基于多层聚焦Inception-V3卷积网络的细粒度图像分类

王 波[1],黄 冕[2],刘利军[1,3],黄青松[1,4],单文琦[1]
(1. 昆明理工大学信息工程与自动化学院,云南昆明 650500; 2. 云南国土资源职业学院信息中心,云南昆明 652501; 3. 云南大学信息学院,云南昆明 650091; 4. 云南省计算机技术应用重点实验室,云南昆明 650500)

摘 要: 细粒度图片具有结构多变、背景干扰大、类间差异小、类内差异大等特点,准确地定位与提取判别性局部特征至关重要. 本文提出一种多层聚焦卷积网络,通过首层聚焦网络能够准确、有效地聚焦于识别局域并生成定位区域,根据定位区域对原图像分别进行裁剪和遮挡后输入下一层的聚焦网络进行训练分类. 其中单层聚焦网络以Inception-V3网络为基础,通过卷积块特征注意力模块和定位区域选择机制来聚焦有效的定位区域;使用双线性注意力最大池化提取各个局部的特征;最后进行分类预测. 本文在3个常用的细粒度数据集CUB-2011、FGVC-Aircraft以及Stanford Cars上进行了实验验证,分别获得了89.7%、93.6%和95.1%的Top-1准确率. 实验结果表明,本模型的分类准确率高于目前主流方法.

关键词: 多层聚焦卷积网络; Inception-V3网络; 注意力机制; 双线性注意力最大池化
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2022)01-0072-07
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20200443

Multi-Layer Focused Inception-V3 Models for Fine-Grained Visual Recognition

WANG Bo[1], HUANG Mian[2], LIU Li-jun[1,3], HUANG Qing-song[1,4], SHAN Wen-qi[1]
(1. *Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan* 650500, *China*;
2. *Information Center, Yunnan Land and Resources Vocational College, Kunming, Yunnan* 652501, *China*;
3. *School of Information, Yunnan University, Kunming, Yunnan* 650091, *China*;
4. *Yunnan Key Laboratory of Computer Technology Applications, Kunming, Yunnan* 650500, *China*)

Abstract: Fine-grained pictures are characterized by variable structure, large background interference, small inter-class difference and large intra-class difference, so accurate positioning and extraction of discriminant local features are crucial. In this paper, a multi-layer focused convolution network is proposed, which can accurately and effectively focus on identifying local areas and generating locating regions through the first-layer focused network. According to the positioning area, the image is cropped and dropped, and then the focus network of the next layer is input for training and classification. The single-layer focused network is based on the Inception-V3 network and focuses the effective location area through the convolutional block feature attention module, and location area selection mechanism. Bilinear attention maximum pooling was used to extract the features of each part. Classification prediction is made. Experimental verification was carried out on three commonly used fine-grained data sets CUB-2011, Fgvc-Aircraft and Stanford Cars the accuracy of top-1 was obtained at 89.7%, 93.6% and 95.1%, respectively. Experimental results show that the classification accuracy of this model is higher than that of the current mainstream methods.

Key words: multilayer focused convolution network; inception-V3; attention mechanism; bilinear attention maximum pooling

收稿日期:2020-05-09;修回日期:2020-10-10;责任编辑:孙瑶
基金项目:国家自然科学基金(No.81860318,No.81560296);云南省计算机技术应用重点实验室开放基金(No.2020106)

