DOI:10.19651/j.cnki.emt.2209822

一种融合间接注意力的自适应特征提取方法

张书恒 李 军 张礼轩 王子文
(南京理工大学自动化学院 南京 210094)

摘 要: 针对基于 ViT 模型的细粒度图像识别算法存在特征提取不全面、参数选取不具普适性等问题,提出一种融合间接注意力的自适应特征提取方法(AFEIA)。首先,对于目标对象的特征提取,采用改进后的自然断点分类算法将特征分为最相关、次相关、不相关三种,对不同的输入样本可以自适应地提取最具辨别性特征,保证了特征提取的准确性;然后,利用注意力权重矩阵,获取被忽略特征中与目标对象间接相关的特征,以获取各对象之间细微的差异,保证了特征提取的全面性。实验表明,使用 AFEIA 方法的 ViT 模型在两个细粒度数据集 CUB-200-2011、Stanford Dogs 上分别达到 91.6%、91.5%的预测准确率,通过可视化方法和消融实验,验证了 AFEIA 方法的有效性。

关键词: 细粒度图像识别;注意力机制;特征提取
中图分类号: TP391 **文献标识码:** A **国家标准学科分类代码:** 520.20

A method of adaptive feature extraction with indirect attention

Zhang Shuheng Li Jun Zhang Lixuan Wang Ziwen
(School of Automation, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: The fine-grained image recognition algorithm based on the ViT model has some problems, such as feature extraction is not comprehensive and parameter selection is not universal. To solve these problems, this paper presents an Adaptive Feature Extraction method with Indirect Attention (AFEIA). Firstly, to classify the characteristics of the object as most relevant, less relevant, and irrelevant, the improved natural breakpoint classification algorithm is used. This method can extract the most discriminative features adaptively for different input samples, which ensures the accuracy of feature extraction. Secondly, the attention weight matrix is used to obtain the features that are indirectly related to the object. This method acquires subtle differences between objects and ensures comprehensive feature extraction. Experiments show that the ViT model using the AFEIA method achieved 91.6% and 91.5% prediction accuracy on two fine-grained datasets CUB-200-2011, and Stanford Dogs, respectively. Visualization methods and ablation experiments verified the effectiveness of the AFEIA method.

Keywords: fine-grained image recognition; attention mechanism; feature extraction

0 引 言

细粒度图像识别(fine-grained image recognition, FGIR)作为图像识别领域的一个重要分支,是对一个大类别中的子类进行分类识别,如识别出各种鸟的种类[1-3]、花的品种[4-6]、车的款式[7-8]等,在工业界和实际生活中有着广泛的业务需求和应用场景。

由于细粒度图像具有较大的类内差异和较小的类间差异,相对于全局信息来说,有辨别性的局部信息更为重要,因此,细粒度图像识别的核心思想是定位出图片中最具辨别性的区域。Transformer[9]是基于自注意力机制(self-attention mechanism)的一种全新架构,为图像识别领域提供了新思路。Dosovitskiy 等[10]基于 Transformer 架构提出 Vision Transformer(ViT)模型用于图像分类任务,在细粒度图像识别领域取得了优秀的效果。

为了获取最具辨别性区域的特征,Wang 等[11]提出从每个 Transformer 层提取重要标记来指导网络选择有辨别性的区域,以弥补局部、低层次和中层次的信息。Zhang 等[12]提出一种包含选择性注意力收集模块(SACM)的模型 AFTrans,对不同层次的特征进行提取,来判断每个输入图像块序列的重要性。Chou 等[13]提出了一种嵌入模块 PIM,可以集成到许多常见的骨干网络中,包括基于 CNN 和基于 Transformer 的网络。

现有的大多数工作通过从网络模型中提取固定数量的

收稿日期:2022-05-03

特征来提高模型性能，同时只选择了目标对象最明显的特征。这种策略存在下面两个方面的问题：第一，对于不同类型和尺度的输入样本采用同样的超参数不能保证结果的一致性。第二，由于只考虑了样本最明显的特征，提取的特征会忽略各样本之间细微的差异。

