doi:10.3969/j.issn.1672-1292.2023.03.005

一种新的基于通道-空间融合注意力及 SwinT 的细粒度图像分类算法

姜　昊,凌　萍,陈寸生保

(江苏师范大学计算机科学与技术学院,江苏 徐州 221116)

[**摘要**]　细粒度图像分类是计算机视觉领域的一大分类任务,其难点在于如何通过类别监督信息自主地找到判别性区域. 提出一种新的通道-空间融合注意力模块,基于该模块设计了一种新的 Swin Transformer 算法 SwinT-NCSA(a Swin Transformer based on a novel channel-spatial attention module),分别从通道维和空间维同时提取特征,再将其融入到 Swin Transformer 模型中以提高其小尺度中多头注意力信息的提取能力. SwinT-NCSA 算法特别关注了对分类有用的区域,同时忽视对分类无用的背景区域,以此在细粒度图像分类任务中达到较高的分类准确率. 在 FGVC Aircraft 飞机数据集、CUB-200-2011 鸟类数据集和 Stanford Cars 车类数据集 3 个公共数据集上的实验表明,SwinT-NCSA 算法可以分别取得 93.3%、88.4%和 94.7%的准确率,优于同类算法.

[**关键词**]　细粒度图像分类,Swin Transformer,通道-空间融合注意力模块,深度学习,弱监督学习
[**中图分类号**]TP183　[**文献标志码**]A　[**文章编号**]1672-1292(2023)03-0036-07

A New Fine-grained Image Classification Algorithm Based on Channel-Space Fusion Attention and SwinT

Jiang Hao, Ling Ping, Chen Cunshengbao

(School of Computer Science and Technology, Jiangsu Normal University, Xuzhou 221116, China)

Abstract: Fine-grained image classification is a major classification task in the computer vision field. Its difficulty lies in how to automatically find the discriminant regions through category supervision information, for which this paper proposes a novel channel-spatial fusion attention module, and based on it, designs a new Swin Transformer algorithm (a Swin Transformer based on a novel channel-spatial attention module, SwinT-NCSA). The proposed algorithm simultaneously extracts features from the channel dimension and spatial dimension, and then integrates them into the Swin Transformer model to improve the extraction ability of multi-head attention information in its small scale. The SwinT-NCSA algorithm pays a particular focus on regions useful for classification, while ignoring background regions useless for classification, to achieve high classification accuracy in fine-grained image classification tasks. Experiments on the FGVC Aircraft aircraft dataset, Caltech-UCSD Birds-200-2011 dataset and the Stanford Cars vehicle class dataset public dataset show that the SwinT-NCSA algorithm can achieve 93.3%, 88.4% and 94.7% accuracy respectively, outperforming peer algorithms.

Key words: fine-grained image classification, Swin Transformer, channel-spatial fusion attention module, deep learning, weak supervised learning

细粒度图像分类是计算机视觉领域图像分类任务中一项富有挑战性的任务,其目标是对图像中的物体在同一大类下的许多子类中进行正确分类[1],因此也被称为子类别图像分类[2]. 细粒度图像分类存在类内相似度小且类间相似度大的分类难点[3],研究者对此提出了诸多解决算法.

Lin 等[4]提出了包含 2 个 VGG 网络的双线性卷积神经网络(bilinear CNN),2 个 VGG 网络可分别用于检测图像的目标区域和从检测出的目标区域类提取出有用的特征信息,该模型的双线性池化(bilinear

收稿日期:2023-02-21.
基金项目:国家自然科学基金面上项目(61872168)、江苏师范大学研究生科研与实践创新计划项目(2022XKT1534).
通讯作者:凌萍,博士,副教授,研究方向:计算智能、数据挖掘、支持向量机. E-mail:6020000012@jsnu.edu.cn

pooling)[5]提供了比普通的线性模型更好的特征融合效果,但双线性模型的维度太高,很难泛化使用. Wang 等[6]通过在 CNN 网络中学习一个卷积过滤器库来提取各个种类辨别性块,从而不需要额外的边界框注释,即可达到提高细粒度图像分类任务的效果. Yang 等[7]从目标检测中的 FPN 寻找灵感[8],设计了 Navigator、Teacher、Scrutinizer 3 个网络来分别生成局部判别性候选区域、候选区域的得分、原始特征和 Teacher 网络得到的具有丰富信息的区域特征融合,这 3 个网络共同完成细粒度图像分类任务.

