网络 ResNet50 中第五阶段第二层卷积(*C5_conv2*)为例,假设输入图像为 $I(x)$,经过一次前向传播,将 *conv2*(第二层卷积)中 512 个卷积核输出的特征图 $F(x)$ 保存下来,用作后续的可解释性计算. $F(x)$ 中包含 512 个特征图,每个特征图对应一个卷积核的语义分布. 对于 $F_k(x)$(k 为卷积核索引),利用分割阈值 g 来过滤弱语义信息,保留强语义信息作为该卷积核的语义特征. 本文改进传统网络剖析中利用概率分布计算阈值的方法,如式(8)所示:

$$g = \frac{1}{H\times W}\sum_{i=1}^{H\times W}p_i \tag{8}$$

其中 H 与 W 表示 $F_k(x)$ 的高度和宽度值,p_i 表示第 i 个像素点的值. 通过计算激活值的平均值作为阈值 g,是因为高于平均值更能表现其卷积核关注的语义. 过滤后的强语义特征图为 $T_k(x)$,过滤方式如式(9)所示:

$$T^{\,i}_{\,k}(x)=\begin{cases}T^{\,i}_{\,k}(x), & \text{if } T^{\,i}_{\,k}(x)\geqslant g\\ 0, & \text{otherwise}\end{cases} \tag{9}$$

$T^{\,i}_{\,k}(x)$ 表示特征图 $T_k(x)$ 上第 i 个点的像素值. 对比每一个卷积核的 $T_k(x)$ 与标注好的语义掩膜,首先对 $T_k(x)$ 进行一次二值化预处理,将保留下来的特征激活值与过滤掉的弱语义信息差异化,得到一张二值语义图 $M_k(x)$. 然后对其进行上采样操作,便于该语义图与语义掩膜进行计算. 利用文献[4]中提出的 IoU 计算方

式,对于不同语义 C 的掩膜 $L_c(x)$,得到的 IoU 值就是卷积核 k 与语义 c 的可解释性评分. 最终得到该层所有卷积核对于不同语义概念的所有可解释性评分. 本文利用整体的平均水平来作为评分阈值 f,具体计算方法如式(10)所示:

$$f=\frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C}\frac{|M_k(x)\cap L_c(x)|}{|M_k(x)\cup L_c(x)|} \tag{10}$$

其中 K 表示该层中卷积核总数. 通过阈值 f 可以获取每一个卷积核评分大于该阈值的语义概念.

