基于因果推断的弱监督野生动物识别算法研究

郭亦嘉[1],沈苏彬[2]
(1. 南京邮电大学 物联网学院,江苏 南京 210023;
2. 南京邮电大学 通信与网络技术国家工程研究中心,江苏 南京 210023)

摘　要:全监督野生动物识别算法的数据集需要大量标注数据,耗时耗力且容易引入噪声。弱监督学习方法只需图像级别的标注信息,但是由于动物的生活习惯和栖息环境等影响,动物通常与特定的背景共同出现,会导致模型受到背景特征的干扰,即“共现混淆”问题。针对该问题,该文提出了一种基于因果推断的识别算法 CI-ResNet,通过因果干预消除混淆因子的影响,提升目标识别和定位精度。研究构建了混淆背景集模块,收集不同类别的背景特征,利用后门调整消除伪相关性,从而实现对特征的因果干预。实验在 CUB-200-2011 数据集和 AwA2 数据集上进行,与现有深度学习方法和公开基准方法相比,在识别准确率和定位精度方面分别提高了 1.92 百分点、2.11 百分点和 1.73 百分点、2.15 百分点。

关键词:因果推断;弱监督学习;后门调整;图像识别;因果图
中图分类号:TP391.3　　**文献标识码**:A　　**文章编号**:1673-629X(2025)05-0158-08
doi:10.20165/j.cnki.ISSN1673-629X.2024.0394

Weakly Supervised Wildlife Recognition Algorithm Based on Causal Inference

GUO Yi-jia[1],SHEN Su-bin[2]
(1. School of Internet of Things,Nanjing University of Posts and Telecommunication,Nanjing 210023,China;
2. National Engineering Research Center on Communication and Networking,Nanjing University of Posts and Telecommunication,Nanjing 210023,China)

Abstract:Fully supervised wildlife recognition algorithms require extensive annotated data,which is time-consuming,labor-intensive,and prone to noise. Weakly supervised learning methods only need image-level annotations. However,due to the influence of animals´ living habits and habitats,animals often co-occur with specific backgrounds,leading to the "co-occurrence confusion" problem,which causes the model to be interfered with by contextual features. To address this issue,we propose a recognition algorithm based on causal inference,CI-ResNet,which removes the influence of confounding factors through causal intervention,thereby improving target recognition and localization accuracy. The study constructs a confounder context set module to collect background features of different categories and uses backdoor adjustment to eliminate spurious correlations,thus achieving causal intervention on features. The experiments were conducted on the CUB-200-2011 dataset and the AwA2 dataset. Compared with the existing deep learning methods and public benchmark methods,the recognition accuracy and localization precision were respectively increased by 1.92 percentage points,2.11 percentage points and 1.73 percentage points,2.15 percentage points.

Key words:causal inference;weakly-supervised learning;back-door adjustment;image recognition;causal graph

0　引　言

保护野生动物,维护生物多样性和生态链完整性,促进人与自然和谐共生,是生态文明建设的重要任务之一。在过去的几年间,科研人员利用相机采集到的动物图片,可以很好地了解动物的习性、物种分布以及动物大致数量,从而更好地为保护野生动物制定相关措施。然而,传统的人工图片识别检测需要专业知识,这就需要一种更加高效、准确的方法来实现野生动物图像的精准识别。通过深度学习的方法,可以实现对野生动物图像的自动化识别,无需人工驻守监视,极大地削减了人工成本开支和时间开支。

使用完全监督的深度学习模型进行野生动物的识

收稿日期:2024-10-07　　**修回日期**:2025-02-11
基金项目:国家自然科学基金(62002174)
作者简介:郭亦嘉(2000-),男,硕士研究生,通讯作者,研究方向为计算机视觉;沈苏彬(1963-),男,博导,研究员,CCF 高级会员(E200005482S),研究方向为物联网及其应用、未来网络及其应用。

别,需要对大量图片进行像素级的标注,即每张图片中对每个出现的目标都进行类别及位置(边界框)的标注。在实际的应用中,全监督目标检测需要大量的完整标注信息的数据。数据集的标注需要大量的时间和人力成本[1],且标注过程中不可避免地带来噪声(标注错误等),这些因素使得大量的实际应用很难直接采用全监督学习方式来解决。

近年来越来越多的工作开始专注于弱监督目标检测(Weakly Supervised Object Detection, WSOD)[2-5],其使用的训练数据只带有图像级别(Image-Level)的标注信息,极大减少了标注数据集的时间成本和人力开销。故采用弱监督野生动物图像识别的方法已经成为一个研究热点。但是由于动物的生活习惯和栖息环境等影响,动物通常与特定的背景共同出现,导致数据集中服从该特征分布的图片较多[6]。例如,在大多数天鹅的图像中,“天鹅”与“水面”同时出现的概率很高,因为天鹅的栖息环境是湖泊,这两个特征将不可避免地纠缠在一起,分类模型将错误地在只有弱监督的情况下将“天鹅”和“水面”之间建立关联。上述现象中,“天鹅”的先验分布引入了“水面”这样的混淆因子,深度学习模型只捕获到它们之间的相关性,忽略了因果关系。具体地说,对于“天鹅”而言,“水面”与它并无因果关系,但由于它们在数据集中的分布呈现正相关(大概率同时出现),导致识别模型将二者建立了联系,模型可能学到了不具备泛化性的关联性,如果训练集和测试集分布差异较大,会导致模型的识别精度以及目标范围方面的定位精度降低[7]。因此,该文将上述“背景-动物类别”之间产生伪相关的现象称为共现混淆(Co-occurrence)。

