ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 34,No. 1,Jan. 2019,pp. 157－166
DOI:10. 16337/j. 1004-9037. 2019. 01. 016

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于多通道视觉注意力的细粒度图像分类

王培森　宋　彦　戴礼荣

(中国科学技术大学语音及语言信息处理国家工程实验室,合肥,230027)

摘　要:视觉注意力机制在细粒度图像分类中得到了广泛的应用。现有方法多是构建一个注意力权重图对特征进行简单加权处理。对此,本文提出了一种基于可端对端训练的深度神经网络模型实现的多通道视觉注意力机制,首先通过多视觉注意力图描述对应于视觉物体的不同区域,然后提取对应高阶统计特性得到相应的视觉表示。在多个标准的细粒度图像分类测试任务中,基于多通道视觉注意的视觉表示方法均优于近年主流方法。

关键词:　图像分类;细粒度图像分析;视觉注意力;图像表示;深度学习

中图分类号:TP391.4　　**文献标志码**:A

Fine-Grained Image Classification with Multi-channel Visual Attention

Wang Peisen, Song Yan, Dai Lirong

(National Engineering Laboratory of Speech and Language Information Processing, University of Science and Technology of China, Hefei, 230027, China)

Abstract: Visual attention mechanism has been commonly used in state-of-the-art fine-grained classification methods in recent years. However, most attention-based image classification systems only apply single-layer or part-specified attention feature, with simple multiplication-based attention applying method, which limits the information provided by the attention. This paper presents a multi-channel visual attention based fine-grained image classification system. Multi-channel attention features are extracted from the image and applied to low-level features, with subtraction of mean values corresponding to each layer of attention for high-order representation, making the model an end-to-end optimizable deep neural network architecture. On multiple commonly used fine-grained classification datasets, the presented method outperforms state-of-the-art methods with a large margin.

Key words: image classification; fine-grained image analysis; visual attention; image representation; deep learning

引　　言

细粒度图像分类(Fine-grained image classification)是近年来计算机视觉领域的热点研究问题,其目的是对粗粒度大类别中更加细致的子类别进行划分[1-2]。这些子类别通常具有较小的类间差异,往往需要依靠微小的局部差异对类别进行区分。如鸟类数据集中的环嘴鸥(Ring-billed gull)和加利福尼亚鸥

基金项目:国家自然科学基金(U1613211)资助项目。
收稿日期:2018-04-28;**修订日期**:2018-08-14

(California gull)这两种鸟类非常相似,仅喙部形态有较大不同,对于掌握相关知识的人类也有较高的辨别难度[3]。相比类间差异,细粒度图像分类中通常存在较大的类内差异,包括物体姿态、尺度、遮挡及背景等。特别地,在每一类别数据量较为有限,且没有对于物体部位额外的人工标注信息的情况下,实现基于弱监督信息的细粒度图像分类,是一项极具挑战性的任务。

针对细粒度图像分类的特点和难点,引入视觉注意力[4]机制,突出图像中具有区分性的重点部位,是近年来细粒度图像分类研究中的常见思路。例如,Jaderberg等人[5]提出了空间变换网络(Spatial transform network),利用软性注意力(Soft attention)在特征图上进行采样,得到经形态变换的特征,相比经典卷积网络,能够更有效地提取空间特征信息。Xiao等人[6]提出的两级注意力模型(Two-level attention model)应用了物体级(Object-level)和部位级(Part-level)两种注意力,使用卷积网络得到物体级信息,再使用聚类的方法得到重点局部区域,从而更为精确地利用多层次信息。Zhang等人提出的SPDA-CNN[7]利用CUB鸟类数据集中的部位标注(Part annotation)训练检测网络,得到对应于数据集中鸟类7个不同部位的硬性注意力(Hard attention),将特征在相应位置进行切割后用于图像分类。Fu等人[8]将视觉注意力与递归结构相结合,在递归网络的每一层级对特征和注意力权重进行融合,从而在模型中结合了多个尺度的关键区域特征。

