

基于干预注意力的细粒度图像识别

陈建锟, 王永雄, 潘志群

(上海理工大学 光电信息与计算机工程学院, 上海 200093)

摘要: 注意力机制在细粒度识别任务中具有关键作用。为了让模型更加关注判别性区域, 提出一种新的基于干预注意力的方法, 为监督注意力学习特征提供关键线索。具体地, 在训练过程中加入干预注意力, 并将注意力应用于数据裁剪和擦除过程, 进而提高模型的学习效率。同时, 将融合注意力应用于特征提取网络, 帮助网络学习更加有效的特征。此外, 引入标签平滑损失函数以及中心正则化损失函数, 有效地提升分类精度。实验表明, 提出的方法具有优异性能, 分别在 CUB-200-2011、Stanford Cars 和 FGVC Aircraft 数据集上实现了 89.8%、95.7% 和 94.7% 的分类准确率, 相比多个细粒度分类算法具有更好的分类效果。

关键词: 细粒度图像识别; 干预注意力; 数据增强; 融合注意力; 标签平滑

中图分类号: TP 391.4 **文献标志码:** A

Fine-grained image recognition based on interventional attention

CHEN Jiankun, WANG Yongxiong, PAN Zhiquan

(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Attention plays a key role in fine-grained image recognition tasks. In order to make the model pay more attention to discriminative regions, a new method based on interventional attention was proposed to provide key clues for supervising attention to learn features. Specifically, the interventional attention was added to the training process, and the attention mechanism was applied to the process of data cutting and dropping to guide the model to improve the learning efficiency. At the same time, the fused attention was applied to the feature extraction network to help the network learn more discriminable features. In addition, label smoothing loss function and center regularization loss function were introduced into the objective function, which effectively improved classification accuracy.

收稿日期: 2024-01-08

基金项目: 上海市自然科学基金资助项目(22ZR1443700)

第一作者: 陈建锟(1998—), 男, 硕士研究生。研究方向: 计算机视觉、人工智能。E-mail: 3159833875@qq.com

通信作者: 王永雄(1970—), 男, 教授。研究方向: 计算机视觉、人工智能。E-mail: wyxiong@usst.edu.cn

引文格式: 陈建锟, 王永雄, 潘志群. 基于干预注意力的细粒度图像识别[J]. 上海理工大学学报, 2025, 47(2): 209-219.

Citation: CHEN Jiankun, WANG Yongxiong, PAN Zhiquan. Fine-grained image recognition based on interventional attention[J]. Journal of University of Shanghai for Science and Technology, 2025, 47(2): 209-219.

Experimental results show that the proposed method has excellent performance, achieving 89.8%, 95.7% and 94.7% classification accuracy on CUB-200-2011, Stanford Cars and FGVC Aircraft dataset respectively. In comparison with the other mainstream fine-grained classification algorithms, the proposed method achieves better classification results.

Keywords: *fine-grained image recognition; interventional attention; data augmentation; fused attention; label smoothing*

近几年，细粒度图像识别^[1]一直是研究热点和难点之一。细粒度图像识别是子类的识别，类别间的差异经常很小，尤其在图像中光照、背景、姿态等因素的干扰下，细粒度识别具有更大的挑战性。早期方法^[2]常以部件标注和局部信息区域为代价获取关键局部信息，但这种方式常需要昂贵的人工与数据标注成本。因此，仅利用分类标签，自动获取图像判别性区域就显得十分重要。

注意力机制是计算机系统模拟人类的搜索识别方式，找到图像中的有辨识性区域，并减少由于不同的视觉外观而造成的杂乱背景、遮挡、姿势变化等。尽管目前注意力机制的应用十分广泛，但是如何学习更加有效的注意力仍然是一个值得研究的问题。传统的基于概率的方法^[3]仅通过损失函数监督最终预测，却常常忽略预测和注意力之间的因果关系。与此同时，数据增强^[4-5]是一种常用的增加数据量及多样性的策略。以往的细粒度识别中通常采用随机数据增强的方法对训练数据进行预处理。然而，由于裁剪区域是随机采样且大部分样本包含大量背景噪声，这些因素会降低训练效率并影响提取特征的质量。

基于上述问题，本文研究如何利用注意力学习更加有效的特征提取，并利用注意力机制引导数据裁剪和擦除的数据增强过程。训练过程仅依靠类别标签，就能够有效提升注意力学习的质量，准确获取图像关键性区域并取得较为理想的分类精度。该方法的优势如下：

- a. 采用干预注意力学习方法加强注意力引导学习的因果推理过程，利用注意力引导数据增强。同时，将通道注意力机制与空间注意力机制进行融合，帮助网络提取关键性区域，有效改善注意力机制性能，计算效率较高。
- b. 采用标签平滑方法降低模型过拟合，并利用中心正则化损失矫正注意力区域的位置，从而

增强对关键性区域的关注能力，加强注意力模块的提取特征能力。

c. 在 3 个经典细粒度图像分类数据集 CUB-200-2011^[6]、Stanford Cars^[7]和 FGVC Aircraft^[8]上进行实验，结果表明本文方法具有较明显的性能提升。

1 相关工作

1.1 细粒度图像识别

在细粒度识别中，现有大多数方法都采用带注意力机制的深度学习网络。Liu 等^[9]提出了全卷积注意力网络，采用强化学习方法获得视觉注意力，更好地适应了不同细粒度域的局部判别区域。Lin 等^[10]提出了双流架构的 B-CNN，有效地将图像表示为两个 CNN 特征的池化外积架构。Wang 等^[11]进一步提出基于双线性卷积神经网络的包含自聚焦模块和稀疏缩放因子的稀疏聚焦框架，其中，每个通道的稀疏缩放因子用来评估特征通道的重要性，并将其用于通道剪枝。谭润等^[12]提出，利用双线性注意力池化和卷积块注意模块构建注意力学习模块和信息增益模块，分别获取目标局部细节信息和目标重要轮廓的数据。

