# 面向细粒度图像分类的双线性残差注意力网络

王阳，刘立波*
宁夏大学信息工程学院，宁夏 银川 750021

**摘要** 细粒度图像之间具有高度相似的外观，其差异往往体现在局部区域，提取具有判别性的局部特征成为影响细粒度分类性能的关键。引入注意力机制的方法是解决上述问题的常见策略，为此，在双线性卷积神经网络模型的基础上，提出一种改进的双线性残差注意力网络：将原模型的特征函数替换为特征提取能力更强的深度残差网络，并在残差单元之间分别添加通道注意力和空间注意力模块，以获取不同维度、更为丰富的注意力特征。在 3 个细粒度图像数据集 CUB-200-2011、Stanford Dogs 和 Stanford Cars 上进行消融和对比实验，改进后模型的分类准确率分别达到 87.2%、89.2%和 92.5%。实验结果表明，相较原模型及其他多个主流细粒度分类算法，本文方法能取得更好的分类结果。

**关键词** 图像处理；细粒度图像分类；注意力机制；残差网络；通道注意力；空间注意力
**中图分类号** TP391.4 **文献标志码** A **doi**：10.3788/LOP57.121011

# Bilinear Residual Attention Networks for Fine-Grained Image Classification

Wang Yang, Liu Libo*
*School of Information Engineering, Ningxia University, Yinchuan, Ningxia* 750021, *China*

**Abstract** Fine-grained images have a highly similar appearance, and the differences are often reflected in local regions. Extracting discriminative local features plays a key role in fine-grained classification. Attention mechanism is a common strategy to solve the problems above. Therefore, we propose an improved bilinear residual attention network based on bilinear convolutional neural network model in this paper: the feature function of the original model is replaced by deep residual network with a stronger feature extraction capability, then channel attention module and spatial attention module are added between the residual units respectively to obtain different dimensions and richer attention features. Ablation and contrast experiments were performed on three fine-grained image datasets CUB-200-2011, Stanford Dogs, and Stanford Cars, the classification accuracy of the improved model reached 87.2%, 89.2% and 92.5%, respectively. Experimental results show that our method can achieve better classification results than the original model and other mainstream fine-grained classification algorithms.

**Key words** image processing; fine-grained image classification; attention mechanism; residual network; channel attention; spatial attention
**OCIS codes** 100.4996; 100.5010; 100.3008; 100.2960

## 1 引 言

近年来，细粒度图像分类已成为计算机视觉领域一项研究热点，它区别于传统的图像分类范畴，旨在区分同一大类物体下的不同子类，如不同科、目的鸟，以及不同款型的汽车等，因此也称子类别图像分类[1]。同时，它也是一项极具挑战性的难题，其难度主要体现在以下两方面：1）类间差异小。不同子类间差异微小，不易区分，以两种杜鹃鸟为例，黑嘴美洲鹃与红树美洲鹃仅下喙部和眼部周围颜色有所不同，其余部位十分相似，非相关专家很难区分。2）类内差异大。同一子类的不同个体乃至同一个体受年

**收稿日期**：2019-08-19；**修回日期**：2019-10-28；**录用日期**：2019-11-02
**基金项目**：国家自然科学基金(61862050)、西部一流大学科研创新项目(ZKZD2017005)
* **E-mail**：liulib@163.com

龄、动作姿态、拍摄角度及背景干扰等因素影响，存在较大的类内差异。

针对以上问题，研究人员提出基于强监督信息的细粒度分类算法，具有代表性的有 PB R-CNN[2]、PS-CNN[3]、HSnet[4]、Mask-CNN[5] 等。这类算法虽然分类精度较高，但依赖额外的人工标注信息，这些标注一般由专家给出，成本偏高，影响了算法的实用性。近年来，基于弱监督信息的分类算法无需物体标注框和部件位置等额外标注信息，大大降低了人工成本，已成为细粒度图像分类研究的趋势。其中，由 Lin 等[6] 提出的双线性卷积神经网络（B-CNN）是一个具有代表性的分类模型，它由两路 VGGNet 构成特征提取函数，并将两个网络提取的卷积特征进行双线性组合，实现了端到端的弱监督分类。B-CNN 模型虽取得了较好的分类结果，但其中仍存在一些限制：模型使用 VGGNet 作为特征提取函数，虽具有一定特征表示能力，但未能充分关注物体的判别性部位对分类的影响，细粒度图像内包含冗余的背景信息，影响分类结果的往往是物体的判别性部位而非所处背景；此外，B-CNN 将两路输出特征通过外积进行组合，生成高维的双线性特征，这些特征同样存在高度冗余[7]，包含许多与分类无关的特征通道。

综上，本文在 B-CNN 模型的基础上，提出一种改进的双线性残差注意力网络（BRAN）以实现弱监督分类模型，主要贡献如下：1）采用两路 ResNet-34 取代 B-CNN 中的 VGG-M[8] 和 VGG-16 作为特征提取函数，在不增大输出特征维度的基础上，增加网络的深度以提高细粒度特征提取能力；2）在网络中引入一种多维注意力机制，分别在两路 ResNet-34 的残差单元之间，添加通道注意力和空间注意力模块，获取更为丰富的多维注意力特征，聚焦特征图中的局部特征通道和空间响应部位，降低特征冗余，同时进一步提升细粒度特征学习能力。在 3 个公开的细粒度图像数据集[9-11] 上进行训练和测试，改进模型 BRAN 的分类准确率分别达到 87.2%、89.2% 和 92.5%。实验结果表明，本文方法能有效提升原模型的分类精度，并优于多数近年来主流的弱监督分类方法。