1 引言

随着深度学习、卷积网络技术的不断发展,深度学习网络在计算机视觉领域得到广泛的应用,如图像检索[1]、场景解析[2]和目标跟踪[3]等. 在细粒度图像识别领域,深度卷积网络也得到了广泛的研究与应用. 由于在细粒度图像识别中,类内差异容易受姿态、视角与位置等因素影响,类间具有相似性,且手工标注位置不稳定且耗费人力,因此细粒度识别任务更具有挑战性. Zhang等人[4]提出强监督细粒度图像分类模型(Part-based Region Convolutional Neural Network, Part-based R-CNN),借助Bounding Box和Part Annotation标签进行对象部件定位等操作得到对象与部件的图像块,最后将对象部件等特征级联之后进行分类. Part-based R-CNN分类在准确率方面取得了不错的效果,但存在算法速度慢、过于依赖标签、定位检测不稳定等问题. 因此Wei等人[5]提出Mask-CNN模型,在训练时仅需要part annotations和image-level标签,同时借助(Fully Convolutional Networks,FCN)学习Part Mask来进行对象部件定位等操作得到对象与部件的图像块,最后将对象部件等特征级联之后进行分类. Mask-CNN取得很好的定位效果以及分类准确率,并且大大降低了对标记的依赖性,但是仍避免不了强监督标记分类. Lin等人提出高阶特征编码双线性卷积网络(Bilinear-CNN, B-CNN)[6]和改进的双线性卷积网络[7],通过对特征图进行外积操作建立了特征图中通道之间的线性相关,并进行端到端的联合优化学习,在细粒度分类任务上取得了优异的性能,但是无法捕捉特征图中通道之间的非线性关系. 于是Ge等人[8]提出一种核化的双线性卷积网络,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,改进了高阶特征编码方法,提升了特征的表达能力. 但是该方法存在外积导致特征的维度增大为原来的平方等缺点. Zheng等人[9]提出(Multi-Attention Convolutional Neural Network, MA-CNN)模型,抛弃手工标记对象部件的方法,采用弱监督学习方法,同时定位多个对象部件,提出信道分组损耗,通过聚类产生多个部分,将这些部分分别与特征进行点乘得到局部精细化的特征分别进行分类,取得了很好的分类准确率. 但是该方法中对象的部件数量有限(2个或4个),因此会限制分类的准确性. 先前工作通常采用定位或者分割来解决类内的差异性,但是类间相似性依然影响特征的学习. 针对该问题,Dubey等人[10]提出细粒度图像识别(Pairwise Confusion,PC),以解决类别间相似性问题,该研究将成对的混淆损失和交叉熵损失结合,使用交叉熵函数强迫神经网络挖掘较高置信度的特征,从而减少损失从而提高分类准确率,提供了很好的解决方法. Yang等人[11]提出一种新的训练模式,使Navigator能够在Teacher的指导下检测到最具判别性的区域,Scrutinizer仔细检查这些区域并做出预测. 这种多主体合作的方式相互促进、相互收益,取得了较好的分类准确率,非常值得借鉴.

针对上述分析,本文提出多层聚焦网络模型. 本文模型的创新性和贡献之处如下.

(1)根据贪心算法思想设计多层聚焦网络模型架构,模型能够有效地定位识别对象位置,同时又能找出更多的识别对象位置. 从而达到更好的分类效果.

(2)提出卷积块特征注意力模块,能在空间和通道对特征进行注意力特征的提取,同时进行降维处理,增强网络对对象有效特征的提取,又降低计算的复杂度.

(3)提出定位区域选择机制,能有效地选择具有代表性的定位区域,同时也能提高模型的泛化能力.

(4)提出双线性注意力最大池化,增强特征的表达,降低维度和减少由卷积层参数误差造成的估计均值的偏移误差,提高模型的鲁棒性.

模型架构如图1所示,本模型由三层聚焦网络、裁剪模块以及遮挡模块构成. 模型设计思想主要源于贪心算法. 贪心算法(又称贪婪算法)是指在对问题求解时,总是做出在当前看来是最好的选择. 通过首层聚焦网络(图2)找到最优的识别位置和特征矩阵并输出定位区域和特征与注意力矩阵积矩阵,其余两层聚焦网络只输出特征与注意力矩阵积矩阵. 裁剪模块根据定位区域对原图像进行裁剪得到裁剪图像,输入下一层聚焦网络进行训练. 裁剪后的图像能更大面积地覆盖识别对象,能够起到非常好的识别分类效果. 但是贪心算法仍然存在问题,它没有从整体最优上加以考虑,所做出的仅是在某种意义上的局部最优解. 同样本模型也存在局部最优、过拟合问题. 所以加入遮挡模块与之对抗来解决局部最优问题,能够促进模型寻找更多具有识别对象的特征信息. 遮挡模块根据定位区域对原图像进行遮挡得到遮挡图像,输入下一层聚焦网络进行训练. 该模型各层之间相互对抗、相互合作、相互受益、共同进步.

