DOI:10.19651/j.cnki.emt.2314646

基于改进的 Transformer 细粒度图像识别算法研究*

李冰锋 刘 帅 杨 艺
(河南理工大学电气工程与自动化学院 焦作 454000)

摘 要:针对细粒度图像识别存在类间差异小、难以区分等问题,本文通过提升网络对图像细节特征的表达能力,来改善这一问题。为此,设计了一种基于改进的 Transformer 细粒度识别算法。首先,可变形卷积令牌嵌入通过自适应调整采样点的位置,来改变卷积操作范围及其卷积核的形状,从而增强网络模型对空间信息的感知能力,以获取更为精准的空间信息;其次,高效相关通道注意力机制通过对通道的自动选择,将通道注意力的计算从通道相邻转换成语义相似,来捕获语义相似的通道信息。而精准的空间信息和语义相似的通道信息将有效提升网络模型局部特征感知能力。实验结果表明,与基线算法相比,本文方法在 CUB-200-2011、Stanford Cars 和 Stanford Dogs 三个数据集上的识别结果分别提升了 1.5%、2.4%、1.5%。结果表明,本文提出的方法通过提升细粒度图像细节特征的表达能力,从而有效提高了细粒度图像识别的有效性。

关键词:细粒度图像识别;Transformer;可变形卷积
中图分类号:TP391 **文献标识码**:A **国家标准学科分类代码**:520.6

Research on improved Transformer fine-grained image recognition algorithm

Li Bingfeng Liu Shuai Yang Yi
(School of Electrical Engineering and Automation, Henan Polytechnic University, Jiaozuo 454000, China)

Abstract: To address the issues of small inter-class differences and difficulty in distinguishing fine-grained images, this paper proposes a method that improves the network's ability to express image detail features, aiming to alleviate this problem. To achieve this, an improved Transformer-based algorithm for fine-grained recognition is designed in this study. Firstly, deformable convolutional token embedding adjusts the sampling points adaptively to modify the convolution operation range and the shape of its kernel, enhancing the network's perception of spatial information for more accurate spatial details. Secondly, an efficient correlation channel attention mechanism automatically selects channels to transform the computation from neighboring channels to semantically similar channels, capturing semantic-related channel information. The precise spatial information and semantically related channel information effectively enhance the network's perception of local features. Experimental results demonstrate that compared to the baseline algorithms, the proposed method improves recognition results by 1.5%, 2.4%, and 1.5% respectively on the CUB-200-2011, Stanford Cars, and Stanford Dogs datasets. These results indicate that the proposed approach effectively enhances the effectiveness of fine-grained image recognition by improving the expression capability of image detail features.

Keywords: fine-grained image recognition; Transformer; deformable convolution

0 引 言

细粒度图像识别是指对具有相似外观但属于不同子类的图像做出更为精细的分类[1]和识别。跟传统的图像识别方法不同,细粒度图像识别可以学习到图像中更为细致的差异,在日常生活中也具有更为广泛的应用,例如,识别不同种类的鸟、不同类型的汽车的识别等。因此,自出现伊始,细粒度图像识别便迅速成为了计算机视觉领域的热点研究问题。

在细粒度图像识别中,由于不同子类之间的图像具有

收稿日期:2023-09-23
*基金项目:河南省科技攻关项目(222102210230)、河南理工大学博士基金(B2018-33)项目资助

高相似的外观和特征,因此,图像的可判别信息往往只存在于图像中很小的区域,其提取和捕获难度极大,极具挑战性[2]。由于卷积神经网络(convolution neural network,CNN)[3]可以很好的提取图像的局部特征,因此其与细粒度图像识别的任务场景高度契合。最近几年,随着 CNN 的蓬勃发展,细粒度图像识别研究也获得了突破性发展[4-5]。目前,主流的细粒度图像识别的方法大体上可分为强监督的和弱监督两类。基于强监督的细粒度图像识别方法需要首先对图像中目标的可判别区域进行准确定位,并在此基础上,依靠可判别区域特征进行细粒度目标的识别,但该类方法除了图像的类别标签外,还需要标注出图像中的可判别区域,当处理大规模数据集或者复杂分类任务时,获取这类标注信息的代价通常较高;而基于弱监督的图像识别方法,仅依靠图像的类别标签便可完成模型训练。相对于强监督细粒度图像识别,弱监督的图像识别方法人工介入较少,目前已经成为细粒度图像识别的主要研究方向,如 Huang 等[6]提出了密集连接 CNN 模型,将每个层的输出都与当前层的输入连接在一起,该方法可以充分融合之前层学到的特征,从而增强网络的特征表达能力;Wang 等[7]提出了通道注意力多分支网络模型,通过在模型中增加多重注意力来提高模型的特征提取能力;Xiao 等[8]提出两级注意力模型,物体级注意力,用于识别图像的最显著区域;部位级注意力用于关注这些显著区域内的相关区域,从而更为精确地利用多层次信息。