针对基于 ViT 模型的细粒度图像识别算法存在特征提取不全面、参数选取不具普适性等问题，提出一种融合间接注意力的自适应特征提取方法（adaptive feature extraction with indirect attention，AFEIA）。AFEIA 方法包含自适应特征提取和融合间接注意力矩阵两部分：采用改进后的自然断点分类算法实现自适应特征提取，保证了特征提取的全面性；融合间接注意力的方法保留了各特征之间存在的联系，保证了特征提取的全面性。

1　Vision Transformer 模型

1.1　注意力机制

由于细粒度图像存在细微的差异，为了提升细粒度图像的识别精度，要求网络模型能够学习到准确而丰富的特征。ViT 模型基于注意力机制，将有限的注意力集中到重点信息上，从而节约资源，快速获取最有效的信息。

注意力机制首先通过快速扫描全局图像，获取需要重点关注的目标区域，然后对这一区域投入更多的注意力资源，以获得更多的细节信息，并抑制其它无用信息[14]。注意力机制是广义上的 CNN，它可以考虑输入序列的全局信息。ViT 模型采用的注意力机制如式(1)所示。

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}=softmax(\boldsymbol{A})\boldsymbol{V} \tag{1}$$

其中，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 是原始特征向量 $\boldsymbol{Z}$ 使用 3 个不同的变换矩阵 $\boldsymbol{W}_q$、$\boldsymbol{W}_k$、$\boldsymbol{W}_v$ 进行线性变换得到的，如式(2)所示。

$$(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=(\boldsymbol{W}_q,\boldsymbol{W}_k,\boldsymbol{W}_v)\boldsymbol{Z} \tag{2}$$

注意力权重矩阵 $\boldsymbol{A}$ 是 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 矩阵的点积，表示 $\boldsymbol{Q}$ 多大程度上能够匹配 $\boldsymbol{K}$，代表两个向量的相似度。为了解决归一化后出现的梯度消失问题，将 $\boldsymbol{Q}$、$\boldsymbol{K}$ 矩阵点积的结果除以 $\sqrt{d_k}$，其中 d_k 表示 $\boldsymbol{K}$ 矩阵的纬度[15]。通过注意力矩阵可以得到任意特征之间的注意力分数，获取各特征之间的联系。

1.2　Vision Transformer 模型

受到 Transformer 在 NLP 领域成功应用的启发，Dosovitskiy 等[10]在视觉领域摆脱了对 CNN 框架的依赖，提出了 ViT 架构，由于 ViT 模型采用了自注意力机制，相比于 CNN 可以考虑到更多的全局信息，同时更加的灵活。ViT 模型框架如图 1 所示。

ViT 模型将输入图像表示为图像块序列，对一个分辨率为 $H\times W$ 像素的输入图像处理步骤如下：首先，将图片分为 N 个固定大小的图像块(patch)序列 $\boldsymbol{X}_p$，其中 $N=\lfloor H/P\times W/P\rfloor$，$P$ 是每一个图像块的大小；然后，将输入序列 $\boldsymbol{X}_p$ 线性投影到 D 维的嵌入空间；最后，添加一个可学习的嵌入向量 $\boldsymbol{X}_{class}$ 来表示图像类别，以及一个与嵌入向量相同纬度的位置向量 $\boldsymbol{E}pos$ 来表示每一个图像块的位置信息。$\boldsymbol{X}_p$ 的嵌入过程如式(3)所示。

