第 19 卷　第 36 期　2019 年 12 月
1671—1815(2019)036-0240-07

科　学　技　术　与　工　程
Science Technology and Engineering

Vol. 19　No. 36　Dec. 2019
© 2019　Sci. Tech. Engrg.

引用格式: 蓝　洁, 周　欣, 何小海, 等. 基于跨层精简双线性网络的细粒度鸟类识别[J]. 科学技术与工程, 2019, 19(36): 240-246
Lan Jie, Zhou Xin, He Xiaohai, et al. Fine-grained bird recognition based on cross-layer compact bilinear network[J]. Science Technology and Engineering, 2019, 19(36): 240-246

基于跨层精简双线性网络的细粒度鸟类识别

蓝　洁[1]　周　欣[1,2]　何小海[1*]　滕奇志[1]　卿粼波[1]
(四川大学电子信息学院[1], 成都 610065; 中国信息安全测评中心[2], 北京 100085)

摘　要　细微的类间差异和显著的类内变化使得细粒度图像分类极具挑战性。为了对鸟类图像进行细粒度识别,提出一种基于跨层精简双线性池化的深度卷积神经网络模型。首先,根据 Tensor Sketch 算法计算出多组来自不同卷积层的精简双线性特征向量; 其次,将归一化后的特征向量级联送至 softmax 分类器; 最后,引入成对混淆对交叉熵损失函数进行正则化以优化网络。提出的模型无需额外的部件标注,可进行端到端的训练。结果表明,在公开的 CUB-200—2011 鸟类数据集上,该模型取得了较好的性能,识别正确率为 86.6%,较 BCNN 提高 2.5%。与多个先进细粒度分类算法的对比,验证了提出模型的有效性和优越性。

关键词　鸟类识别　精简双线性变换　跨层特征融合　成对混淆　细粒度图像分类
中图法分类号　TP391.41;　　**文献标志码**　A

细粒度图像分类(fine-grained image classification, FGIC)旨在区分同一基础类别下不同的从属类别,如不同种类的鸟、汽车和飞机等,细粒度图像分类的难点在于,首先,不同子类别间具有高度相似的视觉内容,其次,即使是同一子类别间,由于姿态、尺度、背景等因素的干扰,变化也十分显著。为了更精准地对细粒度图像进行识别,判别性特征的获取往往需要图像的标注,如物体标注框(object bounding box)和部件标注点(part annotation),以避免复杂模型在小型数据集上的过拟合。早期细粒度分类方法[1,2],在训练时依赖额外的局部标注信息,进行强监督学习以获得更高的分类精度。例如,Part-RCNN[1]首先利用几何约束下的 R-CNN(Region-CNN)完成对象级和局部级(鸟的头、翅膀)的目标检测,再对整体区域和各局部区域提取卷积特征用于最终的分类。然而,局部标注通常需要相应领域的专家才能完成,其人工参与程度高,限制了上述方法在实际场景中的应用[3]。因此,仅需提供图像类别标签的弱监督学习方法应运而生。

2019 年 5 月 8 日收到　国家自然科学基金(61871278)、成都市产业集群协同创新项目(2016-XT00-00015-GX)、四川省科技计划(2018HH0143)和四川省教育厅项目(18ZB0355)资助

第一作者简介: 蓝　洁(1994—),硕士研究生。研究方向: 计算机视觉与模式识别。E-mail: 997815469@qq.com。
* 通信作者简介: 何小海(1964—),教授,博士研究生导师。研究方向: 图像处理与信号处理。E-mail: hxh@scu.edu.cn。

主流的基于弱监督信息的细粒度分类方法主要有两种类型。

第一种类型采用“定位”子网络辅助分类主网络的结构,通过带注意力机制的“定位”网络提供的局部信息,来增强分类网络学习细粒度特征的能力。这类方法的共同思路是不再依赖局部标注,而是利用诸如特征谱聚类[4]、空间变换[5]、循环注意力机制[6,7]等回归出具有判别性的部件位置,再将各部件送入对应的分类网络。这种基于部件的方法充分利用局部级判别性特征,在细粒度数据集上取得了较高的分类准确率。然而,这类方法在实际训练过程中通常会涉及两个网络的交替优化或者需要单独训练两个网络,然后再进行联合调整。这种交替、多阶段的训练策略使整合网络变得复杂[8]。