因果推断是用于数据分析的强大建模工具,可以帮助发现数据中客观存在的因果关系,实现稳定预测,且因果关系也能为数据模型提供较强的可解释性。针对上述问题,该文提出使用因果推断理论抑制由混杂因子带来的伪相关性,提升目标物体的识别与定位精度。同时,通过构建混淆背景库,不断地收集各个类的背景特征,通过后门调整消除无关特征与目标类别之间的伪相关性以达到对分类器的因果干预。基于这两个思想,以 CUB-200-2011 数据集为例,构建了一种基于因果推断的野生动物识别算法 CI-ResNet(Causal Inference Resnet)。主要贡献如下:

(1)提出了一种基于因果推断的弱监督野生动物识别算法,通过对特征进行因果干预,解决野生动物识别领域的“共现混淆”问题,提高弱监督下野生动物的识别与定位准确率。

(2)为了能够消除混杂因子对目标物体的干扰,提出了一个混淆背景库模块,通过因果推断的方式建模待识别目标在所有混杂因子下的分布情况,消除特定混杂因子的干扰。

1　相关工作

1.1　弱监督目标识别算法

目前主流的 WSOD 算法主要是基于 CAM 的网络结构。2016 年在文献[8]中首次有研究者提出了 CAM 的概念。这项工作主要证明了两个结论:一是卷积神经网络提取的特征含有位置信息,尽管在训练的时候没有标记位置信息,但 CAM 最初是用来可视化模型特征图,以便观察模型是通过图像中的哪些区域特征来区分物体类别的。二是用这些位置信息,由于弱监督场景下只有图像级别的标签,CAM 可以作为一种弱监督的目标识别方法,这对后续的弱监督目标识别的发展带来了巨大的改变。

由于 CAM 容易偏向于目标最具判别性的部分,而不是整体目标,因此目前大多数方法的研究重点是如何提高目标定位的精度。这些方法大致可以分为两类:基于擦除的 CAM 方法和基于线性融合的 CAM 方法。

(1)基于擦除的 CAM 方法。文献[9]提出了一种随机擦除方法 Has,擦除策略为在原图上划定多个网格,然后将随机选择部分格子进行擦除,进一步将擦出后的图片输入网络中进行分类训练,迫使网络学习到更多的特征信息。文献[10]提出的 Acol 在基于 Has 上做了进一步改进和提升,通过两个对抗性分类器以弱监督的方式有效地挖掘不同的判别区域。文献[11]提出的 ADL 方法在 Acol 的基础上又进一步探索擦除思路,擦除策略为在中间卷积层上选择激活值较大的区域进行擦除,然后将擦除和未擦出的特征图随机选择一个输入后续的分类网络中,最后使用 CAM 方法实现定位图的提取。

(2)基于线性融合的 CAM 方法。文献[12]提出的 CCAM 方法将 CAM 方法生成的多个激活图进行融合而不是在 CAM 方法中使用最高概率类的激活图来生成最终的定位图。文献[13]提出的 DANet 方法相较于 CCAM 要复杂的多,其将特征图在多个维度进行平均,然后再输入到全局平均池化层(Global Average Pooling, GAP)中进行分类。其中,激活图的融合是在多个维度上进行的。

这些方法通过多种操作让分类网络感知到更多的区域,但随着激活区域变大,模型会学习到过多的无关背景特征,更重要的是上述方法都忽略了“共现混淆”问题,由于动物的生活习惯和栖息环境等影响,动物通常与特定的背景共同出现,“共现混淆”问题在野生动物识别场景下尤为严重,会误导模型学习到与目标没

有因果关系但呈正相关出现的背景特征,那么基于 CAM 的方法得到的激活图就会受到模糊边界的困扰,对后续的目标定位产生不良影响,并且如果训练集和测试集的分布不一致,则会导致模型识别精度降低。

1.2 因果推理

导致“共现混淆”的根本原因是目前的深度学习是学习数据间的统计相关性[14]。以视觉识别模型为例,模型在训练时,通过对形式为 $D=\{(x_i,y_i)\}_{i=1}^{N}$ 的数据集进行训练,通过梯度下降策略,利用反向传播机制更新和优化模型参数来学习条件分布 $P(Y\mid X)$,其中 X 和 Y 分别表示骨干网络提取的特征和图像类别。如果识别动物与某个背景经常共同出现,则在模型拟合输出结果与真实值的过程中,会将待识别目标和背景特征共同作为识别结果的依据,这显然是不符合因果关系的,同时也不符合人类视觉识别的过程。背景偏差将模型的注意力误导到共现背景上而不是对象上,从而导致识别准确率降低。