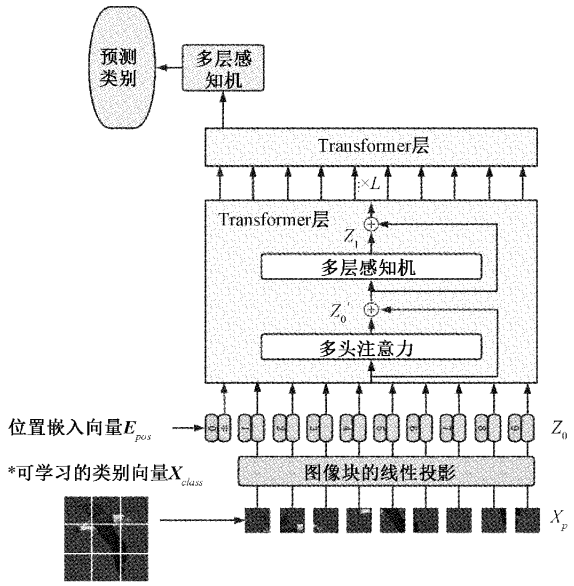


图 1　ViT 模型框架

$$\boldsymbol{Z}_0=[\boldsymbol{X}_{class};\boldsymbol{X}_p^1\boldsymbol{E};\boldsymbol{X}_p^2\boldsymbol{E};\cdots;\boldsymbol{X}_p^N\boldsymbol{E}]+\boldsymbol{E}_{pos},\boldsymbol{E}\in\boldsymbol{R}^{(P^2\cdot C)\times D},\ \boldsymbol{E}_{pos}\in\boldsymbol{R}^{(N+1)\times D} \tag{3}$$

图像经过式(3)处理过后得到的线性嵌入序列(tokens) $\boldsymbol{Z}_0$，作为第一个 Transformer 层的输入。ViT 由 L 层 Transformer 组成，每层由多头注意力机制(multi-head self-attention)和多层感知机(multilayer perceptron)组成，第 $l-1$ 层的输入 $\boldsymbol{Z}_{l-1}$ 经过式(4)和(5)的运算后会变成下一个 Transformer 层的输入。

$$\boldsymbol{Z}_l'=MSA(LN(\boldsymbol{Z}_{l-1}))+\boldsymbol{Z}_{l-1},l=1\cdots L \tag{4}$$

$$\boldsymbol{Z}_l=MLP(LN)\boldsymbol{Z}_l'))+\boldsymbol{Z}_l',l=1\cdots L \tag{5}$$

多头注意力机制包含多个自注意力，输出是由多个自注意力机制输出矩阵拼接后经线性变换得到的。多层感知机包含一个输入层、一个输出层、一个隐含层，相邻层神经元之间使用全连接方式进行连接。

2　模型设计

2.1　自适应特征提取

不同训练样本的特征具有差异性，在现有工作中，特征提取的参数需要由实验经验来确定，不具有普适性。本文提出了一种自适应特征提取方法，使得模型能够根据注意力权重矩阵 $\boldsymbol{A}$ 自动进行特征提取，以获取最具辨别性区域的特征。每个 Transformer 层的注意力权重矩阵 $\boldsymbol{A}$ 如式(6)所示，类别标签与图像序列之间的注意力权重向量 $\boldsymbol{a}^0$ 如式(7)所示。

$$\boldsymbol{A}=[\boldsymbol{a}^0;\boldsymbol{a}_1;\boldsymbol{a}_2;\cdots;\boldsymbol{a}_n] \tag{6}$$

$$\boldsymbol{a}^0 = [a_{0,0}, a_{0,1}, a_{0,2}, \cdots, a_{0,N}] \tag{7}$$

注意力权重较高的特征被认为是与主体对象相关的特征,注意力权重较低的特征被认为是背景信息的特征。为了实现自适应特征提取目的,引入了自然断点分类算法,该算法的核心思想是使每一组内部的相似性最大,组与组之间的相异性最大。通过该算法选择的图像块序列更符合人类的直觉,能够定位出数据中出现明显断层的位置,更加准确地定位出主体对象的特征。同时,为了尽可能地扩大最具辨别性区域与背景区域的差异,对自然断点分类算法进行改进,改进后算法的原理公式为:

$$SDAM = \frac{1}{N}\sum_{i=1}^{N}(a_{0,i} - \overline{\boldsymbol{a}^0})^2 \tag{8}$$

$$SDCM = \sum_{j=1}^{k}\frac{1}{N_j}\sum_{i=1}^{N_i}\left(a_{i,j}^0 - \frac{1}{j}\sum_{m=1}^{j}\overline{a_m^0}\right)^2 \tag{9}$$