但基于 CNN 的分类方法存在感受野有限的缘故,导致模型提取全局特征的能力有限。为此,Dosovitskiy 等[9]提出了视觉 Transformer(vision transformer,ViT)模型,首次将 Transformer 应用到计算机视觉领域中的图像识别任务。用全局自注意力模块(self-attention,SA)的方式将感受野扩大至整张图像,显著提升了分类准确度。然而,全局自注意力本质上是一种全连接操作,这会导致网络模型计算量较大,使得网络模型对硬件要求较高。2021 年,Wu 等[10]提出卷积视觉 Transformer(convolutional vision transformer,CVT)把深度可分离卷积运用到 Transformer 当中,不仅大幅度减少了网络计算量,其性能也有了进一步的提升。

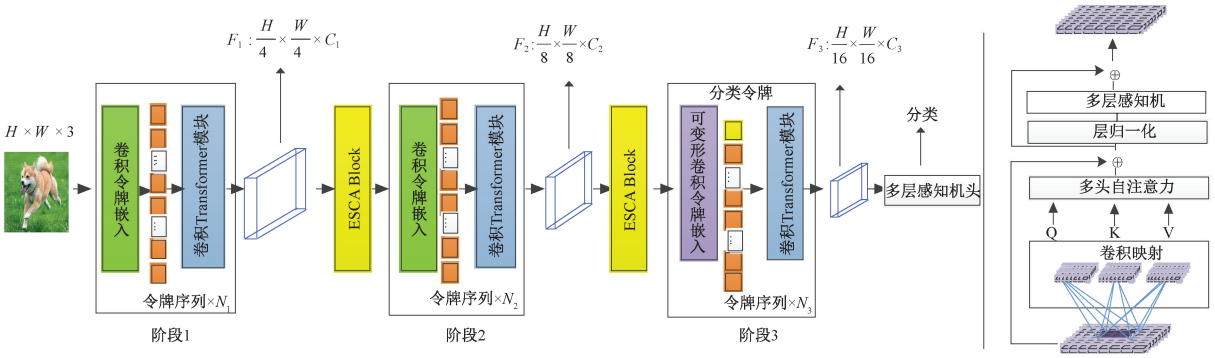
但 CVT 模型也存在一些缺陷,一方面该模型虽然减少了计算量,但其对复杂形变物体的空间感知能力不足;另一方面,Transformer 模块的全局自注意力机制只考虑了空间维度的信息,并未考虑其他维度信息,这将导致模型的特征提取能力有限。

为了进一步提高细粒度图像识别的性能,文章以 CVT 网络为基础,把可变形卷积的思想引入到了 CVT 网络当中,提出了一种基于改进的 Transformer 细粒度图像识别算法(improved transformer fine-grained image recognition algorithm,TransFR)。首先,本文用可变形卷积令牌嵌入(deformable token embedding,DTE)来代替 CVT 的卷积令牌嵌入模块,提高模型对形变较复杂物体的空间感知能力;其次,为了有效捕捉通道之间的局部关联性,以获取更为相似的通道信息,本文在通道注意力的基础上引入了一个高效可选择通道注意力模块,进一步提升模型的识别效果。通过在 3 个公共细粒度数据集上展开实验,实验结果表明本文提出的方法具有良好的分类结果和性能。