因果推理[15-16]是一个跨多个领域的重要研究课题,它不仅仅是一个解释框架,还提供了通过追求因果效应来实现预期目标的解决方案。近年来,越来越多的研究将因果推理应用在计算机视觉领域中[17-19],Judea Pearl 作为因果革命的先驱者,提出了因果关系的方法论,引入了结构因果模型(Structural Causal Model,SCM)、后门调整(Backdoor Adjustment)、do 算子(Do-operator)等概念[20]。这些研究在 Judea Pearl 因果理论的指导下,重新审视并构建视觉识别场景下的因果图(Causal Graph),对混淆因子去偏以达到鲁棒的视觉识别。具体而言,文献[17]通过分析弱监督语义分割场景下各变量之间的因果关系并构建因果图,指出传统的语义分割模型中边界扩张操作会受到背景的影响,比如受到图片中的其他物体、背景等混淆因子的影响,无法准确地得到像素与类别之间的正确因果关系。因此提出了基于因果推断的弱监督语义分割算法(CONTA),该算法的核心是构建混杂因子集进行后门调整,将当前候选区域与所有类别平均区域的相似度求一个加权平均以去除目标与背景之间的混杂影响,得到新的特征图,从而完成对目标像素的因果干预,在弱监督语义分割任务中取得了不错的效果。文献[18]主要解决的问题是目标检测任务中,数据集中某些物体经常共同出现让模型造成错误关联导致模型在测试集中的表现造成负面影响。作者将因果推理的干预思想融入到 Faster R-CNN 目标检测框架中,提出了包含所有潜在干扰因素的混淆因子字典(Confounder Dictionary)模块。该模块由数据集中各个类的平均感兴趣区域(Region Of Interest,ROI)组成,在模型堆输入图像进行检测时,利用混淆因子字典对提取的特征进行后门调整,减轻共现的背景特征对待识别目标的干扰。

上述两种算法均通过构建混杂因子集,然后对骨干网络提取的特征进行重加权的操作实现因果干预,并取得了不错的效果。但是上述算法中混淆因子集的构建是需要额外进行处理的,如果更换数据集,那么混杂因子集也需要重新构建,不够灵活。该文提出的 CI-ResNet 方法受上述思想启发的同时,对混淆因子集的构建作了一些优化。利用 CAM 能够凸显目标区域的同时也能保留图像的背景信息这一优势,采用双分支结构的神经网络,上分支负责收集图像的背景信息存入混淆背景库中,实现对混淆因子的收集,值得注意的是,该部分的操作是在模型训练时就进行的,无需额外的操作。下分支网络利用混淆背景库中收集的背景信息对特征图进行重加权,完成对目标特征的因果干预,实现鲁棒的预测结果。

2 基于因果推断的野生动物识别模型

2.1 结构因果模型

前文分析了弱监督野生动物识别场景下的因果关系,分析“共现混淆”问题出现的原因,讨论了造成偏差的因素,以及该场景下各因素之间的因果关系;本节将采用 SCM 描述图像特征 X、背景混淆因子(Context Confounder) C 和对应类别 Y 之间的因果关系,如图 1 所示,其中箭头表示两个节点之间的因果关系:原因→结果。

$X\rightarrow Y$:此链接表示野生动物图像识别模型中最基本的因果关系,即通过图像特征得到对应的动物类别。

$C\rightarrow X$:此链接表示骨干网在背景特征的影响下生成特征图 X。虽然模型通过 $P(Y\mid X)$ 学习条件分布并利用混杂因子 C 更好地将图像特征 X 和标签 Y 关联起来,例如,当看到一个“水面”区域时,它很可能是一个“天鹅”,但是,$P(Y\mid X)$ 错误地将非因果但正相关的像素与标签关联起来,例如,“水”区域错误地属于“天鹅”。这就是弱监督目标识别不准确的原因之一。该文将在后续章节通过在因果干预中使用因果背景池来避免它。

$X\rightarrow M\leftarrow C$:$M$ 在该因果图中是一个中介变量,是前景 X 和背景 Y 的特定表示。例如,在识别狮子的图像时,M 表示狮子(前景)在场景(背景)中的位置,在没有对 X 和 C 施加干预的情况下,M 会受到 X 和 C 的影响产生偏置特征,从而造成特定背景与类别之间的虚假关联。注意,M 并不等同于输入的图像,而是出现在 CNN 网络中更高层的特征表示。

$X\rightarrow Y$、$M\rightarrow Y$:这些链接表明,最终的预测效应可

以分解为两种方式:直接效应和中介效应。$M \rightarrow Y$ 很容易理解:背景信息对类别的预测有相当大的影响。物体本身和它所处的环境共同影响着人们对它的认识。特别是在共现混淆的影响下,环境可能成为决定预测的主要因素。然而,决定一个物体是否出现的,是这个物体实际出现在图像中,而不是"它应该出现",这促使模型去减轻背景的影响。没有了背景的负面影响,预测更加稳健和可靠,因此,需要模型被迫学习对象本身的特征。

