计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(10):232−239 [doi: 10.15888/j.cnki.csa.008141]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

# 基于注意力机制的弱监督细粒度图像分类①

李文书, 王志骁, 李绅皓, 赵　朋

(浙江理工大学 信息学院, 杭州 310018)
通讯作者: 李文书, E-mail: charlie@zstu.edu.cn

**摘　要**: 针对细粒度图像分类任务中难以对图中具有鉴别性对象进行有效学习的问题, 本文提出了一种基于注意力机制的弱监督细粒度图像分类算法. 该算法能有效定位和识别细粒度图像中语义敏感特征. 首先在经典卷积神经网络的基础上通过线性融合特征得到对象整体信息的表达, 然后通过视觉注意力机制进一步提取特征中具有鉴别性的细节部分, 获得更完善的细粒度特征表达. 所提算法实现了线性融合和注意力机制的结合, 可看作是多网络分支合作训练共同优化的网络模型, 从而让网络模型对整体信息和局部信息都有更好的表达能力. 在 3 个公开可用的细粒度识别数据集上进行了验证, 实验结果表明, 所提方法有效性均优于基线方法, 且达到了目前先进的分类水平.

**关键词**: 细粒度图像分类; 双线性网络融合; 注意力机制; 弱监督学习

引用格式: 李文书,王志骁,李绅皓,赵朋.基于注意力机制的弱监督细粒度图像分类.计算机系统应用,2021,30(10):232–239. http://www.c-s-a.org.cn/1003-3254/8141.html

## Weakly Supervised Fine-Grained Image Classification Based on Attention Mechanism

LI Wen-Shu, WANG Zhi-Xiao, LI Shen-Hao, ZHAO Peng

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

**Abstract**: Fine-grained image classification is challenging due to the difficulty in the effective learning of discriminative objects in images. Therefore, this study proposes a weakly supervised fine-grained image classification algorithm based on the attention mechanism. This algorithm can accurately locate and identify the semantically sensitive features in fine-grained images. First, on the basis of the classic convolutional neural network, the overall information of an object can be expressed by the linear fusion of features. Then, the discriminative details of the features are further extracted through the visual attention mechanism to obtain a more complete fine-grained feature expression. The proposed algorithm combines linear fusion with the attention mechanism and it can be regarded as a network model of multi-network-branch cooperative training and joint optimization. Thus, the network model can better express the overall and local information. Experiments on three publicly available fine-grained identification datasets show that the proposed method is superior to the baseline method and achieves the advanced classification level.

**Key words**: fine-grained image classification; bilinear network fusion; attention mechanism; weakly supervised learning

## 1　引言

近些年来图像分类技术通过对深度卷积神经网络的运用取得了长足的发展, 然而其不足也依然存在. 在当下这个智能化要求又不断提高的大环境下, 简单的“语义级”的图片分类已经不足以满足用户的需求, 因而更加精细化的细粒度图像分类更加值得关注,例如在

① 基金项目: 国家科技部重点研发计划 (2018YFB1004901); 浙江省技术厅重点项目 (2019C25014); 浙江省基金 (LY17C090011)
Foundation item: National Key R & D Program of the Ministry of Science and Technology of China (2018YFB1004901); Key Program of Zhejiang Provincial Department of Science and Technology (2019C25014); Fund of Zhejiang Province (LY17C090011)
收稿时间: 2020-12-31; 修改时间: 2021-01-29; 采用时间: 2021-02-26

生态保护场景中识别不同种类的珍稀鸟类, 水稻种植生产中识别不同种类的虫害, 新零售场景下对同类食品的细分类等等. 利用计算机视觉方法识别细粒度类别 (如鸟类[1,2]、花卉[3,4]、狗类[5,6]、车型[7] 等) 的技术已引起研究者们的广泛关注[8–10]. 其中能够准确定位和表示类别中细微视觉差异的细粒度图像识别技术是非常具有挑战性的.

### 1.1 细粒度图像分类的研究历史与现状

随着计算机硬件算力的提升, 深度学习技术被广泛用于解决复杂图像分类的问题. 其中, 卷积神经网络 (Convolutional Neural Network, CNN) 是深度学习解决分类问题的代表性网络之一. 2015 年, 何恺明等提出的残差神经网络 (Residual Network, ResNet)[11] 采用了更深的网络层数, 并且引入了残差处理单元解决网络退化的问题, 取得了极佳的效果. 2017 年, Google 团队设计一种具有优良局部拓扑结构的网络 Inception-V3[12], 即对输入图像并行地执行多个卷积运算及池化操作, 并将所有输出结果拼接为一个非常深的特征图, 取得了优异的效果. 但是经典卷积神经网络聚焦于类间分类的问题, 并不能有效解决细粒度图像分类 (类内分类) 的分类问题. 细粒度图像的类别精度更加细致, 类间差异更加细微, 往往只能借助于微小的局部差异才能区分出不同的类别.