1 本文方法

1.1 TransFR 模型整体结构

TransFR 整体框架如图 1(a)所示,主干网络由三阶段金字塔结构的子网络构成,每个子网络由卷积令牌嵌入、令牌序列和卷积 Transformer 模块三部分构成。



(a) 整体框架 (b) 内部结构

图 1 网络整体架构图

首先,输入图像先经过卷积令牌嵌入层(convolutional token embedding,CTE),其不仅可以利用卷积来捕获图像中的局部特征,而且随着网络深度的增加还能逐步减少令牌的数量;随后将特征图展平为令牌序列,以便与 Transformer 编码器进行兼容,从而能够应用注意力机制。这里,只在 Stage3 将分类令牌 X_{class} 添加到令牌序列首部用于集成全局特征,用于分类,其他阶段不添加;接着将展平后的令牌序列重塑为二维令牌图后送到积 Transformer 模块对图像进行空间上的全局特征提取。卷积 Transformer 模块由卷积映影层、多头自注意力模块和多层感知机(multi-layer perceptron,MLP)组成,其内部结构如图 1(b)所示,卷积映射层用具有深度可分离卷积代替视

觉 Transformer 中使用的原始位置线性映射，来减少多头自注意力操作的计算复杂度；多头自注意力模块用于捕捉令牌间的长距离依赖关系；在最后阶段输出的分类标记上使用 MLP 来预测结果。

与 CVT 原始模型不同的是文中把最后一个阶段的 CTE 模块替换成了 DTE 模块，目的是为了提高网络对形变复杂物体的空间感知能力；其次为了使网络模型能够同时学习到空间和通道维度层面的特征，获得更全面、更准确的图像特征，在 Stage1 后添加一个高效可选择通道注意力网络(efficient and selectable channel attention network，ESCA-Net)进行通道维度层面的特征提取，以获取更为相似的通道信息。

1.2　可变形令牌嵌入模块

CVT 借鉴了 CNN 的金字塔层级[11]结构，在卷积令牌嵌入层通过改变卷积的步幅大小来调整令牌维度和每个阶段的令牌数量，从而保证每个阶段得到不同尺度的特征图。但由于 CNN 卷积核的几何结构是固定住的，在特征提取时是固定着采样点，不能自适应地对输入进行编码，对于图像中存在形变和尺度变化的目标，其空间信息会被严重破坏。所以本文将可变形思想引入到了细粒度图像识别问题，通过在卷积核中引入可学习的偏移量来修正卷积核的位置和形状，通过自适应的调整采样点位置，来改变采样范围以及卷积核的形状，提高对形变复杂物体的空间信息的感知能力，以获取更为精准的空间信息。

在标准卷积层中，对于输入特征图 $X \in R^{H\times W\times C_{in}}$，卷积权重张量 $W \in R^{K\times K\times C_{in}\times C_{out}}$，$K$ 为卷积核大小，C_{in} 和 C_{out} 代表输入和输出通道。对于 $K\times K$ 个采样位置的卷积核，输出特征图上的每个位置 P 可以表示为：

$$Conv(X)_{p,:}=\sum_{k\in[K\times K]}X_{p+g(k),:}W_{g(k),:,:,:} \tag{1}$$

式中：$g:[K\times K]\rightarrow\Delta K$ 是采样索引在预定偏移量 ΔK 上的双客观映射。比如，令 $\Delta K=\{(-1,-1),(-1,0),\cdots,(0,1)\}$ 则代表 3×3 卷积核且其扩张率为 1，则 $g(0)=(-1,-1)$ 为首个采样偏移量。当 $K=1$ 时，权重张量 W 则与矩阵等价，使得 $W\in R^{C_{in}\times C_{out}}$。

受可变形卷积的启发[12]，本文用一个可变形令牌嵌入模块来学习偏移网络，从而自适应地采集包含信息较多的令牌，在可变形卷积中引入偏移量 $\Delta g(k)$。可变形卷积可表示为：

$$DC(X)_{p,:}=\sum_{k\in[K\times K]}X_{p+g(k)+\Delta g(k),:}W_{g(k),:,:,:} \tag{2}$$