上文所述方法将注意力机制应用于细粒度分类中取得了良好的效果,但其中注意力的作用仍有一定的限制:

(1)对于每次注意力与特征融合过程,注意力权重图均为通道数为1的特征图,没有使用多维度的注意力特征,这限制了对于重点区域分布较为复杂的图像特征提取能力。近年来,注意力机制在计算机视觉领域外的其他领域得到广泛的应用,其中,Vaswani等人提出的多端注意力(Multi-head attention)机制[9]并行地产生多个注意力权重图并同时与特征进行融合,使模型获得对应于输入不同位置的注意力,这一方法在机器翻译等任务上超越了此前基于复杂模型的方法,证明了多通道注意力可以提供更为有效全面的信息。

(2)注意力权重图与图像特征融合的方法较为简单,一般采用的是将注意力权重图与特征图按位置将对应元素相乘的方法,这样一方面无法提取对于分类更为有效的高阶信息,另一方面难以适应形式更为复杂的多通道注意力特征。

基于上述分析,本文提出一种基于多通道注意力的细粒度图像分类模型:提出一种基于神经网络的多通道注意力生成方法,通过提取多通道注意力权重图,得到丰富的空间注意力信息;同时提出了一种新的注意力与特征融合方法,通过提取图像特征对应于注意力的高阶信息,获得对于图像更具描述能力的高层特征。最终构成可端对端训练的深度神经网络模型。在CUB-200-2011、FGVC-Aircraft和斯坦福Cars等常用细粒度图像分类数据集上的实验中,相比于近年主流细粒度图像分类方法,本文模型获得的分类精确度有显著提升。

1 注意力机制作用原理

1.1 注意力提取

注意力的作用可视为从输入信息中选择一部分和任务相关的信息的过程,注意力权重即为这些信息的索引[4]。用$\boldsymbol{x}_{1:N}=[\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N]$表示N个输入信息,注意力变量$z\in[1,N]$可用于表示被选择信息的索引位置,即$z=i$表示选择了第$i$个输入信息。对于使用软性注意力的情况,可令$\alpha_i$表示在给定当前输入信息$\boldsymbol{x}_{1:N}$情况下,选择第$i$个输入信息的概率,即需要提取的注意力权重[9],则有

$$\alpha_i = p\left(z = i|x_{1:N}\right) = \mathrm{softmax}\left(s\left(\boldsymbol{x}_i\right)\right) = \frac{\exp\left(s\left(\boldsymbol{x}_i\right)\right)}{\sum_{j=1}^{N}\exp\left(s\left(\boldsymbol{x}_j\right)\right)} \tag{1}$$

式中:$s\left(\boldsymbol{x}_i\right)$ 为注意力打分函数,可根据实际任务和情形选择相应的模型。例如,采用点积模型

$$s\left(\boldsymbol{x}_i\right) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W} \tag{2}$$

式中 $\boldsymbol{W}$ 为可学习的网络参数。

1.2 注意力与特征的融合

注意力权重作用于特征的方法可视为在一种信息选择机制下对输入信息进行编码的过程[4],对于单一维度的软性注意力权重图,最为普遍的注意力与特征融合方式是将注意力权重以点积的形式将对应位置元素相乘

$$\mathrm{attention}\left(x_{1:N}\right) = \sum_{i=1}^{N}\boldsymbol{\alpha}_i \cdot \boldsymbol{x}_i = E_{z\sim p(z|x_{1:N},\, q)}[x] \tag{3}$$

2 基于多通道视觉注意力的细粒度图像分类

本文所述深度神经网络模型可分为特征提取、注意力权重图生成、注意力权重与特征融合、分类器等多个部分。其中,特征提取模块利用全卷积网络将输入图像转化为低层特征;注意力权重图生成模块输入图像特征得到多通道注意力权重;融合模块将注意力权重与图像低层特征进行融合,得到特征向量作为图像的高层表示;分类器将注意力融合后的特征向量转化为对应于数据集每一类别的概率,从而得到分类结果。以上各部分构成了可端对端训练的图像分类模型框架,如图 1所示。