在上述各种方法中,卷积神经网络(convolu-tional neural network,CNN)是主流的方法之一. CNN 在带有强监督信息的环境和带有弱监督信息的环境下均可进行细粒度图像分类任务. 一般来说,基于强监督信息的方法普遍比基于弱监督信息的方法分类效果好,但强监督信息的收集成本过高,不利于实际应用,因此目前主流的研究方向是基于弱监督信息的方法. 然而,一般的 CNN 在弱监督信息的环境下很难做到更多地注意类间相似度的同时,尽可能忽略类内相似度,因此只利用卷积操作实现局部特征提取的方式在该领域的任务中表现一般. 同时,各种已有的为解决细粒度图像分类的 CNN 网络模型有的设计结构较为复杂,从而带来巨大的算法代价,有的在实验中的表现并未达到预期,进而给出欠佳的实验结果. 所以,CNN 在细粒度图像分类领域中不占优势.

可见,一个能够较好地完成细粒度图像分类任务的网络模型,需提取出对分类有用的类间特征,同时忽略对分类无用的类内及背景特征. 这一理念恰与自然语言处理领域中的 Attention 机制[9]基本一致,基于该思路的研究成果近几年也颇丰. 2015 年 Xiao 等[10]提出了两级注意力模型(two level attention),随后多级注意力模型也被提出,实验表明此类模型的注意力机制对于细粒度图像的局部特征提取有出色的效果. 随着深度学习领域研究的不断推进,2020 年 Google 团队提出用于图片分类的 Vision Transformer 网络模型[11],首次完成了在计算机视觉领域任务中只用 Transformer 不用卷积网络的先例,该模型在整张特征图上做多头注意力运算,使得网络模型计算量极大,不利于网络训练. 李佳盈等[12]提出了一种基于 Vision Transformer 的细粒度图像分类方法,但该方法无法从多尺度解析图片的深层信息. Liu 等[13]提出了 Swin Transformer 网络模型,该网络模型相比于 Vision Transformer,只在相互独立的窗口内而非整张特征图上计算注意力,通过滑动窗口实现不同窗口间的信息交互,最终达到并行训练、抽取全局特征的目的[14],大大减少了计算量的开销.

Swin Transformer 是一个基于多尺度的特征提取网络,结构简单,在计算机视觉领域的众多任务中[15-18]表现均很出色. 在细粒度图像分类领域,目前大多数是基于 CNN 来设计网络模型展开研究,鲜少有将 Swin Transformer 直接应用于此领域. 字节跳动团队曾基于 Vision Transformer 设计了应用在细粒度图像分类领域的模型 TransFG,但该模型未使用移动窗口机制以至于不利于训练高分辨率的图片.

本文提出一种新的通道-空间融合注意力模块,并基于此模块设计了一种新的 Swin Transformer 算法 SwinT-NCSA(a Swin Transformer based on a novel channel spatial attention module). 该算法分别从通道维和空间维同时提取特征,再将其融入到 Swin Transformer 模型中以提高对小尺度中多头注意力信息的提取能力. SwinT-NCSA 算法特别关注了对分类有用的区域,同时忽视对分类无用的背景区域. 在公共数据集上的实验表明,SwinT-NCSA 相比于 Swin Transformer 及同类网络,在预测准确率方面有更优的表现.

1　相关工作

Swin Transformer 网络使用了类似卷积神经网络中的层次化构建方法,即每通过一个 Stage 下采样一次来实现多尺度,使得 Swin Transformer 可实现计算机视觉领域的多种视觉任务.

Swin Transformer 使用了 Windows Multi-Head Self-Attention(W-MSA)的概念,在下采样中,将特征图划分成多个互不相交的区域窗口,且 Multi-Head Self-Attention(MSA)只在每个窗口内进行. MSA 会划分多个 head 分别计算,最后将所有 head 进行融合得到最终的结果,用公式可表示为:

$$\mathrm{head}_i = \mathrm{Attention}(\boldsymbol{Q}\boldsymbol{W}_{i_Q}, \boldsymbol{K}\boldsymbol{W}_{i_K}, \boldsymbol{V}\boldsymbol{W}_{i_V}), \tag{1}$$

$$\mathrm{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{Concat}(\mathrm{head}_1, \cdots, \mathrm{head}_h)\boldsymbol{W}^O, \tag{2}$$

式中,$\boldsymbol{W}_{i_Q}, \boldsymbol{W}_{i_K}, \boldsymbol{W}_{i_V}$分别是 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 的生成矩阵. 相对于 Vision Transformer 中直接对整个特征图(feature map)进行 Multi-Head Self-Attention 计算,这样做的目的是为了减少计算量,尤其是在浅层特征图很大的时候.