图 2 是可变形卷积的基本构架，具体流程如下；

1)对于输入特征图 $U\in R^{B\times C\times H\times W}$，偏移量 Δp_n 先由一个常规卷积求得，输出大小为 $V\in R^{B\times 2C\times H\times W}$，偏移量是 $2C$，表示每个通道 x，y 方向的偏移量，用公式化可以表述为

$$V:(B,2C,H,W)=Standardconv(U) \tag{3}$$

2)将 1 中的 V(相对偏移坐标)和 U(原坐标)相加，便可得到各个像素偏移后的坐标值 position。$position\in R^{B\times 2C\times H\times W}$，但其是一个 float 类型的坐标值，最终需要的是像素，所以这里需要用到双线性插值的方法。

3)求得 position 的所有像素后，便组成了一个新特征图，其进行下一层的前向传递。

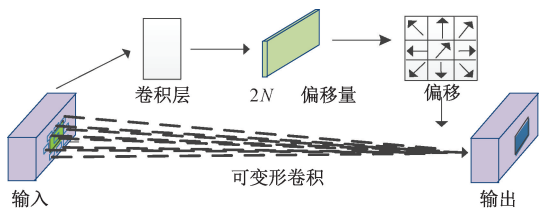


图 2　可变形卷积架构

与普通卷积操作不同，DC 在每个预先指定的偏移量 $g(k)$ 的基础上再学习一个偏移量 $\Delta g(k)$。$\Delta g(k)$ 由输入特征映射 X 经过一个单独的卷积层学习得到。可描述为：

$$DTE(X)=GELU(LN(Flatten(DC(X)))) \tag{4}$$

DTE 通过引入可学习的偏移量，能够根据输入特征图的局部结构动态调整卷积核的采样位置。此外，与基线中的常规网格采样相比，DTE 模块凭借其引入了极少的参数量的优势使其成为当前层级 ViT 的即插即用模块。

1.3　高效可选择通道注意力网络

在视觉任务中，注意力机制可以帮助聚焦模型对当前任务更为关键的信息，降低其他信息的关注度。如 Hu 等[13]提出一种采用挤压激活的通道注意力网络(squeeze and excitation Network，SE-Net)，捕获不同通道之间的关联性，但同时也带来较高的模型复杂度。因此，Wang 等[14]设计一种局部跨通道交互的高效通道注意力机制(efficient channel attention network，ECA-Net)，将跨通道交互范围由全局转化成了局部，显著降低了模型的复杂度。

但 ECA-Net 中的每个通道是由不同的卷积核经过卷积得到，不同的卷积核提取物体不同的特征信息，相邻通道并无明显的关联性。为此，李冰锋等[15]提出了 ESCA-Net，用一维可变形卷积有效地弥补传统卷积操作中忽视通道之间相关性的缺陷，将通道注意力的计算从通道相邻转换成语义相似，以便更好的捕捉通道之间的局部关联性。其网络的结构如图 3 所示。

对于输入特征图 $D\in R^{C\times H\times W}$，$H\times W$ 表示其尺寸，C 为通道数。先在空间上对特征图进行全局平均池化，接着执行降维和转置操作，得到 $R_s\in R^{1\times C\times C}$，如式(5)所示。

$$R_s=f_2(f_1(GAP(D))) \tag{5}$$

式中：$f_1()$代表降维操作，$f_2()$为转置操作，GAP()表示全局平均池化。

R_s 包含了输入特征图 D 中各个通道的全局空间信息，但每个通道是由不同的卷积核经过卷积得到，不同的卷积核提取物体不同的特征信息，所以通道相邻并不能说明其关联性很强，不同的通道之间包含的信息不同，盲目地在