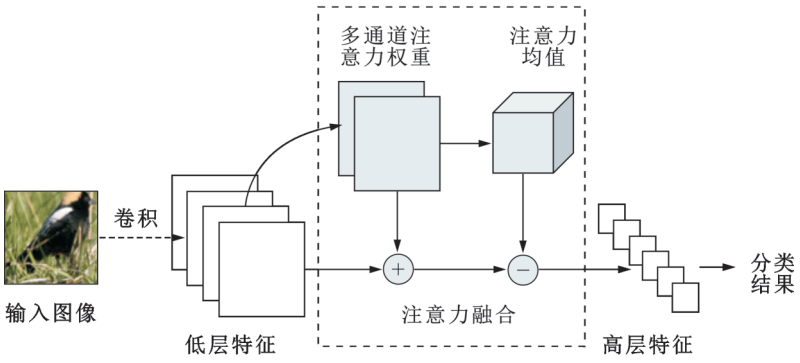

图1 基于多通道视觉注意力的细粒度图像分类模型结构图

Fig.1 Structure of multi-channel attention based fine-grained image classification model

2.1 特征提取

特征提取部分包含多个卷积层,这些卷积层可由预训练的网络转换而来。对于输入的二维图像,最后的卷积层的输出为 $H \times W \times D$ 的特征图,可看作 D 组特征,每一组包含 $N = H \times W$ 个信息,分别对应于相应的空间位置,低层特征可表示为 $\boldsymbol{X} \in \mathbf{R}^{(N\times D)} = \left\{\boldsymbol{x}_{i,d}\right\}$, $i \in \{1, \cdots, N\}$;$d \in \{1, \cdots, D\}$。

2.2 多通道注意力权重的产生

在网络模型中,低层特征 $\boldsymbol{X}$ 如上文所示,K 维度的多通道注意力对应于 K 个对输入信息的选择过程,对应的注意力权重为 $N \times K$ 的特征图,表示为 $\boldsymbol{A} \in \mathbf{R}^{(N\times K)} = \left\{\alpha_{i,k}\right\}$, $i \in \{1, \cdots, N\}$; $k \in \{1, \cdots, K\}$,

其中 K 为注意力权重图的数量。

多通道注意力相当于应用于卷积输出的二维特征的多端注意力[9]。在多通道注意力作用于模型中时，注意力对应于多个单独的对输入信息的选择过程，它们平行地作用于输入特征，即

$$\alpha_{i,k}=p\left(z_k=i|x_{1:N}\right)=\mathrm{softmax}\left(s_k\left(\boldsymbol{x}_i\right)\right) \tag{4}$$

式中：$s_k\left(\boldsymbol{x}_i\right)$ 为对应于第 k 个注意力的打分函数，为保证不同通道的注意力权重关注于特征图中不同空间位置，使用softmax函数作用于注意力权重的通道数 1 ： K。

对于 k 通道的输入 s，作用于其上的softmax 函数表示如下

$$\mathrm{softmax}\left(s\right)=\frac{e^{s_j}}{\sum_{k=1}^{K}s_k} \tag{5}$$

本方法采用的注意力打分函数 $s_k\left(\boldsymbol{x}_i\right)$ 基于注意力机制应用中最为常用的点积模型，并针对输入图像特征的特点对其进行归一化处理，表示如下

$$s_k\left(\boldsymbol{x}_i\right)=\left|\boldsymbol{x}_i\right|^{\mathrm{T}}\boldsymbol{W}_k \tag{6}$$

式中 $\left|\boldsymbol{x}_i\right|$ 表示对输入的低层特征 $\boldsymbol{x}_i$ 对应其维度 D 进行了L2归一化处理，这一处理有助于产生更为稳定的注意力权重；$\boldsymbol{W}_k\in\mathbf{R}^D$，将 D 维度输入转换为对应于 k 通道的输出。

