

基于加强图像块相关性的细粒度图像分类方法

王 坤, 朱子奇

(武汉科技大学 计算机科学与技术学院, 湖北 武汉 430065)

摘 要:在细粒度图像分类任务中,提取出具有区分性的局部特征对识别图像之间的微小差异非常重要。基于 ViT(vision transformer)框架的算法模型在计算机视觉各个研究领域取得了优异的表现。针对基于 ViT 框架的细粒度图像分类模型对图片局部区域关注度低的问题且为进一步加强图像块特征的上下文联系,提出了一种基于加强图像块相关性的细粒度图像分类方法。首先,提出了赋予图像块相关性权重的方法,并嵌套应用于不同层编码器中丰富不同层次特征信息,解决了 ViT 对图像局部特征关注不够的问题;其次,结合图像块的位置信息加强了局部特征上下文的联系,同时减少了噪声信息带来的干扰;最后,提出相似损失函数来学习细粒度图像中微小特征的差异性,优化模型的分类效果。在两个公开数据集 CUB-200-2011 和 Stanford Dogs 上进行实验分别取得了 91.33%、92.15% 的准确率,提出的方法分别比基准模型 ViT 网络提升了 0.63、0.45 百分点,有效提升了细粒度图像分类效果,验证了方法的有效性。

关键词:ViT;细粒度图像分类;局部特征;相关性;图像块特征;编码器

中图分类号:TP391.4

文献标识码:A

文章编号:1673-629X(2023)05-0056-06

doi:10.3969/j.issn.1673-629X.2023.05.009

Fine Grained Image Classification Method Based on Enhanced Patch Correlation

WANG Kun, ZHU Zi-qi

(School of Computer Science & Technology, Wuhan University of Science and Technology, Wuhan 430065, China)

Abstract:In the fine-grained image classification task, it is crucial to extract distinctive local features to identify small differences between images. The algorithm model based on ViT (vision transformer) framework has achieved excellent performance in various research fields of computer vision. Aiming at the problem that the fine-grained image classification model based on ViT framework pays little attention to the local area of the picture and to further strengthen the context connection of patch features, a fine-grained image classification method based on enhancing the correlation of patch is proposed. Firstly, a method of assigning correlation weights to patches is proposed, and nested application is used in different layer encoders to enrich different layer feature information, which solves the problem that ViT does not pay enough attention to local features of images. Secondly, combining the position information of the patch, the local feature context is strengthened, and the interference caused by the noise information is reduced. Finally, the similarity loss function is proposed to learn the difference of minute features in fine-grained images and optimize the classification effect of the model. Experiments on two public data sets, CUB-200-2011 and Stanford Dogs, have achieved an accuracy of 91.33% and 92.15%, respectively. The proposed method improves the benchmark model ViT network by 0.63 and 0.45 percentage points respectively, effectively improving the fine-grained image classification effect, and verifying the effectiveness of the method.

Key words:vision transformer; fine grained image classification; local features; correlation; patch features; encoder

0 引 言

细粒度图像分类又称之为具体子类别图像分类^[1],目的是对粗粒度大类别进行更为精细的子类区分,比如识别不同的飞机、鸟、车、狗等^[2-4]。不同于人脸识别等传统对象级分类任务,细粒度图像任务难点

在于类间差异小类内差异大^[5],比如两种外形十分相似的狗属于完全不同的品种,由于存在光照、遮挡、背景干扰等诸多不确定性元素,借助肉眼很难分辨。因此,细粒度图像分类任务相比于传统图像分类任务难度更大。

收稿日期:2022-07-12

修回日期:2022-11-16

基金项目:国家自然科学基金资助项目(61702382)

作者简介:王 坤(1999-),男,硕士研究生,研究方向为计算机视觉;通信作者:朱子奇(1983-),男,博士,副教授,CCF 会员(55349M),研究方向为计算机视觉、模式识别。