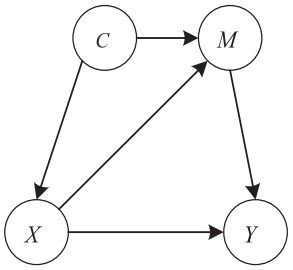

图 1　弱监督野生动物识别场景下的因果图

基于深度学习的野生动物图像识别模型只通过拟合 $P(Y \mid X)$ 发现图像特征与图像类别之间的关联关系,而忽视了后门路径(Back-door Path) $X \leftarrow C \rightarrow M \rightarrow Y$ 对最终结果 Y 的影响,后门路径定义为所有 X 和 Y 之间以指向 X 的箭头为开始的路径,因为这些路径允许 X 和 Y 之间的伪相关信息在通道中流通。在野生动物图像识别场景中动物特征与动物类别的对应关系是可靠的关联关系即因果关系,通过对图 1 分析,发现 C 通过后门通道干扰了动物特征与动物类别的因果效应,导致模型依据背景与目标的共同特征进行分类。接下来,将引入因果干预方法来消除混杂效应。

2.2　弱监督野生动物识别下的因果干预

由图 1 的因果图可知,当训练图像分类器时,提取的视觉特征不可避免地受到视觉背景的影响,故视觉混杂因子 C 会影响区域视觉特征 X,即 $C \rightarrow X$。同样地,视觉背景也会影响分类器的输出概率,即 $C \rightarrow M \rightarrow Y$。因此,神经网络使用弱监督的方式进行训练时,很可能由于混杂因子 C 而学习到 X 和 Y 之间的一些伪相关性,即过度利用视觉背景和类别之间的共现性学习到图像区域视觉表征的偏差。

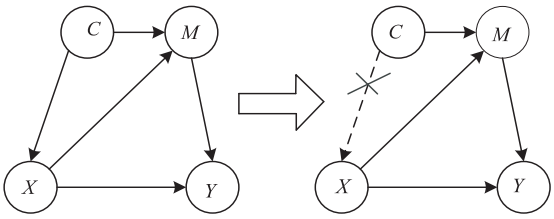

图 2　干预后的因果图

为解决上述问题,该方法使用因果干预似然函数 $P(Y \mid \mathrm{do}(X))$ 作为野生动物图像识别新分类器的训练目标,其中 do 算子的作用是切断 $C \rightarrow X$ 之间的虚假关联。由 Judea Pearl 的因果理论可知,后门调整[21]是对混淆因子"分层"讨论,因此将混淆因子 C 分成 $C = [c_1, c_2, \cdots, c_n]$,其中 n 表示训练集的类别个数,c_i 表示第 i 个类别的混淆背景特征。

如图 2 所示,利用后门调整策略可得下式:

$$P(Y \mid \mathrm{do}(x)) = \sum_c P(Y \mid X = x, M = f(x, C = c)) P(C = c) \tag{1}$$

式中,$P(Y \mid \mathrm{do}(X))$ 的本质是迫使 X 公平地"借用"混杂因子中每一个 c,并将它们"放在一起",以便动物类别的预测。M 可以用 $f(x,c)$ 抽象地表示,说明 M 是由图像特征 X 和 c_i 共同组成,具体的计算将在 2.3 小节中定义。这样,分类器能够消除混淆因子的干扰并学习到 X 和 Y 之间真正的因果关系,从而获得高质量的视觉特征。然而,应用于野生动物图像分类任务时,式 1 需要大量的采样来估计 $P(Y \mid \mathrm{do}(X))$,这使得训练时间望而却步。但是,通过使用归一化加权几何平均[22](Normalized Weighted Geometric Mean, NWGM)近似值,可以将 $\sum_c P(c)$ 移至特征层进行计算,式 1 可近似为式 2。

$$P(Y \mid \mathrm{do}(X)) \approx P\left(Y \mid X = x, M = \sum_i^n f(x, c_i) P(c_i)\right) \tag{2}$$

这样神经网络只需要前向传播一次而不是 n 次就可以得到目标特征与动物类别的因果效应,由于数据集中每个类的数量大致相同,那么 $P(c_i) = \frac{1}{n}$,可以进一步得到式 3。

$$P(Y \mid \mathrm{do}(X)) \approx P\left(Y \mid x \odot \frac{1}{n} \sum_i^n f(x, c_i)\right) \tag{3}$$

其中,$\odot$ 是哈达玛积(Hadamard product),也就是张量的逐元素乘积,将因果背景库中的特征映射到骨干网络提取的特征图上,完成一次后门调整。至此,共现混淆问题已经转移到计算 $\sum_i^n f(x, c_i)$ 中。在 2.3 节中将引入一个因果背景库来表示 $\sum_i^n f(x, c_i)$ 并详细介绍该方法的具体实现。