注意力权重 $\alpha_{i,k}$ 可按如下方法进一步得到

$$\alpha_{i,k}=\mathrm{softmax}\left(\left|\boldsymbol{x}_i\right|^{\mathrm{T}}\boldsymbol{W}_k\right) \tag{7}$$

以上注意力权重的产生过程可利用神经网络中的卷积层和softmax层等常用操作实现，这保证了模型整体可端对端训练的特性。

2.3 多通道注意力权重图的作用

对于空间维度一致，通道数分别为 D 和 K 的低层特征和注意力权重，按照式(3)所示注意力的融合方法，两者的作用过程可写作

$$\mathrm{attention}_k\left(x_d\right)=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\alpha}_{i,k}\cdot\boldsymbol{x}_{i,d} \tag{8}$$

以矩阵的形式将低层特征表示为 $\boldsymbol{X}$，注意力权重表示为 $\boldsymbol{A}\in\mathbf{R}^{(N\times K)}=\left\{\boldsymbol{\alpha}_{i,k}\right\}$，则上述操作可表示为

$$\mathrm{attention}\left(\boldsymbol{X},\boldsymbol{A}\right)=\frac{1}{N}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A} \tag{9}$$

在操作后得到维度为 $D\times K$ 的高层特征。

2.4 注意力均值的引入

进一步地，在本文模型中引入了对应于每一组注意力权重的特征均值。注意力均值是网络参数，代表了所有数据对应的低层特征对应于每一通道注意力的均值。这一操作与图像表示中的VLAD特征的提取过程具有相似之处，VLAD被证明是一种有效的图像表示。特征均值的引入可以提取与类别更为相关的高阶特征，提高输出的融合结果的表达能力，提升分类效果。

对于上文中特征 $\boldsymbol{x}_{i,d}$ 与注意力权重 $\boldsymbol{\alpha}_{i,k}$，注意力均值可表示为 $\boldsymbol{\mu}_{d,k}$。则上文注意力融合方法可改写为

$$\text{attention}_k\left(x_d\right)=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\alpha}_{i,k}\cdot\left(\boldsymbol{x}_{i,d}-\boldsymbol{\mu}_{d,k}\right) \tag{10}$$

将注意力均值以矩阵的形式写为 $\boldsymbol{M}\in\mathbf{R}^{(D\times K)}=\left\{\boldsymbol{\mu}_{d,k}\right\}$,则注意力融合方法可表示为

$$\text{attention}\left(\boldsymbol{X},\boldsymbol{A}\right)=\frac{1}{N}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}-\boldsymbol{A}^{\mathrm{T}}\odot\boldsymbol{M}\right) \tag{11}$$

式中 $\odot$ 代表按 K 对应的维度点积的操作。

式(11)中,注意力融合的作用过程主要由矩阵和向量的乘法构成,可以方便地实现反向操作

$$\begin{aligned}\frac{\partial\left(\text{attention}\left(\boldsymbol{X},\boldsymbol{A}\right)\right)}{\partial\boldsymbol{X}}&=\frac{1}{N}\left(\boldsymbol{A}^{\mathrm{T}}\cdot\frac{\partial\boldsymbol{A}^{\mathrm{T}}}{\partial\boldsymbol{X}}-\boldsymbol{M}\odot\frac{\partial\boldsymbol{A}^{\mathrm{T}}}{\partial\boldsymbol{X}}\right)\\\frac{\partial\left(\text{attention}\left(\boldsymbol{X},\boldsymbol{A}\right)\right)}{\partial\boldsymbol{M}}&=\frac{1}{N}\frac{\partial\boldsymbol{A}^{\mathrm{T}}}{\partial\boldsymbol{X}}\end{aligned} \tag{12}$$

在进行包含减去注意力均值的融合操作后,输出的高层特征维度仍为 $D\times K$。

2.5 注意力权重参数及均值的初始化

上文中注意力权重的产生参数 $\boldsymbol{W}=\left\{\boldsymbol{W}_k\right\}$是网络模型中的关键参数,这一参数可按传统方法随机初始化,但引入与图像类别无相关性的类别信息对于得到更具描述性的注意力权重图具有很强的促进作用,同时,对这一参数进行初始化有助于加快网络收敛,可以减少训练时间。因此,引入一定的外部类别信息,采用聚类的方式对参数 $\boldsymbol{W}$ 进行初始化。