解决细粒度图像分类问题的关键是对目标对象进行有效检测,并从中提取出具有区分性的局部特征。最近,随着 Transformer 架构在 NLP 的机器翻译^[6]等相关研究领域取得显著成果,许多研究者将 Transformer 架构逐渐迁移到计算机视觉任务上,比如图像分类^[7]、目标分割^[8]等。Alexey 等人^[7]将 Transformer 架构直接应用到图像分类任务上提出 ViT 模型,显著提高了传统图像分类任务的性能。作者思路是直接吧图像分割成固定大小的图像块序列,然后通过线性变换得到图像块嵌入向量,这也就类比于 NLP 中的词序列和词向量,然后将图像块嵌入向量和一个空白的分类标记向量^[7] (class token) 直接送入多层编码器进行特征提取分类。

由于 Transformer 的自注意力机制在整合全局信息方面比 CNN 更有优势,能够获得特征长距离依赖,因此基于 ViT 框架的一系列方法在计算机视觉领域的一些任务上表现出色。但 ViT 基于其自注意机制对图像关注区域并非总是有效,结合注意力层捕捉到的注意力图的可视化效果,发现还是会存在捕捉区域与目标区域重叠度低的现象^[9]。

同时 ViT 也有其固有缺陷:(a)对于图片局部区域特征的关注度不够且容易受光照等不确定因素干扰^[9];(b)切割图片带来的关键特征不完整表达^[10];(c)其注意力计算方式需要综合全局信息导致计算量非常大^[10]。考虑到细粒度图像分类任务中图片局部区分性特征的表征能力对分类效果十分关键,许多研究工作都围绕着如何让 ViT 模型提取出更具区分性的特征而展开,但是对图像块输入特征做处理的工作非常少。而且编码其中的多头注意力机制主要通过建立所有图像块向量联系来发挥作用,但所有的图像块起到的作用并非相等。

针对以上问题且结合提取图像区分性局部特征对细粒度图像分类任务至关重要的实际情况,在现有 ViT 工作的基础上,该文提出一种基于加强图像块相关性的细粒度图像分类方法。首先,通过赋予图像块相关性权重系数并对图像块相关度进行评价差异化,加强对局部区分性特征的关注。其次,为了降低分割图像对某些区分性图像块造成的特征不完整表达,引入图像块位置信息加强图像块特征上下文信息的联系。最后,在交叉熵损失函数基础上增加相似损失函数,更有利于细粒度图像分类任务中降低相同子类别差异性。整个方法模块以嵌套方式应用于不同层次的编码器中,融合了不同层次特征信息。结合实验结果证明了该方法思路可行,进一步提升了 ViT 框架在细粒度图像分类任务上的表现效果。

1 相关工作

对于现阶段的细粒度图像分类模型可以按照模型使用了多少辅助信息分为强监督分类模型和弱监督分类模型^[11]。同时,最近基于 ViT 框架的分类模型在许多视觉任务上取得了良好效果,这类模型具有弱监督的分类思想,许多研究工作基于此展开。

1.1 强监督分类模型

在细粒度分类任务上采用强监督分类模型主要是通过数据集提供的额外标注信息训练出一个网络模块,这个网络可以检测出目标物体边框以及部分物体部位的标注框,然后将这个网络获取到的特征信息在主干网络上进行特征调整、融合等操作,最后通过训练分类在细粒度图像分类任务上取得了较好的效果。比较有代表性的工作有 Part-based R-CNN^[12]、Pose-normalized CNN^[13]等。

1.2 弱监督分类模型

基于弱监督学习的细粒度分类模型在不借助标注信息的情况下,也实现了对全局特征和局部特征的较好捕捉。Lin 等人^[8]提出的双线性卷积神经网络模型(B-CNN),通过两个 VGG-Net^[6]网络提取特征,并在特征送进全连接层前对两个分支提取出的特征进行双线性融合操作,有效提升了特征表征能力。但网络模型参数量太大导致训练效率非常低,且 VGG-Net 特征提取网络对目标物体区分性局部部位的关注度不够等问题对最后分类效果产生了一定影响。Xiao 等人^[13]提出的两级注意力模型,首先在图像上生成大量候选区域并过滤保留包含前景物体的候选区域,然后利用网络的两个特征分支分别对物体级特征和部位级特征进行提取并融合,最后进行 SVM^[10,14]分类。Fu 等人^[15]提出的循环注意卷积神经网络(RA-CNN)提出使用注意力区域网络(APN)定位出图像中的目标物体区域,并进行裁剪放大操作,然后将其送进下一层网络获取物体部位级别的图像,在最后一层网络融合物体级和部位级特征进行预测分类。