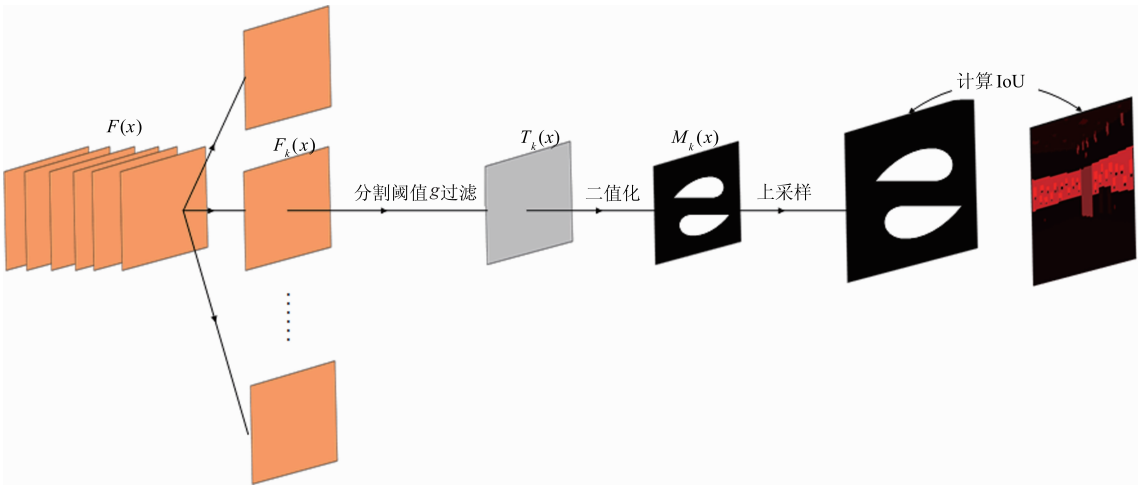


图4 网络剖析计算流程

统计量化可解释性的结果,单一卷积核可能关注多个语义概念,因此选用得分最高的进行统计. 而在整个卷积层中所有卷积核得分最高的语义概念按照高级、低级划分,统计两个级别的卷积核个数. 如果高级语义卷积核的个数和低级的个数相差较少(一般认为小于50)那么可以认为该卷积层既包含了低级语义信息也包含了高级语义信息;如果高级语义卷积核个数远超于低级,则认为该卷积层主要关注于高级语义概念;反之亦然.

4 实验

本文的实验环境为一张 32GB 内存的 Tesla P100 GPU,编程语言为 TensorFlow. 实验基于 DOTA 数据集以及一个包含可解释性标注的 Broden 数据集进行.

4.1 实验数据集

目标检测实验采用 DOTA 数据集,DOTA 是目前遥感图像目标检测的基准数据集,包含 2806 张来自不同传感器和平台的遥感图像. 图像的尺寸从 800×800 到 4000×4000 不等,其中包含了 188282 张不同尺度、不同角度方向的目标个体. 这些个体被标注为 15 个类别:足球场、直升机、泳池、环岛、大型汽车、小型汽车、桥梁、港口、田径场、篮球场、网球场、棒球场、储油罐、轮船、飞机. 对训练集 DOTA_train 中的 1412 张图像进行剪裁,得到 20691 张 600×600 的图像块. 验证集为 DOTA_eval,包含 459 张任意尺度的遥感图像.

可解释性实验采用数据集为 Broden,该数据集从 ADE[26]、OpenSurfaces[27]、Pascal-Context[28]、Pascal-Part[29] 和 Describable Textures Dataset[30] 采样,Broden 的具体情况如表 1 所示.

表 1 Broden 数据集

可解释类别	目标类别	原始数据集	平均采样
场景	468	ADE	38
目标	584	ADE,Pascal	8.35
部件	234	ADE,Pascal	854
材质	32	OpenSurface	1703
纹理	47	DTD	140
颜色	11	自动生成	59250

本文可解释性实验使用的数据随机采集自上述 Broden 数据集,其中采集数据的来源情况如表 2 所示,采集数据的可解释性类别分布情况如表 3 所示.

表 2 本文采集数据的情况

原始数据集	采集图像数量
ADE	513
Pascal	240
OpenSurface	630
DTD	116

表 3 本文采集数据的标注可解释类别分布

可解释类别	采集图像数量
场景	460
目标	754
部件	489
材质	460
纹理	631
颜色	1500

4.2　评估标准

在遥感图像目标检测领域常用的评估标准为是平均准确率(Average Precision,AP),与平均准确率均值(Mean Average Precision,mAP). 首先分别计算模型在DOTA 上 15 个类别的 AP 值,再根据 AP 计算模型的 mAP 数值.

可解释分析实验分为不同模型主干网络可解释性剖析结果的对比实验以及同一模型不同层级的可解释性剖析结果对比实验,实验采用不同语义划分的卷积核个数作为评估标准.

4.3　实验配置

实验训练的迭代次数为 540k 次,基础学习率设置为 10^{-4},每 27k 次迭代保存一次模型,共保存 20 次模型,在迭代 324k 次后学习率下降为 10^{-5},432k 次后下降为 10^{-6}. 由于遥感图像分辨率过大,因此训练批处理大小(Batch Size)设置为 1. 为了更好的对密集目标采样,实验中金字塔锚点框采样的初始尺寸设置为[32,64,128,256,512],锚点框尺度和缩放比设置为[20,21/3,22/3]和[1,1/2,2,1/3,3,5,1/5],每个特征点产生 21 个锚点框. 设置 IoU 阈值大于 0.5 为正样本,小于 0.4 为负样本. 实验设置角度回归的范围为[-90°,0],NMS 阈值为 0.1,损失的平衡权重 λ_1、λ_2、λ_3分别为 1,0.2,0.2.