第二种类型是端到端的特征编码[9—11],通过对特征映射后的高阶统计量进行编码,以获得图像的鲁棒性表示。这类方法首先提取图像的 SIFT(scale invariant feature transform)、HOG(histogram of oriented gradients)等人工特征或深度卷积特征,再利用 VLAD(vector of locally aggregated descriptors)、Fisher 矢量等编码模型对其进行编码。端到端的特征编码以平移不变的方式模拟局部特征交互,这特别适用于纹理和细粒度识别任务。这类方法的优点是简单,不依赖手工标记,但由于对部件的描述不够精准,其性能稍亚于基于部件的方法。

类型二中，Lin 等[9]提出的双线性网络(bilinear convolutional neural network，BCNN)备受关注，它包含两个独立的卷积神经网络，分别模拟图像的位置和外观，通过外积运算融合两组特征图谱后再送入分类器，其分类准确率能与基于部件的模型媲美。然而，经外积操作后，BCNN 的特征维度变为原始维度的平方，通常高达数十万甚至数百万，这极大地增加了后续分析的难度。因此 Gao 等[10]提出了一种精简双线性池化(compact bilinear pooling，CBP)方法，用多项式核函数近似来实现一种轻量级、低维度的双线性表示，在大幅降低特征维度的同时，CBP 的分类准确率仍能与 BCNN 持平。

研究现状表明，更强大的特征描述和特征编码方式对细粒度图像分类至关重要。为此，以鸟类为研究对象，在精简双线性变换的基础上，提出跨层双线性特征融合方法，并引入成对混淆防止过拟合，以期提升细粒度鸟类识别率。

1　精简双线性特征

BCNN 由两个基于 CNN 的特征提取器组成，一个作为部件检测器，另一个为局部特征提取器。记上述两个特征提取器得到的特征图为 $\boldsymbol{F}_1$、$\boldsymbol{F}_2$，且 $\boldsymbol{F}_1$、$\boldsymbol{F}_2 \in R^{c\times H\times W}$，$H$、$W$、$c$ 对应特征图的长、宽和通道数，即在位置(h,w)处分别存在 c 维的特征向量 $\boldsymbol{f}_1(h,w)$，$\boldsymbol{f}_2(h,w) \in R^c$，其中 $h \in [1,H]$，$w \in [1,W]$。

BCNN 首先计算两组特征图在每个位置的外积，再对所得结果求和池化，即可得到整个图像的双线性特征，可表示为

$$\boldsymbol{B}(f_1,f_2) = \sum_{(h,w)} \boldsymbol{f}_1(h,w) \otimes \boldsymbol{f}_2(h,w) \tag{1}$$

式(1)中：$\otimes$ 为外积运算，可等价为 $\boldsymbol{f}_1^{\mathrm{T}}\boldsymbol{f}_2$；$\boldsymbol{B}(\boldsymbol{f}_1,\boldsymbol{f}_2)$ 为 $c\times c$ 的特征矩阵，可展开成维度为 c^2 的特征向量。若 $c=512$，其维度高达 26 万，计算较为复杂。

基于双线性特征的图像分类器通常使用支持向量机或逻辑回归来实现，这可被视为一种线性核机制。因此，CBP 将双线性变换看成一种多项式核函数，根据 Tensor Sketch 算法近似地计算 d 维精简双线性特征向量，其原理如图 1 所示。

Tensor Sketch 算法主要步骤如下。

首先，利用 Count Sketch 函数 $\boldsymbol{\Psi}$ 将特征向量 $\boldsymbol{f}_k \in R^c$ 映射到特征空间，$k=1,2$。定义两个随机向量 $\boldsymbol{s}_k \in \{-1,1\}^c$，$\boldsymbol{h}_k \in \{1,2,\cdots,d\}^c$，$s_k(i)$ 和 $h_k(i)$ 的初始化服从均匀分布，且在后续运算中两者的值固定不变。$\boldsymbol{h}_k$ 用于寻找 $\boldsymbol{f}_k$ 的第 i 个元素 $f_k(i)$ 在