Chen 等^[13]提出了一种图像分割操作作用于细粒度分类，将图像分割成局部补丁，随机混洗重建新图像进行训练，从而迫使模型关注局部细节信息而非全局信息。Li 等^[14]提出一种多分支通道增强网络，对原始图像进行多模式打乱操作，然后将不同混淆程度的图像组合成 3 对信息作为输入，利用打乱操作迫使通道寻找判别性区域。Du 等^[15]进一步应用多尺度拼图生成器，以渐进式训练策略捕获交叉粒度信息。还有 Zheng 等^[16]和 Sun 等^[17]也提出基于注意力的细粒度分类方法，并以自下而上的方式设计注意力模型，在细粒度识别任务中取得了较为优异的结果。

1.2 数据增强

数据增强是对输入图像的预处理, 常用的方式为对图像进行空间水平的变换。在以往的方法^[4-5, 18]中, 都采用基于图像裁剪和图像擦除的数据增强方式, 包含对图像进行几何变换的数据增强操作。这对于提升深度学习模型的鲁棒性, 避免过拟合及提升模型的泛化能力都具有积极作用。

Devries 等^[5]和 Singh 等^[19]的思路都是随机采用一些方形区域对原始输入图像进行一定掩蔽, 以增强卷积神经网络的鲁棒性。但是, 传统数据裁剪和擦除的区域具有一定的随机性, 常常会将需要检测的目标或者关键性区域遮蔽, 从而影响增强数据的可利用性。同时, 随机的掩码过程容易产生不受控制的噪声数据, 影响增强数据的区域分布, 降低增强数据的质量。

基于上述问题, Gong 等^[20]提出不同于随机选取掩码的增强方式, 该方法先取得图像的显著性图, 再从中分析取得图像区域的重要程度, 从而选取相应区域进行图像裁剪或擦除操作。Cubuk 等^[21]通过设计一个特定的方式, 创建数据增强方式的搜索空间, 以获得优秀的目标数据集的精度。但是这类方法相比而言成本大, 需要时间比较长, 产生的增强数据较为复杂。本文采用的注意力引导的数据裁剪和擦除方式更为简单, 可生成更优质的增强数据以提升性能。

1.3 干预推理

以深度学习为基础的方法在特征提取能力方面有着比较优秀的效果, 但是深度神经网络在训练过程中有一定的不确定性。为了能够揭开黑箱, 理解事实的逻辑因果, 并获得真正的因果影响因素, 需要将学习到的伪逻辑关系进行解耦合。因此, 将因果推理理论应用到深度学习模型中的分析方法就显得至关重要。该方法有助于揭开黑箱, 减少对于伪逻辑关系的错误学习, 有效提升模型的泛化能力及鲁棒性。

在图像分类任务中, 常使用因果关系作为工具降低偏差数据的影响。由因果推断理论得知, 混杂因子造成的伪逻辑关系对提升模型在测试集上的准确性具有不利影响。为找到训练过程中的混杂因子, 增强学习稳定性, 减少训练集中伪逻辑关系的不利影响, 可以利用干预推理方法进行训练。

干预推理^[22]是指通过对某个事件发生的某一

先决条件进行确定性假设, 然后通过该假设对事件发生结果进行重新表征, 从而得到以此为先决条件的推理结果。具体而言, 干预操作方法通过确定一个变量的值, 限制该变量随其他自然变量变化的自然趋势。在图模型中, 干预操作会删除所有指向该变量的边。以此为依据, 得到干预条件下的推理结果, 并将其与原推理结果进行对比, 得出最后结果。基于因果推理中的干预理论, 对网络学习到的视觉注意力进行干预操作, 即为干预注意力学习方法。

2 算法原理

2.1 网络结构

细粒度图像分类研究的主要问题是关键性区域的提取与学习能力的加强, 所以训练过程中常需要采取一些方式加强网络对细节的关注。一是增强网络对区分性特征的提取能力, 从而使最终提取特征更加有效; 二是利用注意力机制将网络学习的重点放在更加需要关注的地方, 并提升目标关键区域的模型表征能力。网络训练流程如图 1 所示。

在特征提取阶段, 首先, 利用融合注意力机制对特征提取网络所得到的特征图进行通道和空间上的加强, 得到相应特征图。然后, 将所得特征图输入卷积层获得自然注意力图, 该注意力图可以显示目标的局部显著特征。同时, 采用均匀分布注意力为干预手段, 产生与自然注意力图同一维度的干预注意力图。最后, 将特征图分别与自然注意力图及干预注意力图进行双线性注意力池化, 该操作具体原理如图 2 所示。

具体地, 将特征图分别与 n 张自然注意力图、干预注意力图进行按位相乘的操作, 从而得到特征融合后的特征图 $F_1, F_2, \dots, F_n$ 及干预特征图 $F'_1, F'_2, \dots, F'_n$ 。然后, 对所得特征图进行全局平均池化, 获得相应特征向量 $G_1, G_2, \dots, G_n$ 及相应干预特征向量 $G'_1, G'_2, \dots, G'_n$ 。最后, 将这些特征进行拼接从而得到相应特征矩阵, 再通过线性层得到相应注意力特征输出向量 pp 与 pp' 。最终, 由自然注意力特征输出向量 pp 减掉干预注意力特征输出向量 pp' 得到预测结果向量 p 。