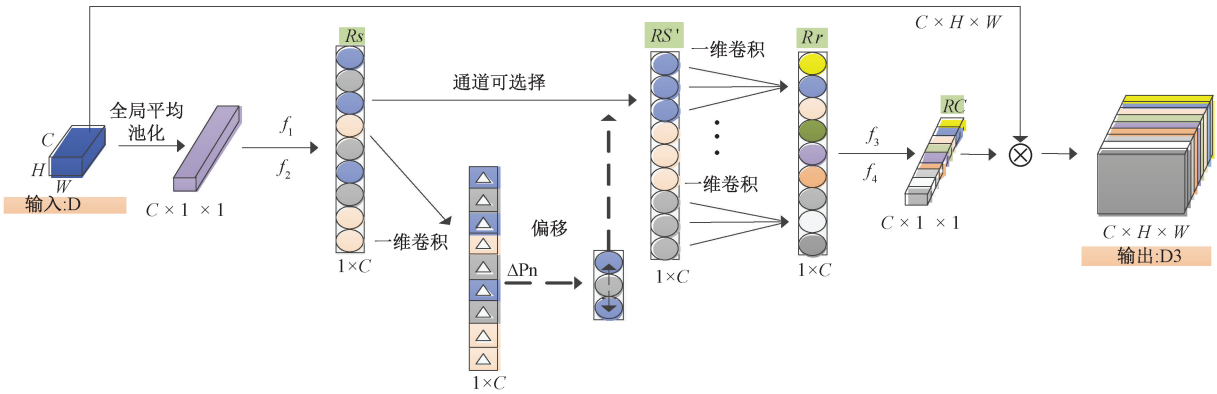


图 3 高效可选择通道注意力网络

相邻通道间执行一维卷积操作,可能会导致重要的信息丢失。因此,本文提出了一个高效可选择的通道选择网络,通过通道的自动选择,将通道注意力的计算从通道相邻转换成语义相似,在建立通道相关联的基础上,对数据进行一维卷积运算,其公式化描述如式(6)所示。

$$R_r(p_0)=\sum_{p_n\in K}W(p_n)R_s(p_0+p_n+\Delta p_n) \tag{6}$$

式中:$R_r(p_0)$ 为以 p_0 为通道中心的通道选择后一维卷积的输出,$W(p_n)$、$R_s(p_0+p_n+\Delta p_n)$ 分别为一维卷积及其输入向量,p_n 为卷积核在特征图上对应的位置索引,$K=[-i,i],i\in N^+$ 为卷积核索引组成的集合。(p_0+p_n) 为输入向量的原始位置索引,Δp_n 是 p_n 的偏移量,Δp_n 由一维卷积学习得到,其值的大小决定了通道选择特征偏离原始特征的程度。

通常,偏移量 Δp_n 并非整数,从而导致输出结果 $R_r(p_0)$ 也非整数,就会导致在 R_s 当中无法取到相匹配的值。为解决这个问题,本文对 Δp_n 进行上下取整,得到偏移量 $\Delta p_c=\lceil\Delta p_n\rceil$ 和 $\Delta p_f=\lfloor\Delta p_n\rfloor$。$p_c$ 与 p_f 为经过通道选择后的索引通道,如式(7)。

$$\begin{aligned}p_c&=p_0+p_n+\Delta p_c\\p_f&=p_0+p_n+\Delta p_f\end{aligned} \tag{7}$$

通过线性差值,可算出 $R_s(p_0+p_n+\Delta p_n)$ 的值,如式(8)所示。

$$\begin{aligned}R_e&=[R_s(p_c)-R_s(p_f)](\Delta p_n-\Delta p_f)\\R_s(p_0+p_n+\Delta p_n)&=R_s(p_f)+R_e\end{aligned} \tag{8}$$

通过 1×1 卷积,学习不同通道之间的重要性后再执行维度转换和升维操作,最后用激活函数获得新特征图的通道权重分布 $R_c\in R^{C\times1\times1}$,如式(9)所示。

$$R_c=sigmoid(f_4f_3(R_r))) \tag{9}$$

$f_3()$是维度转换操作,$f_4()$是升维操作。

最终得到 ESCA-Net 输出表达式,如式(10)所示

$$D_3=D\otimes R_c \tag{10}$$

其中,$D_3\in R^{C\times H\times W}$,$\otimes$表示加权乘积操作。

2 实验结果和分析

2.1 实验数据集及划分

为了公平、公正地验证本文方法的性能,文中所有实验都是在 CUB-200-2011[16]、Stanford-Dogs[17]、Stanford-Cars[18]3 个常见公用细粒度图像数据集上进行的,每个数据集的类别数及训练集、测试集的划分情况如表 1 所示。