## 2 B-CNN 模型概述

B-CNN 模型由 Lin 等[6] 于 2015 年提出，该模型主要由两个并列的 CNN 完成特征提取过程，原文分别选用 VGG-M[8] 和 VGG-D（即 VGG-16）作为特征函数，并将两路 VGG 最后的全连接层和 Softmax 层替换为双线性池化层，对两路特征函数的输出结果进行双线性组合和池化后得到最终的双线性特征表示向量。B-CNN 模型利用了图像二阶统计信息，以平移不变的特性对局部特征间的组合交互关系进行建模，在只有图像类别标签的情况下实现了弱监督分类。同时，B-CNN 简化了梯度计算，使得其端到端的模型更加容易被训练，该模型架构如图 1 所示。

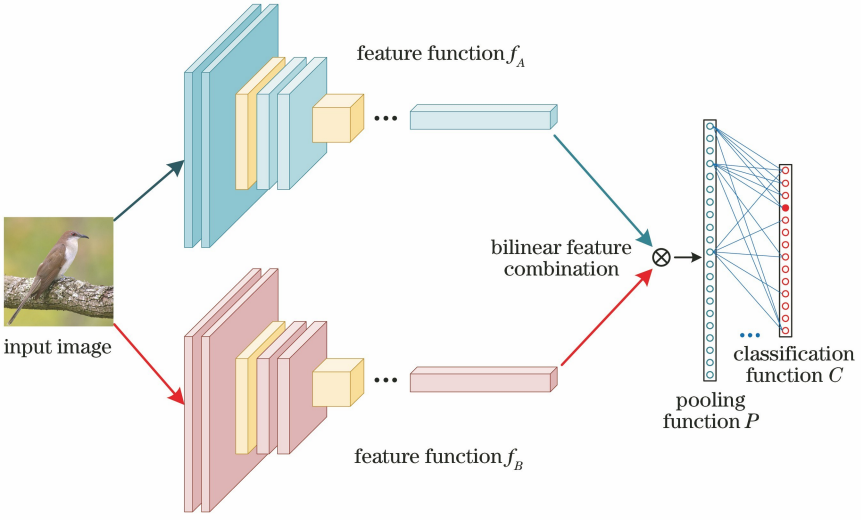


图 1 双线性卷积神经网络模型架构

Fig. 1 Architecture of B-CNN model

B-CNN 模型可由一个四元函数 $B=F(f_A, f_B, P, C)$ 表示，其中 $f_A$ 和 $f_B$ 为特征函数，$P$ 为池化函数，$C$ 为分类函数。特征函数 $f_A$ 和 $f_B$ 表示一种映射关系 $f: \boldsymbol{I}\times\boldsymbol{L}\rightarrow\mathbf{R}^{K\times T}$，其中 $\boldsymbol{I}$ 代表输入图像，

$\boldsymbol{L}\in\mathbf{R}^K$ 表示输入图像的位置范围，$f$ 将二者映射为 $K\times T$ 维的特征图，其中 $K$ 表示特征图的空间分辨率大小，$T$ 表示特征通道维数。设 $\boldsymbol{m}$ 和 $\boldsymbol{n}$ 分别代表特征函数 $f_A$ 和 $f_B$ 输出的特征向量，它们具有相同大小的空间分辨率 $K$，而通道维数 $T$ 不必相同，设 $K=w\times h$，则 $\boldsymbol{m}\in\mathbf{R}^{w\times h\times t_1}$，$\boldsymbol{n}\in\mathbf{R}^{w\times h\times t_2}$，这里 $w\times h$ 表示特征向量的空间分辨率大小，$t_1$ 和 $t_2$ 表示特征通道维数。将特征向量 $\boldsymbol{m}$ 和 $\boldsymbol{n}$ 通过外积运算（这里指线性代数中的张量积[12]）进行双线性组合，得到双线性特征

$$\boldsymbol{b}(\boldsymbol{l},\boldsymbol{I},f_A,f_B)=f_A(\boldsymbol{l},\boldsymbol{I})\otimes f_B(\boldsymbol{l},\boldsymbol{I})=\boldsymbol{m}^{\mathrm{T}}\boldsymbol{n},\tag{1}$$

式中：双线性特征 $\boldsymbol{b}\in\mathbf{R}^{w\times h\times t_1\times t_2}$；$\boldsymbol{l}\in\boldsymbol{L}$；$\boldsymbol{L}\in\mathbf{R}^K$；$\otimes$ 表示向量的外积运算；T 表示对向量 $\boldsymbol{m}$ 的转置。可以看出，双线性组合不会损失输出特征的空间维度大小，而通道维度变为特征向量 $\boldsymbol{m}$ 和 $\boldsymbol{n}$ 通道维数的乘积 $t_1\times t_2$，双线性特征组合示意图如图 2 所示。

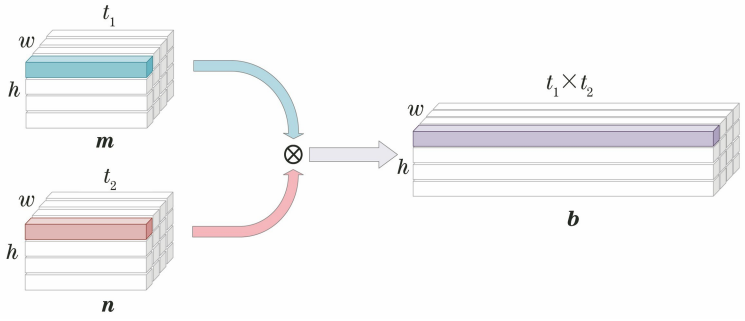


图 2 双线性特征组合示意图

Fig. 2 Schematic of bilinear feature combination

为了进一步获得图像描述子，池化函数 $P$ 对图像中各位置的双线性特征进行聚合以获取图像的全局表示，一种池化方式是将所有的双线性特征进行累加求和，即

$$\varphi(\boldsymbol{I})=\sum_{l\in L}\boldsymbol{b}(\boldsymbol{l},\boldsymbol{I},f_A,f_B)=\sum_{l\in L}\boldsymbol{m}^{\mathrm{T}}\boldsymbol{n},\tag{2}$$