在可解释性实验中,对本文检测网络进行一次前向传播即可得出结果. 实验中计算出的分割阈值为 0.04,评分阈值为 0.05. 实验设置批处理大小设置为 16 提高我们的分析效率.

4.4　遥感图像目标检测实验

4.4.1　对比实验

比较本文遥感图像目标检测模型与其他 5 种经典目标检测算法,其中包含 Faster R-CNN[8],R-FCN[31],YOLOv2[13],SSD[11] 以及 RetinaNet[2]. Faster R-CNN 是 DOTA 原数据集的基准模型 Faster R-CNN Oriented[3](FR-O). YOLO 中采用的主干网络为 DarkNet19,其余对比算法的主干网络与本文一致皆采用 ResNet50. 将 DOTA 训练集中图像剪裁为 600×600 作为对比实验采用的训练数据. 实验对比模型 540k 迭代后的 mAP 值以及 DOTA 中 15 个类别的 AP 值. 最终实验结果如表 4 所示,表中的英文缩写是类别的英文简写,“BD”为棒球场(Baseball Diamond),“GTF”为田径场(Ground Track Field),“SV”为小型车辆(Small Vehicle),“LV”为大型车辆(Large Vehicle),“TC”为网球场(Tennis Court),“BC”为篮球场(Basketball Court),“ST”为油罐(Storage Tank),“SBF”为足球场(Soccer Ball Field),“RA”为环岛(Roundabout),“SP”为游泳池(Swimming Pool),“HC”为直升机(Helicopter).

表 4　对比 DOTA 数据集任务——定向目标检测 OBB 实验

模型方法	mAP(%)	Plane	BD	Bridge	GTF	SV	LV	Ship	TC	BC	ST	SBF	RA	Harbor	SP	HC
SSD[10]	10.59	39.83	9.09	0.64	13.18	0.26	0.39	1.11	16.24	27.57	9.23	27.16	9.09	3.03	1.05	1.01
YOLOv2[13]	21.39	39.57	20.29	36.58	23.42	8.85	2.09	4.82	44.34	38.25	34.65	16.05	37.62	47.23	25.50	7.45
R-FCN[31]	26.79	37.80	38.21	3.64	37.26	6.74	2.60	5.59	22.85	46.94	66.04	33.37	47.15	10.60	26.19	17.96
FR-O[3]	52.93	79.09	**69.12**	17.17	63.49	34.20	37.16	36.20	89,19	69.60	58.96	49.40	52.52	46.69	44.80	46.30
FR-H[3]	36.29	47.16	61.00	9.80	51.74	14.87	12.80	6.88	56.26	59.97	57.32	47.83	48.70	8.23	37.25	23.05
RetinaNet-O	56.43	**89.55**	65.13	32.72	56.29	39.13	46.86	**54.81**	90.51	**62.64**	81.85	62.31	58.76	50.74	42.26	12.87
IENet[32]	57.14	80.20	64.54	39.82	32.07	**49.71**	**65.01**	52.58	81.45	44.66	78.51	46.54	56.73	**64.40**	**64.24**	**36.70**
Ours	**57.74**	89.52	65.66	32.60	**57.76**	41.10	47.29	53.75	**90.55**	58.02	**83.18**	**68.06**	**62.47**	52.52	45.40	18.25

实验结果表明:一些在自然图像上表现较好的网络并不适用于遥感图像检测,比如 SSD、YOLO、R-FCN 等. 并且双阶段网络表现要优于单阶段网络,这是因为双阶段网络在训练中相对于单阶段网络更加复杂且不存在类别不平衡等问题. 分析各个类别的 AP 值我们发现传统网络对于目标尺度较小的类别,比如小型车辆(Small Vehicle,SV),检测的效果相较更差,这是因为其缺乏特征金字塔来强化小目标的特征,以及缺少多尺度检测. RetinaNet-O(RetinaNet Oriented)是我们在 RetinaNet 中加入角度回归的结果,由于其网络损失函数中设计了解决类别不平衡问题的方法 Focal Loss,使得它的效果优于所有传统网络. 本文的方法利用弱语义注意力网络分支过滤掉了背景噪声,加强了目标的特征,因此各个类别的 AP 值都有一定的提升并且在检测效果中是最优的.