表 1 细粒度图像数据集详细信息

数据集	类别	训练集	测试集
CUB-200-2011	200	5 994	5 794
Stanford Dogs	120	12 000	8 580
Stanford Cars	196	8 144	8 041

此外,为避免过拟合情况的出现,本文还采取水平翻转、垂直翻转和 AutoAugment[19]的措施来扩充数据样本,部分数据集展示如图 4 所示。

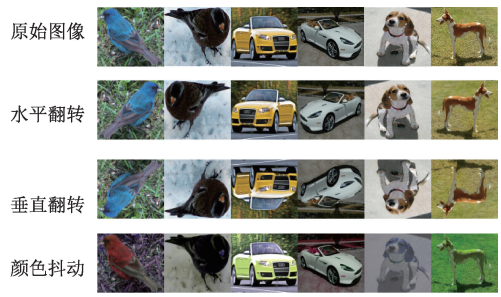


图 4 数据增强部分数据集效果图

2.2 实验环境及参数设置

本文所有实验均在 Ubuntu18.04 操作系统下进行,硬件平台为:CPU 为 Intel® Core™ i7-12700,GPU 为 NVIDIA GeForce RTX 3080Ti(共 1 块,显存为 16 GB)。软件配置为:Cuda11.3,Cudddn7.6.5,Python3.9.13。深度学习框架采用 PyTorch。

本文选用通用的准确率(Accuracy)作为细粒度图像分类评价指标,即:

$$Accuracy = \frac{I_{ac}}{I_{total}} \tag{11}$$

式中：I_{ac} 为测试样本中分类正确的图片样本数量，I_{total} 为测试集图片总样本。

此外，为防止 TransFR 模型因计算量过大导致模型在训练时收敛困难，统一把图像处理成 224×224 的尺寸；在训练阶段，还需加载 ImageNet 数据集在 CVT 模型上进行训练的预训练权重作为初始值，以便 TransFR 模型更快地完成训练。模型参数的优化方法使用具有解耦权重衰减的 Adam 方法（Adam with decoupled weight decay，AdamW），动量（momentum）数值统一为 0.9，由于 Stanford Dogs 数据集样本量跟其余两个数据集相比较大，所以在 Stanford Dogs 上训练时 batch size 设定为 16，其余数据集上时 batch size 设置为 8，迭代次数（epoch）为 100，训练阶段学习率初始值设定为 0.001，选择余弦退火（cosine annealing）的方式来掌握学习率的降落幅度。

2.3　消融实验

为了验证本文的 DTE 模块和 ESCA-Net 的有效性，本文在 CUB-200-2011、Stanford-Dogs、Stanford-Cars 数据集上进行了消融实验。实验结果如表 2 所示。

表 2　可变形令牌嵌入和高效可选择通道注意力网络的消融实验分析

模型	参数量/M	准确率/%		
		CUB	Dogs	Cars
CVT（baseline）	20.0	74.0	74.6	88.6
CVT＋DTE	20.8	74.7	75.4	89.7
CVT＋ESCA-Net	20.0	75.0	76.6	89.6
TransFR	20.8	75.5	77.0	90.1

从表 2 实验结果可以看出，在未引入任何模块的情况，基准骨干网络在 CUB-200-2011、Stanford Dogs 和 Stanford Cars 上分别实现 74.0%、74.6%和 88.6%的分类准确率，在此基础上，在既替代卷积令牌嵌入模块的同时又增加 ESCA-Net 注意力模块的情况下，模型的分类精度可以达到 75.5%、77.0%和 90.1%，在 3 个数据集上分别实现了 1.5%、2.4%和 1.5%的性能提升。特别地，当仅加入 DTE 模块后，增强了模型对复杂形变物体的空间感知能力，分类准确度相比于基准模型在 CUB-200-2011、Stanford Dogs 和 Stanford Cars 三个数据集上分别提升了 0.7%、0.7%和 0.9%；在引入 ESCA-Net 注意力模块后，增强了模型捕捉通道维度的信息特征能力，分类准确度在 CUB-200-2011、Stanford Dogs 和 Stanford Cars 三个数据集上分别提升了 1%、2%和 1%；实验结果表明，将 DTE 模块与 ESCA-Net 注意力模块并行组合可以进一步带来性能上的收益，其能捕获图像中更具判别性是区域，进而提高模型的分类性能。