$$GVF = 1 - \frac{SDCM}{SDAM}, GVF \in (0,1) \tag{10}$$

其中,式(8)中 SDAM 表示 $\boldsymbol{a}^0$ 的总方差。式(9)为该算法主要改进的公式,SDCM 表示每类特征的方差和;k 表示特征的类别数,为了提取目标对象的特征同时排除背景噪声的干扰,自适应地筛选出最相关、次相关、不相关的特征,将 k 设置为 3;$\overline{\boldsymbol{a}_m^0}$ 表示每种特征的聚类中心,为了扩大背景和目标对象之间的方差,将前 j 个类别聚类中心的平均值作为第 j 类的聚类中心,增大了次相关、不相关特征的方差和,减少了最相关特征中的干扰。式(10)得到的 GVF 表示特征提取的准确度,其数值范围在 0～1,GVF 的值越大表明特征提取越准确。

自然断点分类算法实现自适应特征提取的步骤为:首先,分别计算每类的方差和,用方差和的大小来比较分类的好坏;然后,通过不断迭代每个特征对应的类别使 GVF 最大;最后,对最大类别边界点之间的特征进行提取,以实现自适应特征提取的目的。

通过本文提出的自适应方法,模型能够更加精确地提取主体对象的特征,减少背景区域的干扰。同时,针对不同训练样本不再需要实验经验来确定提取特征的数量,降低了调试的难度,保证了网络模型的普适性。

2.2 融合间接注意力

注意力矩阵 $\boldsymbol{A}$ 表示任意特征之间的关联程度,其中 $\boldsymbol{a}^0$ 是类别标签与图像序列之间的注意力权重向量,本文将 $\boldsymbol{a}^0$ 定义为“直接注意力”,将通过自适应特征提取方法获取到的特征定义为“直接特征”。但矩阵 $\boldsymbol{A}$ 中除 $\boldsymbol{a}^0$ 以外的向量同样包含丰富的信息,本文将这些注意力向量定义为“间接注意力”,通过间接注意力提取到的特征定义为“间接特征”。

由于训练样本不仅仅包含类别标签与图像之间的对象级关系,同时图像中各特征之间也存在细微的联系。如 CUB-200-2011 细粒度图像数据集中冠毛小海雀的头部是它最具辨别性的特征,但是它的羽毛和栖息环境也是不可忽略的重要特征信息。使用上述自适应特征提取方法,可以使模型获取到与类别标签最直接相关的特征,但是会忽略每个特征之间存在的联系,因此本文提出了一种融合间接注意力的方法来保证特征提取的全面性。

间接注意力包含了各特征之间的关系,与主体对象间接相关的特征对提高模型预测准确率具有巨大的作用。融合间接注意力的方法如式(11)所示。

$$\boldsymbol{Z}_{indirect} = \boldsymbol{Z}_l[Max(\boldsymbol{a}_{indirect})] \tag{11}$$

首先,通过间接注意力矩阵 $\boldsymbol{a}_{indirect}$ 找到每个向量中权重最大值对应特征的索引,根据这些索引,对原始特征矩阵 $\boldsymbol{Z}_l$ 进行切片并拼接,作为提取到的间接特征。通过融合间接注意力的方法可以找到与直接特征存在关联的间接特征,保留了各特征之间的联系。间接特征作为对直接特征的补充,提高了特征提取的全面性。准确而全面的特征提取可以提升模型的性能,使得模型能够发现类别之间细微的差异。

2.3 模型框架

本文采用 ViT 模型作为主干网络,提出了一种融合间接注意力的自适应特征提取方法 AFEIA,采用 AFEIA 方法的 ViT 模型整体框架如图 2 所示。

AFEIA 模块融合了自适应特征提取方法和间接注意力两部分。原始的注意力矩阵 $\boldsymbol{A}_l$ 作为 AFEIA 模块的输入,AFEIA 模块首先通过自适应方法得到直接注意力;然后,将直接注意力与原始注意力进行残差连接,残差连接的结果作为间接注意力模块的输入;最后,将间接注意力模块的输出 $\boldsymbol{Mask}_l$ 与原始特征 $\boldsymbol{Z}_l$ 进行哈达玛积,哈达玛积的结果 $\boldsymbol{Z}_l^{selected}$ 为通过 AFEIA 方法最终提取到的特征。