同时, 通过利用注意力裁剪和注意力擦除法对图像进行数据增强, 并将增强图像与原始图

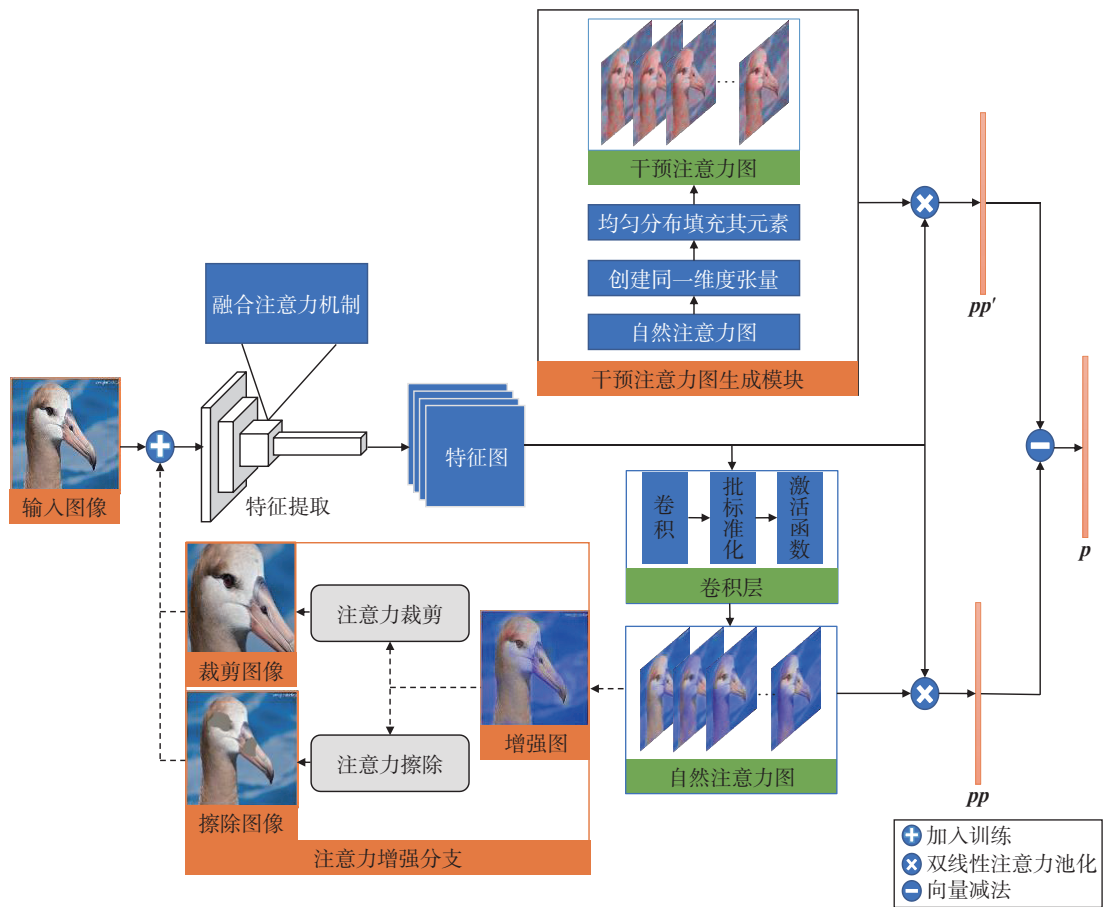


图1 网络训练流程图

Fig.1 Flow chart of the network training

像一起输入网络中，有效增强模型对于待分类目标局部显著特征的关注。对于每一幅训练图像，注意力增强分支将随机选择一张自然注意力图作为增强图来指导数据增强过程。该分支通过选择其中显著特征部分将其擦除得到擦除图像，同时将相关显著特征部分剪裁下来得到裁剪图像，进而帮助后续网络提取更加显著的特征部分。

2.2 融合注意力

传统通道注意力模块常采用 SENet^[23]，该网络的基本原理是，通过获取不同特征图的关键程度，为学习特征图赋予不同权重值，迫使网络对不同特征图采用不同的关注程度。但该网络的全连接层会获得所有通道的依赖关系，这会增加计算量并对预测结果产生负面影响。本文方法将空间注意力机制与通道注意力机制加以融合。具体地，将通道注意力网络中的全连接层改为 1×1 卷积层，并通过动态选取卷积核，保持通道特征维度一致，从而有效减少参数量。为了提升特征提取网络对有辨别性区域的关注，将通道注意力机

制中单一的全局平均池化 (global average pooling, GAP)改成采用全局最大池化和全局平均池化融合的方式，将所得结果相加并进行激活得到输出结果，将所得结果通过激活函数得到通道注意力特征。将该特征输入空间注意力网络中：首先，将其沿通道维度进行最大值池化和平均值池化操作；然后，将池化结果进行拼接并进行卷积操作；最后，将所得结果通过激活函数得到输出结果。该模块通过嵌入到总框架中特征提取模块的骨干网络中发挥作用。由于融合注意力模块具有即插即用的特性，因此，以本文实验用到的 ResNet 骨干模型为例，可将该模块插入到 ResNet 每一个残差网络的最后输出位置，从而增强骨干网络的特征提取效率与质量。融合注意力模型如图 3 所示。