本方法使用正交匹配追踪(OMP-k)算法[10]对参数 $\boldsymbol{W}$ 进行初始化,求取如下运算的最小值

$$\begin{aligned}&\min_{W,s}\sum_{i=1}^{N_{\text{data}}}\left\|\boldsymbol{W}\boldsymbol{s}\left(\boldsymbol{x}_i\right)^{\mathrm{T}}-\boldsymbol{x}_i\right\|\\&\text{s.t.}\left\|W_j\right\|_2^2=1,\ \forall j\\&\left\|s\left(\boldsymbol{x}_i\right)\right\|_0\leqslant k,\ \forall i\end{aligned} \tag{13}$$

式中:$\left\|s\left(\boldsymbol{x}_i\right)\right\|_0$ 为 $s\left(\boldsymbol{x}_i\right)$ 中非零元素的个数;N_{data} 对应于所有用于初始化的数据。

当 k 取为1时,OMP-1算法也被称为增益形状向量量化(Gain-shape vector quantization)或球形 k 均值算法,可看作 k 均值算法的一种特殊形式。这一算法起到对归一化的特征进行聚类的作用。

在对模型参数进行初始化时,将训练数据输入网络卷积层,经网络前向运算收集对应的输出特征,再按照OMP-1算法计算得到权重。同时,数据均值 $\boldsymbol{M}=\left\{\boldsymbol{\mu}_{d,k}\right\}$也可使用由初始化数据收集到的特征与注意力权重进行初始化,表达式为

$$\mu_{d,k}=\frac{1}{N}\sum_{i=1}^{N_{data}}\boldsymbol{x}_{i,d}\cdot\boldsymbol{\alpha}_{i,k} \tag{14}$$

3 实验结果及分析

3.1 实验模型设置

本文提出模型的特征提取部分的卷积网络可由预训练模型得到,在实验中选择了在ImageNet数据集中预训练的VGG-16网络[11]作为这一部分的基础网络。经预训练的VGG-16网络可以有效地得到丰富的卷积特征,在很多深度神经网络模型中被作为基础。在本文模型中,截取预训练网络最后的卷积

层输出,即conv5_3,作为模型中的低层特征,特征维数为512。

在实验中选取网络卷积层输入图像尺寸为 512 像素 × 512 像素。图像在输入网络之前经过了几项常用的数据增强操作,包括以 224/256 的比例切取部分图像、随机对图像进行镜像、减去图像均值等。

如上文所述,注意力权重图生成部分可由卷积核大小为 1 × 1 的卷积层和softmax等操作实现,卷积层的参数由OMP-1方法初始化。在实验中,注意力权重图的通道数 K 是网络模型的关键参数,可通过实验确定其取值和分类精度的关系。在将基本的注意力权重图通道数设为 $K=32$ 的情况下,从卷积层输出的低层特征的通道数为 512,在注意力与低层特征融合后输出的高层特征为一长向量,其维数为 $32 \times 512 = 16\,384$。为增强这一高层特征作为图像表示的稳定性,对其进行L2归一化处理得到最终的高层特征。

注意力与特征融合得到的高层特征输入全连接层,输出维度与数据集对应的类别相对应,再经softmax层后可得到每一类别的概率输出。这一分类器中全连接层的输入维度较高,为加速网络训练速度,可收集训练图像对应高层特征向量训练线性SVM分类器,用SVM模型参数对全连接层的参数进行初始化。

3.2 数据集

为全面评测本文方法用于细粒度图像分类的性能,使用了CUB-200-2011鸟类数据集[3]、FGVC-Aircraft 飞行器数据集[12]和斯坦福Cars汽车数据集[13]等多个细粒度图像分类中常用的数据集进行评测。