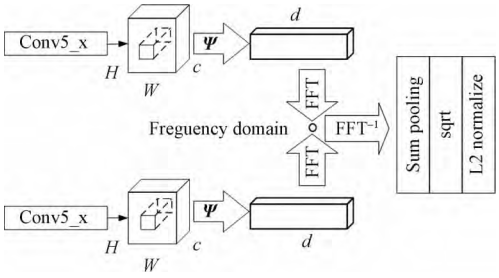

图 1　精简双线性模块

Fig. 1　Compact bilinear module

特征空间中对应的索引 $j=h_k(i)$，则有：

$$\boldsymbol{\Psi}(\boldsymbol{f}_k,\boldsymbol{h}_k,\boldsymbol{s}_k) = \{Q_1,Q_2,\cdots,Q_d\} \tag{2}$$

$$Q_j = \sum_{i,j=h_k(i)} s_k(i) f_k(i) \tag{3}$$

式(3)中：$i \in \{1,2,\cdots,c\}$；$j \in \{1,2,\cdots,d\}$。

其次，Tensor Sketch 算法指出，可通过计算两个向量 Count Sketch 的卷积得到两个向量外积的 Count Sketch，可表示为

$$\boldsymbol{\Psi}(\boldsymbol{f}_1\otimes\boldsymbol{f}_2,\boldsymbol{h}_k,\boldsymbol{s}_k) = \boldsymbol{\Psi}(\boldsymbol{f}_1,\boldsymbol{h}_1,\boldsymbol{s}_1) * \boldsymbol{\Psi}(\boldsymbol{f}_2,\boldsymbol{h}_2,\boldsymbol{s}_2) \tag{4}$$

同时，卷积定理指出，时间域中的卷积等价为频率域中的乘积。于是，式(4)可表示为

$$\boldsymbol{\Psi}(\boldsymbol{f}_1\otimes\boldsymbol{f}_2,\boldsymbol{h},\boldsymbol{s}) = F^{-1}\{F[\boldsymbol{\Psi}(\boldsymbol{f}_1,\boldsymbol{h},\boldsymbol{s})]^{\circ}F[\boldsymbol{\Psi}(\boldsymbol{f}_2,\boldsymbol{h},\boldsymbol{s})]\} \tag{5}$$

式(5)中：F 表示快速傅里叶变换，F^{-1} 表示傅里叶逆变换，$^{\circ}$ 表示逐元素相乘。

最后，对双线性特征向量 $\boldsymbol{x}=\boldsymbol{\Psi}(\boldsymbol{f}_1,\boldsymbol{f}_2)$ 进行归一化。即通过开符号平方根($\boldsymbol{y} \leftarrow \operatorname{sign}\boldsymbol{x}\sqrt{|\boldsymbol{x}|}$)后，再 L_2 规范化$\left(z \leftarrow \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|_2}\right)$。

由文献[10]可知，当 d 取值 8 192 时，CBP 能达到和 BCNN 相当的表征能力，并大幅降低了运算量。

2　跨层精简双线性网络

2.1　跨层特征融合

由于深度学习的成功，CNN 已被广泛用作视觉识别的通用特征提取器。在卷积神经网络中，较深卷积层中的每个空间单元对应于特定的感受野，图像中的判别语义信息往往出现在不同的尺度上。因此一些工作致力于挖掘不同卷积层特征的有效性。例如，Long 等[12]在全卷积网络中结合了中层和高层卷积特征，以提供更精细的细节和更高级别的语义，实现了更精准的图像分割。

如图 2 所示，BCNN[D，D]采用了卷积层参数完全共享的模式[13]。即 F_1、F_2 均只采用 VGG-16 第 5 段卷积层 conv5_3 输出的特征图谱，然后再进行双线性变换。