式中：池化函数 $P$ 通过 $\varphi(\boldsymbol{I})$ 将双线性特征 $\boldsymbol{b}$ 转化为一个 $t_1t_2\times1$ 维的列向量，记为 $\boldsymbol{x}$，将 $\boldsymbol{x}$ 进行带符号的开平方运算 $\mathrm{sign}(\boldsymbol{x})\sqrt{|\boldsymbol{x}|}$ 得到向量 $\boldsymbol{y}$，对其进行 L2 正则化约束 $\boldsymbol{y}/\|\boldsymbol{y}\|_2$ 后得到最终表示向量 $\boldsymbol{z}$，最后将 $\boldsymbol{z}$ 输入到分类函数 $C$ 中完成分类。

## 3 本文方法

### 3.1 特征函数选取

上述 B-CNN 模型的特征函数，即骨干网络部分主要由两路 VGGNet 组成，作为通用的分类网络，VGGNet 虽然具有一定特征表示能力，但在细粒度图像分类问题上，对于判别性局部特征的提取存在一定局限性[13]；此外，VGGNet 较多的网络参数耗费了大量计算资源，导致更高的内存占用率，使得模型在速度和精度方面都有所限制，影响了实用性。

随着深度卷积神经网络的发展，网络深度对图像分类准确率有着重要影响。通常情况下，当网络层数较少时，增加深度可以得到更好的特征提取效果，提升分类精度；但当层数较多时（如超过 30 层），继续增加深度会带来更高的训练和测试误差，使得网络训练时难以收敛，反而降低准确率[14]。误差升高的主要原因是增加层数时会出现梯度消失和梯度爆炸[15-16]现象，特别是梯度消失的问题，使得梯度在反向传播时无法有效地更新至浅层网络进行权重调整。针对以上问题，He 等[17]提出一种深度残差网络（ResNet），相较于其他卷积神经网络，ResNet 采用一种残差学习结构将原始输入信息通过跳跃连接方式直接传输至下一层网络，同时梯度在反向传播时，也是通过跳跃连接直接传递至上一层。残差网络的基本组成结构为残差单元，图 3 为残差单元结构示意图。

设 $x$ 为残差单元的输入，$H(x)$ 为残差单元的期望输出，若将 x 直接传至输出部分作为初始结果，则此时网络只需要学习 $F(x)=H(x)-x$ 即可，这就是 ResNet 的一个基本残差单元。通过这种残差单元结构，ResNet 相当于将学习目标由完整的输出值 $H(x)$ 改变为输出值与输入值之差 $H(x)-x$，简化了网络学习目标，降低了学习难度。ResNet 的提出有效地解决了深层网络梯度消失的问题，使得分类准确率大幅提升，并具有良好的可移植性。相比 B-CNN 模型的特征函数 VGG-M 和 VGG-16，ResNet 具有更深的网络结构，能更加精细化地学习细粒度图像中的局部特征，提升分类精度。

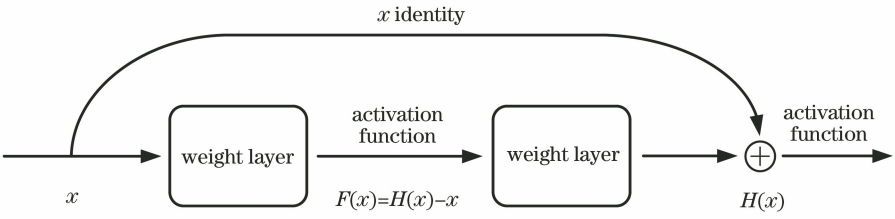


图 3 残差单元结构

Fig. 3 Structure of residual unit

因此，本文选取 ResNet-34 取代 B-CNN 中的两路 VGGNet 作为改进后模型的特征函数部分，ResNet-34 包含 conv1～conv5 五组卷积块共 33 个卷积层和 1 个全连接层，一共 34 层。将两路 ResNet-34 去掉最后的全连接层作为模型的骨干网络，网络最后一层卷积层的输出特征维数为 512，相较原 B-CNN 模型，改进网络在增加深度的同时保持了相同的输出特征维度，避免了双线性组合后特征维度的成倍递增。

### 3.2 引入注意力机制

注意力机制是图像分类与识别领域常见的特征强化策略，源于对人类大脑特有的视觉信号处理机制的模拟。人们在观察和识别物体时，会有针对性地将注意力集中在目标的显著部位而忽略一些全局和背景信息，这种选择性关注的机制恰与细粒度图像分类任务中依赖判别性部位的特点相一致，近年来已在细粒度图像分类领域得到了广泛应用。因此，为了进一步提取判别性部位特征，在 3.1 节改进后的网络中引入一种多维注意力机制——采用 CBAM(Convolutional Block Attention Module)算法[18]在骨干网络的两路特征函数中分别提取通道和空间两个维度的注意力权重图，并将权重分布于原特征图进行特征融合，将融合后的通道注意力与空间注意力模块分别添加至第一路网络 conv4 和 conv5 与第二路网络 conv2 和 conv3 卷积块之间，以获取不同维度、更为丰富的注意力特征。