Caltech-UCSD Birds-200-2011细粒度图像数据集,简称CUB-200-2011,是现阶段细粒度图像分类研究中最经典也最常用的数据集。CUB-200-2011数据集包含200种北美鸟类的图像,共计11 788张,按照数据集提供的划分,训练图像5 994张,测试图像5 794张。这一数据集具有类别间差异小、图像中鸟类姿态位置多样、训练数据量有限等具有挑战性的特点。

FGVC-Aircraft细粒度图像分类数据集包含102种不同型号的飞行器图像,每一种型号包含100张图像,共计10 200张图像,其中约三分之一作为测试。这一数据集中图像里的主要物体是不同型号的飞机,由于数据集许多飞机型号间划分非常详细,部分类别间相似性非常高;而飞机涂装和所处环境不同导致类别内部有较大变化,使得FGVC-Aircraft成为具有挑战性的细粒度图像数据集。

斯坦福Cars细粒度图像分类数据集包含196种不同型号的小汽车图像,共计16 185张,其中8 144张图像作为训练,其他作为测试。Cars数据集很多类别对应的汽车型号具有相同的车型和制造商,而同一类别汽车的视角、涂装又具有较大变化,具有很强的细粒度图像分类特点。

对于这几种细粒度图像分类数据集,本文均按照数据集提供的标准训练与测试划分使用,两者间没有重复的数据,保证了模型的有效性,同时便于与其他方法进行对比。

3.3 实验结果及分析

实验1 CUB-200-2011数据集分类结果

按照上文所述模型配置,本文基于多通道注意力的细粒度分类模型在CUB-200-2011数据集中得到了87.5%的分类精确度,如表1所示。对照方法中一部分使用了图像类别之外的额外监督信息,包括数据集提供的包围盒(Bounding box)和部位标注。在对照方法中,SPDA-CNN,Mask-CNN,CB-CNN,B-CNN和RA-CNN均和本文方法一样使用了VGG-16作为基础网络,这更有助于比较模型在图像低层特征的基础上提取分类有效信息的能力。从表1中实验结果可知,本方法相较于此前不使用额外标注的弱监督分类方法,分类精确度有显著提升;同时相较使用部位等数据集标注的方法,本方法达

到了同一水平的分类精确度。这一结果证明了基于多通道注意力的模型具有有效提取分类相关特征、对细粒度图像进行有效区分的能力。

表 1 不同方法在CUB-200-2011数据集中的分类精确度

Tab. 1 Classification accuracies of different methods on CUB-200-2011 dataset

方法	使用标注	分类精确度/%
PB R-CNN[14]	包围盒,部位	73.9
SPDA-CNN[7]	包围盒,部位	85.1
Mask-CNN (VGG-16)[15]	部位	85.4
Mask-CNN (ResNet-50)	部位	87.3
Two-level[6]		77.9
CB-CNN[16]		84.0
B-CNN[17]		84.1
ST-CNN[5]		84.1
PDFS[18]		84.5
RA-CNN[8]		85.3
本文方法		**87.5**

实验2 FGVC-Aircraft数据集和Cars数据集分类结果

按照上文所述配置,本文基于多通道注意力的细粒度分类模型在FGVC-Aircraft数据集中得到了88.4% 的分类精确度;在Cars数据集中得到了92.5% 的分类精确度。表2中展示了不同方法在这两个数据集中的比较结果,其中,B-CNN [D, D]作为基于深度神经网络的方法使用了VGG-16网络作为基础网络,与本文方法相同;B-CNN [D, M]则结合了VGG-16和VGG-M[21]两种网络提取的特征。从表中结果可知,本文方法相较于此前方法,在分类精确度上有显著提升。同时,结合网络模型的复杂度进行比较,可知在使用规模相同或更小的基础模型时,本文方法所采用的多通道注意力模型能够更为有效地提取与细粒度图像分类相关的特征。

表 2 不同方法在FGVC-Aircraft数据集和Cars数据集中的分类精确度

Tab. 2 Classification accuracies of different methods on FGVC-Aircraft and Cars datasets