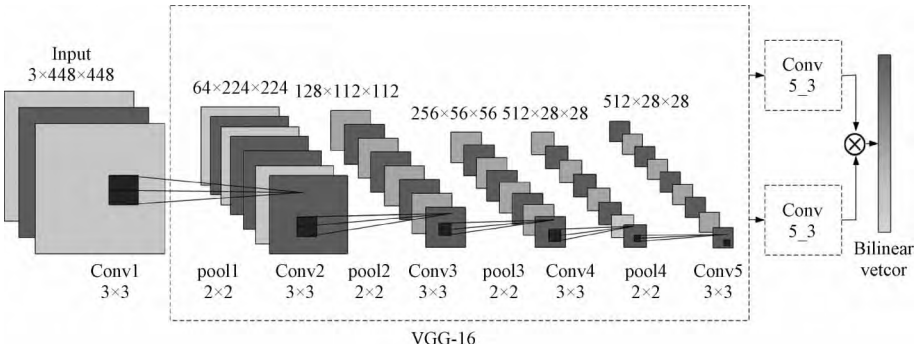

图 2 BCNN[D,D]的参数完全共享模式

Fig. 2 Parameter fully-shared mode of BCNN[D,D]

然而,忽略中间卷积层特征,仅仅利用最后单个卷积层输出的特征图谱,会导致细粒度图像判别性信息的丢失。为研究 CNN 内不同卷积层特征的有效性,采用 Grad-CAM[14] 方法对 VGG-16 模型中 conv5_1、conv5_2 及 conv5_3 的响应热力图进行了可视化,结果如图 3 所示。选取第 5 组卷积是因为与较低卷积层相比,它们在捕获语义部件信息方面具有良好的表现力。

由图 3 可观察到,不同卷积层对输入图像中各部件的判别性各不相同,携带着不同的语义信息。如图 3 中第一行图片,conv5_1 对黑脚信天翁的尾部、头部和翅膀均有较强的响应,而 conv5_3 仅保留了对头部的激活响应。受此观察的启发,为更好地捕获层间特征关系,提出了一种跨层双线性特征融合方法,如图 4 所示。将 conv_i,conv_j,…,conv_n 层的特征图分别记为 $\boldsymbol{F}_i$,$\boldsymbol{F}_j$,…,$\boldsymbol{F}_n$,则融合后的特征 $\boldsymbol{F}$ 可表示为

$$\boldsymbol{F} = \text{concat}(\boldsymbol{F}_i \odot \boldsymbol{F}_j, \boldsymbol{F}_i \odot \boldsymbol{F}_n, \boldsymbol{F}_j \odot \boldsymbol{F}_n, \cdots) \quad (6)$$

式(6)中:⊙表精简双线性运算。这种融合方式充分考虑了层间特征交互,利用来自多个卷积层的特征图谱,提取多组双线性特征。在最终分类之前对提取到的特征进行级联,以增强特征提取器的表示能力。

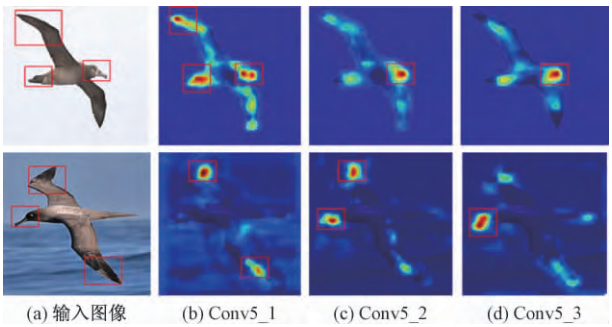

图 3 各高层卷积的不同激活响应

Fig. 3 Different activation responses for each high-level convolution layers

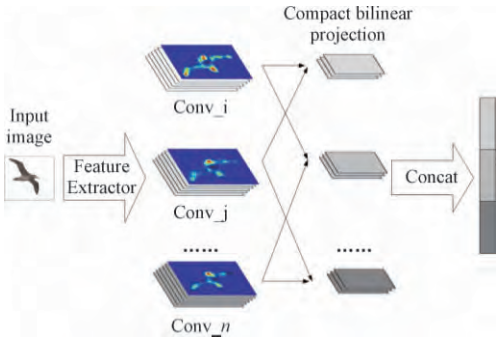

图 4 跨层特征融合示意图

Fig. 4 Cross-layer feature fusion diagram

2.2 跨层精简双线性网络结构

基于 2.1 节所提特征融合方式,构建了一种跨层精简双线性池化网络(cross-layer compact bilinear pooling,CL-CBP)用于细粒度鸟类识别。与仅利用来自单个卷积层特征的 BCNN 和 CBP 相比,CL-CBP 将卷积层视为部分属性提取器,利用了来自多个层的层间特征交互。提出的 CL-CBP 结构组成如图 5 所示,选取 VGG-16 作为特征提取器,并选取了来自 Conv5 的特征图谱,通过精简双线性变换后得到三组维度为 8 192 的特征向量,再将其级联后送入 softmax 分类器。