原始的 ViT 模型更关注于高层的全局信息,对于局部的信息关注较少。为了充分地融合高层的语义信息和低层的位置信息,模型将每个 Transformer 层的注意力权重向量 $\boldsymbol{A}_l$ 通过 AFEIA 模块进行特征提取,并将提取到的高、中、低层特征进行融合,作为模型最后一个 Transformer 层的输入。

特征融合有多种组合方式,准则是保持各层特征的准确性和多样性,与单独的特征相比,融合后的特征能够提高模型的预测准确率。这就要求特征融合方法能够扬长避短,使不同层的特征相互独立且优势互补。本文采用的特征融合方法如式(12)所示。

$$\boldsymbol{Z}_{final} = \sum_{l=1}^{L-1}\boldsymbol{Z}_l \cdot \boldsymbol{Mask}_l + \boldsymbol{Z}_{L-1}^0 \tag{12}$$

其中,$\boldsymbol{Z}_l$ 是第 l 层的原始特征矩阵;$\boldsymbol{Mask}_l \in \boldsymbol{R}^{(N+1)\times D}$ 矩阵是由 AFEIA 方法得到的二元矩阵,矩阵中元素为 1 的位置表示对应的特征被选择,为 0 则表示对应的特征被舍弃,如式(13)所示。

$$m_{i,j} = \begin{cases} 1, & i \in \boldsymbol{Z}_{direct}, \boldsymbol{Z}_{indirect} \\ 0, & i \notin \boldsymbol{Z}_{direct}, \boldsymbol{Z}_{indirect} \end{cases} \tag{13}$$

通过 $\boldsymbol{Z}_l$ 与 $\boldsymbol{Mask}_l$ 做哈达玛积后会对目标对象的特征

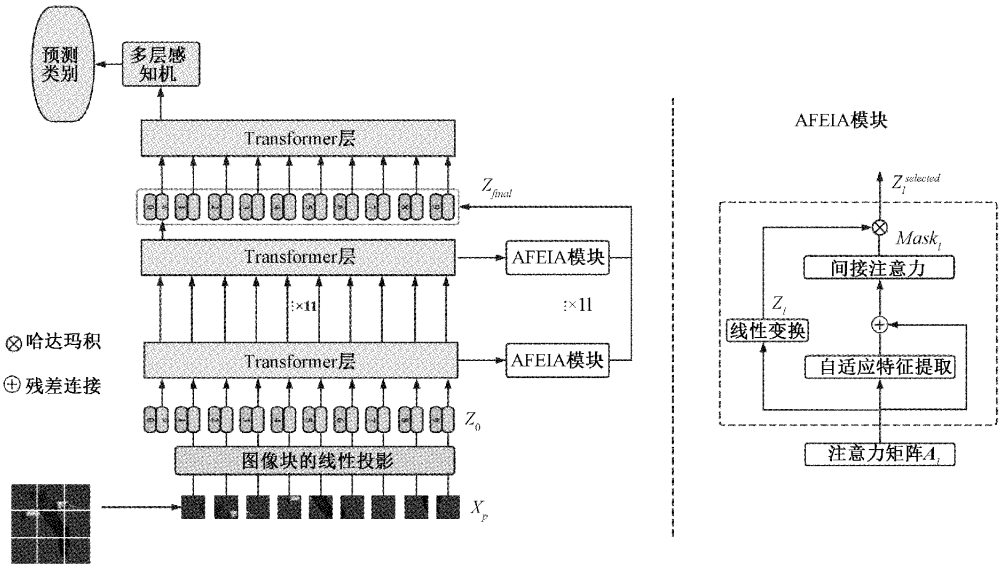


图 2　模型整体框架

进行提取，同时减少背景的干扰；$\boldsymbol{Z}_{L-1}^{0}$ 表示第 $L-1$ 层的类别标签，它是一个可学习的嵌入向量，表示此图像对应的类别信息，拼接 $\boldsymbol{Z}_{L}^{0}$ 以指导分类器输出准确的预测结果。