细粒度图像分类发展初期仍依靠人工注释的边界框/部件注释. 大量的人工参与使得部分定义和注释变得昂贵且主观, 这对于细粒度的识别任务都不是最优的 (文献 [13,14] 表明边界框/部件注释依赖于人的注释, 由此带来主观性强和成本昂贵等问题). 越来越多的算法倾向于不再依赖人工标注信息, 而使用类别标签来完成分类任务.

Lin 等提出了一种端到端的双线性网络[15], 通过对卷积层输出的特征进行外积操作, 能够建模不同通道之间的线性相关, 从而增强了卷积网络的表达能力. Ge 等基于双线性网络提出一种核化的双线性卷积网络[16], 通过使用核函数的方式有效地建模特征图中通道之间的非线性关系, 进一步增强卷积网络的表达能力. 此模型能够融合不同通道的信息, 但是并没有有效地提取出具有鉴别性的局部特征.

一种采用锚框的机制[17,18] 可以有效地定位信息区域而无须边界框并挖掘概率较高的含更多的对象特征语义的区域, 从而增强整个图像的分类性能, 但是对特征全局定位有所欠缺. 除了使用锚框定位局部特征, 注意力机制也被应用于细粒度图像分类, Fu 等[19] 提出了一种注意力网络, 利用两个任务之间的联系, 相互增益彼此的精度, 在多尺度上递归地学习区分度大的区域以及多尺度下的特征表达. 该方法较好地提取了局部的特征信息, 但是对全局信息捕捉较弱. 为了进一步在图像中同时产生多个注意位置, 基于提取多个局部特征的注意力方法[20,21] 相继提出,但有限的注意力个数并不能充分表达图像的特征.

### 1.2 计算机视觉中注意力机制的研究历史与现状

人类视觉系统中存在一种现象, 当人眼在接受外部信息时对每个区域的关注度存在差异, 例如人眼在看一幅图像时会聚焦在感兴趣的目标身上而忽略背景图像, 这就是人类视觉系统中的注意力机制, 这一机制也被应用于计算机视觉中. 近几年, 计算机视觉中注意力机制发展迅速[22], 出现了很多基于注意力机制提出的深度学习网络, 其主要实现方式是通过为特征图添加掩码 (*mask*) 的形式, 即通过为特征图添加权重, 将有用的特征标识出来. 从注意力域的角度可以将注意力实现方式分为空间域和通道域.

空间域是从特征图的空间位置关系出发, 不区分通道带来对分类性能的影响. 其中, 细粒度图像分类任务中, 2020 年 Yan 等[23] 提出的网络模型, 在细粒度图像分类中, 不同空间位置能够获得不同的关注点, 不同大小的空间特征图能够获得递进的特征信息, 所以作者提出了一种空间转换器, 对图像做空间变换将关键信息提取出来.

通道域是对不同通道的加权, 不考虑通道中每个像素点的位置差异. 在卷积神经网络中, 每张图像初始都有 RGB 三个通道. 卷积层的卷积操作变换图像的通道, 其等价于对原图像进行了分解. 每个通道都是原图在不同卷积核上的分量. 虽然每个通道都是原图的分量, 但在具体任务中并不是每个通道都发挥着相同的作用. 基于此, Li 等[24,25] 提出了 SKnet 网络模型, SKnet 通过对每个通道加权的方式标注出对结果贡献较大的通道, 具体做法是通过对每个特征图做全局平均池化将 $H\times W\times C$ 的特征图挤压到长度为 $C$ 的一维向量, 然后通过激励函数以获得每个通道的权重, 最后对原始特征图上的每个像素点加权.