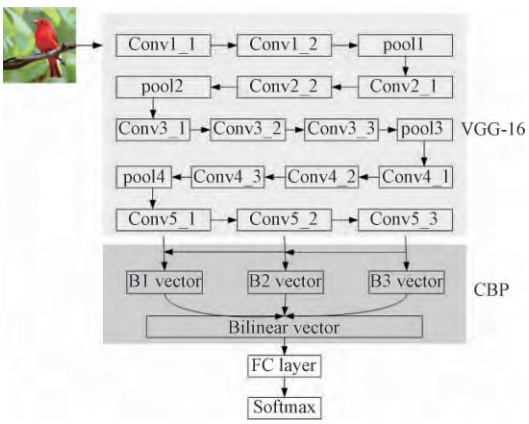

图 5 CL-CBP 网络架构

Fig. 5 Architecture of CL-CBP network

3　改进的损失函数

与 LSVC(large scale visual classification) 任务类似，FGIC(fine-grained image classification) 任务通常使用交叉熵损失作为目标函数。交叉熵损失在 LSVC 任务中的优良表现得益于 ImageNet 数据集中显著的类间差异，这使得神经网络能够在大量数据中学习到广义判别性特征。

然而，如图 6 所示，由于各子类别类间差异微小、类内差异较大，交叉熵损失函数对 FGIC 任务可能并不理想。假定训练集中的两个样本具有非常相似的视觉内容但却具有不同的类别标签，最小化交叉熵损失将迫使神经网络去挖掘能以高置信度区分这两个图像的特征，如背景的树木，这非常容易导致模型对数据集的过拟合。

图 6　CUB-200-2011 类内差异和类间差异

Fig. 6　Differences between intra-class and inter-class of CUB-200—2011

为解决上述问题，引入一种成对混淆(pairwise confusion，PC) [15] 的方法，在最小化交叉熵损失的同时，增加随机样本对之间预测概率向量的欧氏距离作为惩罚项，以制约不同类别图像特征向量之间的距离。这可被视为一种正则化方法，引导神经网络去学习训练集样本所共有的特征，能有效提升模型泛化能力。

具体地，对于输入的两组训练样本 (x_1, y_1)，(x_2, y_2)，CL-CBP 网络的损失函数 L_{total} 定义为

$$L_{\text{total}}(x_1, x_2, y_1, y_2; \theta) = \sum_{i=1}^{2} L_{\text{CE}}[\boldsymbol{p}_\theta(y|x_i), y_i] + \lambda\gamma(y_1, y_2) D_{\text{EC}}[\boldsymbol{p}_\theta(y|x_1), \boldsymbol{p}_\theta(y|x_2)] \tag{7}$$

$$\gamma(y_1, y_2) = \begin{cases} 1, & y_1 \neq y_2 \\ 0, & y_1 = y_2 \end{cases} \tag{8}$$

式中：θ 为 CL-CBP 网络模型参数；λ 为超参数，经实验验证，$\lambda = 20$ 时性能最佳；$\boldsymbol{p}_\theta(\boldsymbol{y}|\boldsymbol{x})$ 为 softmax 分类器输出的预测概率分布向量；L_{CE} 表示交叉熵损失：

$$L_{\text{CE}}[\boldsymbol{p}_\theta(y|x_i, y_i)] = -y_i \lg \frac{\boldsymbol{p}_\theta(y|x_i)}{y_i} \tag{9}$$