3　实验结果与分析

3.1　实验结果

本文实验在 CUB-200-2011[15] 和 Stanford Dogs[16] 两个细粒度图像数据集上进行。CUB-200-2011 数据集包含了 200 种鸟的类别，有 5 994 张训练样本和 5 794 张测试样本。Stanford Dogs 数据集包含了 120 种狗的图像，有 12 000 张训练样本和 8 580 张测试样本。实验采用 ViT 作为主干网络，采用 ViT-B 16 作为预训练模型。

训练样本需要预处理后才能作为模型的输入。首先将输入图片统一缩放为 600×600 像素大小，然后随机裁剪为 448×448 像素大小，最后进行水平翻转以及改变图像亮度、对比度、饱和度和色调等属性，以此增加样本的多样性和数量，提高模型的泛化能力。将经过预处理的图像分隔成 16×16 像素的小图像块并展平，每个图像块的纬度为 16×16×3=768，图像块的数量为 $(448/16)^2+1=785$，则输入矩阵 $\boldsymbol{Z}_0 \in \boldsymbol{R}^{785\times768}$。

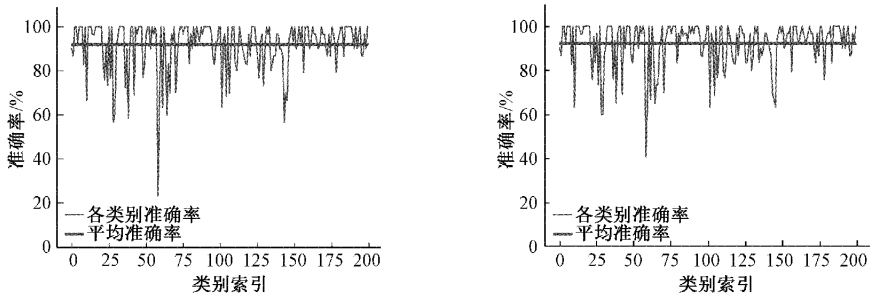
实验环境基于 linux 操作系统的服务器，CPU 为 Intel Xeon Gold 5218R，GPU 为 4 块 GeForce RTX 2080 Ti。深度学习框架选择 Pytorch1.8.1，编程环境选择 Python3.8.10，CUDA 驱动版本选择 11.1，batch_size 设置为 8，learning_rate 设置为 0.02。由于刚开始训练时模型权重是随机初始化的，此时选择一个较大的学习率可能会带来模型的不稳定，因此本文采用 cosine decay 方法[17] 来对学习率进行预热，并采用 SGD 算法[18] 对模型参数进行更新。

如表 1 所示，分别在 CUB-200-2011 和 Stanford Dogs 数据集上比较了 AFEIA 方法与其它方法的实验结果。AFEIA 方法在 CUB-200-2011 数据集上达到了 91.6% 的预测准确率，相较于同样使用 ViT 网络的 FFVT 和 AFTrans，分别增加了 0.4% 和 0.1% 的预测准确率；在 Stanford Dogs 数据集上达到了 91.5% 的预测准确率，相较于 FFVT 和 AFTrans，分别增加了 0.1% 和 0.2% 的预测准确率。AFEIA 方法在同样采用 ViT 作为主干网络的模型中取得了最好的结果，验证了 AFEIA 方法的有效性。

表 1　各方法在 CUB-200-2011 和 Stanford Dogs 数据集上的对比结果

细粒度图像识别方法	主干网络	CUB-200-2011	Stanford Dogs
RA-CNN[19]	VGG-19	85.3	87.3
MaxEnt[20]	DenseNet-161	86.6	83.6
DFL-CNN[21]	ResNet-50	87.4	84.9
API-Net[22]	DenseNet-161	90.0	90.3
StackedLSTM[23]	GoogleNet	90.4	—
DeepFVE[24]	InceptionV3	90.7	—
TransFG[25]	ViT-B 16	90.9	90.4
FFVT[11]	ViT-B 16	91.2	91.4
AFTrans[12]	ViT-B 16	91.5	91.3
AFEIA	**ViT-B 16**	**91.6**	**91.5**