现有的方法着重挖掘图像的细节特征, 但是没有将细节特征更好地融合到全局特征. 另一方面, 如何将注意力网络和双线性网络融合值得关注. 本文的贡献

如下:

1) 通过线性融合不同通道的特征来建模不同通道之间的线性相关, 从而增强了卷积网络的表达能力;

2) 通过注意力机制提取显著特征中具有鉴别性的细节部分放入网络中训练, 进一步挖掘具有鉴别性的特征, 以提升细粒度图像的识别能力;

3) 通过主网络和注意力网络分合作训练及共享训练参数, 在挖掘表征细粒度图像视觉差异的细节特征的同时, 兼顾全局特征的学习.

## 2 基于注意力机制的细粒度图像分类模型

在图像细粒度分类中, 如何获取物体整体与局部信息是一个难点. 针对这一难点, 本文提出了一种基于注意力机制的弱监督细粒度图像分类 (ATtention mechanism Convolutional Neural Networks, AT-CNN) 的学习方法用以自动定位和学习细粒度图像中语义敏感对象. 该方法首先采用经典的卷积网络方法 (ResNet[11], Inception-Net[12] 等) 提取图像的特征图. 主网络通过双通道融合网络表达细粒度特征的整体信息; 然后通过弱监督学习的方式, 将特征图通道进行排序并筛选显著的特征放入注意力网络中获取对细节特征的表征能力; 最后通过主网络和注意力网络共享网络参数, 共同训练, 增强网络对细粒度图像中具有代表性特征的表达能力. 本文方法不依赖于边界框/零件标注, 可以实现对细粒度图像物体位置的追踪及分类, 实现了细粒度端到端的弱监督分类任务, 总体分类网络结构如图 1 所示.

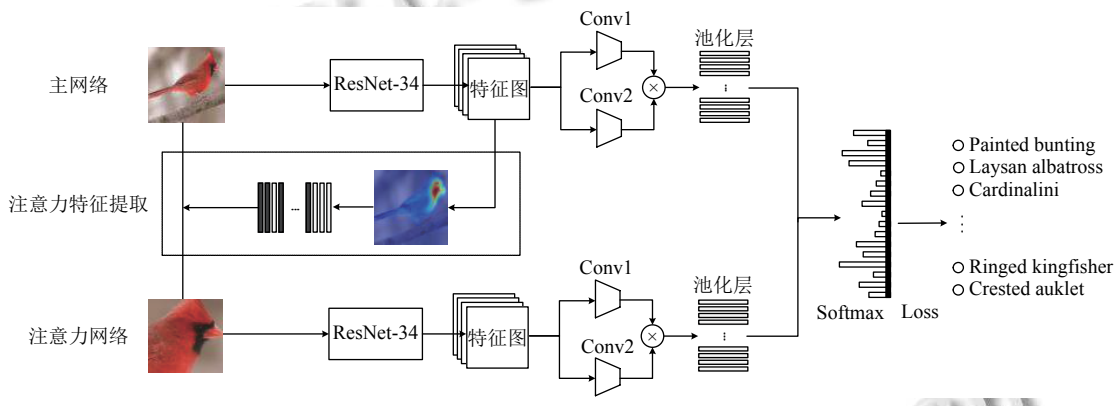



图 1 总体分类网络结构

### 2.1 主网络分支

主网络分支作为基础架构, 首先通过卷积神经网络 (ResNet, Inception-V3) 作为基础网络 (BaseNet) 来提取图像的特征. 然后再通过两个特征提取器 (Conv1, Conv2) 组成, 它们的输出使用外部乘积相乘, 获得图像表示, 如图 2 所示.

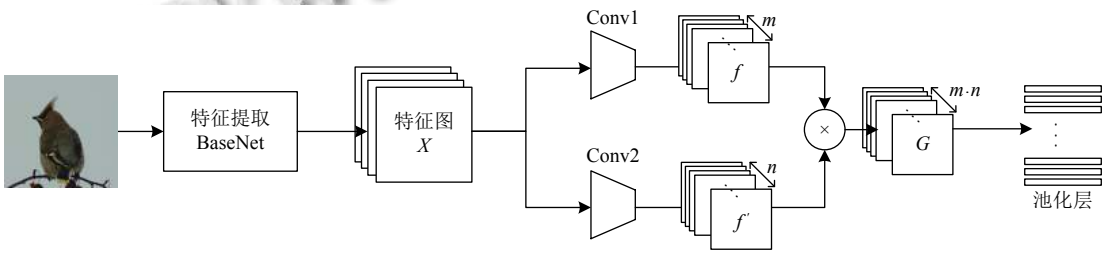



图 2 主网络结构

设 $I$ 表示为输入的训练图片, $X$ 表示为卷积网络提取的特征图. 则有:

$$X = BaseNet(I) \tag{1}$$

其中, $f$ 表示特征图 $X$ 通过卷积层Conv1展开后的特征矩阵, $f \in \mathbb{R}^{m \times k}$, 其中, $m$ 表示特征图的通道数目, $k$ 表示特征图中包含的特征数目; $f'$ 表示输入的特征图 $X$ 通

过卷积层Conv2得到展开后的特征矩阵$f' \in \mathbb{R}^{n\times k}$.