虽然模型在基准网络上得到了更优的效果,但相较于目前一些基于 DOTA 数据集的方法存在精度上的差距. 这是因为模型采用的主干网络为 ResNet50,并且本文没有着重于利用训练上的技巧提高效果,而是通过弱语义注意力网络来提升基准网络的性能. 实验发现,在 RetinaNet 上加入注意力网络后可以显著地提高模型的效果,印证文章提出的方法是有效的. 图 5 中展示了模型检测的部分结果.

检测结果可以看到对于密集排列的车辆,模型可

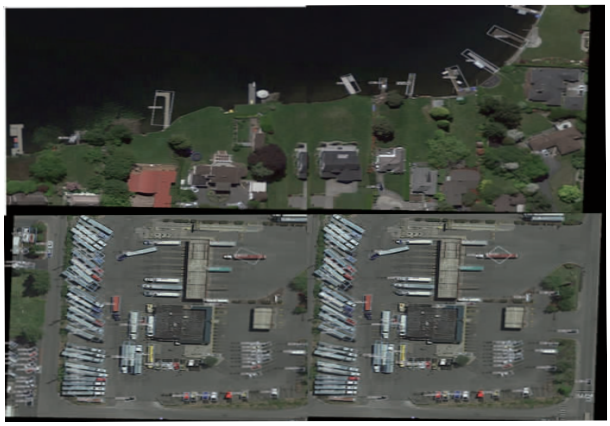


图5 部分检测结果

以利用倾斜检测框得到较好的检测效果;对于处于复杂背景中的港口、停车场等目标,模型也可以准确获取到目标的位置.

4.4.2 消融实验

为了判别网络中模块的有效性,实验以 RetinaNet 为基准,分别对特征金字塔模型、弱语义注意力模块的进行消融实验. 评估特征金字塔模型和注意力机制模块的时候,依旧按照 mAP 值进行精度比较,探究各个模块对模型效果的贡献,消融实验的结果如表 5 所示.

表 5 消融实验

基准 RetinaNet	特征金字塔	注意力机制	mAP(%)
√			55.32
√	√		56.43
√	√	√	**57.74**

特征金字塔有效性:实验中采用的特征金字塔模型与传统的 RetinaNet 采用的金字塔模型不相同. 实验中利用特征金字塔三层特征图,对其进行卷积操作生成五层金字塔特征. 虽然这种方法增加了网络的复杂性,但是遥感图像目标多样,这种方法可以提取到更加丰富的语义信息. 经过消融实验发现特征金字塔可以提高模型约 1.1% 的 mAP 值.

注意力机制的有效性:利用弱语义注意力网络生成目标置信度,进而强化目标特征. 实验中对不同层级特征进行强化,实验结果证明弱语义注意力网络可以提升模型 1.3% 的 mAP 值.

在 RetinaNet 基准上分别加入本文的特征金字塔结构和注意力网络,有效的增强了检测效果,通过消融实验也验证了各个模块在模型精度上的提升,因此证明了本文提出方法的有效性.

4.5 可解释性实验及分析

4.5.1 分类任务与检测任务可解释性剖析结果对比实验

本文用于目标检测任务的模型中包含主干网络 ResNet50,而 ResNet50 也可作为分类任务的特征提取网络. 本节实验主要对不同任务下的 ResNet50 进行可解释性对比. 首先对不同任务下的特征提取网络进行网络剖析,然后对比网络剖析结果中的可解释性语义,最终通过对比发现检测任务和分类任务网络关注的语义信息究竟有何异同. 实验中选取 ResNet50 第五阶段第二层(C5_layer2)为例,C5_layer2 中共 512 个卷积核. 对两个网络剖析的结果如表 6 所示.