单头 Self-Attention 的公式如下：

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{SoftMax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V}. \tag{3}$$

对于普通的 MSA 模块,特征图中的每个像素点都要通过 $\boldsymbol{W}_Q,\boldsymbol{W}_K,\boldsymbol{W}_V$ 生成对应的 query($\boldsymbol{q}$),key($\boldsymbol{k}$)及 value($\boldsymbol{v}$). 假设 $\boldsymbol{q},\boldsymbol{k},\boldsymbol{v}$ 的向量长度与 feature map 的深度 C 保持一致,则对应所有像素生成 $\boldsymbol{Q}$ 的过程如下：

$$\boldsymbol{A}^{hw\times C}\times\boldsymbol{W}_Q^{C\times C}=\boldsymbol{Q}^{hw\times C}, \tag{4}$$

式中,$\boldsymbol{A}^{hw\times C}$为所有像素拼接在一起得到的一个 $h\times w$ 行 C 列的矩阵($h\times w$ 个像素点,深度为 C),$\boldsymbol{W}_Q^{C\times C}$ 为生成 query 的变换矩阵,$\boldsymbol{Q}^{hw\times C}$为 feature map 中的每个像素点通过 $\boldsymbol{W}_Q^{C\times C}$ 得到的 query 再拼接在一起得到的矩阵.

根据矩阵运算的计算量公式可知生成 $\boldsymbol{Q}$ 的计算量为 $hw\times C^2$,同理可得生成 $\boldsymbol{K},\boldsymbol{V}$ 的计算量均为 $hw\times C^2$,则计算量的总和为 $3\times hw\times C^2$. 然后是 $\boldsymbol{Q}$ 与 $\boldsymbol{K}^{\mathrm{T}}$ 相乘,由于 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 维数相同,故计算量为$(hw)^2\times C$,若忽略$\sqrt{d}$以及 SoftMax 的计算量,最后乘 $\boldsymbol{V}$ 的计算量仍是$(hw)^2\times C$,全部求和就是 $3hwC^2+2(hw)^2C$,这是单头的 Self-Attention 的计算量,多头注意力模块相比单头注意力模块的计算量仅多了最后一个融合矩阵的计算量 hwC^2,故多头的 Self-Attention 的计算量为 $4hwC^2+2(hw)^2C$.

对于 W-MSA 而言,首先要将 feature map 分割成多个窗口,假设每个窗口的高宽都为 S,则总共会有$(h/S)\times(w/S)$个窗口,然后对每个窗口内部独立地使用多头注意力模块,由前面计算的高为 h、宽为 w 的 feature map 的多头注意力的计算量可知,由于此时窗口的高和宽均为 S,故每个窗口的计算量为 $4(SC)^2+2(S)^4C$,又因为有$(h/S)\times(w/S)$个窗口,所以总的计算量为$(h/S)\times(w/S)\times(4(SC)^2+2(S)^4C)=4hwC^2+2S^2hwC$. 因为 $S<h$,$h=w$,所以 $S^2<hw$,即 W-MSA 拥有更少的计算量,且划分的窗口越多越节省计算量.

这样做虽然减少了总体的计算量,但会阻隔不同窗口之间的信息传递,所以 Shifted Windows Multi-Head Self-Attention(SW-MSA)的概念又被提出,其核心思想是通过平移特征图后再分割窗口以实现不同窗口之间的信息交互. 将 W-MSA 和 SW-MSA 当作组合成对使用,构成 Swin Transformer Block 模块,最终可实现 Swin Transformer 网络. 两个连续的 Swin Transformer Block 模块结构如图 1 所示.