Conv1和Conv2卷积层设计在 3.1 节详细阐述. 通过对$f$和$f'$进行外积聚合得到图像的表达, 公式如下:

$$G = \frac{1}{mn} f \cdot f' = \frac{1}{mn}\begin{bmatrix} \langle x_1, x_1' \rangle & \langle x_1, x_2' \rangle & \cdots & \langle x_1, x_n' \rangle \\ \langle x_2, x_1' \rangle & \langle x_2, x_2' \rangle & \cdots & \langle x_2, x_n' \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle x_m, x_1' \rangle & \langle x_m, x_2' \rangle & \cdots & \langle x_m, x_n' \rangle \end{bmatrix} \tag{2}$$

式 (2) 中, 外积聚合矩阵$G \in \mathbb{R}^{m\times n}$, $x_i$表示$f$中的第$i$行(即特征图中第$i$个通道), $i \in 1,2,\cdots,m$, $x_j'$表示$f'$中的第$j$行 (即特征图中第$j$个通道), $j \in 1,2,\cdots,n$. 从式中可以看出, 外积聚合得到的矩阵$G$中的每个元素均为通道之间的内积, 从而可以捕捉到特征图中通道之间的线性关系.

为了使特征矩阵具有较好的分布, 需要对特征进行归一化. 首先将得到的聚合矩阵展开成向量$g$, 再进行带符号的平方根正则化和归一化, 具体公式如下:

$$g = vec(G) \tag{3}$$

$$s = \text{sign}(g)\sqrt{|g|} \tag{4}$$

$$c_{\text{main}} = s/\|s\| \tag{5}$$

最后将特征$c_{\text{main}}$作为图像的最终表达, 送至Softmax中进行端到端的联合训练.

### 2.2 注意力机制

经典的注意力网络[23,26] 可以有效提取图像中关注物体的空间位置, 但是细粒度图像分类任务需要定位到具有鉴别性特征的空间位置. 为此, 本文注意力机制分为注意力特征提取和注意力网络两个步骤. 注意力特征提取通过对通道的筛选, 提取出关注物体中最显著的局部特征. 注意力网络将提取出的局部特征, 放入网络中继续训练, 提高网络对细粒度对象具有鉴别性特征的表达能力.

#### 2.2.1 注意力特征提取

随着训练次数的增加, 输入图像通过卷积得到特征的感受野随网络深度变化而变化, 网络可以逐步定位到关注区域的位置.

假设输入图像通过一系列的卷积层及池化层得到大小为$C\times H\times W$的特征图, 然后特征图上每个$C\times 1\times 1$的跨通道的向量, 随着训练次数的增加, 能在固定的空间位置上表示原始图像中的对应位置. 对应的热力图, 如图 3 所示. 基于分类神经网络能够对图像中关注对象的自动聚焦的特性, 本文设计的注意力机制网络可以将聚焦的位置信息反馈到最初的图像上. 通过类别信息学习到显著特征的位置信息以实现网络的弱监督学习.

(a) 原始图像 (b) 特征图的热力图

图 3 训练后卷积网络得到的热力图

特征图的热力图可以随着训练或者物体的整体空间位置. 为得到物体细粒度对象的空间位置, 需要通过计算特征图不同通道的掩码, 对图像通道进一步筛选提取.

首先通过主网络分支可以得到 $n$ 个特征图$p_1, p_2, \cdots, p_n$(即 $n$ 个通道):

$$p_1, p_2, \cdots, p_n = BaseNet(I) \tag{6}$$

对 $n$ 通道分别取平均, 然后依次求得每个通道相对的权重, 记为$w_i, i \in (0,n)$:

$$q_i = mean(p_i),\ i \in (0,n) \tag{7}$$

$$w_i = \frac{q_i}{\sum_{i=1}^{n}\sqrt{q_i}} \tag{8}$$

权重$w_i$越大意味着该通道的特征越显著. 通过取前 $m$ 个权重作和得到注意力掩码 ($mask$), 从而在显著物体中得到更细粒度的显著特征. 特征进一步提取示意图见图 4.