单层聚焦网络流程图如图2所示. 本网络主要由Inception-V3网络[12]、卷积块特征注意力模块、定位区域选择机制、双线性注意力最大池化构成. 本网络设计思想同样遵循贪心算法思想. 使用Inception-V3作为本网络的基础网络,用来提取基础特征. 卷积块特征注意力模块借鉴卷积块注意力模块(Convolutional Block Attention Module)[13]方法,能在空间和通道进行注意力特征的提取以及进行降维处理,得到特征矩阵和注意力矩阵. 降维处理一方面能够降低计算复杂度,另外一方面是为了选择最能代表识别对象的特征. 定位区域选择机制将从注意力矩阵中选出最佳定位区域,机制加

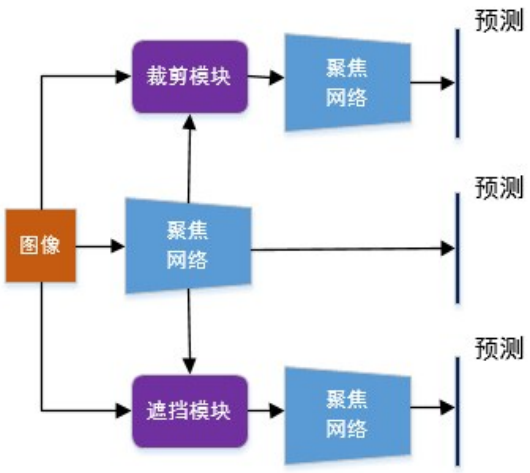

图1 多层聚焦网络架构图

入softmax函数以及均方和等方式计算概率，softmax函数使用指数函数能让大的值更大，让小的更小，增加区分对比度. 同时本文设计了防止过拟合的策略，每次倾向随机选择定位区域，其中概率较大的特征矩阵被选中的优先级相对而言会高一些. 将定位区域输入裁剪和遮挡模块，对原图像进行裁剪和遮挡得到的裁剪图像和遮挡图像，作为下一层聚焦网络的输入. 受双线性模型的影响，本文提出双线性注意力最大池化，对特征矩阵和注意力矩阵进行外积相乘，再配合最大池化提取特征能保留图像更多的纹理信息，同时提高所提取特征的鲁棒性. 双线性注意力最大池化输出特征与注意力矩阵积矩阵去做分类预测. 最后为了解决网络中局部最优过拟合这一问题，并增强模型的泛化能力，加入注意力中心损失，对属于同一对象特征的方差进行惩罚.