#### 3.2.1 通道注意力模块

由特征函数生成的卷积特征图中，包含不同的特征通道，在细粒度图像分类问题中，每个特征通道可能表示图像中不同的信息，其中一些通道包含无关的图像背景信息，存在冗余。因此，将注意力集中在包含判别性部位信息的特征通道上，赋予其更高的权重分布，能够有效地提升细粒度分类效果[19]，本文在 3.1 节改进后网络的第一路特征函数 conv4 和 conv5 两组卷积块之间添加通道注意力模块，通道注意力模块的结构如图 4 所示。

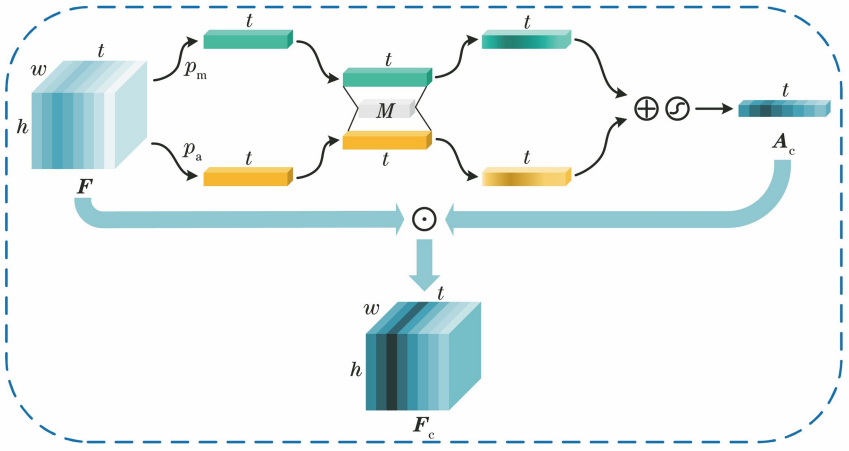


图 4 通道注意力模块

Fig. 4 Channel attention module

通道注意力模块的特征提取与融合步骤如下：

1) 将特征函数 $f_A$ 生成的卷积特征图作为原始输入 $\boldsymbol{F}$，设 $\boldsymbol{F}\in\mathbf{R}^{w\times h\times t}$，其中 $w\times h$ 表示 $\boldsymbol{F}$ 的空间维度大小，$t$ 表示通道数量。为了有效提取通道注意力，将 $\boldsymbol{F}$ 在空间维度上进行压缩，同一通道的特征被压缩为一个实数，这一步可通过池化操作实现。

2) 采取一种多尺度的池化方式，分别使用最大池化函数 $p_m$ 和平均池化函数 $p_a$ 对 $\boldsymbol{F}$ 进行降

维，得到两个 $1\times1\times t$ 大小的特征向量，将两个向量输入同一个共享网络中以获取通道维度的注意力权重分布，共享网络由包含一个隐藏层的多层感知机组成。

3）将重新分配注意力权重后的两个输出向量进行对应元素求和运算，并使用 Sigmod 激活函数对合并后的特征向量进行映射，生成通道注意力权重 $\boldsymbol{A}_c$，$\boldsymbol{A}_c\in\mathbf{R}^{1\times1\times t}$。

4）将注意力权重 $\boldsymbol{A}_c$ 与原特征图 $\boldsymbol{F}$ 进行特征融合，这里采用一种对应元素相乘的融合方法，最终得到融合后的注意力特征图 $\boldsymbol{F}_c$，$\boldsymbol{F}_c\in\mathbf{R}^{w\times h\times t}$，用 $\boldsymbol{F}_c$ 替换 $f_A$ 中的原始输入特征 $\boldsymbol{F}$，实现通道维度的注意力提取。

步骤 1）～4）的通道注意力提取及融合过程可表示为

$$\boldsymbol{A}_c=\sigma\{M[p_m(\boldsymbol{F})]\oplus M[p_a(\boldsymbol{F})]\},\tag{3}$$

$$\boldsymbol{F}_c=\boldsymbol{F}\odot\boldsymbol{A}_c,\tag{4}$$

式中：$M$ 表示包含多层感知机的共享网络；$\oplus$ 表示向量的对应元素求和运算；$\sigma$ 表示 Sigmod 激活函数；$\odot$ 表示向量的对应元素相乘运算。

### 3.2.2 空间注意力模块

不同于通道注意力模块，空间注意力模块更加侧重于关注判别性部位的空间位置信息，是对通道注意力的一种补充，在 3.1 节改进后网络的第二路特征函数 conv2 和 conv3 之间添加空间注意力模块，空间注意力模块的结构如图 5 所示。