通过对测试集中每个类别的预测准确率进行分析,每一类样本的预测准确率如图 3 所示。图 3(a)为原始 ViT 模型各类样本的预测准确率,数据集中第 59 到 66 类物种都同属于鸥科动物,相似度极高,其中对第 59 类测试样本加州海鸥的预测准确率仅有 23.3%,表明原始的 ViT 模型对这几类物种的辨别能力不足。图 3(b)为采用了 AFEIA 方法的模型各类样本预测准确率,对加州海鸥的预测准确率提高到了 40.8%。可以看出,采用了 AFEIA 方法的模型由于提取到了样本中细微的差别,对难以辨别的对象有较好的识别准确率。



(a) 原始ViT模型 (b) 采用了AFEIA方法的模型

图 3 各类样本的预测准确率

3.2 消融实验

为了进一步验证 AFEIA 方法的有效性,在 CUB-200-2011 数据集上进行了消融实验,实验结果如表 2 所示,同样的实验结果也可以在其它数据集上得到验证。

表 2 在 CUB-200-2011 数据集上的消融实验结果

模块	准确率/%
ViT	90.3
ViT+自适应特征提取	91.5
ViT+自适应特征提取+间接注意力	91.6

通过表 2 可以看出,自适应特征提取方法相对于直接使用全部特征的主干网络,网络模型的预测准确率从 90.3%提高到了 91.5%,提高了 1.2%。在应用自适应特征提取方法的同时,又加入间接注意力提取的特征,可以进一步提高 0.1%的预测准确率,达到了 91.6%,验证了 AFEIA 方法的有效性。

利用 Grad-CAM 方法[26]对模型每层的特征进行可视化分析,同时跟主干网络进行对比。图 4 为模型各层提取特征的可视化对比结果,第 1 行图像为原始 ViT 模型的热力图,第 2 行是添加了间接注意力的热力图,第 3 行为融合间接注意力的自适应特征方法提取特征的热力图。

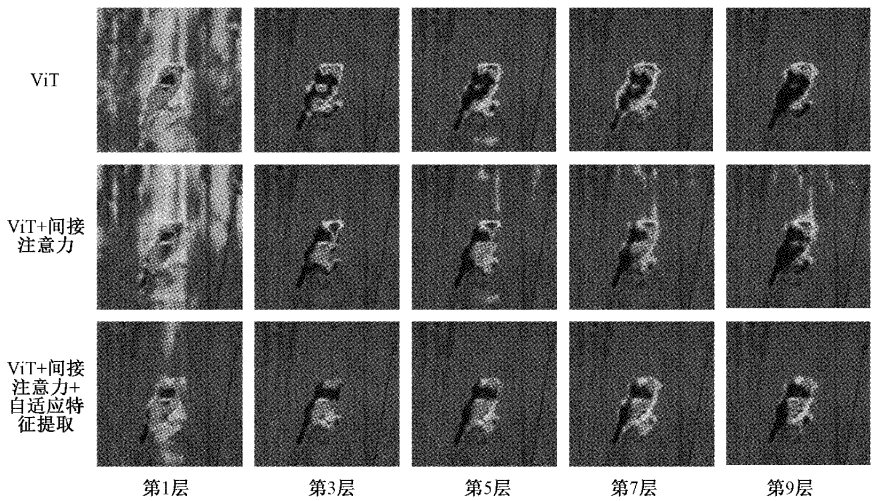


图 4 特征可视化热力图

通过对比可以看出,添加了间接注意力的模型相比原始 ViT 模型关注的特征更全面,如第 3 层、第 5 层、第 7 层注意到了原始 ViT 模型未考虑到的鸟腹部位置的信息,但是考虑间接注意力可能会引入背景的干扰。通过观察图 4 第三行可以发现,添加自适应的方法后,模型保留了腹部信息的同时消除了背景信息的干扰。实验表明 AFEIA 方法提取的特征相比于原始 ViT 模型更加准确和全面。

4 结 论

本文基于 ViT 模型提出了一种融合间接注意力的自

适应特征提取方法 AFEIA，增强了细粒度图像识别模型的预测准确率。针对不同训练样本特征提取参数不具普适性的问题，AFEIA 方法采用改进后的自然断点分类算法自适应地提取最具辨别性的特征，保证了特征提取的准确性。同时由于间接特征对于模型也是非常重要的，因此 AFEIA 方法融合了间接注意力，使得模型对特征的提取更加全面。从实验结果可以看出，AFEIA 方法在同样使用 ViT 作为主干网络的模型中取得了最好的结果，表明了采用 AFEIA 方法的模型具有较好的性能。通过消融实验，验证了 AFEIA 方法的有效性。特征可视化表明通过 AFEIA 方法提取的特征更加准确和全面。

参考文献

[1] WAH C, BRANSON S, WELINDER P, et al. The caltech-UCSD birds-200-2011 dataset[J]. California Institute of Technology, 2011:1-8.