1.3 基于 ViT 架构的分类模型

ViT 框架在传统图像分类任务上表现很出色,主要因为其工作机制就是利用自注意力机制捕获全局上下文的特征信息。许多工作基于 ViT 展开,如 DeiT^[16-17]进一步探索了 ViT 架构如何保证数据高效训练和特征提取。CrossViT^[18-19]探索了 ViT 架构在多尺度情况下的表现,将不同尺度的图像块的特征提取并进行有效融合,取得了非常理想的分类效果。TransFG^[20]是第一个将 ViT 架构应用在细粒度图像分类任务上并取得优异表现的网络。主要工作是在最后一层编码器之前,提出部分选择模块 PSM 来选择与物体目标区域相关度大的图像块向量,然后和分类标记

向量一起作为最后一层编码器的输入进行后续预测分类。

许多研究工作都围绕着如何让 ViT 模型提取出更具区分性的特征而展开,但是对图像块输入特征做处理的工作非常少。并且编码器中的多头注意力机制主要通过建立所有图像块向量联系来发挥作用,但所有的图像块起到的作用并非相等^[20]这一特性在之前的工作中研究并不多。因此,该文是基于 ViT 架构如何学习到图像块作用的量化值,使后续阶段的多头注意力机制学习到比较重要的图像块特征。同时为了学习到相同类别之间的差异性,提出相似损失函数来优化模型,提升任务效果。

2 文中方法

2.1 赋予图像块相关性权重

自注意力机制捕捉全局上下文信息在视觉信息中就是计算图像块向量的注意力;由于这个过程容易忽略图像局部区域信息,而对于局部上下文信息,即图像块之间的联系,会想到利用卷积操作的特性来使模型可以学习到不同图像块特征的重要程度,这样就从全局和局部两种互补角度使模型更适合细粒度图像分类任务。对于每一组输入向量 $A \in R^{(N+1) \times D}$, N 代表图像块向量个数, D 代表向量维度,通过池化操作,学习到代表每个向量的感受野值:

$$w_i = f_{\text{pooling}}(A_i), i \in (0, 1, \dots, N) \quad (1)$$

其中, f_{pooling} 为平均池化函数。

然后,经过全连接层和激活函数层来捕获图像块向量的相关性,得到和图像块特征数量相等的一组特征权值。

$$\partial_{\text{patch}} = \sigma(W_{\text{patch}}^p \otimes w_i) = \begin{cases} 0, i = 0 \text{ and } p = 1 \\ 1, i = 0 \text{ and } p = 2, 3 \\ \sigma(W_{\text{patch}}^p \otimes w_i), i > 0 \end{cases} \quad (2)$$

其中, σ 代表 Sigmoid 激活函数, \otimes 代表卷积操作, W_{patch}^p 代表第几个模块的权值矩阵, $p = (1, 2, 3)$ 。 w_i 中的 i 代表是否为分类标记向量,如果 $i = 0$ 代表是分类标记向量,否则为图像块向量;由于最后取分类标记向量作为分类评价标准,在初始化分类标记向量时也就是嵌入的第一个加强图像块相关性模块中分类标记向量权值设为 0,嵌入的其他两个模块中的分类标记向量权值设为 1,这样的做法对分类标记向量学习到全局特征更有利。

为了更好地提取到具有区分性的局部特征,引入相关度评价函数 f_{α} 对权值进行过滤,得到一些对细粒度分类任务比较重要的图像块特征。具体做法是首先将上一步得到的图像块特征权值利用排序函数 f_{sort} 进行降序排序返回代表各权值大小序号的向量组,所以图像块特征并没有被打乱顺序,在相关度评价函数 f_{α} 中引入图像相关度因子 α 代表保留的图像块特征数量,经过大量实验证明保留前 70% 的图像块区域向量效果最佳即 $\alpha = 0.7$,剩下的图像块权重值设置为零,这种处理方式抑制了许多的不相关信息表达,同时在一定程度上降低了注意力计算量。最后将所得权值与对应输入向量加权相乘得到加权特征:

$$A_{\text{patch}} = A \odot f_{\alpha}(f_{\text{sort}}(\partial_{\text{patch}}), \partial_{\text{patch}}) \quad (3)$$

$$f_{\alpha}(W_{\text{rank}}, \partial) = \begin{cases} 0, 0 < \partial < f_{ts}(W_{\text{rank}}, \alpha) \\ \partial, \partial \geq f_{ts}(W_{\text{rank}}, \alpha) \end{cases} \quad (4)$$