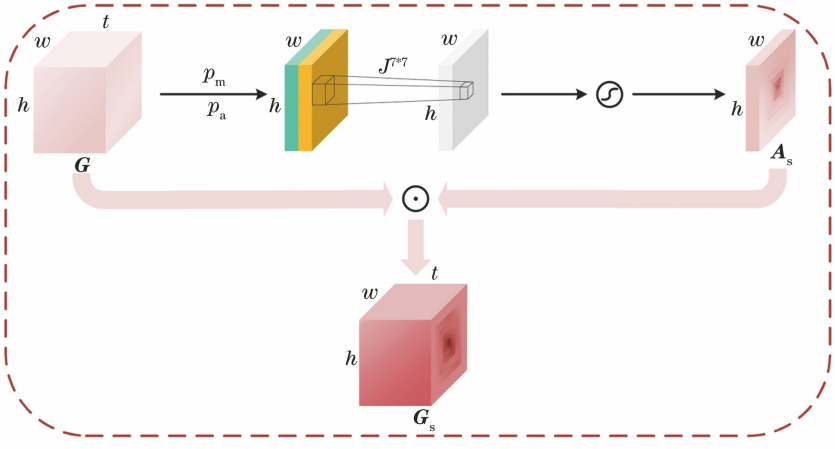


图 5 空间注意力模块

Fig. 5 Spatial attention module

空间注意力模块的特征提取与融合步骤如下：

1）将特征函数 $f_B$ 生成的卷积特征图作为原始输入 $\boldsymbol{G}$，$\boldsymbol{G}\in\mathbf{R}^{w\times h\times t}$，其中 $w\times h$ 表示 $\boldsymbol{G}$ 的空间维度大小，$t$ 表示通道数量，将 $\boldsymbol{G}$ 沿通道轴方向进行压缩提取空间注意力信息，一列通道值被压缩为一个通道，这一步通过通道维度的池化实现。

2）同样采取多尺度的池化方式，用最大池化函数 $p_m$ 和平均池化函数 $p_a$ 对 $\boldsymbol{G}$ 进行降维，得到两个 $w\times h\times1$ 大小的特征图，将两个特征图沿通道轴方向使用对应元素求和方法进行拼接，得到一个 $w\times h\times2$ 大小的新特征图。

3）使用一个 $7\times7$ 的卷积核对拼接后的特征图进行卷积，再次将其大小压缩为 $w\times h\times1$，对卷积后的特征图使用 Sigmod 激活函数进行映射，生成空间注意力图 $\boldsymbol{A}_s$，$\boldsymbol{A}_s\in\mathbf{R}^{w\times h\times1}$。

4）最后将空间注意力图 $\boldsymbol{A}_s$ 与原特征图 $\boldsymbol{G}$ 使用对应元素点乘方法进行特征融合，得到融合后的空间注意力特征图 $\boldsymbol{G}_s$，$\boldsymbol{G}_s\in\mathbf{R}^{w\times h\times t}$，用 $\boldsymbol{G}_s$ 替换 $f_B$ 中的原始输入特征 $\boldsymbol{G}$，实现空间维度的注意力提取。

步骤 1）～4）的空间注意力提取及融合过程可表示为

$$\boldsymbol{A}_s=\sigma\{J^{7*7}[p_m(\boldsymbol{G})\oplus p_a(\boldsymbol{G})]\},\tag{5}$$

$$\boldsymbol{G}_s=\boldsymbol{G}\odot\boldsymbol{A}_s,\tag{6}$$

式中：$J^{7*7}$ 表示使用 $7\times7$ 大小的卷积核进行卷积运算。

在添加两个不同维度的注意力模块后，网络获取了更为丰富的注意力特征。本文改进网络 BRAN 模型的残差注意力结构如图 6 所示。

上述两点改进方法，使得改进后模型 BRAN 在很大程度上解决了原有 B-CNN 的一些限制。相较原模型的特征函数 VGGNet，改进后模型 BRAN 的特征函数 ResNet 具有更强的局部特征提取能力，同时更少的网络参数使得模型更加容易被训练，减少过拟合现象；同时，通过在残差网络中加入不同维度的注意力模块，获取更为丰富的信息，使得改进模