D_{EC} 表示欧式距离：

$$D_{\text{EC}}[\boldsymbol{p}_\theta(y|x_1), \boldsymbol{p}_\theta(y|x_2)] = \sum_{i=1}^{2} [\boldsymbol{p}_\theta(y_i|x_1) - \boldsymbol{p}_\theta(y_i|x_2)]^2 = \|\boldsymbol{p}_\theta(y|x_1) - \boldsymbol{p}_\theta(y|x_2)\|_2^2 \tag{10}$$

CL-CBP 网络的损失函数计算过程如图 7 所示。训练时，将一个训练 batch 中的样本对随机划分为两组，然后分别输入到权重共享的孪生网络，各网络单独计算交叉熵损失，两个网络之间根据 γ 值判断是否存在混淆损失；测试时，仅使用一个网络前向传播。

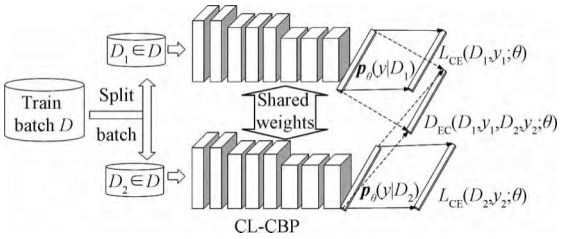

图 7　CL-CBP 网络损失函数计算过程

Fig. 7　Calculation process of loss function for CL-CBP

4　实验结果及分析

4.1　数据集

CUB-200-2011 是细粒度图像分类中极具代表性和挑战性的鸟类图像数据集，该数据集包含 200 种鸟类，比如美洲黑鸦、白冠麻雀等。典型的数据划分方式为：训练集 5 994 张图片，测试集 5 794 张图片，共计 11 788 张图片。

4.2　实验内容

实验环境为 Ubuntu 16.04 操作系统，基于 Caffe 深度学习框架，使用的显卡为 NVIDIA Geforce GTX 1080 TI，显存为 11 GB，使用 CUDA 8.0 和 cuDNN 加速。

为与基准模型做公平对比，CL-CBP 网络特征提取器同样选取 VGG-16，并去掉第五个池化层 pool5 以及 fc6、fc7、fc8 三个全连接层。首先对数据集做预处理，将其按长宽比例缩放为 $512 \times S$。训练阶段对图片进行打乱、水平翻转和随机裁剪，输入图片尺寸为 448×448，batchsize 设为 12；测试阶段仅对图片进行中心裁剪，batchsize 设为 2。

相比于 ImageNet 数据集，CUB-200-2011 属于小型数据集。为避免样本数量过少导致模型过拟合，本文的训练过程分为两步。

步骤 1　加载在 ImageNet 上的预训练模型初始

化 VGG-16 权重参数。各超参数设置如下: 基础学习速率为 1, gamma 值为 0.25, 权重衰减为 0.000 05,动量 0.9。迭代次数为 60 000 次,采用 SGD 优化器,每隔 10 000 次迭代调整一次学习率。仅用训练集学习精简双线性层和最后一层全连接,待训练损失收敛后,保存模型参数。

步骤 2 加载步骤 1 中保存的模型初始化整个 CL-CBP 网络,再用训练集以 0.001 的学习速率微调整个网络。各超参数设置如下: gamma 值为 0.5,权重衰减 0.005,动量 0.9。采用 SGD 优化器,迭代至损失收敛后,在测试集上得到最终的性能。

为验证所提跨层双线性特征融合的有效性,开展了不同的特征融合方法的对比实验,表 1 给出了各融合方式及其在测试集上的正确率。

表 1 不同的特征融合方式在测试集上的正确率

Table 1 The accuracy of different feature fuison methods on test data set

特征融合方式	网络名称	正确率/%
conv5_1⊙conv5_2 + conv5_2⊙conv5_3 + conv5_1⊙conv5_3	CL-CBP-a	84.96
conv5_1⊙conv5_1 + conv5_2⊙conv5_2 + conv5_3⊙conv5_3	CL-CBP-b	85.48
conv5_1⊙conv5_3 + conv5_2⊙conv5_3 + conv5_3⊙conv5_3	CL-CBP-c	85.87

从表 1 可知,提出的特征融合方式,也即 CL-CBP-c 网络,能够在测试集上取得最高正确率 85.87%。相比于其他两种方式,其正确率分别提高了 0.91% 和 0.39%。