表 6 分类主干网络和检测主干网络结果对比情况

任务类型	任务设置	可解释语义概念	卷积核个数
分类任务	训练集:ImageNet[33]	场景	211
		目标	109
	模型:ResNet50	部件	167
		材质	5
	卷积层:C5_layer2	纹理	20
		颜色	0
检测任务	训练集:DOTA	场景	209
		目标	38
	模型:ResNet50	部件	54
		材质	0
	卷积层:C5_layer2	纹理	211
		颜色	0

实验结果显示,通过网络剖析,在检测任务中共有 209 个卷积核关注于场景信息、211 个卷积核关注于纹理信息,其余包含 92 个目标、部件卷积核而对于颜色和材质的语义卷积核则为 0. 按照本文提出的高级、低级划分方式得出结论,该层中既包含了高级的语义信息也包含了低级的纹理信息. 而在分类任务中有 211 个卷积核关注于场景信息、109 个卷积核关注于纹理信息、167 个卷积核关注于部件信息,其余包含 25 个关注材质、纹理卷积核,关注颜色语义卷积核个数则为 0. 因此,分类任务的模型在该层只包含高级语义信息的卷积层. 可视化实验结果分别如图 6、图 7 所示.

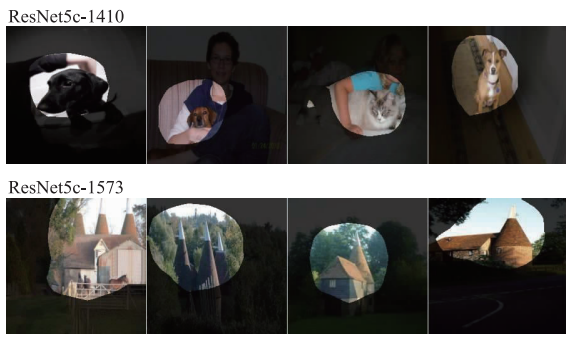


图6 分类模型经过算法得到部分结果的可视化

实验表明,检测模型中主干网络的关注信息与分类网络不同. 分类网络主要通过卷积层获取输入图像的语义信息进行分类任务,而检测模型的主干网络需

要提取既包含语义信息也包含底层特征信息. 因此得出结论,在分类任务中,需要主干网络提取数据语义特征,而定位任务需要获取上下文位置特征. 实验结论符合人类认知.