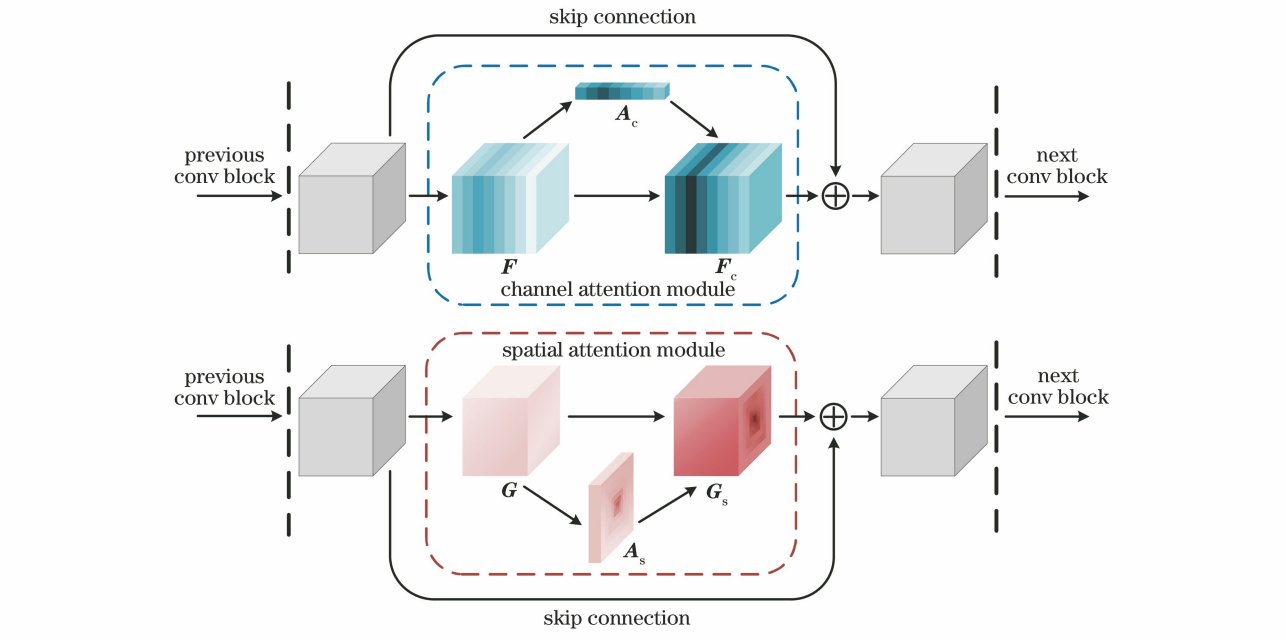


图 6 BRAN 模型的残差注意力结构

Fig. 6 Residual attention structure of BRAN model

型 BRAN 相较原模型能够关注和学习细粒度图像中更具有判别性的局部特征，而这正是影响细粒度分类性能的关键。

## 4 实验与分析

### 4.1 数据集选取与预处理

为了评估本文改进方法的有效性，选取 3 个常用的细粒度图像公开数据集进行实验，分别是加州理工学院鸟类数据库 CUB-200-2011[9]、斯坦福大学狗类数据集 Stanford Dogs[10] 和斯坦福大学汽车数据集 Stanford Cars[11]。其中 CUB-200-2011 是细粒度图像分类领域最常用的数据集，它包含 200 类共 11788 张北美鸟类图像，Stanford Dogs 和 Stanford Cars 则分别包含 120 类和 196 类不同的狗类和汽车图像。3 个数据集的原始信息如表 1 所示（Train 和 Test 分别表示不同数据集包含的初始训练及测试图像数量）。

表 1 三个细粒度图像数据集的详细信息

Table 1 Detailed statistics of three fine-grained image datasets

| Dataset | Class | Train | Test | Total |
|---|---|---|---|---|
| CUB-200-2011 | 200 | 5994 | 5794 | 11788 |
| Stanford Dogs | 120 | 12000 | 8580 | 20580 |
| Stanford Cars | 196 | 8144 | 8041 | 16185 |

由于 CUB-200-2011 鸟类数据集中每一类包含的原始训练图像只有 50～60 张，为了提高模型的泛化能力和鲁棒性，对训练图像进行数据增强，主要包括旋转、翻转和随机裁剪等。为了保证随机裁剪不会影响到原图中的判别性特征，通过分析验证，选取 5% 为阈值将训练集图像分别从上、下、左、右进行裁剪。经由旋转、翻转及随机裁剪后的图像数量约为原训练集的 7 倍，再选取其中约 1/7 作为模型的验证集，用来训练模型的超参数、预估模型的泛化能力。图 7 展示了训练集数据增强示例（示例中图像来源 CUB-200-2011 中的黑嘴美洲鹃）。

### 4.2 模型训练

本文使用开源深度学习框架 PyTorch[20] 作为平台，在 3 个细粒度图像数据集上使用 2 个 NVIDIA Quadro P5000 GPU 通过随机梯度下降法进行并行训练。由于细粒度图像数据集的规模较小，训练和测试图像有限，直接在 3 个细粒度图像数据集上进行训练可能会导致网络无法收敛，因此使用 ImageNet 数据集上预训练的 ResNet-34 网络参数进行初始化，再在 3 个细粒度图像数据集上进行模型微调。模型使用 Adam[21] 优化器来训练和优化网络，学习率 $\alpha$ 设置为 0.001，一阶和二阶矩估计指数衰减率分别设置为 0.9 和 0.99，训练批尺寸设置为 64。

在测试阶段，采用分类准确率作为结果评价指标，它是图像分类最常用的评价指标之一，定义为正确分类的图像占总数的比例，

图 7 训练集数据增强示例

Fig. 7 Example of training data augmentation

$$A_{\text{accuracy}}=\frac{I_{\text{c}}}{I}, \tag{7}$$

式中：$I_{\text{c}}$ 代表分类正确的图像数量；I 代表测试图像总数。

### 4.3 实验与结果分析

为了综合验证本文方法的有效性，能够较好地提升 B-CNN 模型及其他主流细粒度分类算法的分类结果，在本节进行了如下几个实验。

#### 4.3.1 消融实验

在本节的消融实验方案中，对比了以下几个双线性网络结构：1）用 B-CNN（baseline）表示原双线性网络模型；2）用 B-CNN（resnet×2）代表 3.1 节替换特征提取函数后的双线性网络；3）用 BRAN（cha. attention）表示在 2）中仅添加 3.2.1 节通道注意力模块后的双线性网络；4）用 BRAN（spa. attention）表示在 2）中仅添加 3.2.2 节空间注意力模块后的双线性网络；5）用 BRAN（cha.& spa attention）表示同时在两路特征函数中分别添加通道注意力模块和空间注意力模块后的双线性网络。实验结果如表 2 所示（Backbone 表示不同方法所使用的基础网络）。

表 2 本文方法在 CUB-200-2011 数据集上的消融实验分析

Table 2 Ablation experiment and analysis of proposed method on CUB-200-2011 dataset

| Approach | Backbone | Accuracy /% |
|---|---|---|
| B-CNN(baseline) | VGG-M+VGG-D | 84.1 |
| B-CNN(resnet×2) | ResNet-34×2 | 85.0 |
| BRAN(cha. attention) | ResNet-34×2 + channel attention | 86.2 |
| BRAN(spa. attention) | ResNet-34×2 + spatial attention | 85.5 |
| BRAN(cha.& spa. attention) | ResNet-34×2 + cha. & spa. attention | 87.2 |