2.3　野生动物识别算法 CI-ResNet

在本节中,将以 CUB-200-2011 数据集和 AwA2 数据集为例,构建了一种基于因果推断的野生动物识别算法 CI-ResNet(Causal Inference Resnet)。如图 3 所示,其核心是因果背景库。因果背景库的主要思想是累积每个类的所有背景,然后将背景重新投射到如式 3 所示的卷积层的特征映射中,以追求原因 X 和结果 Y 之间的纯粹因果关系。CI-ResNet 采用双分支结构,上分支主要负责生成激活图并更新到因果混淆因

子集中,下分支负责对特征图进行因果干预并产生更鲁棒的预测结果。模型主要包含三个部分:特征提取模块、CAM 模块、因果背景库模块。

(1)特征提取模块。该模块选择的是 ResNet50[23] 模型作为骨干网络。具体而言,将 224 * 224 * 3 大小的野生动物图像输入骨干网络,经过一系列的卷积、池化,最后得到特征图 $X \in R^{c\times h\times w}$,其中 c 为特征图的通道数, h 和 w 分别表示特征图的高和宽。

(2)Grad-CAM 模块。该模块负责产生对应类别的类激活图,其过程如图 3 所示。具体而言,将特征图 X 送入分类器后产生预测概率 $P=[p_1,p_2,\cdots,p_n]$,然后通过 Grad-CAM[24] 方法产生概率最大的类别的激活图 M ,Grad-CAM 方法相比于需要将全连接层改为全局平均池化层的传统 CAM 方法,其通过反向传播过程中的梯度信息来计算激活图,所以不需要改变网络结构就可以得到对应类的激活图。具体的计算公式如式 4 和式 5 所示。

$$a_c^k=\frac{1}{Z}\sum_i\sum_j\frac{\partial y^c}{\partial A_{ij}^k} \tag{4}$$

$$M=\mathrm{ReLU}\left(\sum_k\alpha_k^cA^k\right) \tag{5}$$

其中, A_{ij}^k 代表特征层 A 的第 k 个通道上坐标 (i,j) 的数值, y^c 代表前向传播得到的类别 c 所对应的预测分数,因此对 A_{ij}^k 和 y^c 求偏导可得到类别 c 在特征层 A 上反向传播得到的梯度信息。Z 为特征层的大小。然后将计算得到的梯度在宽度 i 和高度 j 的维度上进行全局平均池化,得到重要性权重 a_k^c ,最后将权重与对应通道的特征图进行加权求和,并通过 ReLU 函数激活后输出即可得到激活图 M 。

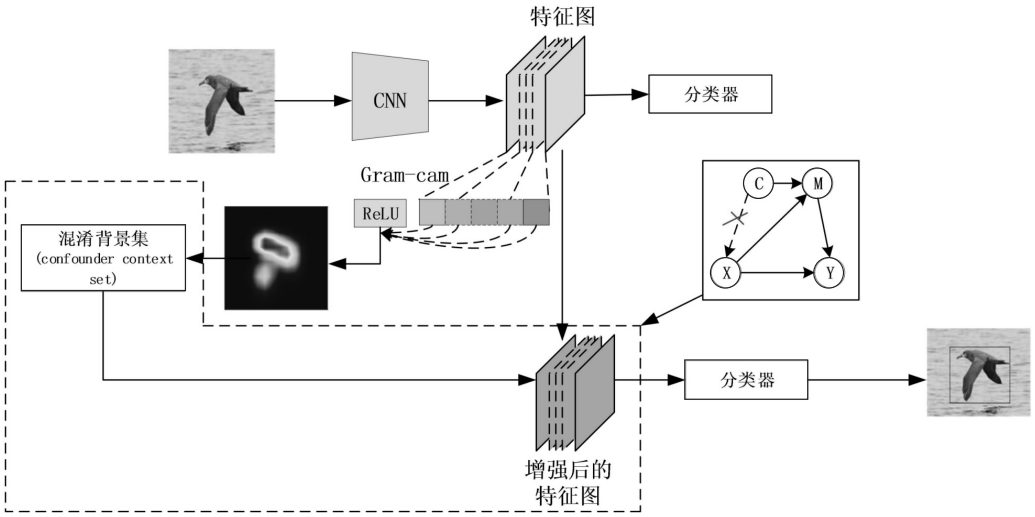

图 3 CI-ResNet 网络结构

(3)因果背景库模块。在提出的 CI-ResNet 网络结构中使用特定类别的激活图来近似构建因果背景集 $C^{n\times h\times w}=[c_1,c_2,\cdots,c_n]$,其中 n 为数据集中类的数量, $c_i\in R^{h\times w}$ 对应第 i 个类图像的平均激活图,该激活图由 Grad-CAM 模块产生。混淆背景集通过累积最高概率类的激活图,不断存储每个类的所有背景特征,因果背景集的具体更新方式如式 6 所示。

$$C_\lambda=\mathrm{BN}(C_\lambda+\mathrm{BN}(M_\lambda)) \tag{6}$$

其中, $\lambda=\mathrm{argmax}(p_1,p_2,\cdots,p_n)$,BN 表示批量归一化。将每个类的所有背景特征投影到最后一个卷积层的特征映射 X 上,生成增强的特征映射 X^I ,以此完成对特征的后门调整,具体如式 7 所示。该模块不仅可以消除纠缠背景对图像特征映射的负面影响,而且可以突出图像特征映射的积极区域,在有效解决共现混淆问题的同时也能扩大激活图定位目标的范围。