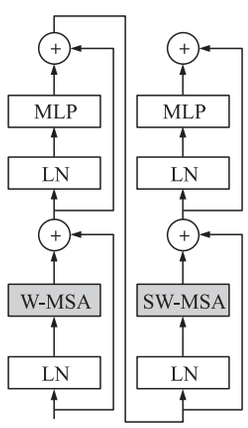

图 1　两个连续的 Swin Transformer Blocks

Fig. 1　Two consecutive Swin Transformer Blocks

2　SwinT-NCSA 算法

2.1　SwinT-NCSA 算法过程

SwinT-NCSA 网络模型的主干是目前较为流行的 Swin Transformer 模型,但会在原有模型的多个 Stage 模块后各增加一个新的通道-空间融合注意力模块,来保证在下采样之前学习到更多图片上的局部细节特征,也即在除最后两个 Stage 模块以外的每个 Stage 模块后添加一个通道-空间融合注意力模块,以此来更好地提取图像特征. SwinT-NCSA 模型结构如图 2 所示,输入网络的是经过重塑(reshape)后得到的 224×224 的 RGB 三通道图像,黑白图像需要先转为 RGB 图像后才可输入. 图像首先经下采样后得到 56×56×96 的特征图,将其输入 Stage1 模块(图 1 中的×2 和×6 均表示在 Stage 中堆叠 Swin Transformer Block 模块的个数),经过 Swin Transformer Block 模块后先经通道-空间融合注意力模块继续充分提取图像局部细小特征,随后再进行下采样得到 28×28×192 的特征图. 然后将其输入到 Stage2 模块,Stage2 模块后进行同样的设计,Stage3、Stage4 后则不添加该模块,最终从 Stage4 后添加的线性层得到分类结果. 该模型实际是在 Transformer 的基础上融合卷积层和线性层来共同提取图片特征,以期望达到更好的效果.

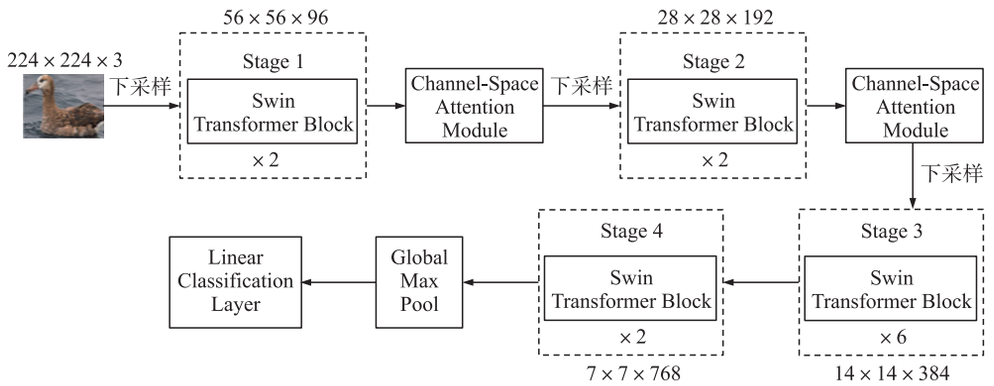

图 2　模型整体结构图

Fig. 2　Model overall structure diagram

2.2　通道-空间融合注意力模块

经典 Swin Transformer 中的 MLP(多层感知机)模块只使用了非线性层提取注意力(Attention)模块后的信息,而通道-空间融合注意力模块在 MLP 模块后通过在通道维继续使用非线性层提取更深层通道信息,并利用卷积同时并行地提取空间维上的局部特征,以此来更准确地捕获经过 Attention 计算后图片局部细节上的特征. 通道-空间融合注意力模块并不会改变特征图的高、宽和通道的维数,该模块结构如图 3 所示.