2.3 干预注意力

深度学习训练算法常通过最小化损失函数进行参数优化，这种方法将训练过程看作黑箱，忽略网络内部的训练结果，不利于对内部变量进行

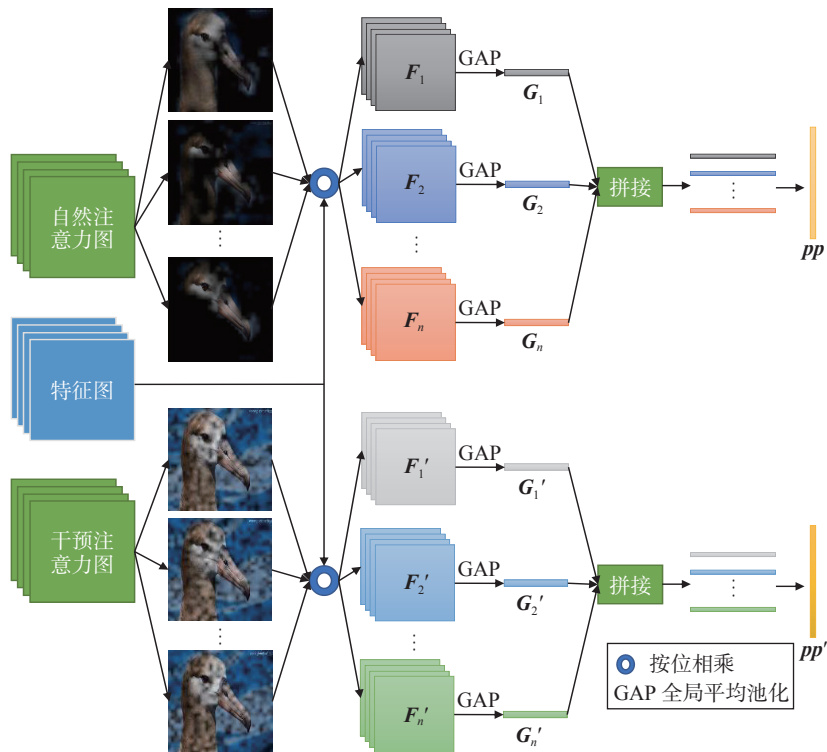


图 2 双线性注意力池化原理图

Fig.2 Schematic diagram of bilinear attention pooling

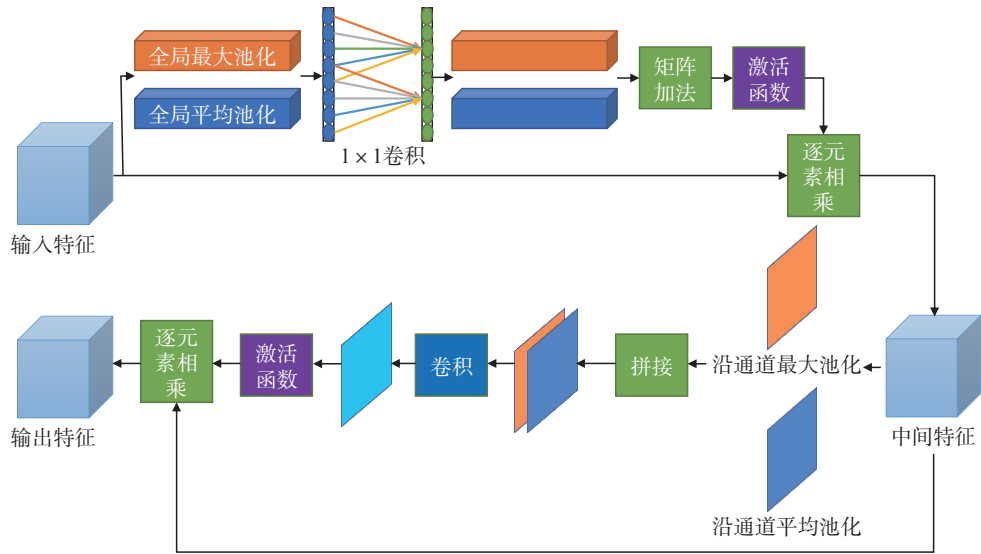


图 3 提出的融合注意力模块

Fig.3 Proposed fusion attention module

直观控制。干预注意力学习将因果推理理论引入到注意力机制中,通过直接控制内部变量使模型学习到更加有效的注意力图,从而实现注意力质量的提升。

在干预注意力学习中,通过引入因果图的方式,操作内部变量的值分析因果关系,并且查看其效果,该操作在常见的因果推理文献中被称为

干预,用公式可以表示为 $\text{do}(\cdot)$ 。通过一个现实中的干预机制调整各变量的状态,从而达到干预的目的。利用预先设计的干预方法,观察自然注意力和干预注意力对预测结果的影响,促使网络更好地学习有效的注意力分布,并抑制数据集中的混杂因素对训练的负面影响,增强深度模型的可解释性。同时,干预注意力帮助模型理解不同注

注意力区域对预测结果的影响关系,从而更好地学习图像关键区域的规律和特征,并揭示图像中的潜在信息和重要特征。干预注意力学习的原理如图4所示。

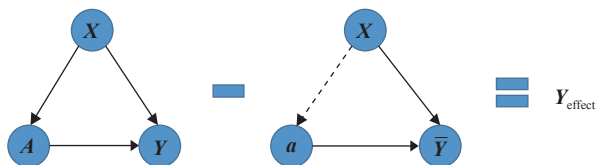


图4 干预注意力原理图

Fig.4 Schematic diagram of interventional attention

图4中: X 代表特征图; A 代表自然注意力图; a 代表干预注意力图; Y 代表自然注意力预测; $\bar{Y}$ 代表干预注意力预测; Y_{effect} 代表注意力影响输出预测。

在实验过程中,进行干预 $\text{do}(A = a)$,需先在网络中生成一个干预注意力图 a ,从而替换学习到的注意力图。注意力图选取要具有真实合理性,要在合理的选择范围内,即干预注意力图的维度大小应该与自然注意力图 A 保持一致,从而保证干预注意力的真实性。在实际操作中,可应用均匀分布注意力或随机分布注意力当作干预方式。干预过程需保持特征图 X 不变,并保证固定干预注意力图 a ,目的是断开从特征图生成注意力图因果推理支路,从而完成干预注意力操作。根据注意力机制的公式,完成干预后,得到干预注意力预测 $\bar{Y}$,其数学表达式为

$$\bar{Y} = \bar{Y}(\text{do}(A = a), X = X) = \mathbb{C}([\phi(X * a_1), \dots, \phi(X * a_n)]) \quad (1)$$