$$mask = w_1 + w_2 + \cdots + w_m \tag{9}$$

#### 2.2.2 注意力网络

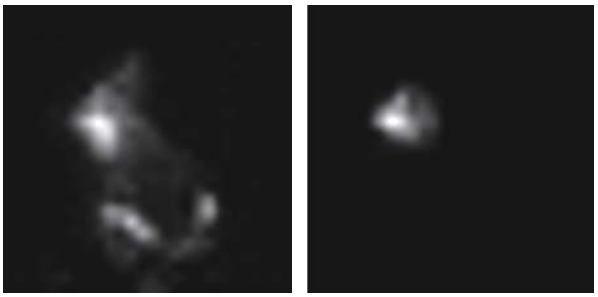
从特征图中提取到多个最显著特征的位置信息后, 通过掩码 ($mask$) 操作将前景的物体 (较显著特征) 在原图中提取出来, 记为 $I_{\text{attention}}$.

$$I_{\text{attention}} = I \cdot mask \tag{10}$$

$$c_{\text{attention}} = conv(I_{\text{attention}}) \tag{11}$$

再将提取到的较显著特征放入网络训练, 可以得

到显著图的特征图$c_{attention}$, 增强神经网络对显著特征的表达能力. 注意力网络与主网络共用网络结构, 通过共同训练参数, 实现网络对整体图像类别和鉴别性的特征都有较好的表达能力. 注意力网络如图 5 所示.



(a) 卷积得到特征图 (b) 筛选过后的特征图

图 4 特征图通过筛选通道的前后对比图

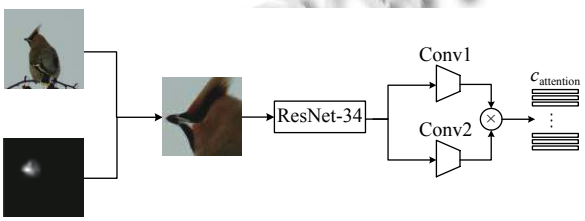



图 5 基于注意力机制提取的特征

### 2.2.3 损失函数

总损失函数结合主网络的损失函数和注意力网络的损失函数, 达到合作训练、共同优化.

$$Loss(x,I)=Loss_{\text{main}}(x,c_{\text{main}})+Loss_{\text{attention}}(x,c_{\text{attention}}) \tag{12}$$

其中, $x$ 表示图像类别信息, $c_{\text{main}}$、$c_{\text{attention}}$ 分别表示主网络分支, 注意力网络分支预测的类型信息.

分类损失使用的是 Softmax 函数, *Loss* 的计算公式如下:

$$Loss(p,q)=-\sum p(x_i)\log_2(q(x_i)) \tag{13}$$

$$q(x_i)=\frac{\exp(x_i)}{\sum_{i=1}^{n}\exp(x_i)} \tag{14}$$

其中, $p(x_i)$为$x_i$的目标值.

模型在梯度下降方法上采用 Adam 和 SGD 联合训练的方式. Adam 与 AdaDelta 方法收敛速度更快, 但是由于细粒度图像的特征较难学习, 导致模型的准确率往往达不到最优解, 而 SGD 的收敛速度较慢, 但最终得到的预测效果要优于 Adam. 因此为了较快收敛同时避免出现局部最优的现象, 最终选定先用 Adam 训练,等准确率不再继续提高的时候,保存网络权重,再采用 SGD 方法微调.

整体算法总结如算法 1.

算法 1. Adam 和 SGD 联合训练算法

输入: full image *I*
输出: predict probability *P*

for *t*=*1*,*T* do
  Take full image= *I* /*输入原始图像*/
  *F*, *F'*=BaseNet(*I*) /*主网络分支*/
  Calculate the weight *w*[*i*] of each channel of *F*/*计算通道权重*/
  for *s*=1, *m* do
    *mask* = *mask* + *w*[*s*] /*生成注意力掩码*/
  end for
  $I_{attention}$= *I* * *mask* /*通过掩码对原图提取*/
  $F_{attention}$,$F'_{attention}$=BaseNet($I_{attention}$)/*放入注意力网络进一步训练*/
  $c_{main}$=*F*⊗*F'*/*主网络特征线性融合*/
  $c_{attention}$=$F_{attention}$⊗$F'_{attention}$/*注意力网络特征线性融合*/
  Loss(*x*,*I*)=$Loss_{main}$(*x*,$c_{main}$) +$Loss_{attention}$(*x*,$c_{attention}$)/*损失函数计算*/
  BP($Loss_{total}$) get gradient w.r.t.*Net Param* /*反向传播*/
  Update *Net Param* using Adam/SGD
End

## 3 实验分析

本文主要通过 3 个典型细粒度图像数据集 (包括 CUB-200-2011 鸟类数据集, FGVC 飞机数据集, 斯坦福狗数据集) 进行验证, 如表 1 所示.