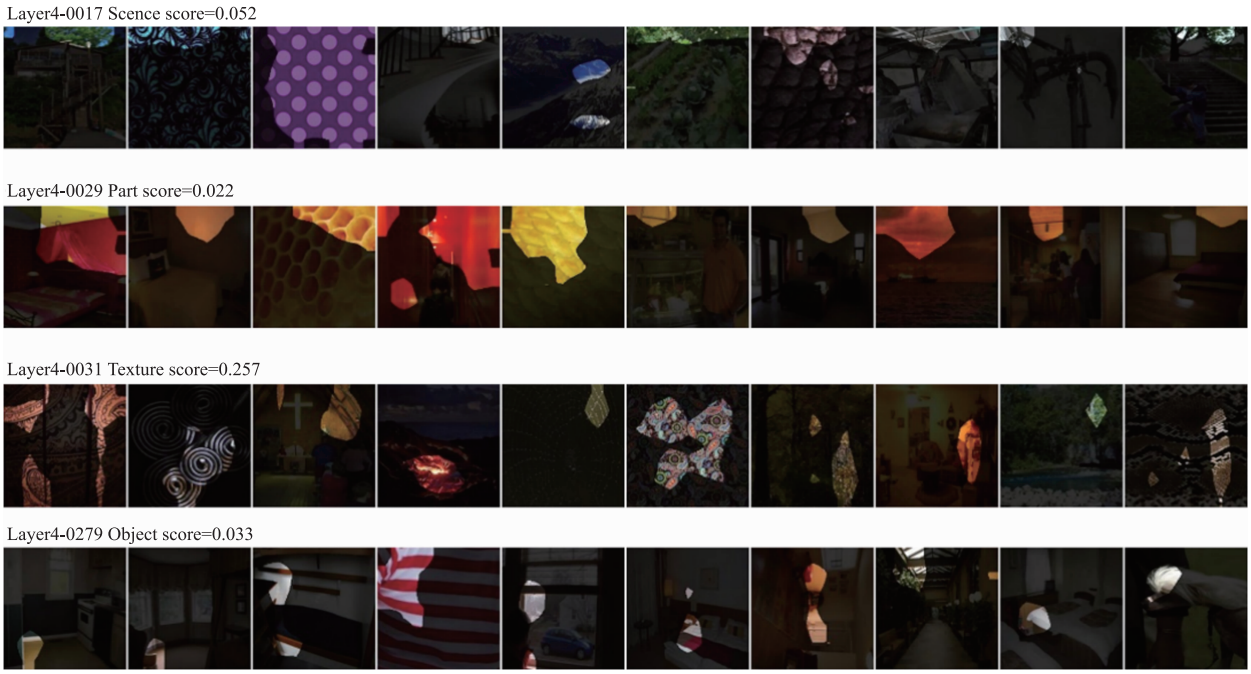


图7　检测模型经过剖析算法得到的部分可视化结果

4.5.2　检测模型不同层级可解释性剖析结果对比实验

对 ResNet 整个 C5 块中的前三个卷积层进行实验,实验结果如表 7 所示.

表 7　ResNet 第五阶段前三层卷积剖析结果

ResNet-C5Block 层数	可解释含义
Layer1	目标、材质
Layer2	场景、纹理
Layer3	目标

实验结果显示 Layer1 关注的语义信息为目标与材质、Layer2 关注的语义为场景与纹理,都为既包含高级语义信息又包含低级语义信息的层,印证了对比实验中主干网络需要提供给检测网络上下文和语义信息的结论. Layer3 中对目标激活的语义较多,说明检测模型顶层关注的是目标语义信息.

本节实验除了进一步印证了 4.5.1 节实验中得出的结论外,Layer1 与 Layer3 作为 ResNet 最后阶段的卷积层都包含了目标的语义信息,因此实验表明目标检测模型的主干网络要提取目标语义的特征信息.

4.5.3　实验结果总结与分析

结合 4.5.1 节与 4.5.2 节中的对比实验以及对其实验结论分析,探究卷积神经网络隐藏层表征信息的可解释性含义. 根据卷积神经网络的原理得知,神经网络的隐藏层表征信息是通过卷积的线性运算、激活函数非线性运算将输入信息映射到特征空间后的结果. 由于人类无法对网络的特征映射结果做出解释,从而导致卷积神经网络是黑盒模型.

网络剖析的算法可以将特征映射的结果(表征信息)与一系列人类可解释性语义相关联,并且通过计算关联程度,量化可解释性. 以 4.5.1 节对比实验为例,C5_layer2 通过网络剖析算法,得到 209 个定义为场景语义的卷积核,211 个纹理卷积核,38 个目标语义卷积以及 54 个部件卷积核. 在 209 个场景语义卷积核中随机选中第 63 个卷积核,并对其结果进行分析,结果如表 8 所示.

表 8　C5_layer2 层第 63 个卷积核网络剖析结果

可解释类别	标注像素点个数	网络激活像素点个数	IoU
场景	12544	8930	0.21342600
目标	57150	13359	0.05435595
部件	1368	705	0.05603668
纹理	25088	61765	0.19802199
颜色	829402	79289	0.01363349

输入数据通过 63 号卷积核映射到特征空间. 由于人类无法理解卷积映射的结果,因此,本文实际上是根据可解释语义对特征空间做出解释. 63 号卷积核网络剖析的结果表明映射到特征空间的语义为场景.

对于卷积层来说,4.5.2 节实验结果发现输入信息在特征空间的语义多为一个或者两个. 进一步分析,训练过程中采用的优化算法本质上是根据目标任务逐渐

调整线性变换的权重从而调整特征空间的语义,更有效地获取完成目标任务所需要的特征信息. 例如,优化后的 C5_layer2 将输入信息映射到场景和纹理语义的特征空间以此完成目标检测任务. 通过 4.5.1 节实验结果也可以发现,对于不同任务目标,优化算法调整权重值后,相同网络的结构也获取到了不同的特征信息,因此可知根据目标任务不同,卷积网络提取的特征语义也是不同的.