$$f(x,c_i)=X\odot\mathrm{Conv}_{1\times1}(C_\lambda) \tag{7}$$

式中, $\odot$ 表示哈达玛积,即张量对应像素的点积,将因果干预后的特征映射 X^I 送入下分支中,得到更鲁棒的预测结果。

$$X^I=X+\frac{1}{n}\sum_i^nf(x,c_i) \tag{8}$$

至此完成了基于因果干预的弱监督图像识别与定位。

3 实验与分析

3.1 数据集介绍

该文在两个公开数据集 CUB-200-2011[25] 和 AwA2[26] 上进行实验,CUB-200-2011 是一个常用的野生鸟类识别的数据集,其中包含 200 种不同种类的野生鸟类,每个种类包含 60 张图像,总共包含 11 788 张图像,在训练集中只有图像级别的类别标注。AwA2 数据集常用于零样本学习,该数据集包含 50 种野生动物,总共包含 37 322 张图像,为了将该数据集用在本次实验中,对数据集进行了图像级别的标注。这两个野生动物数据集中包含大量“共现混淆”的图片,野生动物与特定背景出现的频率较高,符合条件。

3.2　实验环境与参数设置

采用在 ImageNet 数据集上预训练的 ResNet50 网络作为主干。实验使用的服务器 GPU 为 Nvidia GeForceRTX4080 16 GB,模型训练平台选用基于 Python 编程语言的 PyTorch 深度学习框架。输入图像分辨率为 224 × 224,训练过程中,使用自适应矩估计[27](Adaptive Momentum Estimation,Adam)优化器加速神经网络模型的训练过程,其中 Adam 优化器的两个参数分别设置为 $\beta_1 = 0.9$, $\beta_2 = 0.99$。实验中,设置学习率为 0.000 5,批大小为 12,训练轮数为 200。

3.3　评价指标

识别准确率方面,采用图像分类中最常用的 Top-1 Cls acc(Top-1 Classify Accuracy)、Top-5 Cls acc(Top-5 Classify Accuracy)作为性能评价指标。Top-1 Cls 是指在分类问题中,模型预测的最高概率类别与实际类别相符的比率,即公式 9。

$$\mathrm{Acc} = \frac{I_{\mathrm{ac}}}{I_{\mathrm{total}}} \tag{9}$$

其中,I_{ac} 为正确分类的图像数量,I_{total} 为测试集图像总量。Top-5 Cls 是在 Top-1 Cls 的基础上,对比概率最大的前 5 个类是否包含图片真正对应的类。

定位准确率方面,采用 Top-1 Loc acc(Top-1 Localization Accuracy)、Top-5 Loc acc(Top-5 Localization Accuracy)作为评价指标。这里需要使用边界框交并比(IOU),其定义如下:

$$\mathrm{IOU} = \frac{\mathrm{gt-bbox} \cap \mathrm{pre-bbox}}{\mathrm{gt-bbox} \cup \mathrm{pre-bbox}} \tag{10}$$

其中,gt-bbox ∩ pre-bbox 表示两边框的交集,gt-bbox ∪ pre-bbox 表示两边框的并集。IOU 主要评价两个边框之间的差异,IOU 越大,表示两个边框的重合度越高,说明预测越准确。Top-1 Loc 是指分类网络的 Top1 类别正确并且预测框与标注框的 IOU 大于 0.5 时,则 Top-1 Loc 正确。Top-5 Loc 表示的是分类网络的 Top5 预测是正确的并且定位 IOU 是大于 0.5 时,则 Top-5 Loc 正确。

3.4　实验结果分析

将提出的模型在 CUB-200-2011 数据集和 AwA2 数据集上进行评估,选取 ADL[11]、ACoL[10]、Rethinking-CAM[28]、NL-CCAM[12]、GRCAM[29]、SACM[30] 和 DA-WOSL[31] 等先进且典型的方法作为对比。为了实验的公平性,对比方法都采用 ResNet50 作为骨干网络。分类和定位精度结果如表 1 和表 2 所示,黑体数据为最优结果。可以看到 CI-ResNet 方法在所有评估指标下都可以与现有方法相提并论。具体地说,CI-ResNet 在 CUB 数据集上的分类准确率为 85.5%,与目前最先进的 DA-WOSL 方法相比提升了 1.92 百分点,在 AWA2 数据集上的分类准确率为 92.15%,与目前最先进的 NL-CCAM 方法相比提升了 1.73 百分点。在定位指标下,CI-ResNet 在 CUB 数据集和 AWA2 数据集上的定位准确率分别为 64.51% 和 79.73%,与目前最先进的 DA-WOSL 方法相比分别提高了 2.11 百分点和 2.32 百分点。综上所述,在弱监督野生动物图像识别中引入因果推断,克服共现混淆问题是非常有效果的。