表 1 3 种常用细粒度分类数据集介绍

| 数据集 | 类别 | 训练集数量 | 测试集数量 |
|---|---|---|---|
| CUB-200-2011 | 200 | 5994 | 5794 |
| Stanford Dogs | 120 | 12000 | 8580 |
| FGVC Aircraft | 100 | 6667 | 3333 |

实验是在 Linux 下的 Python 3.6.6、TensorFlow 1.12.0 和 2 块 16 GB NVIDIA Tesla GPU 下进行的. 使用 Inception-V3 预训练模型, 训练时 batchsize为 16, weight decay 为 0.0001, 初始的学习率为 0.001, 后续采用指数型衰减法, 逐步计算学习率.

### 3.1 线性融合结构设计

为了保证特征融合网络结构中, 卷积层Conv1和卷积层Conv2分别得到的特征图包含的特征数目相同 (即特征图的尺度不变), 而特征矩阵的通道不同, 我们分别设计了以下不同的双线模块:

图 6(a) 为输入特征图分别通过卷积核为 1×1×*n*

的卷积层Conv1和卷积核为 1×1×$m$ 的卷积层Conv2, 再经过外积聚合得到特征通道为 $m$×$n$ 的特征图. 这种方式聚合过于单一, 融合得到特征表达能力较弱.

图 6(b) 为输入特征图分别通过卷积核为 1×1×$n$ 的卷积层Conv1和卷积核为 3×3×$m$, 填充值 (*padding*) 为 1 的卷积层Conv2, 再经过外积聚合得到融合的特征图. 通过 3×3 的卷积核, 需要加入填充值才能维持特征尺寸的不变. 额外加入的填充值会加入额外不必要的信息, 不利于特征的线性融合.

图 6(c) 为输入特征图分别通过卷积核为 1×1×$n$ 的卷积层Conv1和卷积核为 1×3×$m$, 3×1×$m$ 的卷积层Conv2, 再经过外积聚合得到融合的特征图. 综合了以上两种方法, 经实验验证, 通过两种卷积对特征具有较好的表达.

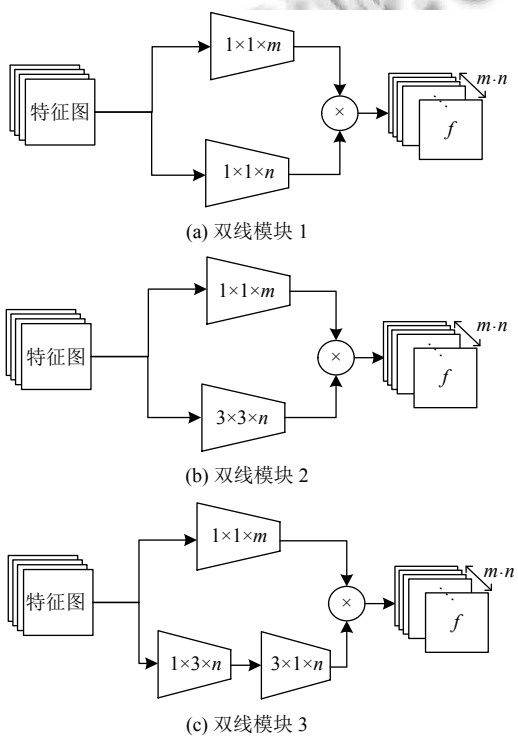



图 6 不同的特征融合网络结构

通过当网络结构只有主网络分支, 分别采用 4 种不同卷积核在数据集 CUB-200-2011 上的分类实验, 以验证不同卷积核对分类结果的影响. 实验结果如表 2 所示.

比较这 3 种方案对分类结果的影响, 可以得出以图 6(c) 方式在单主网络分支中进行特征融合取得最高的准确率, 其能更好地计算特征图在某一区域的响应.