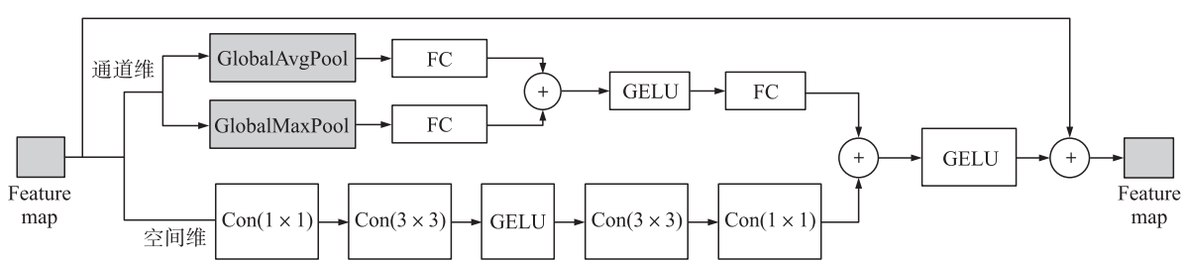

图 3　通道-空间融合注意力模块

Fig. 3　Channel-spatial fusion attention module

具体来说,在经典的 Swin Transformer 网络模型结构中将 Stage1 后得到的特征图经过 reshape 后得到 $F=F_{\text{Stage1}}\in\mathbf{R}^{H'\times W'\times C'}$,其中,$H'$,$W'$,$C'$分别表示特征图的高、宽和通道数. 接着该特征图 F 分别进入两个不同的路径,即通道注意力路径和空间注意力路径,可用 $M_C(F)\in\mathbf{R}^{C'}$ 和 $M_P(F)\in\mathbf{R}^{H'\times W'}$ 分别表示通道注意图和空间注意力图. 为了有效计算通道注意力,需要对输入特征图的空间维度进行压缩,对空间信息聚合的常用方法是平均池化,而最大池化收集了独特的物体特征,可以推断更细的通道上的注意力. 因此,平均池化和最大池化后得到的特征是同时使用的. 通道注意力路径 $M_C(F)$ 的计算可表示为:

$$M_C(F)=\sigma(FC(\text{GlobalAvgPool}(F))+FC(\text{GlobalMaxPool}(F))),\tag{5}$$

式中,GlobalAvgPool 表示全局平均池化;GlobalMaxPool 表示全局最大池化;FC 表示全连接层;σ 表示 GELU 激活函数. 式(5)表示将特征图 F 分别同时经过平均池化层和最大池化层后再经过全连接层,然后对两者进行按元素相加,再通过 GELU 激活函数进行非线性映射,最终得到通道维的输出.

空间注意力路径 $M_P(F)$ 的计算可表示为:

$$M_P(F)=\text{Conv}^{1\times1}(\text{Conv}^{3\times3}((\text{Conv}^{3\times3}(\text{Conv}^{1\times1}(F))))),\tag{6}$$

式中,Conv 表示卷积操作,1×1 和 3×3 表示卷积核大小分别为 1 和 3;σ 表示 GELU 激活函数. 将特征图 F 经一系列卷积操作来提取局部特征后将 $M_C(F)$ 升维,再和 $M_P(F)$ 按元素相加后经 GELU 激活函数,并从残差连接[19]中得到启发,为了信息前后向传播更加顺畅,低层特征可直接传递到高层网络,再和原特征图按元素相加得到最终的模块输出 F_{CSAM},该计算过程可表示为:

$$F_{\text{CSAM}}=F+\sigma(M_C(F)+M_P(F)).\tag{7}$$

通过对通道-空间融合注意力模块结构的分析可知,对通道维单独地进行特征提取可以更好地将色

彩信息学习到模型中,而对空间维的单独学习可以让网络更专注于纹理特征的学习,最后将两者融合起来得到与原先特征图一致的形状.

3　实验与结果

本节通过消融实验论证所提出的通道-空间融合注意力模块的技术可行性及有效性,并在 3 个标准细粒度图像分类数据集上通过对比实验比较 SwinT-NCSA 与同类算法分类的准确率.

3.1　实验数据集

如表 1 所示,本文选用了 3 个公开数据集进行实验验证,分别是 CUB-200-2011 鸟类数据集[20]、FGVC Aircraft 飞机数据集[21]和 Stanford Cars 车类数据集[22].

表 1　细粒度图像分类数据集

Table 1　Datasets of fine-grained image classification

数据集	类别数	训练集	测试集
CUB-200-2011	200	5 994	5 794
FGVC Aircraft	100	6 667	3 333
Stanford Cars	196	8 144	8 041

3.2　实验平台及参数

SwinT-NCSA 是在 Windows10 操作系统下基于 Python3.8 和 Pytorch1.7.0 实现的. SwinT-NCSA 使用了在 Image Net 数据集上预训练的 Swin Transformer Tiny 网络参数作为主干网络的预训练权重,通过微调技术将其应用在本文的改进模型中,以便 SwinT-NCSA 能更快收敛. 输入的训练图像经随机裁剪为 224×224 的图像并进行随机水平翻转,模型采用 AdamW 优化器,损失函数为交叉熵损失函数,初始学习率为 0.001,采用指数型衰减法,后续逐步更新学习率,batch size 为 16.