表 1　各方法的识别准确率　%

模型	CUB-200-2011		AwA2	
	Top-1 Cls	Top-5 Cls	Top-1 Cls	Top-5 Cls
ADL[11]	79.20	90.68	85.28	93.57
GRCAM[27]	80.47	92.14	82.04	95.39
Rethinking-CAM[28]	77.42	90.32	79.61	87.94
SACM[29]	82.34	93.18	88.48	96.30
ACoL[10]	71.90	91.57	89.51	94.42
NL-CCAM[12]	78.43	92.17	90.42	96.02
DA-WOSL[30]	83.58	93.23	88.30	95.51
CI-ResNet(ours)	**85.50**	**94.36**	**92.15**	**98.74**

表 2　各方法的定位准确率　%

模型	CUB-200-2011		AwA2	
	Top-1 Loc	Top-5 Loc	Top-1 Loc	Top-5 Loc
ADL[11]	60.29	69.25	70.84	77.42
GRCAM[27]	61.01	70.18	73.29	81.50
Rethinking-CAM[28]	60.30	69.72	72.13	80.23

续表 2

模型	CUB-200-2011		AwA2	
	Top-1 Loc	Top-5 Loc	Top-1 Loc	Top-5 Loc
SACM[29]	62.02	71.73	77.58	85.76
ACoL[10]	50.18	59.62	76.93	87.84
NL-CCAM[12]	52.40	67.47	74.95	82.48
DA-WOSL[30]	62.40	72.50	77.41	86.30
CI-ResNet(ours)	**64.51**	**73.80**	**79.73**	**88.01**

3.5 消融实验

为了更好地了解因果背景库模块的有效性,同时避免一定的偶然性,分别使用 VGG-16[32]、ResNet50[23]、InceptionV3[33]为骨干网络在 CUB-200-2011 数据集和 AwA2 数据集上进行了多次消融研究。CI-ResNet 的核心是因果背景库模块,因此在不同的骨干网络中分别比较是否包含因果背景库对于分类和定位准确率的影响。消融研究的结果见表 3。在三种不同的骨干网络中加入因果背景库模块均可以在分类和定位准确率上有所提升。综上所述,在 CUB-200-2011 数据集和 AwA2 数据集上,使用因果背景池可以同时提高分类和定位的准确性。

表 3 消融实验 %

数据集	骨干网络	是否包含因果背景库	Top-1 Cls	Top-5 Cls	Top-1 Loc	Top-5 Loc
CUB-200-2011	VGG16		74.43	91.85	57.16	70.12
		√	**78.56**	**92.03**	**63.39**	**72.54**
	ResNet50		78.62	92.56	60.79	69.58
		√	**85.50**	**94.36**	**64.51**	**73.80**
	InceptionV3		76.28	90.47	59.24	68.43
		√	**79.68**	**92.50**	**61.40**	**71.08**
AwA2	VGG16		85.83	93.01	76.10	83.56
		√	**88.37**	**94.86**	**78.45**	**87.42**
	ResNet50		89.46	95.27	75.02	86.61
		√	**92.15**	**98.74**	**79.73**	**88.01**
	InceptionV3		86.21	95.38	75.89	84.82
		√	**91.47**	**97.94**	**77.39**	**86.34**

4 结束语

传统的弱监督图像识别算法大多只考虑了识别中最具辨别力的部分造成的定位不良问题,往往忽视了目标特征与其特定背景之间的虚假关联,即该文提出的共现混淆问题,该问题在野生动物识别场景下尤为严重。针对这一问题,构建弱监督野生动物识别场景下的因果图并指出背景先验是该因果图中的一个混杂因素。通过分析因果图,尝试使用后门调整的因果干预方法来消除混杂因素,提出了基于因果推断的弱监督野生动物识别新框架 CI-ResNet,由于混杂因子的不可观测性,采用特定类别的激活图来近似构建因果背景集,然后将融合的背景重新投影到卷积层的特征映射中,实现对目标特征的因果干预。在公开数据集 CUB-200-2011 和 AwA2 上,将 CI-ResNet 和现有的弱监督识别算法进行大量对比实验和因果背景集模块的消融实验,CI-ResNet 在各方面的指标均有提升。实验结果证明了因果推断在解决共现混淆问题时的有效性。

该文仅针对的是野生动物图像中单目标识别的背景干扰问题。未来将探究使用因果干预解决多目标识别的背景干扰问题,在多目标识别的场景下,背景干扰更为复杂,混淆因子的构建将更具有挑战性。并且通过因果干预解决了野生动物场景下的共现混淆问题,后续将进一步研究如何加入虚事实推理,将因果推理从干预上升到虚事实层级。

参考文献:

[1] SHARIF M H,JIAO L,OMLIN C W. CNN-ViT supported weakly-supervised video segment level anomaly detection[J]. Sensors,2023,23(18):7734.

[2] 齐建东,马鐘添,张德怀. 基于 BS-ResNeXt-50 的密云地区野生动物图像识别[J]. 林业科学,2023,59(8):112-122.