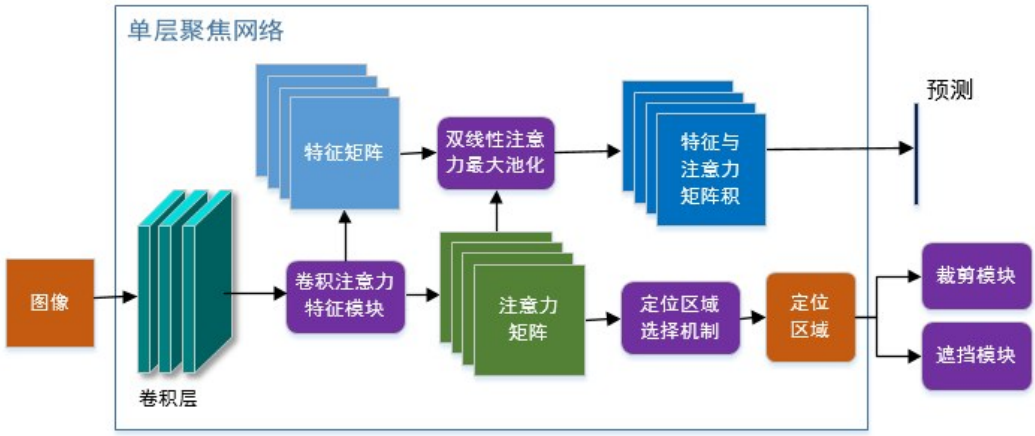

图2 单层聚焦网络流程图

2 多层聚焦卷积网络

2.1 聚焦卷积网络

2.1.1 卷积块特征注意力模块

卷积块注意力模块(Convolutional Block Attention Module，CBAM)是一种结合空间和通道的注意力机制. 在分类和检测模型上被广泛使用. 其优势不言而喻，但仍然存在以下不足：缺乏通道与通道之间的交互，无法捕捉通道之间的非线性关系，过于关注特征表达从而导致过拟合. 为改进CBAM存在的这些不足，提出卷积块特征注意力模块(Convolutional Block Feature Attention Module，CBFAM). 该模块是在CBAM原有的基础上加入多层1*1卷积核对特征进行降维处理，实现跨通道的交互和信息整合，增加网络的非线性能力，提高网络的表达能力并且减少计算量，进行通道数随机截取以此提高泛化能力，最后输出的是特征图 $\boldsymbol{F}$ 和注意力图 $\boldsymbol{A}$. 将从Inception-V3网络中提取图像I的特征图 $\boldsymbol{F}\in\mathbf{R}^{C\times H\times W}$ 输入CBFAM模块中得到特征图 $\boldsymbol{F}_a\in\mathbf{R}^{C\times H\times W}$ 和注意力图 $\boldsymbol{A}\in\mathbf{R}^{M\times H\times W}$ 计算公式如式(1)所示.

$$\begin{aligned}
&M_c(\boldsymbol{F})=\sigma\left(W_1\left(W_0\left(\boldsymbol{F}_{\text{avg}}^c\right)\right)+W_1\left(W_0\left(\boldsymbol{F}_{\text{max}}^c\right)\right)\right)\\
&M_S(\boldsymbol{F})=\sigma\left(\boldsymbol{f}^{7*7}\left[\boldsymbol{F}_{\text{avg}}^c;\ \boldsymbol{F}_{\text{max}}^c\right]\right)\\
&M(\boldsymbol{F})=M_s(M_c(\boldsymbol{F})\otimes\boldsymbol{F})\otimes M_c(\boldsymbol{F})\\
&\boldsymbol{F}_a=\boldsymbol{f}^{1*1}\left(M(\boldsymbol{F})\right)\\
&\boldsymbol{A}=\bigcup_{k=1}^{M}\boldsymbol{f}^{1*1}\left(\boldsymbol{F}_a\right)
\end{aligned}\tag{1}$$

其中，A 表示物体的某个部分；C、H、W 分别表示特征图的通道数、高度、宽度；$\boldsymbol{F}_{\text{avg}}^c$、$\boldsymbol{F}_{\text{max}}^c$ 分别代表经过全局平均池化层和全局最大池化层计算后的特征值；W_0 和 W_1 代表的是多层感知机模型中的两层参数；σ 表示的是sigmoid激活函数；$\boldsymbol{f}^{7*7}$ 表示卷积层使用7*7的卷积核，$\boldsymbol{f}^{1*1}$ 表示卷积层使用1*1的卷积核；M 是注意力图的数量.

2.1.2 定位区域选择机制

在注意力图$\boldsymbol{A}\in\mathbf{R}^{M\times H\times W}$中有$M$个特征矩阵表示图像中的不同对象部件. 如何选择其中一个特征矩阵来代表最具有判别性的对象部件尤为重要. 本文希望模型在找到最佳部分特征的同时也能够找到更多具有判别性的部分特征,从而提高模型泛化能力. 使用softmax函数处理注意力图$\boldsymbol{A}$,将注意力图中的值映射成为(0,1)的值,这些值的累和为1(满足概率的性质). softmax函数加入幂函数使这些值两极化:正样本的结果将趋近于1,而负样本的结果趋近于0. 根据softmax函数的性质进一步去除噪音. 计算M个特征矩阵被选中的概率p_M,其计算公式如式(2)所示,在参考概率p_M的同时采用随机的方式从M个特征矩阵中选出一个特征矩阵$\boldsymbol{A}_k\in\mathbf{R}^{1\times H\times W}$,其中概率大的特征矩阵优先选中.

$$\begin{aligned}\boldsymbol{A}&=\frac{\mathrm{e}^{(\boldsymbol{A})}}{\sum_j \mathrm{e}^{(\boldsymbol{A})}}\\ p_M&=\frac{\sqrt{\overline{\boldsymbol{A}_i}}}{\sum_M\sqrt{\overline{\boldsymbol{A}_i}}}\end{aligned}\tag{2}$$