表 2 不同特征融合网络对分类结果的影响

| Conv1, Conv2 | 1×1, 1×1 | 1×1, 3×3 | 1×1, 1×3 3×1 |
|---|---|---|---|
| 准确率 (%) | 84.8 | 85.2 | 86.4 |

### 3.2 消融实验

如上所述分类算法由基础网络、线性融合、注意力网络 3 部分组成. 通过在 CUB-200-2011 数据集中进行实验探索每个组件可以做出的贡献, 如表 3 所示.

表 3 不同组件组合对分类结果的影响

| 结构组件 | 准确率 (%) |
|---|---|
| ResNet-101 | 83.5 |
| ResNet-101 + 线性融合 | 85.8 |
| ResNet-101 + 线性融合 + 注意力网络 | 87.2 |
| Inception-V3 | 83.7 |
| Inception-V3 + 线性融合 | 86.4 |
| Inception-V3 + 线性融合 + 注意力网络 | 87.8 |

通过 BaseNet 选用 Inception-V3 作为基线方法, 在 CUB-200-2011 数据集中网络的预测准确率如图 7 所示.

图 7 展示了训练次数 (epoches) 对分类精准度的影响. 可以发现随着迭代次数的增加, 混合模型准确度有所提高, 但当训练次数达到一定值 (4 万次左右), 准确度会保持稳定.

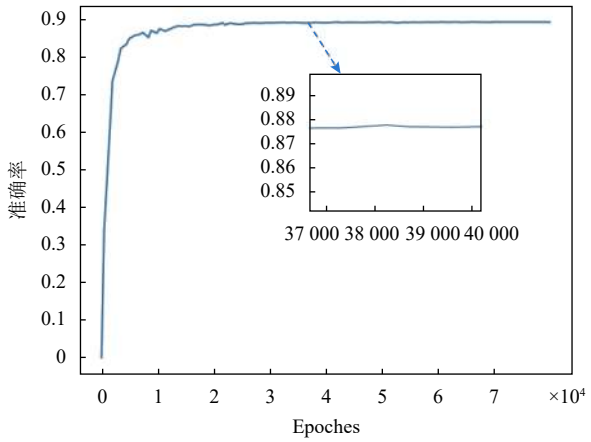



图 7 模型在数据集 CUB-200-2011 的准确率图

### 3.3 与先进分类方法的比较

为了验证模型的有效性, 我们分别在 CUB-200-2011、FGVC-Aircraft 和 Stanford-Dogs 三个细粒度经典数据集上进行实验, 结果如表 4、表 5 和表 6 所示.

由实验结果可知, 本文方法在 3 个细粒度图像识别数据库上均优于基线方法 (BCNN, Inception), 如在

CUB-200-2011 数据上比 Inception-V3 高出 4%, 比 BCNN 高出 3.7%. 并在 3 个细粒度数据库上与先进分类方法相比, 均取得了领先的水平.

表 4 相关分类方法在 CUB 200-2011 的准确度

| Method | Top-1 准确率 (%) |
|---|---|
| VGG-16[22] | 77.8 |
| ResNet-101[11] | 83.5 |
| Inception-V3[12] | 83.7 |
| BCNN[15] | 84.1 |
| RA-CNN[19] | 85.4 |
| MA_CNN[20] | 86.5 |
| DFL-CNN[17] | 87.4 |
| **AT-CNN** | **87.8** |

表 5 相关分类方法在 FGVC-Aircraft 的准确度

| Method | Top-1 准确率 (%) |
|---|---|
| VGG-16[22] | 80.5 |
| ResNet-101[11] | 87.2 |
| Inception-V3[12] | 87.4 |
| BCNN[15] | 84.1 |
| RA-CNN[19] | 88.4 |
| MA_CNN[20] | 89.9 |
| DFL-CNN[17] | 92.0 |
| **AT-CNN** | **92.0** |

表 6 相关分类方法在 Stanford Dogs 的准确度

| Method | Top-1 准确率 (%) |
|---|---|
| VGG-16[22] | 76.7 |
| ResNet-101[11] | 81.2 |
| Inception-V3[12] | 81.5 |
| BCNN[15] | 83.2 |
| RA-CNN[19] | 87.3 |
| MA_CNN[20] | 87.6 |
| DFL-CNN[17] | 88.3 |
| **AT-CNN** | **88.5** |

## 4 结论与展望

本文提出了一种基于注意力机制的弱监督细粒度图像分类. 该算法针对细粒度图像类别中细微的视觉差异, 设计了基于线性融合网络、注意力网络同步训练的网络模型用于提取细粒度图像中鉴别性强的特征. 经实验论证, 所提方案可行且有效, 进一步提升了细粒度分类的准确性.