式中: $\mathbb{C}$ 表示分类器; $a_1, \dots, a_n$ 表示 $1 \sim n$ 张注意力干预图; ϕ 表示全局平均池化; $*$ 表示按位相乘,这就是干预注意力学习的基本思路。学习到的注意力影响输出预测 Y_{effect} 可用 $Y(A = A, X = X)$ 和 $\bar{Y}(\text{do}(A = a), X = X)$ 之差的期望表示,其数学表达式为

$$Y_{\text{effect}} = E[Y(A = A, X = X) - \bar{Y}(\text{do}(A = a), X = X)] \quad (2)$$

式中, E 表示期望。

在细粒度图像分类任务中,复杂的背景、遮挡以及姿态变化往往会对网络的分类结果造成巨大影响。将干预理论引入注意力机制中,可以有效提升干预对象的学习质量,将网络学习重点放在更加有区分度的目标上,同时减少网络学习冗余信息,有助于增强模型训练的鲁棒性,并有效

提升模型泛化能力。与现有方法相比,干预注意力有效提升了注意力学习的质量,有助于解决细粒度分类任务中需要更好关注图像中细微差异的问题,从而提升了网络对于图像中关键细节的分辨能力,进而有效提升其在细粒度图像分类问题上的分类准确率。

2.4 损失函数

当遇到多类分类问题时,预测正确类别标签通常为1,其余位置为0。但这种硬标签的预测策略会使网络训练结果过于武断,不利于训练过程中对噪声抵抗能力以及网络泛化能力的提升。当引入标签平滑策略时,模型不要求正确类别预测为1,而仅要求正确类别预测为 $1 - \varepsilon$,并将其他类别的预测标签设为 ε ,其中, ε 表示与1相比极小的正实数。对于细粒度图像分类任务而言,区分与表达类别内差异是其要解决的关键性问题。因此,在损失函数中引入中心正则化损失函数以缩短类内距离,使其更多反应和表达类内相似性。标签平滑交叉熵损失函数 $L_{\text{ce}}(Y_{\text{effect}}, y')$ 和中心正则化损失函数 L_{center} 表示为

$$L_{\text{ce}}(Y_{\text{effect}}, y') = (1 - \varepsilon)L_{\text{ce}}(Y_{\text{effect}}, y) + \varepsilon L_{\text{ce}}(Y_{\text{effect}}, u) \quad (3)$$

$$L_{\text{center}} = \frac{1}{2} \sum_{i=1}^m \sum_{n=1}^K \|g_n(I_i) - c_n(y_i)\|_2^2 \quad (4)$$

式中: y' 表示平滑分类标签; y 表示分类标签; $L_{\text{ce}}(\cdot)$ 表示标准交叉熵的损失函数; u 表示均匀分布; m 表示图像数量; $g_n(I_i)$ 表示第 i 张图像 I_i 的局部响应特征; $c_n(y_i)$ 表示第 i 张图像分类标签 y_i 的部位特征中心; K 表示自然注意力图通道数量,其取值应兼顾网络训练时精度与速度的要求。若 K 取值过小,则难以达到训练精度要求;相反,若其取值过大,则会过度消耗计算成本,且对训练精度的提升效果有限。经验证, K 为32时更能满足相关要求。同时,将注意力质量当作一个监督信号,指导注意力的学习过程。所以总的目标函数为

$$L = L_{\text{ce}}(Y_{\text{effect}}, y') + \alpha L_{\text{center}} \quad (5)$$

式中, α 是两种损失函数的线性组合系数。

通过优化新的目标函数,期望注意力发现最具判别性的区域,避免次优结果。其次,希望利用干预注意力监督预测结果,迫使分类器根据主要线索来作决定,而非偏见信息,降低训练集当

中偏见成分的影响。与现有方法相比, 该损失函数的设计能够达到更好地修正卷积神经网络参数权重的目的, 使提取的图像特征得以更有效地进行图像分类。

3 实验及结果分析

为了评估模型的有效性, 本文将提出的学习方法在几个常用的细粒度图像分类数据集上进行实验, 通过消融实验验证所加模块对于基线模型准确度的提升效果, 并与其他先进的方法进行比较, 验证方法的有效性。下面描述具体实验数据集、实施设置细节及实验结果。

3.1 数据集

实验采用 3 个常用的经典公开细粒度图像数据集, 分别是 CUB-200-2011、Stanford Cars 和 FGVC Aircraft, 3 个数据集的类别数量、训练集和测试集的图像数量如表 1 所示。

表 1 数据集信息
Tab.1 Information of datasets

数据集	类别数量	训练集图像数量	测试集图像数量
CUB-200-2011	200	5 994	5 794
Stanford Cars	196	8 144	8 041
FGVC Aircraft	100	6 667	3 333

3.2 实验设置

针对每一张训练图像, 利用本文数据增强方法将训练数据集扩充 3 倍, 本文代码基于 Pytorch 深度学习框架实现, 以 ResNet 系列和 VGG19 作为骨干模型, 同时采用注意力机制加强特征提取能力。注意图是在 1×1 卷积后得到的。在 GPU 上使用动量为 0.9, epoch 数为 160, 权重衰减为 0.000 01, 最小批量大小为 4 的随机梯度下降 (stochastic gradient descent, SGD) 来训练模型。将初始学习速率设置为 0.001, 在每 2 个 epoch 之后进行 0.9 的指数衰减。其中, 在 CUB-200-2011 数据集上以 VGG19 为骨干网络的对比实验中采用的最小批量大小为 8。为了有效兼顾网络训练速度与精度要求, 使其更好展现所提方法的有效性及合理性, 本文参数选取考虑了模型的复杂度、数据集的大小和分布, 并评估了模型的准确性、泛化能力和稳定性, 同时考虑了可用的计算资源, 从而确定此参数选取策略。本文采用分类准确度检