为验证引入成对混淆方法的有效性,对带成对混淆损失(with PC) 和不带成对混淆损失(without PC) 的 CL-CBP-c 网络分别进行了训练。实验结果如表 2 所示。由表 2 可知,成对混淆对 CL-CBP 网络具有正向调节作用,在网络权重参数共享的前提下,CL-CBP(with PC) 在 CUB-200-2011 测试集上的准确率较 CL-CBP(without PC) 提升了 0.76%。

表 2 CL-CBP 是否带成对混淆在测试集上的正确率

Table 2 The accuracy of CL-CBP with or without pairwise confusion on the test set

网络	正确率/%
带成对混淆	85.87
不带成对混淆	86.63

4.3 对比实验

为充分验证本文模型的有效性,选取并对比了各双线性模型,各模型在训练和测试阶段均无需提供额外的部件标注,其指标对比如表 3 所示。由表 3 可知,提出的跨层精简双线性网络在成对混淆损失作为惩罚项的弱监督学习下,在 CUB-200-2011 测试集上取得了 86.6% 的准确率,比仅使用单层卷积作双线性操作的 BCNN 提高了 2.5 个百分点,比仅带成对混淆的 PC-BCNN 提高了 1 个百分点,这证明了跨层特征交互的有效性和成对混淆对网络的正向调节。

表 3 进一步从维度、参数存储量和计算复杂度三个方面比较了不同的双线性模型。表 3 中各指标均基于 CUB-200-2011 的 200 路分类问题,在尺寸为 $c \times H \times W$ 的特征图上计算双线性特征,故 N 取 200。对于尺寸为 448 × 448 的输入图像,使用 VGG-16 为特征提取器,有 $c = 512$ 和 $H = W = 28$。LRBP[17] 参数 m、r 分别设置为 100、8,本文方法中参数 d 设置为 8 192。从对比结果可知,BCNN[9] 和 IBP[16] 特征维度极高、分类器参数量极大,LRBP 分类器参数量少但计算复杂度比本文方法高。综合考量,本文方法具有特征维度较低、计算复杂度低、参数量较少的优点,并在测试集上取得较高正确率,平均预测时间为 0.038 s/per image,具有实际应用价值。

为充分验证本文模型的优越性,还与基于部件的分类网络进行了对比,结果如表 4 所示。综合前述实验结果和表 4 的对比结果可知,CL-CBP 网络能进行端到端训练,无须显式定位目标部件和复杂的两级训练策略仍能取得较高识别正确率。这证明了提出的 CL-CBP 网络的优越性。

表 3 与各双线性模型在正确率、特征维度、参数存储量、计算复杂度上的对比

Table 3 Comparison with accuracy, feature dimension, parameters memory, and computational complexity of each bilinear model

网络	正确率/%	特征维度/k	特征计算复杂度	分类器计算复杂度	特征参数存储/KB	分类器参数存储/MB
BCNN[9]	84.1	256	$O(HWc^2)$	$O(Nc^2)$	0	200
IBP[16]	85.8	256	$O(HWc^2 + c^2)$	$O(Nc^2)$	0	200
LRBP[17]	84.2	10	$O(HWcm + HWm^2)$	$O(Nrm^2)$	200	0.6
Kernel-Activation[18]	85.3	–	–	–	–	–
PC-BCNN[15]	85.6	–	–	–	–	–
本文方法	86.6	24	$O(HW(c + d\lg d))$	$O(Nd)$	4	19

表 4　与基于部件的分类网络正确率对比

Table 4　Accuracy comparison with part-based classification networks

网络	特征提取器	正确率/%
Part-RCNN[1]	AlexNet	76.4
PA-CNN[2]	VGG-16	82.8
ST-CNN[5]	Inception net	84.1
RA-CNN[6]	VGG-19	85.3
MAMC[19]	Resnet-101	86.5
本文方法	VGG-16	86.6

5　结论

为了对鸟类图像进行细粒度识别，提出了一种跨层精简双线性网络，充分利用来自不同卷积层特征图谱的层间特征相关性和交互性，根据 Tensor Sketch 近似计算出图像高辨别度的双线性特征，并加以成对混淆对交叉熵损失函数正则化以防止过拟合。该网络弥补了单一卷积层得到的双线性特征的不充分性，网络结构简单、特征维度低、计算复杂度低，能实现端到端的训练，且在 CUB-200-2011 数据集上取得较高的识别率，具有实际应用价值。