相关度评价函数 f_{α} 里 W_{rank} 代表权值向量组里面的排序情况, ∂ 代表具体权值, f_{ts} 函数确定保留图像块区域的权值阈值。模型框架如图 1 所示。

该文选择的基准模型是 ViT/Base 模型,编码器数量为 12,多头注意力机制中的自注意力层数为 12。其中 MLP Head 层同 ViT 模型中一致为一个全连接分类层。由于注意力在不同层具有相似性^[21]也被证明是可能的,尤其在相邻层;同时 ViT 架构中浅层和深层编码器学习到的局部特征也是不同的^[22]。因此,为了减

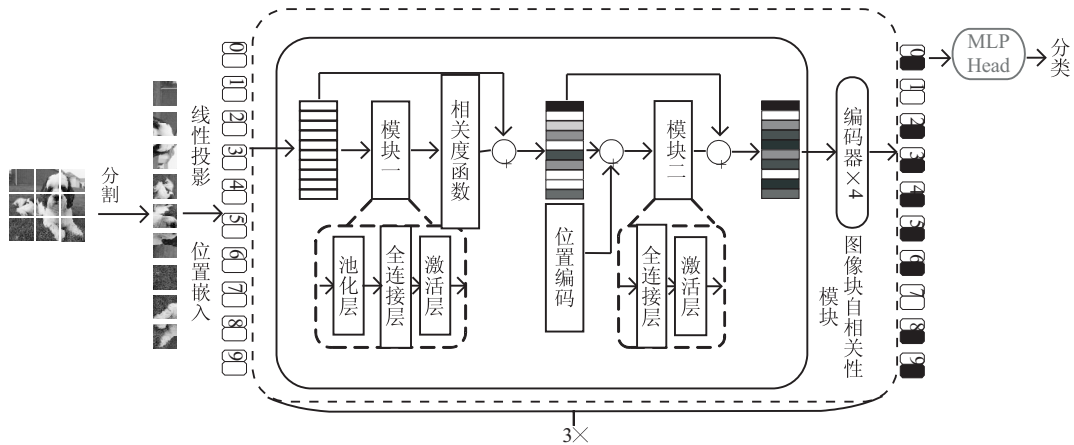


图 1 模型框架

少注意力相似性带来的不必要的注意力计算量,同时将提出的加强图像块相关性模块尽可能应用到不同层次的编码器中来丰富特征表达,选择将提出的方法模块与编码器以 $1 : \frac{12}{m}$ 的方式嵌套叠加,经过大量实验证明, m 值为 3,即以 $1 : 4$ 的方式嵌套组合效果最佳。

2.2 嵌入位置信息加强局部区域特征表达

由于图片分割成一组图像块的操作会使局部区域的特征表达不完整^[20],特别是当细粒度图像分类任务所需的类间区分性特征因分割操作受到影响时,就会降低分类精度。为了降低这种影响,可以选择把图片分割成互相有重叠区域的图像块^[20],即:

$$N = \left\lceil \frac{H - P + S}{S} \right\rceil * \left\lceil \frac{W - P + S}{S} \right\rceil \quad (5)$$

其中, H 、 W 代表原始图像的分辨率, P 代表分割后图像块的分辨率,因为这里分割成的图像块为正方形,所以 P 也就代表分割后的图像块的长度和宽度, S 代表卷积核的滑动窗口大小,也就是实现了重叠的图像块划分,得到的 N 就是一张图片被分割成多少个图像块。但这也会造成计算量的增加,特别是在分辨率较高的数据集上,会对实验环境产生很高的要求。