细粒度图像分类任务为了识别图中具有鉴别性的对象并对其进行有效的学习, 往往模块较多且网络层次较深, 这导致模型较大难以部署. 下一步, 我们关注如何在保证精准度的情况下压缩网络模型, 满足移动端的性能要求.

## 参考文献


1 Branson S, Van Horn G, Belongie S, *et al*. Bird species categorization using pose normalized deep convolutional nets. arXiv: 1406.2952, 2014.

2 Zhang XP, Xiong HK, Zhou WG, *et al*. Picking deep filter responses for fine-grained image recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 1134–1142.

3 Nilsback ME, Zisserman A. A visual vocabulary for flower classification. Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2006. 1447–1454.

4 Reed S, Akata Z, Lee H, *et al*. Learning deep representations of fine-grained visual descriptions. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 49–58.

5 Khosla A, Jayadevaprakash N, Yao BP, *et al*. Novel dataset for fine-grained image categorization. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, IEEE. 2011.

6 Krause J, Jin HL, Yang JC, *et al*. Fine-grained recognition without part annotations. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 5546–5555.

7 Zhao B, Wu X, Feng JS, *et al*. Diversified visual attention networks for fine-grained object classification. IEEE Transactions on Multimedia, 2017, 19(6): 1245–1256. [doi: 10.1109/TMM.2017.2648498]

8 李彦冬, 郝宗波, 雷航. 卷积神经网络研究综述. 计算机应用, 2016, 36(9): 2508–2515, 2565. [doi: 10.11772/j.issn.1001-9081.2016.09.2508]

9 李旭冬, 叶茂, 李涛. 基于卷积神经网络的目标检测研究综述. 计算机应用研究, 2017, 34(10): 2881–2886, 2891. [doi: 10.3969/j.issn.1001-3695.2017.10.001]

10 罗建豪, 吴建鑫. 基于深度卷积特征的细粒度图像分类研究综述. 自动化学报, 2017, 43(8): 1306–1318.

11 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.

12 Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for computer vision. Proceedings of

the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2818–2826.

13 Zhang N, Donahue J, Girshick R, *et al*. Part-based R-CNNs for fine-grained category detection. Proceedings of the 13th European Conference on Computer Vision. Cham: Springer, 2014. 834–849.

14 Zhang H, Xu T, Elhoseiny M, *et al*. SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 1143–1152.

15 Lin TY, RoyChowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 1449–1457.

16 葛疏雨, 高子淋, 张冰冰, 等. 基于核化双线性卷积网络的细粒度图像分类. 电子学报, 2019, 47(10): 2134–2141. [doi: 10.3969/j.issn.0372-2112.2019.10.015]

17 Yang Z, Luo TG, Wang D, *et al*. Learning to navigate for fine-grained classification. Proceedings of the 15th European Conference on Computer Vision. Cham: Springer, 2018. 420–435.

18 Wang YM, Morariu VI, Davis LS. Learning a discriminative filter bank within a CNN for fine-grained recognition. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4148–4157.

19 Fu JL, Zheng HL, Mei T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 4476–4484.

20 Zheng HL, Fu JL, Mei T, *et al*. Learning multi-attention convolutional neural network for fine-grained image recognition. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 5219–5227.

21 Ji JS, Jiang LF, Zhang T, *et al*. Adversarial erasing attention for fine-grained image classification. Multimedia Tools and Applications, 2020, (9): 1–23.

22 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556, 2015.

23 Yan YC, Ni BB, Wei HW, *et al*. Fine-grained image analysis via progressive feature learning. Neurocomputing, 2020, 396: 254–265. [doi: 10.1016/j.neucom.2018.07.100]

24 Hu J, Shen L, Albanie S, *et al*. Squeeze-and-Excitation Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011–2023. [doi: 10.1109/TPAMI.2019.2913372]

25 Li X, Wang WH, Hu XL, *et al*. Selective kernel networks. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2020. 510–519.

26 Liang L, Cao JD, Li XY, *et al*. Improvement of residual attention network for image classification. In: Cui Z, Pan JS, Zhang SS, *et al*. eds. Intelligence Science and Big Data Engineering. Visual Data Engineering. Cham: Springer, 2019. 529–539.