根据p_M采用概率和随机方式得到定位区域注意力矩阵$\boldsymbol{A}_k\in\mathbf{R}^{1\times H\times W}$进行标准化处理,计算公式如式(3)所示.

$$\boldsymbol{A}_k=\frac{\boldsymbol{A}_k-\min\left(\boldsymbol{A}_k\right)}{\max\left(\boldsymbol{A}_k\right)-\min\left(\boldsymbol{A}_k\right)}\tag{3}$$

2.1.3 双线性注意力最大池化

双线性卷积网络通过对卷积层输出的特征图进行外积操作建模了特征图中通道之间的线性相关,在细粒度图像识别任务上取得了优异的性能,但是无法捕捉特征图中通道之间的非线性关系,并且导致特征的维度增大为原来的平方. 因而提出双线性注意力最大池化(Bilinear Attention Max Pooling)方法,解决无法捕捉通道之间的非线性关系问题并降低维度. 将特征图$\boldsymbol{F}_a\in\mathbf{R}^{C\times H\times W}$与每个注意力图$\boldsymbol{A}\in\mathbf{R}^{M\times H\times W}$相乘,生成$M$个部分的特征图$\boldsymbol{F}_k\in\mathbf{R}^{C\times H\times W}$,加入非线性激活函数、最大池化层,计算公式如式(4)所示.

$$\boldsymbol{F}_k=\sum_i^k \boldsymbol{A}_i\boldsymbol{F}_a\tag{4}$$

其中,$\boldsymbol{F}_k$为两个张量的元素乘. 通过全局最大池层得到第k个特征与注意力矩阵积矩阵$\boldsymbol{f}_k\in\mathbf{R}^{1\times C}$,计算公式如式(5)所示.

$$\boldsymbol{f}_k=\mathrm{MaxPool}\left(\boldsymbol{F}_k\right)\tag{5}$$

$\boldsymbol{P}_k\in\mathbf{R}^{M\times C}$表示对象不同部分特征矩阵将作为模型输出进行预测,$\boldsymbol{P}_k$由特征$\boldsymbol{f}_k$叠加而成. 计算公式如式(6)所示.

$$\boldsymbol{P}_k=\begin{pmatrix}f_1\\ f_2\\ \vdots\\ f_M\end{pmatrix}\tag{6}$$

2.1.4 注意力中心损失

Wen等人[14]提出的中心损失问题,在配合softmax函数使用时希望学习到的特征具有更好的泛化性和辨别能力. 通过惩罚每个种类的样本和该种类样本中心的偏移,使得同一种类的样本尽量聚合在一起. 为了解决局部优化过拟合这一问题,增强模型的泛化能力,本文中对属于同一对象部件的特征的方差进行惩罚,即部件特征$\boldsymbol{P}_k\in\mathbf{R}^{M\times C}$将趋近于全局特征中心$\boldsymbol{C}_k\in\mathbf{R}^{M\times C}$,在同一对象部件$k_{th}$中反馈注意力图$\boldsymbol{A}$. 其中损失函数$L_C$只在首层聚焦网络和使用裁剪图像作为输入的聚焦网络中使用,计算公式如式(7)所示.

$$L_C=\frac{1}{2}\sum_{k=1}^{M}\left\|\boldsymbol{P}_k-\boldsymbol{C}_k\right\|_2^2\tag{7}$$

$\boldsymbol{C}_k$初始化为(0,0),计算公式如式(8)所示.

$$\boldsymbol{C}_k\leftarrow\boldsymbol{C}_k+\beta\left(\boldsymbol{P}_k-\boldsymbol{C}_k\right)\tag{8}$$

其中,β控制$\boldsymbol{C}_k$的更新速度,中心更新计算只在首层聚焦网络和使用裁剪图像作为输入的聚焦网络中使用.