方法	Aircraft数据集/%	Cars数据集/%
Chai等人[19]	72.5	78.0
Fisher Vector[20]	80.7	82.7
B-CNN[17] [D, M]	83.9	91.3
B-CNN [D, D]	84.1	90.6
本文方法	**88.4**	**92.5**

实验3 注意力权重通道数

对于本文所述多通道注意力模型,式(11)中多通道注意力权重图 $\boldsymbol{A}$ 的通道维度 K 是一项关键参数。注意力权重通道数较低时,可能难以提供足够的注意力信息,影响分类结果;注意力权重通道较多时,会增加模型参数进而增加模型的计算复杂度,同时增加注意力作用后输出的图像表示向量的维数,难以获得紧凑的图像表示。表3展示了注意力权重图通道数逐渐增加,分别取4,8,16,32和64等值时,

按上文所述模型配置在CUB-200-2011数据集中训练得到的模型分类精确度。实验结果中,注意力通道数取为4时,分类精确度与注意力通道数为8时相差较大,达7.2%,这证明此时注意力权重特征不足以提供足够的信息,对分类精确度有较大影响。注意力权重图通道数不小于16时,分类精确度均较为接近,此时模型包含了充分的注意力信息。由实验结果可知,在应用时取注意力权重图通道数为16或32可在分类精确度和模型复杂度间取得较好的平衡。

表 3 采用不同注意力权重图通道数的模型在CUB-200-2011数据集中的分类精确度

Tab. 3 Classification accuracy for the proposed model with different number of channels of the attention weight on CUB-200-2011 dataset

注意力权重图通道数	分类精确度/%
4	78.4
8	85.6
16	87.0
32	87.5
64	**87.6**

实验4 图像表示特征

本文所述模型中,式(11)中注意力作用后输出的高层特征向量可作为输入图像的一种特征表示,此时,以模型的这一层作为输出,得到高层向量作为图像的特征提取器。这一向量的维度和对图像分类的精确度是评估模型性能的关键因素。表4对不同的具有提取图像特征向量能力的图像分类模型作出了比较,以特征向量维度和CUB-200-2011数据集上的分类精确度作为评测结果。其中,本文方法使用了注意力权重通道数分别为16和32两种配置。在对照方法中,CNN-FC使用VGG-16的fc7层4 096维输出作为表示向量,VGG-16也是表中所有方法的基础网络;CNN-IFV从神经网络的fc7和fc8层输出进行降维得到Fisher vector作为图像表示向量;B-CNN使用双线性池化(Bilinear pooling)将两组卷积的512维输出融合得到维数非常高的表示向量;CB-CNN方法对FB-CNN进行改进,在保持分类精确度的同时降低了表示向量维度。从表中结果可知,本方法在保持维度较低的表示向量的同时,在细粒度图像分类任务中取得了更佳的分类结果。这证明了模型中采用的注意力作用方法可以更为有效地提取到有助于分类的重要信息。

表 4 不同图像表示模型的特征向量维度与分类精确度

Tab. 4 Comparison of different models' feature vector length and classification accuracy

方法	特征向量维度	分类精确度/%
CNN-FC	4 096	66.1
CNN-IFV[22]	51 200	64.2
B-CNN[17]	2.6e5	84.0
CB-CNN-RM[16]	8 192	83.8
CB-CNN-TS[16]	8 192	84.0
本文方法($K=16$)	8 192	87.0
本文方法($K=32$)	16 384	**87.5**

4 结束语

本文提出并验证了一种用于细粒度图像分类的深度神经网络模型,这一模型中应用了多通道视觉注意力,并在注意力与图像的融合过程中减去注意力对应均值提取高阶信息,同时提出了对注意力参数进行初始化的方法,构成了一套可端对端训练的图像分类框架,同时可用于提取紧凑的图像表示。在CUB-200-2011等多种细粒度图像分类数据集上的实验证明,相比于传统注意力模型和其它经典细粒度图像分类框架,本文基于多通道视觉注意力的细粒度图像分类模型在分类精度上具有显著优势。

参考文献:

[1] 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[J].自动化学报,2017,43(8):1306-1318.
Luo Jianhao, Wu Jianxin. A survey on fine-grained image categorization using deep convolutional features[J]. Acta Automatica Sinica, 2017, 43(8):1306-1318.