2.4　对比试验

为进一步验证本文方法的优越性，将其与 ECA-Net、SE-Net、CA-Net[20]和 CBAM[21]的等算法在 CUB-200-2011 公用数据集上展开实验，并与本文算法进行了对比。实验结果如表 3 所示。

表 3　不同注意力模型实验结果对比

模型	参数量/M	计算量/M	准确率/%
			CUB
CVT	19.92	4 055.40	74.0
CVT＋DTE＋SE-Net	19.97	4 061.87	74.9
CVT＋DTE＋ECA-Net	19.95	4 061.85	75.0
CVT＋DTE＋CBAM	19.96	4 062.26	75.2
CVT＋DTE＋CA-Net	19.96	4 062.52	74.7
TransFR(ours)	19.95	4 061.85	**75.5**

表 3 展示了本文模型 TransFR 在 CUB-200-2011 数据集上与其他算法的实验对比结果。从测试结果可看出，本文模型在分类准确度明显高于现存的大多数主流注意力机制，展示了最先进的性能。如：本文引入的 ESCA-Net 注意力机制与 SE-Net、CA-Net、CBAM 相比不仅参数量有所减少，分类精度也明显提升。相比于 ECA-Net，用不同数量的卷积核学习不同通道之间的重要性，无法对具有关联性的通道进行组合排序，本文注意力机制可以有效捕捉通道之间的局部关联性，以获取更为相似的通道信息。

与基准骨干网络 CVT 相比，提升了 1.5%的分类性能，且在参数量增加忽略不计的条件下，在 CUB-200-2011 细粒度图像数据集上获得了优异的分类结果，分类准确度超过了 75.5%。实验结果表明，本文在引入了注意力机制和替换特征提取模块后的 TransFR 模型能有效学习到细粒度图像分类的关键特征，从而提高模型的分类性能和泛化能力。

2.5　性能分析

本文模型与原 CVT 模型训练时的损失曲线进行对比，由图 5 可以发现，本文方法损失下降速度更快，收敛更快，也从侧面证明了本文模型的优越性。

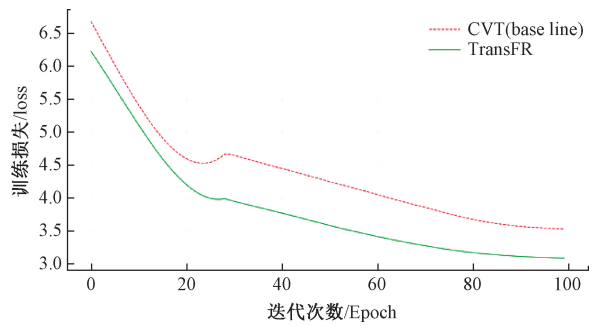


图 5　损失函数曲线

2.6 TransFR 可视化

为了分析本文模型对网络提取特征能力的影响,本文对特征提取网络输出层特征图进行可视化输出。从每个数据集中分别选取 3 个样本图像进行可视化,如图 6 所示。

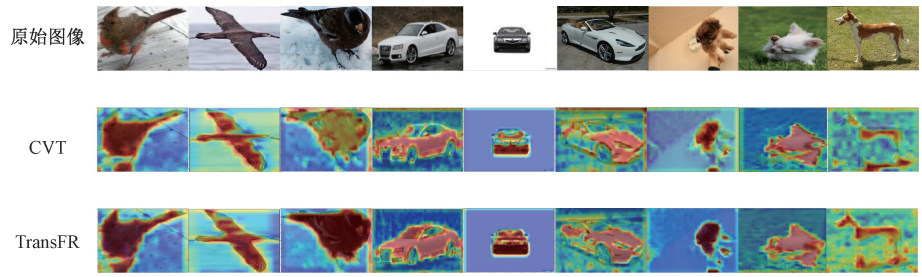


图 6 特征图可视化结果

图中深红色部位为目标检测相关区域。图 6 表明,本文所提算法热力图中的显著性区域更为全面,主要集中在局部更具有辨别性的位置,例如鸟类的翅膀、头部包含较多的权重;狗类的关注点大多聚集在耳部和腿部等地方,而车辆则大多聚集在车框、车灯等区域。这说明本文算法能更好地学习细粒度图像特征,捕捉细节特征的能力更强。