2.2 裁剪模块

根据定位区域注意力矩阵$\boldsymbol{A}_k^*$对原图像进行裁剪,裁剪后的图像能更大面积地覆盖识别对象,去除无关背景,起到非常好的分类效果. 通过设置$\boldsymbol{A}_k^*$大于阈值$\theta_c\in[0.4,0.6]$时等于1,其他为0,得到裁剪边框$\boldsymbol{C}_k$,采取上采样的方式从原图像中放大这个区域得到裁剪图像作为下一层聚焦网络的输入. 由于对象部分的规模增加,因此能更好地对对象提取更细粒度的特性. 该模块在模型训练和测试中使用,计算公式如式(10)所示.

$$\boldsymbol{C}_k(i,j)=\begin{cases}1, & \text{if}\left(\boldsymbol{A}_k^*>\theta_c\right)\\ 0, & \text{otherwise}\end{cases}\tag{9}$$

2.3 遮挡模块

根据定位注意力矩阵$\boldsymbol{A}_k^*$对原图像进行局部遮挡,促进模型找出更多具有代表性的多个有判别性对象部分的定位区域注意力矩阵$\boldsymbol{A}_k^*$. 以此对抗由于裁剪模块产生局部最优过拟合的负面效果,提高分类的稳健性和定位的准确性. 通过设置$\boldsymbol{A}_k^*$小于或等于阈值$\theta_d\in[0.4,0.6]$时为1,其他为0,得到遮挡框$\boldsymbol{D}_k$,将遮挡框与原图像相乘得到遮挡图像作为下一层聚焦网络的输入. 该模块只在模型训练时使用,计算公式如式(10)所示.

$$\boldsymbol{D}_k(i,j)=\begin{cases}1, & \text{if}\left(\boldsymbol{A}_k^*<\theta_c\right)\\ 0, & \text{otherwise}\end{cases}\tag{10}$$

3 实验

3.1 数据集

实验部分在3个细粒度图像识别数据集CUB-200-2011[15]、FGVC-Aircraft[16]以及Stanford Cars[17]上对本文方法进行评估.

CUB-200-2011数据集共包含来自200个鸟类物种的11788张图像,其中5994张训练与验证图像,5794张测试图像. FGVG-Aircraft数据集包括1万张图像共100种飞机型号,其中训练与验证图像共6667张,测试图像为3333张. Stanford Cars数据集具有196个汽车类别共16185张图像,其中训练与验证图像共8144张,测试图像为8041张. 本文实验中均未采用边界框等额外标记信息. 数据集示例实例如图3所示.

图3 数据集示例图

3.2 实验参数

实验的环境配置如下. CPU:Intel(R) Silver;GPU:NVIDIA TITAN XP;内存为128G;操作系统为Ubuntu16.04;采用Inception-V3作为主干,输入图像大小为512×512. 定位区域裁剪和遮挡后的图片大小为512×512. 特征图的通道数目均为768,注意力图通道数均为32. 本文使用带动量的批处理随机梯度下降算法(SGD),动量为0.9,权重衰减为0.00001,批量大小设置为16. 训练次数为80,初始学习率设置为0.001,每隔2个epoch指数衰减0.9. 数据增广方式为图像的随机水平翻转.

3.3 模型组件测试

本模型主要由5个部分组成,包括卷积块特征注意力模块、定位区域选择机制、双线性注意力最大池化、裁剪模块、遮挡模块. 本文在CUB200-2011数据集上测试每个模块对模型的影响.

实验结果如表1所示. 在单层聚焦模型中,相对Inception-V3网络的准确率83.7%,效果提升非常显著,模型准确率提升到86.8%,证明了本文方法的有效性. 在多层聚焦网络中,模型在分别加入裁剪和遮挡模块后均得到提升. 使用全部模块后,模型达到最高的准确率. 实验结果证明了该模型在多层聚焦卷积网络中每个模块的有效性.

3.4 注意力图数量测试

在弱监督模型中,注意力图是进行对象部件定位的重要部分,在本文的M个注意力图中每一个都代表了对象的一个部件,例如鸟的头部、飞机的机翼等. 本文对注意力图M大小的设置进行相应的实验.