从表 2 的结果可以看出，在仅添加通道注意力模块、空间注意力模块和同时添加两个模块后的网络比原双线性网络模型的分类准确率分别提高了 2.1%、1.4%和 3.1%，同时添加两个注意力模块后网络的分类准确率达到了最高。

#### 4.3.2 对比实验

1）不同数据集测试结果对比

选用 4.3.1 节同时添加通道和空间注意力模块后的网络 BRAN（cha. & spa attention）为本文最终的分类模型，并在其他两个数据集上进行测试，在 BRAN（cha. & spa attention）模型两路特征函数提取的注意力特征图进行双线性组合前，分别将其通道注意力模块和空间注意力模块生成的注意力图进行可视化，同时也对基准模型 B-CNN 最后一个卷积层的特征图进行可视化，可视化结果如图 8 所示。

其中，不同行代表了不同的数据集，对于每个数据集，图 8（a）为原始输入图像，图 8（b）为 B-CNN 最后一个卷积层特征图的可视化结果，图 8（c）和（d）分别为 BRAN 模型通道和空间模块的注意力图可视化结果。可以看出，在添加通道注意力模块和空间注意力模块后的网络在不同数据集上均有较好的表现。

2）不同弱监督算法分类结果对比

本文在弱监督分类模型 B-CNN 的基础上进行改进，无需物体标注框和部件位置等额外标注信息，仅使用类别标签实现了基于弱监督信息的分类模型，选取两级注意力（Two-level attention）模型[22]、NAC[23]、B-CNN[6]、ST-CNN[24]、DVAN[25]、RA-CNN[26]、MA-CNN[19] 和 MAMC[27] 等近年来主流弱监督分类算法，将本文方法分别在三个数据集上的实验结果和以上方法进行对比，结果如表 3 所示（Backbone 表示模型所使用的基础网络，Accuracy 代表分类准确率）。

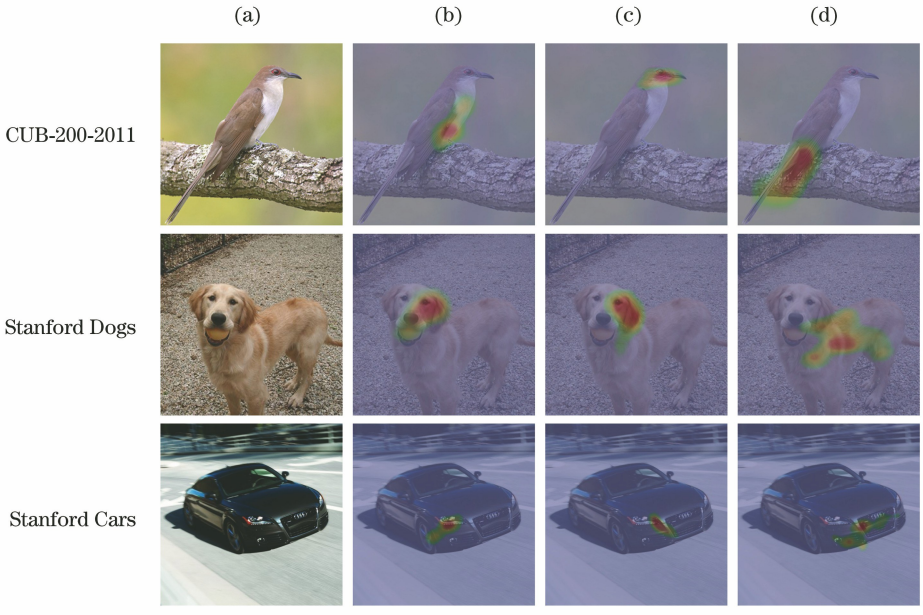


图 8 特征图可视化结果对比。(a)原始输入图像;(b) B-CNN;(c)通道注意力模块;(d)空间注意力模块

Fig. 8 Visualization of different feature maps. (a) Original images; (b) B-CNN; (c) channel attention maps; (d) spatial attention maps

表 3 不同的弱监督算法分类准确率对比

Table 3 Comparison with weakly-supervised methods in terms of classification accuracy

| Approach | Backbone | Accuracy /% | | |
|---|---|---|---|---|
| | | Birds[9] | Dogs[10] | Cars[11] |
| Two-level attention[22] | VGG19 | 77.9 | - | - |
| NAC[23] | VGG19 | 81.01 | 68.61 | - |
| B-CNN[6] | VGG-M+VGG-D | 84.1 | - | 91.3 |
| ST-CNN[24] | Inception-v2×3 | 84.1 | - | - |
| DVAN[25] | VGG-16×3 | 79.0 | 81.5 | 87.1 |
| RA-CNN[26] | VGG-19×3 | 85.3 | 87.3 | 92.5 |
| MA-CNN[19] | VGG-19×3 | 86.5 | - | 92.8 |
| MAMC[27] | ResNet-101 | 86.5 | 85.2 | **93.0** |
| BRAN | ResNet-34×2 | **87.2** | **89.2** | 92.5 |

由表中结果可以看出,本文方法在 CUB-200-2011 和 Stanford Dogs 两个数据集上的分类效果均优于近年来主流的弱监督方法,在 Stanford Cars 上的分类准确率接近 MAMC 和 MA-CNN 模型,与 RA-CNN 模型持平,达到了 92.5%,相较原 B-CNN 模型提高了 1.2%。结果表明,本文在添加通道注意力和空间注意力模块后的网络 BRAN 能聚焦细粒度图像中的判别性部位,提升局部特征的提取能力,并在多个细粒度图像数据集上取得了良好的分类效果。