通过分析实验的结果,进一步了解卷积神经网络处理视觉信息的方式,有助于理解卷积神经网络的工作原理. 后续工作可以根据上述实验结果及结论对模型进行优化,即通过强化表征信息中特定的语义信息,加强特征提取的有效性.

5 结论

本文提出了一个基于弱语义注意力的遥感图像可解释目标检测算法,针对遥感图像目标检测中尺度变化大、背景复杂、目标定向以及缺乏可解释性的问题,改进特征金字塔网络,利用多层级特征进行多尺度预测,提升模型的检测性能. 此外,本文提出了弱语义注意力网络,通过弱语义标注分割目标与背景特征以此达到强化目标特征、过滤背景噪声的目的,提高模型对目标的识别与检测效果. 为了更好地实现检测框与目标的匹配,引入角度值的回归对水平框进行几何变换,解决了目标定向、密集排列的问题. 最终,改进网络剖析算法,量化主干网络卷积层及卷积核的语义信息并进行可解释性分析,使模型易于理解. 未来工作将利用网络剖析算法得到的结论进一步弥补模型在可解释性上的不足,并改进弱语义注意力网络结构,提高模型的准确性.

参考文献

[1] Zhang Q, Zhu S C. Visual interpretability for deep learning: a survey[J]. Frontiers of Information Technology & Electronic Engineering, 2018, 19(1): 27 – 39.

[2] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Los Alamitos: IEEE Computer Society Press, 2017. 2980 – 2988.

[3] Xia G S, Bai X, Ding J, et al. DOTA: A large-scale dataset for object detection in aerial images[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Los Alamitos: IEEE Computer Society Press, 2018. 3974 – 3983.

[4] Bau D, Zhou B, Khosla A, et al. Network dissection: Quantifying interpretability of deep visual representations[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Los Alamitos: IEEE Computer Society Press, 2017. 6541 – 6549.

[5] 张顺,龚怡宏,王进军. 深度卷积神经网络的发展及其在计算机视觉领域的应用[J]. 计算机学报, 2019, 42(3): 453 – 482.
Zhang S, Gong Y H, Wang J J. The development of deep convolution neural network and its applications on computer vision[J]. Chinese Journal of Computers, 2019, 42(3): 453 – 482. (in Chinese)

[6] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Los Alamitos: IEEE Computer Society Press, 2014. 580 – 587.

[7] Girshick R. Fast r-cnn[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Los Alamitos: IEEE Computer Society Press, 2015. 1440 – 1448.

[8] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[A]. Advances in Neural Information Processing Systems[C]. Cambridge: MIT Press, 2015. 91 – 99.

[9] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Los Alamitos: IEEE Computer Society Press, 2017. 2961 – 2969.

[10] Cai Z, Vasconcelos N. Cascade r-cnn: Delving into high quality object detection[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Los Alamitos: IEEE Computer Society Press, 2018. 6154 – 6162.

[11] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[A]. European Conference on Computer Vision[C]. Cham: Springer, 2016. 21 – 37.

[12] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Los Alamitos: IEEE Computer Society Press, 2016. 779 – 788.

[13] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Los Alamitos: IEEE Computer Society Press, 2017. 7263 – 7271.

[14] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. ArXiv Preprint, 2018, arXiv: 1804.02767.

[15] Samek W, Wiegand T, Müller K R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models[J]. ArXiv Preprint, 2017, arXiv: 1708.08296.