验细粒度图像分类模型的训练效果, 该性能指标可直观反映模型的分类性能。分类准确率公式为

$$Z = \frac{T}{W} \times 100\% \tag{6}$$

式中: T 表示所有类别中被正确分类的图像样本数量; W 代表总的样本数量。

3.3 对比实验

在 CUB-200-2011、Stanford Cars 和 FGVC Aircraft 数据集上, 与其他细粒度分类方法进行比较评估, 实验结果如表 2 所示。本文方法使用相同主干网络, 在 3 个广泛使用的数据集集中进行测试, 与同类别先进方法相比, 表现出约 1% 的性能提升。可以观察到, 随着骨干网络层数加深, 本文方法的分类准确率在 3 个数据集上都有 1%~4% 的提升, 说明骨干网络特征提取能力的优劣对于分类准确率有一定影响。同时, 比较本文方法与其他方法在应用不同层数骨干网络时在 3 个数据集上的分类表现, 可以发现, 本文方法在不同数据集上的性能提升程度存在一定区别, 这是由于不同数据集的数据分布差异导致的。例如, 在 Stanford Cars 数据集上, 本文方法在应用较浅层的 ResNet34 作为骨干网络时表现出相对更优的性能。这是由于该数据集的类内差异更小, 通过所

表 2 不同方法的性能对比

Tab.2 Performance comparison of different methods

方法	骨干模型	分类准确率 (%)		
		CUB-200-2011	Stanford Cars	FGVC Aircraft
RA-CNN ^[24]	VGG19	85.3	92.5	88.2
MA-CNN ^[16]		86.5	92.8	89.9
本文方法		87.2	92.9	90.6
BRAN ^[25]	ResNet34	87.2	92.5	
SnapMix ^[26]		87.1	94.0	92.4
本文方法		87.8	95.0	93.1
MAMC ^[17]	ResNet50	86.2	92.8	
DFL-CNN ^[27]		87.4	93.1	91.7
NTS-Net ^[28]	ResNet50	87.5	93.9	91.4
DCL ^[13]		87.8	94.5	93.0
Cross-X ^[29]	ResNet50	87.7	94.6	92.6
S3N ^[30]		88.5	94.7	92.8
ASD ^[31]	ResNet50	88.6	94.9	93.5
本文方法		88.9	95.3	93.9
API-Net ^[32]	ResNet101	88.6	94.9	93.4
本文方法		89.8	95.7	94.7

提出的干预注意力机制等方法可以关注判别性区域，使网络在训练过程中更多学习到有助于分类的特征，从而在该数据集上得到更高的分类准确率。

3.4 消融实验

为检验本文方法关键组件的有效性，在 CUB-200-2011 数据集上利用 5 组消融实验，验证干预推理机制、融合注意力增强和标签平滑损失函数的使用对实验结果的影响。最终结果如表 3 所示。实验 1 中，基线模型采用 ResNet50 作为特征提取网络，同时应用注意力引导数据增强方法，最终预测结果由最后的全连接层获得。实验 2 相比基线模型，在特征提取网络中加入融合注意力分支，其准确率提升约 0.5%，证明该分支加强了网络对特征的提取能力。

表 3 消融实验
Tab.3 Ablation experiment

实验编号	基线模型	融合注意力	干预注意力	标签平滑	分类准确率/%
1	√				87.1
2	√	√			87.6
3	√		√		87.3
4	√			√	88.7
5	√	√	√		88.1

实验 3 和实验 5 分别在实验 1 和实验 2 的基础上加入干预注意力分支，其性能分别提升 0.2% 和 0.5%，证明该分支对于提升注意力学习的质量及模型准确率具有一定的效果，且当特征提取网络的性能提升时，该分支的性能表现越明显。实验 4 在实验 1 的基础上应用标签平滑方法，其分类准确率相较基线模型提升约 1.6%，证明该方法的应用对于减少过拟合，提高模型的置信度有积极作用。仅利用标签平滑对分类准确率的提升效果十分明显，但并不意味着融合注意力与干预注意力对于模型训练的重要程度不高，可能仅在特定数据集与参数配置情况下呈现出这样的现象，也从侧面证明所提各方法对训练都有较好效果。

3.5 损失函数的参数评估

为检验本文方法的中心正则化损失函数对模型的性能影响，对式 (5) 中线性组合系数 α 进行参数评估。实验在 Stanford Cars 数据集上进行，骨干模型采用 ResNet50，最终对比结果如图 5 所示。通过观察可得，在中心正则化损失函数的线性组合系数加大到 2 的情况下，分类准确率提升约

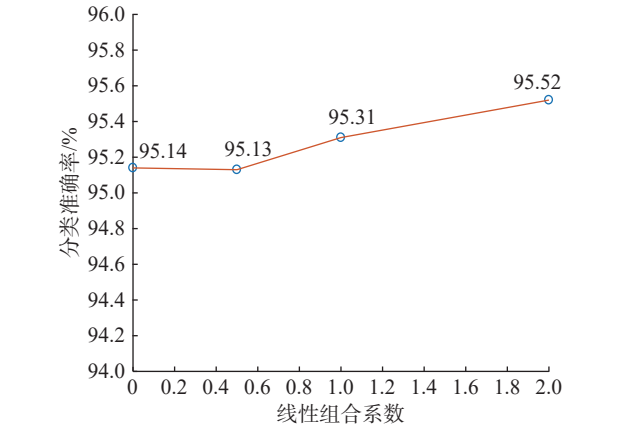


图 5 不同线性组合系数取值下的分类准确率折线图

Fig.5 The line graph of classification accuracy for different linear combination coefficient values

0.4%，其原因在于中心正则化损失将目标各个部位的位置误差计入损失函数，通过矫正各部位的位置信息使得目标定位更精准，加强了提取特征的准确度，从而提升了模型的分类性能。

3.6 数据可视化

为直观展现注意力增强分支的裁剪与擦除图像的应用效果，并展现提出的干预注意力引导的数据增强网络的图像注意力关注区域，对于输入图像的自然注意力图与干预注意力图及本文方法的裁剪图像与擦除图像及其对应的注意力图采用可视化方法。实验在 CUB-200-2011 鸟类公开数据集上进行，骨干模型采用 ResNet50，可视化热力图如图 6 所示。