3.3　消融实验

消融实验是指移除或改变某一模块后观察网络整体性能的变化,从而判定该模块对整体网络的作用. SwinT-NCSA 主要由 Patch Embedding、Swin Transformer Block、通道-空间融合注意力模块及 Patch Merging(下采样)模块组成. 为了验证该模型中每个模块的有效性,实验以 Swin Transformer 模型为基线模型,分别在不同 Stage 模块后添加通道-空间融合注意力模块,观察所得结果,以此验证 SwinT-NCSA 模型设计细节(即在除最后两个 Stage 模块以外的每个 Stage 模块后添加一个通道-空间融合注意力模块)的正确性和有效性. 实验中均使用相同的预训练权重和超参数设置. 在 CUB-200-2011 数据集上的实验结果如表 2 所示.

表 2　消融实验结果

Table 2　Results of ablation experiments

After Stage1	After Stage2	After Stage3	After Stage4	Acc/%
√	×	×	×	87.8
×	√	×	×	88.1
×	×	√	×	87.6
×	×	×	√	87.2
√	√	√	√	87.4
×	√	√	×	88.2
×	×	×	×	87.2
√	√	×	×	88.4

注:√表示添加了 CSAM 模块,×表示未添加 CSAM 模块.

实验结果表明,在 Stage1 和 Stage2 后分别添加通道-空间融合注意力模块可在 CUB-200-2011 数据集上得到 88.4%的分类准确率,且对基线而言有 1.1%的提升,说明通道-空间融合注意力模块和 Swin Transformer 的融合在细粒度图像分类任务中有效. 这是因为该模型是一种多尺度的网络模型,对于 Stage1 和 Stage2 而言可以关注较小的感受野区域,而对于细粒度图像分类,能否正确分类的关键点往往都在图片中局部细小的特征上,所以在 Stage1 和 Stage2 后分别添加通道-空间融合注意力模块会使得网络可以从一开始就更好地关注图片中的局部细小特征,从而达到比较好的效果.

3.4　对比实验

为了观察 SwinT-NCSA 在真正细粒度图像分类任务中的表现,将 SwinT-NCSA 与其他基于弱监督信息的主流网络模型进行了对比实验,为了保证实验结果的可靠性,实验过程所有网络模型的训练集和测试集

数量完全相同,SwinT-NCSA 与其他算法在不同数据集上的分类准确率对比如表 3 所示.

表 3　SwinT-NCSA 与同类算法分类准确率的比较

Table 3　Comparison of classification accuracy of SwinT-NCSA and peer algorithms　%

模型	CUB-200-2011	FGVC Aircraft	Stanford Cars
Bilinear-CNN[23]	84.1	84.1	91.3
DT-RAM[24]	86.0	—	93.1
MA-CNN[25]	86.5	89.9	92.8
Boost-CNN[26]	85.6	88.5	92.1
NTS-Net[27]	87.7	91.4	93.9
WS-DAN[28]	89.4	93.0	94.5
Swin Transformer	87.2	91.9	92.8
SwinT-NCSA	88.4	93.3	94.7

实验结果表明,对比各种 CNN 模型,Swin Transformer 的分类效果较好,这说明视觉 Transformer 在细粒度图像分类问题上确实有良好的表现. SwinT-NCSA 在 3 个数据集上给出了比 Swin Transformer 更高的分类准确率,并在 FGVC Aircraft 飞机数据集和 Stanford Cars 车类数据集上达到了非常好的效果,说明 SwinT-NCSA 与经典 Swin Transformer 相比在局部细小特征的提取上有更好的性能. 但 SwinT-NCSA 在 CUB-200-2011 数据集上效果并不突出,原因可能在于样本数量不足,鸟类这一活物会做出不同的肢体动作,每一类鸟只有 30 张不到的图片作为训练集,使得网络很难学出极好的效果;当训练集和测试集按 8∶2 划分时,SwinT-NCSA 在 CUB-200-2011 数据集上的准确率从 88.4%提升到 89.3%,说明了提高训练样本的数量能极大地提高 SwinT-NCSA 对细粒度图像的分类准确率. 这也是 Swin Transformer 的不足之处,即使用预训练模型来帮助更快收敛,但单类样本数据量不足仍难训练出理想的效果.