[3] 宋宣宣. 基于深度学习的两栖动物细粒度图像识别方法研究与实现[D]. 贵州:贵州大学,2024.

[4] 杨 帆. 基于人工智能下的野生动物识别研究与应用[J]. 中国高新科技,2024,12(16):72-74.

[5] 杨拂晓,费 龙,闫泰辰. 基于深度学习的野生动物图像识别研究综述[J]. 北京测绘,2024,38(9):1237-1242.

[6] 柯　澳,王宇聪,胡博宇,等. 基于图像的野生动物检测与识别综述[J]. 计算机系统应用,2024,33(1):22-36.

[7] LAKE B M,ULLMAN T D,TENENBAUM J B,et al. Building machines that learn and think like people[J]. Behavioral and Brain Sciences,2017,40(6):1440-1448.

[8] ZHOU B,KHOSLA A,LAPEDRIZA A,et al. Learning deep features for discriminative localization[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas:IEEE,2016:2921-2929.

[9] SINGH K K,LEE Y J. Hide-and-seek:forcing a network to be meticulous for weakly-supervised object and action localization[C]//Proceedings of the international conference on computer vision. Venice:IEEE,2017:3544-3553.

[10] CINBIS R G,VERBEEK J,SCHMID C. Weakly supervised object localization with multi-fold multiple instance learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2016,39(1):189-203.

[11] CHOE J,SHIM H. Attention-based dropout layer for weakly supervised object localization[C]//Proceedings of the conference on computer vision and pattern recognition. Long Beach:IEEE,2019:2219-2228.

[12] YANG S,KIM Y,KIM Y,et al. Combinational class activation maps for weakly supervised object localization[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. Seattle:IEEE,2020:2941-2949.

[13] XUE H,LIU C,WAN F,et al. Danet:divergent activation for weakly supervised object localization[C]//Proceedings of the IEEE/CVF international conference on computer vision. Long Beach:IEEE,2019:6589-6598.

[14] RAO Y,CHEN G,LU J,et al. Counterfactual attention learning for fine-grained visual categorization and re-identification[C]//Proceedings of the IEEE/CVF international conference on computer vision. Nashville:IEEE,2021:1025-1034.

[15] PEARL J. Causal inference in statistics:an overview[J]. Statistics Surveys,2009,3(2):96-146.

[16] YAO L,CHU Z,LI S,et al. A survey on causal inference[J]. ACM Transactions on Knowledge Discovery from Data,2021,15(5):1-46.

[17] ZHANG D,ZHANG H,TANG J,et al. Causal intervention for weakly-supervised semantic segmentation[J]. Advances in Neural Information Processing Systems,2020,33(1):655-666.

[18] WANG T,HUANG J,ZHANG H,et al. Visual commonsense R-CNN[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle:IEEE,2020:10760-10770.

[19] ZHANG H,XIAO L,CAO X,et al. Multiple adverse weather conditions adaptation for object detection via causal intervention[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2022,46(3):1742-1756.

[20] PEARL J,MACKENZIE D. The book of why:the new science of cause and effect[J]. Science,2018,361(6405):854-855.

[21] KOCH B J,SAINBURG T,GERALDO BASTÍAS P,et al. A primer on deep learning for causal inference[J]. Sociological Methods & Research,2024,25(6):471.

[22] YANG X,ZHANG H,QI G,et al. Causal attention for vision-language tasks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Nashville:IEEE,2021:9847-9857.

[23] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas:IEEE,2016:770-778.

[24] SELVARAJU R R,COGSWELL M,DAS A,et al. Grad-cam:visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE international conference on computer vision. Venice:IEEE,2017:618-626.

[25] WAH C,BRANSON S,WELINDER P,et al. The caltech-ucsd birds200-2011 dataset[R]. California Institute of Technology,2011.

[26] XIAN Yongqin,LAMPERT C H,BERNT S,et al. Zero-shot learning - a comprehensive evaluation of the good,the bad and the ugly[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2017,41(9):2251-2262.

[27] KINGA D,ADAM J B. A method for stochastic optimization[C]//International conference on learning representations(ICLR). San Diego:Ithaca,2015:5-6.

[28] BAE W,NOH J,KIM G. Rethinking class activation mapping for weakly supervised object localization[C]//Computer vision - ECCV 2020:16th European conference. Heidelberg:Springer,2020:618-634.

[29] HUI W,TAN C,GU G,et al. Gradient-based refined class activation map for weakly supervised object localization[J]. Pattern Recognition,2022,128(11):5-8.

[30] 吴朝捷. 基于背景抑制擦除与伪监督的弱监督目标定位方法研究[D]. 广州:华南理工大学,2023.

[31] ZHU L,SHE Q,CHEN Q,et al. Boosting weakly supervised object localization and segmentation with domain adaption[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2024,46(12):8680-8695.

[32] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv:1409.1556,2014.

[33] SZEGEDY C,VANHOUCKE V,IOFFE S,et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the conference on computer vision and pattern recognition. Las Vegas:IEEE,2016:2818-2826.