参　考　文　献

1　Zhang N, Donahue J, Girshick R, et al. Part-based R-CNNs for fine-grained category detection [C]// European Conference on Computer Vision. Zurich: Springer, 2014: 834-849

2　Krause J, Jin H, Yang J, et al. Fine-grained recognition without part annotations [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 5546-5555

3　黄凯奇，任伟强，谭铁牛. 图像物体分类与检测算法综述 [J]. 计算机学报，2014，37(6)：1225-1240
Huang Kaiqi, Ren Weiqiang, Tan Tieniu. A review on image object classification and detection [J]. Chinese Journal of Computer, 2014, 37(6): 1225-1240

4　Xiao T, Xu Y, Yang K, et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 842-850

5　Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks [C]//Advances in Neural Information Processing Systems. Montreal: IEEE, 2015: 2017-2025

6　Fu J, Zheng H, Mei T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 4438-4446

7　Zheng H, Fu J, Mei T, et al. Learning multi-attention convolutional neural network for fine-grained image recognition [C]//Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 5209-5217

8　Wang Y, Morariu V I, Davis L S. Learning a discriminative filter bank within a CNN for fine-grained recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4148-4157

9　Lin T Y, Roy Chowdhury A, Maji S. Bilinear cnn models for fine-grained visual recognition [C]// Proceedings of the IEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 1449-1457

10　Gao Y, Beijbom O, Zhang N, et al. Compact bilinear pooling [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 317-326

11　Cimpoi M, Maji S, Kokkinos I, et al. Deep filter banks for texture recognition, description, and segmentation [J]. International Journal of Computer Vision, 2015, 118(1): 65-94

12　Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 3431-3440

13　Lin T Y, Roy Chowdhury A, Maji S. Bilinear convolutional neural networks for fine-grained visual recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1309-1322

14　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

15　Dubey A, Gupta O, Guo P, et al. Pairwise confusion for fine-grained visual classification [C]//Proceedings of the European Conference on Computer Vision (ECCV). Munish: Springer, 2018: 70-86

16　Lin T Y, Maji S. Improved bilinear pooling with cnns [C]//Proceedings of the British Machine Vision Conference. London: BMVA Press, 2017

17　Kong S, Fowlkes C. Low-rank bilinear pooling for fine-grained classification [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 365-374

18　Cai S, Zuo W, Zhang L. Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization [C]// Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 511-520

19　Sun M, Yuan Y, Zhou F, et al. Multi-attention multi-class constraint for fine-grained image recognition [C]//Proceedings of the European Conference on Computer Vision (ECCV). Munish: Springer, 2018: 805-821

Fine-grained Bird Recognition Based on Cross-layer Compact Bilinear Network

LAN Jie[1], ZHOU Xin[1,2], HE Xiao-hai[1*], TENG Qi-zhi[1], QING Lin-bo[1]
(College of Electronics and Information Engineering, Sichuan University[1], Chengdu 610065, China;
China Information Technology Security Evaluation Center[2], Beijing 100085, China)

[**Abstract**] Fine-grained image classification is quite challenging due to subtle inter-class differences and large intra-class variations. In order to fine-grain the bird image, a deep convolutional neural network model based on cross-layer compact bilinear pooling was proposed. Firstly, multiple sets of compact bilinear feature vector from different convolutional layers were calculated according to the Tensor Sketch algorithm. And then, the normalized feature vectors were concatenated and sent to the softmax classifier. Finally, the pairwise confusion was used to regularize the cross entropy loss function in order to optimize network. The model can be end-to-end trained without additional part annotation. The result shows that it achieves good performance on the public CUB-200—2011 bird dataset and gets 86.6% recognition accuracy, which surpasses BCNN by 2.5%. The comparison with several advanced algorithms proves the validity and superiority of the proposed model.

[**Key words**] bird species recognition compact bilinear transformation cross-layer feature fusion pairwise confusion fine-grained image classification