表1 组成部分及其组合的贡献

注意力特征	选择机制	双线性注意力	裁剪	遮挡	准确率/%
					83.7
✔					84.6
✔	✔				85.1
✔		✔			85.9
	✔	✔			85.2
✔	✔	✔			86.8
✔	✔	✔		✔	87.5
✔	✔	✔	✔		88.3
✔	✔	✔	✔	✔	89.7

实验结果如表2所示. 当M=4时,对象部件可能分为鸟的头、翅膀、尾巴和脚等4个部分,当M不断增加时,对象部件会定位到鸟喙、眼等部位. 在这一过程中模型将去掉相似性区域、无关区域,找到细微的判别区域,从而使分类准确率不断提升. 当M=32后,准确率到达最高,同时意味着已经分出最佳判别区域对象部件的大小. 之后M再增加时,产生的对象部件将是最佳判别区域部件的子集,模型是通过对判别性区域的裁剪放大后的图像再分类. 因此M再增加,准确率也趋于稳定.

3.5 与当前方法准确率对比

本节实验进一步在三个细粒度图像数据集上对本文方法进行全面的评估,并与当前主流方法进行比较.

实验结果如表3所示[7~12,18~27]. 本模型在当前主流细粒度图像数据集上都达到最佳的准确率.

表2　注意力图数量的影响

注意力图通道数	准确率/%
4	86.8
8	87.9
16	88.5
32	89.7
64	89.7

表3　与当前方法准确率对比

Methods	CUB	Aircraft	Cars
VGG-19[18]	77.8	80.5	85.7
ResNet-101[19]	83.5	87.2	91.2
Inception-v3[12]	83.7	87.4	90.8
B-CNN[7]	84.1	84.1	91.3
ST-CNN[20]	84.1	-	-
核化双线性[8]	86.1	91.3	92.8
PDFR[21]	84.5	-	-
RA-CNN[22]	85.4	88.4	92.5
GP-256[23]	85.8	89.8	92.8
MA-CNN[9]	86.5	89.9	92.8
MAMC[24]	86.5	-	93.0
PC[10]	86.9	89.2	92.9
DFL-CNN[25]	87.5	92	93.8
NTS-Net[11]	87.5	91.4	93.9
MPN-COV[26]	88.7	91.4	93.3
WS-DAN[27]	89.4	93.0	94.5
本文	89.7	93.6	95.1

4　结论

本文提出一种多层Inception-V3聚焦卷积网络,能够提取更多有效的局部特征,聚焦于对象的识别位置,得到更具有判别力的图像表达. 同时多层聚焦卷积网络之间相互对抗、相互合作、相互受益、共同进步,进一步提高网络的性能. 实验表明,本文所提出的多层聚焦网络在3个细粒度图像数据集上均取得了优异的性能. 在未来的工作中,可以将本文方法应用至更多计算机视觉任务中,同时进一步考虑将本文提出的多层聚焦网络应用于其他卷积网络架构中.

参考文献

[1] 柯圣财, 赵永威, 李弼程, 等. 基于卷积神经网络和监督核哈希的图像检索方法[J]. 电子学报, 2017, 45(1): 157-163.
KE S C, ZHAO Y W, LI B C, et al. Image retrieval based on convolutional neural network and kernel-based supervised hashing[J]. Acta Electronica Sinica, 2017, 45(1): 157-163. (in Chinese)

[2] 王泽宇, 吴艳霞, 张国印, 等. 基于空间结构化推理深度融合网络的RGB-D场景解析[J]. 电子学报, 2018, 46(5): 1253-1258.
WANG Z Y, WU Y X, ZHANG G Y, et al. RGB-D scene parsing based on spatial structured inference deep fusion networks[J]. Acta Electronica Sinica, 2018, 46(5): 1253-1258. (in Chinese)

[3] 李康, 李亚敏, 胡学敏, 等. 基于卷积神经网络的鲁棒高精度目标跟踪算法[J]. 电子学报, 2018, 46(9): 2087-2093.
LI K, LI Y M, HU X M, et al. A robust and accurate object tracking algorithm based on convolutional neural network [J]. Acta Electronica Sinica, 2018, 46(9): 2087-2093. (in Chinese)

[4] ZHANG N, DONAHUE J, GIRSHICK R, et al. Part-based R-CNNs for fine-grained category detection[C]//European Conference on Computer Vision. Cham, Switzerland: Springer, 2014: 834-849.

[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). Santiago, Chile: IEEE, 2015: 1449-1457.