4　结论

本文提出一种新的通道-空间融合注意力模块,将其融入 Swin Transformer 中,进而提出 SwinT-NCSA 算法. SwinT-NCSA 旨在提高网络模型对小尺度下局部特征的提取能力,并配合其中的 Attention 机制更好地关注了对分类有用的区域,同时忽视对分类无用的背景区域. 在 CUB-200-2011 鸟类数据集、FGVC Aircraft 飞机数据集和 Stanford Cars 车类数据集 3 个主流数据集上的实验验证了该模型的有效性,证明了 SwinT-NCSA 在实际的细粒度图像分类任务中比同类算法有更优的表现. 但该模型还存在收敛速度慢、参数量增多的问题,将在后续工作中继续完善.

[参考文献](References)

[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.

[2] 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.

[3] WEI X S, WU J X, CUI Q. Deep learning for fine-grained image analysis: a survey[J/OL]. arXiv Preprint arXiv: 1907.03069v1,2019.

[4] LIN T Y, ROYCHOWDHURY A, MAJI S. Bilinear CNN models for fine-grained visual recognition[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015.

[5] LIN T Y, MAJI S. Improved bilinear pooling with CNNs[J/OL]. arXiv Preprint arXiv: 1707.06772, 2017.

[6] WANG Y M, MORARIU V I, DAVIS L S. Learning a discriminative filter bank within a CNN for fine-grained recognition [C]//Proceedings of the 2018 IEEE/CVF conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 4148-4157.

[7] YANG Z, LUO T G, WANG D, et al. Learning to navigate for fine-grained classification[C]//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: ECCV, 2018.

[8] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Honolulu, USA: IEEE, 2017.

[9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International

Conference on Neural Information Processing Systems. Long Beach, USA: NIPS, 2017.
[10] 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]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015.
[11] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J/OL]. arXiv Preprint arXiv: 2010.11929, 2021.
[12] 李佳盈,蒋文婷,杨林,等. 基于 ViT 的细粒度图像分类[J]. 计算机工程与设计,2023,44(3):916-921.
[13] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021.
[14] XU Y F, WEI H P, LIN M X, et al. Transformers in computational visual media: a survey[J]. Computational Visual Media, 2022, 8(1): 33-62.
[15] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020.
[16] MEINHARDT T, KIRILLOV A, LEAL-TAIXE L, et al. Trackformer: multi-object tracking with transformers[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022.
[17] YANG H H, FU Y W. Wavelet U-Net and the chromatic adaptation transform for single image dehazing[C]//Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP). Taipei, China: IEEE, 2019.
[18] MEI J B, WANG M M, LIN Y N, et al. TransVOS: video object segmentation with transformers[J/OL]. (2021-06-01). arXiv Preprint arXiv: 2106.00588, 2021.
[19] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016.
[20] WAH C, BRANSON S, WELINDER P, et al. The caltech-UCSD birds-200-2011 dataset[R]. Pasadena, USA: California Institute of Technology, 2011.
[21] MAJI S, RAHTU E, KANNALA J, et al. Fine-grained visual classification of aircraft[J/OL]. (2013-06-21). arXiv Preprint arXiv: 1306.5151, 2013.
[22] KRAUSE J, STARK M, DENG J, et al. 3D object representations for fine-grained categorization[C]//Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops. Sydney, Australia: IEEE, 2013.
[23] LIN T Y, ROYCHOWDHURY A, MAJI S. Bilinear CNN models for fine-grained visual recognition[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015.
[24] LI Z C, YANG Y, LIU X, et al. Dynamic computational time for visual attention[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). Venice, Italy: IEEE, 2017.
[25] ZHENG H L, FU J L, MEI T, et al. Learning multi-attention convolutional neural network for fine-grained image recognition [C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017.
[26] MOGHIMI M, BELONGIE S, SABERIAN M, et al. Boosted convolutional neural networks[C]//Proceedings of the 2016 British Machine Vision Conference (BMVC). York, UK: BMVA, 2016.
[27] YANG Z, LUO T G, WANG D, et al. Learning to navigate for fine-grained classifycation[C]//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: ECCV, 2018.
[28] HU T, QI H G, HUANG Q M, et al. See better before looking closer: weakly supervised data augmentation network for fine-grained visual classification[J/OL]. (2019-01-26). arXiv Preprint arXiv: 1901.09891, 2019.

[责任编辑:严海琳]