[2] Zhao B, Feng J, 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.

[3] Wah C, Branson S, Welinder P, et al. The caltech-ucsd birds-200-2011 dataset[R]. Technical Report CNS-TR-2011-001. Pasadena, CA, USA: California Institute of Technology, 2011.

[4] Borji A, Itti L. State-of-the-art in visual attention modeling[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1):185-207.

[5] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks[C]//Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015. Montreal, Quebec, Canada: Curran Associates, 2015:2017-2025.

[6] 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]//IEEE Conference on Computer Vision and Pattern Recognition 2015. Boston, MA, USA: IEEE, 2015: 842-850.

[7] Zhang H, Xu T, Elhoseiny M, et al. SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition [C]//IEEE Conference on Computer Vision and Pattern Recognition 2016. Las Vegas, NV, USA: IEEE, 2016:1143-1152.

[8] Fu J, Zheng H, Mei T. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition 2017. Honolulu, HI, USA: IEEE, 2017:4476-4484.

[9] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. Long Beach, CA, USA: Curran Associates, 2017: 6000-6010.

[10] Coates A, Ng A. Neural networks: Tricks of the trade[M]. 2nd ed. Berlin, Heidelberg: Springer, 2012:561-580.

[11] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB / OL]. http://arxiv.org/abs/1409.1556, 2014.

[12] Maji S, Rahtu E, Kannala J, et al. Fine-grained visual classification of aircraft[EB / OL]. http://arxiv.org/abs/1409.1556, 2013.

[13] Krause J, Stark, M, Deng J, et al. 3D object representations for fine-grained categorization[C]//4th International IEEE Workshop on 3D Representation and Recognition. Sydney, Australia: IEEE, 2013.

[14] Zhang N, Donahue J, Girshick R, et al. Part-based R-CNNS for fine-grained category detection[C]//13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014:834-849.

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

[16] Gao Y, Beijbom O, Zhang N, et al. Compact bilinear pooling[C]// IEEE Conference on Computer Vision and Pattern Recogni

tion 2016.Las Vegas, NV, USA:IEEE, 2016:317-326.

[17] Lin T, Chowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition[C]//IEEE International Conference on Computer Vision.Santiago, Chile:IEEE, 2016:1449-1457.

[18] Zhang X, Xiong H, Zhou W, et al. Picking deep filter responses for fine-grained image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition 2016.Las Vegas, NV, USA:IEEE, 2016:1134-1142.

[19] Chai Y, Lempitsky V, Zisserman A. Symbiotic segmentation and part localization for fine-grained categorization[C]// IEEE International Conference on Computer Vision.Sydney, Australia:IEEE, 2013:321-328.

[20] Gosselin P, Murray N, Jégou H, et al. Revisiting the fisher vector for fine-grained classification[J]. Pattern Recognition Letters, 2014, 49:92-98.

[21] Chatfield K, Simonyan K, Vedaldi A, et al. Return of the devil in the details: Delving deep into convolutional nets[C]//British Machine Vision Conference.Nottingham, UK:BMVA Press, 2014.

[22] Cimpoi M, Maji S, Vedaldi A. Deep filter banks for texture recognition and segmentation[C]//IEEE Conference on Computer Vision and Pattern Recognition 2015.Boston, MA, USA:IEEE, 2015:3828-3836.

作者简介:

王培森(1993-),硕士研究生,研究方向:计算机视觉,E-mail: wangps@mail.ustc.edu.cn。

宋彦(1972-),男,副教授,研究方向:音、视频分析和检索,E-mail: songy@ustc.edu.cn。

戴礼荣(1962-),男,教授,研究方向:语音识别和信号处理,E-mail: lrdai@ustc.edu.cn。

(编辑:夏道家)