而图像块区域特征表达相似性与位置信息紧密联系也通过可视化效果可知^[7],受启发于此,为减小分割带来的特征不完整表达影响,加强行列位置信息相同的图像块特征联系,利用图片中每个图像块区域位置信息即图 2 中位置编码信息,记为 $W_{\text{pos}}^i(x_{\text{top}}, y_{\text{left}}, x_{\text{bottom}}, y_{\text{right}})$, $i \in (1, 2, \dots, N)$,与位置嵌入向量矩阵 A_{pos} 相乘并投影为标量来得到一组权重标识不同区域图像块的位置关系的权重:

$$\partial_{\text{pos}} = \sigma(W_1 \otimes (W_{\text{pos}} \otimes A_{\text{pos}})) \quad (6)$$

$$A_{\text{pos}} = W_2 \otimes A_{\text{patch}} \quad (7)$$

其中, N 标识图像块数量; A_{pos} 的维度为 $4 * d$, d 是图像块特征向量维度; $(x_{\text{top}}, x_{\text{bottom}})$ 和 $(y_{\text{left}}, y_{\text{right}})$ 分别代表图像块区域的左上角和右下角坐标。最后学习到的权重信息 ∂_{pos} 与上一阶段的图像块向量相乘作为编码器的输入向量:

$$A_{\text{ecd}} = A_{\text{patch}} \odot \partial_{\text{pos}} \quad (8)$$

2.3 损失函数

为了模型更好地关注到细粒度图像特征,文中方法选择把输出向量中的分类标记向量 z 即第一个向量作为评价标准,考虑到细粒度图像分类任务子类别之间的差异非常小,为了尽可能学习到这个微小特征的差异性,在交叉熵损失函数基础上增加了相似损失函数 L_{sim} ,使相同标签对应的分类标记向量差异性最小化,提升任务效果。

$$L_{\text{sim}} = \frac{1}{N^2} \sum_i \sum_j y_i = y_j (1 - f_{\cos}(z_i, z_j)) \quad (9)$$

其中, f_{\cos} 是计算余弦相似度函数。所以整体模型的损失为:

$$L_{\text{total}} = L_{\text{cross}}(y_i, y_i') + L_{\text{sim}}(z) \quad (10)$$

其中, L_{cross} 为交叉熵损失函数。在 L_{total} 的基础上,模型不断训练优化,最后使得整个网络拟合,模型提取出具有区分性的局部特征的能力显著提升。

3 实验与分析

3.1 数据集介绍

CUB-200-2011 鸟类数据集包括 200 种鸟类,共 11 788 张图像,其中有 5 994 张训练图像和 5 794 张测试图像,每张图像均有图像类标记信息,包括鸟的标记框、鸟的关键部位信息,以及鸟类的属性信息。Stanford Dogs 犬类数据集包括 120 种犬类,共 20 580 张图像,其中 12 000 训练图像和 8 580 张测试图像,每张图像有类标记信息和的标记框,关键特征包括毛发颜色、鼻子。

3.2 实验参数

采用基于 ViT-Base/16 的 ImageNet21k 数据集预训练模型^[7]。首先,将输入的原始图像进行预处理。预处理包括将图像像素大小随机缩放,然后裁剪到 $448 * 448$ 的像素级别,并对裁剪后的图像进行随机旋转,其中对用于训练的数据集进行随机裁剪,对用于测试的数据集进行中心裁剪。实验中统一把图片分割为 $16 * 16$ 大小的图像块,滑动窗口步长大小也为 16, Batchsize 大小设置为 16, Epoch 大小为 100。网络权值更新使用 SGD 优化器,SGD 优化器的动量设置为 0.9。CUB-200-2011 数据集的学习率设置为 0.03, Stanford Dogs 数据集的学习率设置为 0.003。所有实验均使用了三张 NVIDIA GeForce RTX3090 GPU,在 Linux 系统上运行并基于 Pytorch 框架,借助了 Apex 工具。

3.3 消融实验

为证明加入赋予权值模块和嵌入位置编码信息对分类效果的影响,采用 ViT 作为基准模型,分别对不加入位置编码信息和加入位置编码信息的加强图像块相关性两种模块进行消融实验。采用交叉熵损失函数和相似损失函数,对三种模型分别训练 40 次,实验结果如表 1 所示。

表 1 消融实验结果

方法	CUB (准确率/%)	Dogs (准确率/%)
ViT ^[8]	90.7	91.7
文中方法(不嵌入位置编码)	91.15	92.0
文中方法	91.33	92.15