## 5 结 论

本文在 B-CNN 模型的基础上,提出了改进的双线性残差注意力网络模型,通过替换原有的特征提取函数,并在残差单元间添加通道注意力模块和空间注意力模块,实现了多维注意力机制的引入,增强了细粒度图像判别性局部特征的提取能力。在多个细粒度图像数据集上进行消融与对比实验,结果表明,本文方法能有效提升原 B-CNN 模型性能,且优于大多数近年来主流弱监督算法的分类准确率。另一方面,由于双线性特征向量外积的组合方式会极具增大特征维度,耗费计算资源,因此,在尽量不损失分类精度的同时对双线性特征进行降维,提高模型的实用性,是本文后续工作的进展方向。

## 参 考 文 献

[1] Zhao B, Feng J S, Wu X, et al. A survey on deep learning-based fine-grained object classification and

semantic segmentation[J]. International Journal of Automation and Computing, 2017, 14(2): 119-135.

[2] Zhang N, Donahue J, Girshick R, et al. Part-based R-CNNs for fine-grained category detection[M]// Computer Vision - ECCV 2014. Cham: Springer International Publishing, 2014: 834-849.

[3] Huang S L, Xu Z, Tao D C, et al. Part-stacked CNN for fine-grained visual categorization[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016. Las Vegas, NV, USA. IEEE, 2016: 1173-1182.

[4] Lam M, Mahasseni B, Todorovic S. Fine-grained recognition as HSnet search for informative image parts[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017. Honolulu, HI. IEEE, 2017: 2520-2529.

[5] Wei X S, Xie C W, Wu J X, et al. Mask-CNN: localizing parts and selecting descriptors for fine-grained bird species categorization[J]. Pattern Recognition, 2018, 76: 704-714.

[6] Lin T Y, RoyChowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition[C]// 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015. Santiago, Chile. IEEE, 2015: 1449-1457.

[7] Lin T Y, Maji S. Visualizing and understanding deep texture representations[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016. Las Vegas, NV, USA. IEEE, 2016: 2791-2799.

[8] Chatfield K, Simonyan K, Vedaldi A, et al. Return of the devil in the details: delving deep into convolutional nets[C]// Proceedings of the British Machine Vision Conference 2014, Nottingham. British Machine Vision Association, 2014.

[9] Wah C, Branson S, Welinder P, et al. The Caltech-UCSD Birds-200-2011 dataset[R]. Computation & Neural Systems Technical Report, CNS-TR. Pasadena, CA, USA: California Institute of Technology, 2011.

[10] Khosla A, Jayadevaprakkash N, Yao B P, et al. Novel dataset for fine-grained image categorization: Stanford dogs[C]// Proceedings of the 1st Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA: IEEE, 2011, 2(1).

[11] Krause J, Stark M, Jia D, et al. 3D object representations for fine-grained categorization[C]// 2013 IEEE International Conference on Computer Vision Workshops, December 2-8, 2013. Sydney, Australia. IEEE, 2013: 554-561.

[12] Kolda T G, Bader B W. Tensor decompositions and applications[J]. SIAM Review, 2009, 51(3): 455-500.

[13] Zhang Y. Research on the algorithm for fine-grained image classification[D]. Harbin: Harbin Institute of Technology, 2018: 22-23.
张阳. 细粒度图像分类算法研究[D]. 哈尔滨: 哈尔滨工业大学, 2018: 22-23.

[14] Zhang M, Lü X Q, Wu L, et al. Multiplicative denoising method based on deep residual learning[J]. Laser & Optoelectronics Progress, 2018, 55(3): 031004.
张明, 吕晓琪, 吴凉, 等. 基于深度残差学习的乘性噪声去噪方法[J]. 激光与光电子学进展, 2018, 55(3): 031004.

[15] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult[J]. IEEE Transactions on Neural Networks, 1994, 5(2): 157-166.

[16] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[J]. Journal of Machine Learning Research, 2010, 9: 249-256.

[17] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016. Las Vegas, NV, USA. IEEE, 2016: 770-778.

[18] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[M]// Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 3-19.

[19] Zheng H L, Fu J L, Mei T, et al. Learning multi-attention convolutional neural network for fine-grained image recognition[C]// 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017. Venice. IEEE, 2017: 5209-5217.

[20] Paszke A, Gross S, Chintala S, et al. Automatic differentiation in PyTorch[C]// In 2017 Neural Information Processing Systems Autodiff Workshop: The Future of Gradient-based Machine Learning Software and Techniques. Long Beach, CA, USA, December 9, 2017.

[21] Kingma D P, Ba J. Adam: a method for stochastic optimization[EB/OL]. [2019-08-18]. arXiv preprint arXiv: 1412.6980, 2014.

[22] Xiao T J, Xu Y C, Yang K Y, et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification [C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015. Boston, MA, USA. IEEE, 2015: 842-850.

[23] Simon M, Rodner E. Neural activation constellations: unsupervised part model discovery with convolutional networks [C] // 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015. Santiago, Chile. IEEE, 2015: 1143-1151.

[24] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks [C] // Advances in Neural Information Processing Systems (NIPS). Montreal, Canada, 2015: 2017-2025.

[25] Zhao B, Wu X, Feng J S, et al. Diversified visual attention networks for fine-grained object classification[J]. IEEE Transactions on Multimedia, 2017, 19(6): 1245-1256.

[26] Fu J L, Zheng H L, Mei T. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017. Honolulu, HI. IEEE, 2017: 4438-4446.

[27] Sun M, Yuan Y C, Zhou F, et al. Multi-attention multi-class constraint for fine-grained image recognition [M] // Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 834-850.