观察自然注意力图可知，网络更多地将关注集中在鸟的头部和身体等关键性识别区域，这有助于网络学习到更关键的区分区域，可以帮助其理解在处理任务时的关注重点，从而使模型的优化和改进方向更为明确，有助于提升模型性能。观察干预注意力图发现，其关注区域较为分散，其作用在于帮助模型更好地学习图像的规律和特征，去除网络学习的冗余信息，增强注意力学习质量，提升模型的泛化能力和鲁棒性。观察裁剪图像及其注意力图可知，裁剪图像范围更多集中在对分类更有效的目标区域，在注意力图中可以看到网络能够从中找到更重要的特征，将其加入训练过程，有助于网络学习更加细微的特征。观察擦除图像及其注意力图，可以发现擦除图像将图像中比较有辨识性的区域进行擦除，从而促使模型学习到某些更为隐藏性的特征，加强对图像中区分性特征学习的全面性。注意力引导的裁剪与擦除技术可以有效地防止模型过度依赖数据中

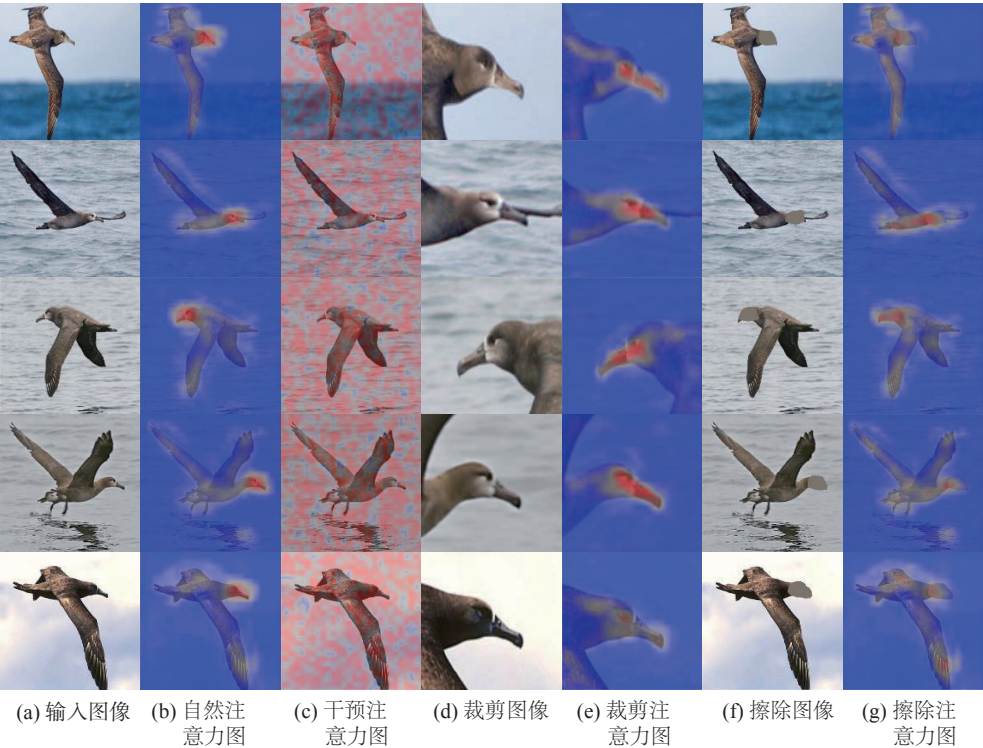


图 6 输入图像及其增强图像可视化展示

Fig.6 Visualization of the input images and their enhanced images

的局部特征，促使模型学习更加全局和抽象的特征表示，从而提升模型的性能。

3.7 可视化图对比

为有效对比本文方法与基线方法，将两种方法在网络关注区域的注意力图进行可视化，实验在 CUB-200-2011 鸟类公开数据集上进行，可视化热力图对比如图 7 所示。本文模型采用以 ResNet50

为骨干模型结合注意力引导数据增强的方法，与基线方法进行比对，可观察到本文方法的注意力图将关注区域更多聚焦到检测目标中，同时将关注区域更分散地体现于目标区域中，可有效降低网络对某些关键性区域的过分依赖，减少复杂背景对检测准确度的影响。最终，该方法帮助网络准确地监测目标特征，取得更加有效的细节区域，为提升网络分类精度提供重要的帮助。

4 结 论

提出了一种新的干预注意力引导的数据增强方法，在细粒度图像分类任务中，干预注意力能够有效引导网络学习图像关键判别性特征，挖掘数据间关联性，提高模型的学习效率和泛化能力，并为人类理解网络训练过程提供良好的实验示例。设计的标签平滑损失函数以及中心正则化损失函数能够有效提升模型分类精度。最终实验结果表明，本文方法相较其他细粒度分类方法，在分类准确率上具有一定提升，但本文方法在训练速度上还存在一定的不足。今后的研究可考虑一些更加轻量化的方法，通过减少参数量，使训练速度得到一定程度的提升。可尝试结合更高级

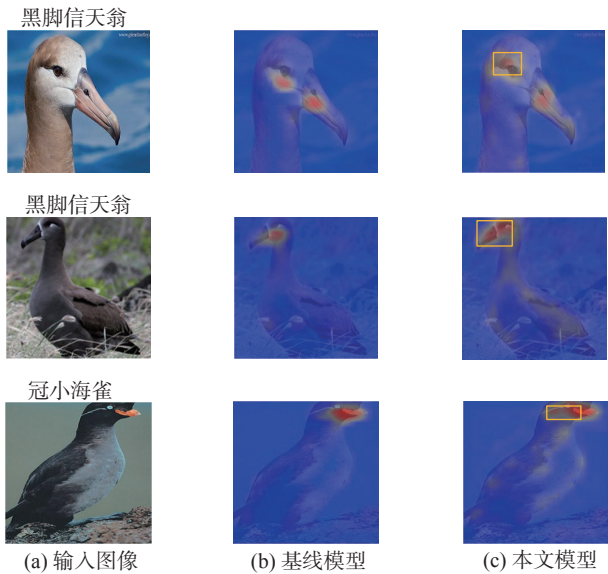


图 7 本文模型与基线模型可视化图展示

Fig.7 Visualization of the proposed model and the baseline model

的可解释性模型,以提高模型解释的准确性和可信度。另外,该方法可用于增强注意力学习质量,降低数据集偏差带来的影响。将其应用于各种细粒度图像识别任务,具有良好的泛化性。