[16] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[A]. European Conference on Com-

puter Vision[C]. Cham:Springer,2014. 818 – 833.
[17] Zhou B,Khosla A,Lapedriza A,et al. Object detectors emerge in deep scene cnns[J]. ArXiv Preprint,2014,arXiv:1412. 6856.
[18] Goyal Y,Mohapatra A,Parikh D,et al. Towards transparent AI systems: Interpreting visual question answering models[J]. ArXiv Preprint,2016,arXiv:1608. 08974.
[19] 张峰,钟宝江. 基于兴趣目标的图像检索[J]. 电子学报,2018,46(8):1915 – 1923.
Zhang F,Zhong B J. Image retrieval based on interested objects[J]. Acta Electronica Sinica,2018,46(8):1915 – 1923. (in Chinese)
[20] Lin T Y,Dollár P,Girshick R,et al. Feature pyramid networks for object detection[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Los Alamitos:IEEE Computer Society Press,2017. 2117 – 2125.
[21] He K,Zhang X,Ren S,et al. Deep residual learning for image recognition[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Los Alamitos: IEEE Computer Society Press, 2016. 770 – 778.
[22] 李宝奇,贺昱曜,何灵蛟,等. 基于全卷积神经网络的非对称并行语义分割模型[J]. 电子学报,2019,47(5):1058 – 1064.
Li B Q,He Y Y,He L J,et al. Asymmetric parallel semantic segmentation model based on full convolutional neural network[J]. Acta Electronica Sinica,2019,47(5):1058 – 1064. (in Chinese)
[23] Long J,Shelhamer E,Darrell T. Fully convolutional networks for semantic segmentation[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Los Alamitos: IEEE Computer Society Press, 2015. 3431 – 3440.
[24] Ma J,Shao W,Ye H,et al. Arbitrary-oriented scene text detection via rotation proposals[J]. IEEE Transactions on Multimedia,2018,20(11):3111 – 3122.
[25] 黄德根,张云霞,林红梅,邹丽,刘壮. 基于规则推理网络的分类模型[J]. 软件学报, 2020, 31(4): 1063 – 1078.
Huang D G,Zhang Y X,Lin H M,Zou L,Liu Z. Rule inference network model for classification[J]. Journal of Software,2020,31(4):1063 – 1078. (in Chinese)
[26] Zhou B, Zhao H, Puig X, et al. Scene parsing through ade20k dataset[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Los Alamitos:IEEE Computer Society Press,2017. 633 – 641.
[27] Bell S,Bala K,Snavely N. Intrinsic images in the wild[J]. ACM Transactions on Graphics,2014,33(4):1 – 12.
[28] Mottaghi R,Chen X,Liu X,et al. The role of context for object detection and semantic segmentation in the wild[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Los Alamitos:IEEE Computer Society Press,2014. 891 – 898.
[29] Chen X,Mottaghi R,Liu X,et al. Detect what you can: Detecting and representing objects using holistic models and body parts[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Los Alamitos:IEEE Computer Society Press,2014. 1971 – 1978.
[30] Cimpoi M,Maji S,Kokkinos I,et al. Describing textures in the wild[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Los Alamitos:IEEE Computer Society Press,2014. 3606 – 3613.
[31] Dai J,Li Y,He K,et al. R-fcn: Object detection via region-based fully convolutional networks[A]. Advances in Neural Information Processing Systems[C]. Cambridge: MIT Press,2016. 379 – 387.
[32] Lin Y,Feng P,Guan J. IENET:Interacting embranchment one stage anchor free detector for orientation aerial object detection[J]. ArXiv Preprint,2019,arXiv:1912. 00969.
[33] Krizhevsky A,Sutskever I,Hinton G E. Imagenet classification with deep convolutional neural networks[A]. Advances in Neural Information Processing Systems[C]. Nevada:Lake Tahoe,2012. 1097 – 1105.

作者简介

周　勇　男,1974 年 9 月出生于江苏省徐州市,中国矿业大学计算机科学与技术学院教授. 主要研究方向为数据挖掘、机器学习和人工智能.
E-mail:yzhou@ cumt. edu. cn

陈思霖　男,1998 年 4 月出生于吉林省吉林市,中国矿业大学计算机科学与技术学院硕士研究生. 主要研究方向为计算机视觉、图像处理、目标检测.
E-mail:silin. chen@ cumt. edu. cn