3 结 论

本文以为 Transformer 基本框架,提出了一种基于改进的 Transformer 细粒度图像识别算法。可变形卷积令牌嵌入通过自适应调整采样点的位置,来改变卷积操作范围和及其卷积核的形状,增强对空间信息的感知能力,以获取更为精准的空间信息;高效相关通道注意力,通过通道的自动选择,将通道注意力的计算从通道相邻转换成语义相似,以获取更为相似的通道信息。而精准的空间信息和语义相似的通道信息将有效提升细粒度图像识别问题的特征感知能力,来提升其识别精度。实验显示,提出的模型在多个细粒度图像分类上都有优异的分类准确度。

参考文献

[1] 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.

[2] 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.

[3] KRIZHEVSKY A,SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [J]. Advance in Neural Information Processing Systems, 2012, 25 (2), DOI: 10.1145/3065386.

[4] 朱阳光,刘瑞敏,黄琼桃. 基于深度神经网络的弱监督信息细粒度图像识别[J]. 电子测量与仪器学报,2020,34(2):115-122.

[5] 齐爱玲,王宣淋. 融合通道与位置信息的 ResNet 细粒度图像识别[J]. 国外电子测量技术,2022,41(12):103-111.

[6] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[J]. IEEE Computer Society, 2016, DOI: 10.1109/CVPR.2017. 243.

[7] 王彬州,肖志勇. 面向细粒度图像识别的通道注意力多分支网络[J]. 激光与光电子学进展,2021, 58(22):164-172.

[8] XIAO T, XU Y, YANG K, et al. The application of two-level attention models in deep convolution-al neural network for fine-grained image classification[J]. IEEE, 2014, DOI:10.1109/CVPR.2015.7298685.

[9] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale[J]. ArXiv Preprint, 2020, ArXiv:2010.11929.

[10] WU H P, XIAO B, CODELLA N, et al. CVT: Introducing convolutions to vision transformers[C]. 2021 IEEE/CVF International Conference on Computer Vision, New York: IEEE Press, 2021: 22-31.

[11] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large scale image recognition [J]. Computer Science, 2014, DOI: 10.48550/arXiv.1409. 1556.

[12] DAI J, QI H, XIONG Y, et al, Deformable networks[C]. 2017 IEEE International Conference on Computer Vision, New York: IEEE Press, 2017:764-773.

[13] HU J, SHEN L, SUN G, et al. Squeeze-and-excitation networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018:7132-7141.

[14] WANG Q, WU B, ZHU P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 11534-11542.

[15] 李冰锋,段鑫鑫,杨艺,等.基于特征增强和损失优化的弱监督目标检测算法[J].兵器装备工程学报,2023,44(6):196-203.

[16] WAH C, BRANSON S, WELINDER P, et al. The caltech-UCSD birds-200-2011 dataset[J]. California Institute of Technology,2011.

[17] ADITYA K,NITYANANDA J,BANGPENG Y,et al. L:Novel dataset for fine-grainedimage categorization[J]. 2013.

[18] KRAUSE J, STARK M, DENG J,et al. 3D object representations for fine-grained categorization[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, 113-123.

[19] CUBUK E D, ZOPH B, MANE D, et al. Autoaugment: Learning augmentation policies from data[J]. ArXiv Preprint,2018, ArXiv:1805.09501.

[20] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[C]. 2021 IEEE Conference on Computer Vision and Pattern Recognition, 2021.

[21] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]. Europe Conference on Computer Vision,2018:3-19.

作者简介

李冰锋,讲师,博士,主要研究方向为迁移学习、计算机觉、目标检测。

E-mail:libingfeng@hpu.edu.cn

刘帅(通信作者),硕士,主要研究方向为计算机视觉、目标检测。

E-mail:shuailiu0317@163.com

杨艺,副教授,博士,主要研究方向为深度学习、强化学习与智能控制。

E-mail:1286535923@qq.com