参考文献:

- [1] 白尚旺,王梦瑶,胡静,等.多区域注意力的细粒度图像分类网络[J].计算机工程,2024,50(1):271–278.
- [2] ZHANG N, DONAHUE J, GIRSHICK R, et al. Part-based R-CNNs for fine-grained category detection[C]//Proceedings of the 13th European Conference on Computer Vision. Zurich: Springer, 2014: 834–849.
- [3] CHEN G Y, LU J W, YANG M, et al. Learning recurrent 3D attention for video-based person re-identification[J]. *IEEE Transactions on Image Processing*, 2020, 29: 6963–6976.
- [4] PARK S, KWAK N. Analysis on the dropout effect in convolutional neural networks[C]//Proceedings of the 13th Asian Conference on Computer Vision. Taipei, China: Springer, 2017: 189–204.
- [5] DEVRIES T, TAYLOR G W. Improved regularization of convolutional neural networks with cutout [DB/OL]. (2017-08-15) [2024-01-08]. <https://arxiv.org/abs/1708.04552>.
- [6] WAH C, BRANSON S, WELINDER P, et al. The Caltech-UCSD birds-200-2011 dataset[R]. Pasadena: California Institute of Technology, 2011.
- [7] KRAUSE J, STARK M, DENG J, et al. 3D object representations for fine-grained categorization[C]//Proceedings of 2013 IEEE International Conference on Computer Vision Workshops. Sydney: IEEE, 2013: 554–561.
- [8] MAJI S, RAHTU E, KANNALA J, et al. Fine-grained visual classification of aircraft [DB/OL]. (2013-06-21) [2024-01-08]. <https://arxiv.org/abs/1306.5151>.
- [9] LIU X, XIA T, WANG J, et al. Fully convolutional attention networks for fine-grained recognition [DB/OL]. (2016-03-22) [2024-01-08]. <https://arxiv.org/abs/1603.06765>.
- [10] LIN T Y, ROYCHOWDHURY A, MAJI S. Bilinear CNN models for fine-grained visual recognition[C]//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 1449–1457.
- [11] WANG Y X, LI G J, MA L. A sparse focus framework for visual fine-grained classification[J]. *Multimedia Tools and Applications*, 2021, 80(16): 25271–25289.
- [12] 谭润,叶武剑,刘怡俊.结合双语义数据增强与目标定位的细粒度图像分类[J].计算机工程,2022,48(2): 237–242,249.
- [13] CHEN Y, BAI Y L, ZHANG W, et al. Destruction and construction learning for fine-grained image recognition[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 5152–5161.
- [14] LI G J, WANG Y X, ZHU F T. Multi-branch channel-wise enhancement network for fine-grained visual recognition[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 5273–5280.
- [15] DU R Y, CHANG D L, BHUNIA A K, et al. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches[C]//Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020: 153–168.
- [16] ZHENG H L, FU J L, MEI T, et al. Learning multi-attention convolutional neural network for fine-grained image recognition[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 5219–5227.
- [17] SUN M, YUAN Y C, ZHOU F, et al. Multi-attention multi-class constraint for fine-grained image recognition [C]//Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018: 834–850.
- [18] 江雨燕,陶承凤,李平.数据增强和自适应自步学习的深度子空间聚类算法[J].计算机工程,2023,49(8): 96–103,110.
- [19] 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 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 3544–3553.
- [20] GONG C Y, WANG D L, LI M, et al. KeepAugment: a simple information-preserving data augmentation approach[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 1055–1064.
- [21] CUBUK E D, ZOPH B, MANÉ D, et al. AutoAugment: learning augmentation strategies from data[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 113–123.
- [22] NAIMI A I, Tyler J VanderWeele. Explanation in causal inference: methods for mediation and interaction[J]. *European Journal of Epidemiology*, 2016, 31(10): 1065–1066.
- [23] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake

- City: IEEE, 2018: 7132–7141.
- [24] FU J L, ZHENG H L, MEI T. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 4476–4484.
- [25] 王阳, 刘立波. 面向细粒度图像分类的双线性残差注意力网络 [J]. 激光与光电子学进展, 2020, 57(12): 121011.
- [26] HUANG S L, WANG X C, TAO D C. SnapMix: semantically proportional mixing for augmenting fine-grained data[C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2021: 1628–1636.
- [27] WANG Y M, MORARIU V I, DAVIS L S. Learning a discriminative filter bank within a CNN for fine-grained recognition[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4148–4157.
- [28] YANG Z, LUO T G, WANG D, et al. Learning to navigate for fine-grained classification[C]//Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018: 438–454.
- [29] LUO W, YANG X T, MO X J, et al. Cross-X learning for fine-grained visual categorization[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 8241–8250.
- [30] DING Y, ZHOU Y Z, ZHU Y, et al. Selective sparse sampling for fine-grained image recognition[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 6598–6607.
- [31] SUN G L, CHOLAKKAL H, KHAN S, et al. Fine-grained recognition: accounting for subtle differences between similar classes[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2020: 12047–12054.
- [32] ZHUANG P Q, WANG Y L, QIAO Y. Learning attentive pairwise interaction for fine-grained classification[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2020: 13130–13137.

(编辑: 董 伟)