可以看出,与基准相比,两种模型的准确率均有提升,且因为加入位置编码信息的模型提升了局部区域上下文信息的利用率,较之没嵌入位置编码信息的模型也有提升效果。说明提出的方法能使模型学习到更多局部区分性特征,并降低了分割图片带来的特征不完整表达影响。

3.4 嵌套比和图像块相关度因子选取实验

文中方法中需要调整的参数有加强图像块相关性模块与编码器嵌套比 $1: \frac{12}{m}$ 中的 m 值和图像块相关度因子 α 。为了保证单一变量原则,首先把加强图像块相关性模块只嵌入到第一层编码器前,然后对 α 进行调整,如表 2 所示。当 α 为 0.7 时,即保留前 70% 的图像块区域参与注意力计算效果最好,当 α 太小,缺少一些图像块区域参与计算对精度有影响,当 α 太大,会存在不相关信息的干扰。

表 2 不同 α 值实验结果

α 值	CUB (准确率/%)	Dogs (准确率/%)
0.5	91.18	91.14
0.6	91.21	91.18
0.7	91.28	92.07
0.8	91.21	92.01

然后调整嵌套比的 m 值,如表 3 所示,当把模块叠加到编码器中时,效果都会提升,当 m 大于等于 3 时效果相当,但为了减少计算量,应尽可能把模块嵌入数量降低,所以 m 取值 3 比较好,也就是嵌入三个模块。

表 3 不同 m 值实验结果

m 值	CUB (准确率/%)	Dogs (准确率/%)
2	91.25	92.10
3	91.33	92.15
4	91.33	92.16
5	91.37	92.17
6	91.38	92.17

3.5 对比实验

为了证明提出方法的有效性,将文中方法与先进的方法进行对比,文中方法的基准网络为 ViT,加入加强图像块相关性模块后,在数据集 CUB-200-2011 和 Stanford Dogs 分别提升了 0.63、0.45 百分点,也证明了文中方法能够有效加强局部信息的表征能力。如表 4 所示,相比于一些基于 CNN 方法的模型,文中方法效果提升明显。Cross-X 网络利用不同图像之间和不同网络层之间的关系对细粒度图像数据集进行多尺度特征学习,也达到了不错的效果。API-Net 通过构建

注意力成对交互网络进行互向量学习区分微小差异,达到了非常好的效果,但对微小差异捕捉能力更强,性能更好。对比 FDL 网络方法,文中方法也表现的非常好。相比于基准网络 ViT 模型,效果也有提升。

表 4 对比实验结果

不同模型	CUB (准确率/%)	Dogs (准确率/%)
Cross-X ^[23]	87.7	88.9
FDL ^[24]	89.1	84.9
API-Net ^[25]	90.0	90.3
ViT ^[8]	90.7	91.7
文中方法	91.33	92.15

4 结束语

在细粒度图像分类任务中,针对 ViT 框架对图像局部区域关注不够的问题且为进一步加强图像块特征的上下文联系,提出一种基于加强图像块相关性的细粒度图像分类方法。赋予图像块相关性权重并对其评价差异化,加强网络对局部区域的关注,引入图像块位置信息加强图像块上下文信息的联系,有效降低了分割图片对图像块造成的特征不完整,整个模块与编码器以嵌套方式组合丰富了不同层次的特征表达,并引入相似损失函数提升任务表现。实验表明,该方法有效提升了细粒度图像分类效果。下一步的研究可以考虑如何充分利用当前图像块与局部相邻图像块区域的联系,进一步加强图像块特征的表征能力,提升分类效果。

参考文献:

- [1] ZHAO Bao, FENG Jiashi, WU Xiao, et al. A survey on deep learning bas-ed fine-grained object classification and semantic segmentation[J]. International Journal of Automation and Computing, 2017, 14(2): 119-135.
- [2] WAH C, BRANSON S, WELINDER P, et al. The Caltech-UCSD birds-200-2011 dataset, CNS-TR-2011-001[R]. Pasadena: California Institute of Technology, 2011.
- [3] MAJI S, RAHTU E, KANNALA J, et al. A fine-grained visual classification of aircraft[J]. Machine Learning, 2013, 15: 23-27.
- [4] KRAUSE J, STARK M, DENG Jia, et al. 3D object representations for fine-grained categorization[C]//Proc of IEEE international conference on computer vision. Washington DC: IEEE, 2013: 554-561.
- [5] 王 阳, 刘立波. 面向细粒度图像分类的双线性残差注意力网络[J]. 激光与光电子学进展, 2020, 57(12): 163-172.
- [6] ZHANG Ning, DONAHUE J, GIRSHICK R, et al. Part-