[7] LIN T Y, MAJI S. Improved bilinear pooling with CNNs [C]//Proceedings of the British Machine Vision Conference 2017. London, UK: British Machine Vision Association, 2017: 117.1-117.12.

[8] 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[J]. 电子学报, 2019, 47(10): 2134-2141.
GE S Y, GAO Z L, ZHANG B B, et al. Kernelized bilinear CNN models for fine-grained visual recognition[J]. Acta Electronica Sinica, 2019, 47(10): 2134-2141. (in Chinese)

[9] 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). Venice, Italy: IEEE, 2017: 5219-5227.

[10] DUBEY A, GUPTA O, GUO P, et al. Pairwise confusion for fine-grained visual classification[C]//European Conference on Computer Vision. Cham, Switzerland: Springer, 2018: 71-88.

[11] YANG Z, LUO T G, WANG D, et al. Learning to navigate for fine-grained classification[C]//European Conference on Computer Vision. Cham, Switzerland: Springer, 2018: 438-454.

[12] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 2818-2826.
[13] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]//European Conference on Computer Vision. Cham, Switzerland: Springer, 2018: 3-19.
[14] WEN Y D, ZHANG K P, LI Z F, et al. A discriminative feature learning approach for deep face recognition[C]// European Conference on Computer Vision. Cham, Switzerland: Springer, 2016: 499-515.
[15] WAH C, BRANSON S, WELINDER P, et al. The Caltech-UCSD Birds-200-2011 Dataset (Technical Report CNS-TR-2011-001) [R]. USA: California Institute of Technology, 2011.
[16] MAJI S, RAHTU E, KANNALA J, et al. Fine-grained visual classification of aircraft[EB/OL]. [2021]. https://arxiv.org/abs/1306.5151.
[17] KRAUSE J, STARK M, JIA D, et al. 3D object representations for fine-grained categorization[C]//2013 IEEE International Conference on Computer Vision Workshops. Sydney, NSW, Australia: IEEE, 2013: 554-561.
[18] YAN Z C, ZHANG H, PIRAMUTHU R, et al. HD-CNN: Hierarchical deep convolutional neural network for large scale visual recognition[EB/OL]. [2021]. https://arxiv. org/abs/1410.0736.
[19] 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). Las Vegas, NV, USA: IEEE, 2016: 770-778.
[20] JADERBERG M, SIMONYAN, ZISSERMAN A, et al. Spatial transformer networks[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: ACM, 2015: 2017-2025.
[21] ZHANG X P, XIONG H K, ZHOU W G, et al. Picking deep filter responses for fine-grained image recognition [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 1134-1142.
[22] 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). Honolulu, HI, USA: IEEE, 2017: 4476-4484.
[23] WEI X, ZHANG Y, GONG Y H, et al. Grassmann pooling as compact homogeneous bilinear pooling for fine-grained visual classification[C]//European Conference on Computer Vision. Cham, Switzerland: Springer, 2018: 365-380.
[24] SUN M, YUAN Y C, ZHOU F, et al. Multi-attention multi-class constraint for fine-grained image recognition [C]//European Conference on Computer Vision. Cham, Switzerland: Springer, 2018: 834-850.
[25] WANG Y M, MORARIU V I, DAVIS L S. Learning a discriminative filter bank within a CNN for fine-grained recognition[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 4148-4157.
[26] LI P H, XIE J T, WANG Q L, et al. Towards faster training of global covariance pooling networks by iterative matrix square root normalization[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 947-955.
[27] HU T, QI H G. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification[EB/OL]. [2021]. https://www.researchgate.net/publication/330726056_See_Better_Before_Looking_Closer_Weakly_Supervised_Data_Augmentation_Network_for_Fine-Grained_Visual_Classification.

作者简介

王 波 男,1995年3月出生于湖南省邵阳市. 昆明理工大学信息工程与自动化学院硕士研究生. 主要研究方向为深度学习和图像处理.
E-mail:251970441@qq.com

黄青松(通信作者) 男,1962年4月出生于湖南省长沙市. 昆明理工大学信息工程与自动化学院副院长、教授、研究生导师. 主要研究方向为智能信息系统.
E-mail:ynkmhqs@sina.com