- based R-CNNs for fine-grained category detection [C]//Proc of the 13th European conference on computer vision. Zurich: Springer, 2014: 834–849.
- [7] QIAN Xuelin, FU Yanwei, XIANG Tao, et al. Pose-normalized image generation for person re-identification [J]. arXiv:1712.02225v4, 2018.
- [8] LIN Di, SHEN Xiaoyong, LU Cewu, et al. Deep LAC: deep localization, alignment and classification for fine-grained recognition [C]//Proc of IEEE conference on computer vision and pattern recognition. Boston: IEEE, 2015: 1666–1674.
- [9] TANG Gongbo, MULLER M, RIOS A, et al. Why self-attention? A targetted evaluation of neural machine translation architectures [J]. International Journal of Automation and Computing, 2018, 145: 40–43.
- [10] DOSOVITSKY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words; transformers for image recognition at scale [J]. International Journal of Automation and Computing, 2021, 156: 48–95.
- [11] 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.
- [12] CHATFIELD K, SIMONYAN K, VEEDALDI A, et al. Return of the devil the details; delving deep into convolutional nets [C]//Proc of the British machine vision conference. [s. l.]: British Machine Vision Association, 2014.
- [13] XIAO Tianjun, XU Yichong, YANG Kuiyuan, et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification [C]//2015 IEEE conference on computer vision and pattern recognition (CVPR). Boston: IEEE, 2015: 842–850.
- [14] CORINNA C, VAPNIK V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273–297.
- [15] FU Jianlong, ZHENG Heliang, MEI Tao. Look closer to see better; recurrent attention convolutional neural network for fine-grained image recognition [C]//Proc of IEEE conference on computer vision and pattern recognition. Honolulu: IEEE, 2017: 4438–4446.
- [16] VASWANI A, SAHAZER N, PARMAR N, et al. Attention is all you need [C]//Proc of the 31st international conference on neural information processing system. Red Hook: Curran Associates Inc., 2017: 6000–6010.
- [17] CHEN Jieneng, LU Yongyi, YU Qihang, et al. TransUNet: transformers make strong encoders for medical image segmentation [J]. arXiv:2102.04306, 2021.
- [18] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention [J]. arXiv:2012.12877, 2021.
- [19] CHEN C R, FAN Quanfu, PANDA R. CrossViT: cross-attention multi-scale vision transformer for image [C]//2021 IEEE/CVF international conference on computer vision (ICCV). Montreal: IEEE, 2021.
- [20] HE Ju, CHEN Jieneng, LIU Shuai, et al. TransFG: a transformer architecture for fine-grained recognition [J]. arXiv: 2103.07976, 2021.
- [21] VIG J, BELINKOV Y. Analyzing the structure of attention in a transformer language model [C]//Proc of the 2nd Black-boxNLP workshop on analyzing and interpreting neural networks for NLP. Los Angeles: IEEE, 2019: 63–76.
- [22] WANG Jun, YU Xiaohan, GAO Yongsheng. Feature fusion vision transformer for fine-grained visual categorization [J]. arXiv:2107.02341, 2022.
- [23] LUO Wei, YANG Xitong, MO Xianjie, et al. Cross-X learning for fine-grained visual categorization [C]//2019 IEEE/CVF international conference on computer vision (ICCV). Seoul: IEEE, 2019: 8241–8250.
- [24] ZHUANG Peiqin, WANG Yali, QIAO Yu. Learning attentive pairwise interaction for fine-grained classification [J]. arXiv:2002.10191, 2020.
- [25] LIU Chuanbin, XIE Hongtao, ZHA Zhengjun, et al. Filtration and distillation: enhancing region attention for fine-grained visual categorization [C]//Proceedings of the AAAI conference on artificial intelligence. Boston: AAAI, 2020: 11555–11562.