[2] ZHUANG P, WANG Y, QIAO Y. Learning attentive pairwise interaction for fine-grained classification[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7):13130-13137.

[3] 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,2020,230(2):115-122.

[4] HAN K, XIAO A, WU E, et al. Transformer in transformer[J]. Advances in Neural Information Processing Systems, 2021, 34:1-12.

[5] KOLESNIKOV A, BEYER L, ZHAI X, et al. Big transfer(bit): General visual representation learning[C]. European Conference on Computer Vision, Springer, Cham, 2020: 491-507.

[6] 任意平，夏国强，李俊丽. 基于花蕊区域定位的花卉识别方法[J]. 电子测量技术，2020(7)：97-102.

[7] ZHUANG P, WANG Y, QIAO Y. Learning attentive pairwise interaction for fine-grained classification[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7):13130-13137.

[8] JIA C, YANG Y, XIA Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision[C]. International Conference on Machine Learning, PMLR, 2021: 4904-4916.

[9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30:1-15.

[10] 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.

[11] WANG J, YU X, GAO Y. Feature fusion vision transformer for fine-grained visual categorization[J]. ArXiv Preprint, 2021, ArXiv:2107.02341.

[12] ZHANG Y, CAO J, ZHANG L, et al. A free lunch from ViT: Adaptive attention multi-scale fusion transformer for fine-grained visual recognition[J]. ArXiv Preprint, 2021, ArXiv:2110.01240.

[13] CHOU P Y, LIN C H, KAO W C. A novel plug-in module for fine-grained visual classification[J]. ArXiv Preprint, 2022, ArXiv:2202.03822.

[14] 王溪波，曹士彭，赵怀慈，等. 双边特征聚合与注意力机制点云语义分割[J]. 仪器仪表学报，2021，42(12)：175-183.

[15] BRITZ D, GOLDIE A, LUONG M T, et al. Massive exploration of neural machine translation architectures[J]. ArXiv Preprint, 2017, ArXiv:1703.03906.

[16] KHOSLA A, JAYADEVAPRAKASH N, YAO B, et al. Novel dataset for fine-grained image categorization: Stanford dogs[C]. Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC), Citeseer, 2011, 2(1):1-10.

[17] HE T, ZHANG Z, ZHANG H, et al. Bag of tricks for image classification with convolutional neural networks[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 558-567.

[18] WOODWORTH B, PATEL K K, STICH S, et al. Is local SGD better than minibatch SGD?[C]. International Conference on Machine Learning, PMLR, 2020: 10334-10343.

[19] FU J, ZHENG H, MEI T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 4438-4446.

[20] DUBEY A, GUPTA O, RASKAR R, et al. Maximum-entropy fine grained classification[J]. Advances in neural information processing systems, 2018, 31:1-11.

[21] WANG Y, MORARIU V I, DAVIS L S. Learning a discriminative filter bank within a cnn for fine-grained recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4148-4157.

[22] ZHUANG P, WANG Y, QIAO Y. Learning attentive pairwise interaction for fine-grained classification[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 13130-13137.

[23] GE W, LIN X, YU Y. Weakly supervised complementary parts models for fine-grained image

classification from the bottom up[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 3034-3043.

[24] KORSCH D, BODESHEIM P, DENZLER J. End-to-end learning of fisher vector encodings for part features in fine-grained recognition [C]. DAGM German Conference on Pattern Recognition Springer, Cham, 2021: 142-158.

[25] HE J, CHEN J N, LIU S, et al. Transfg: A transformer architecture for fine-grained recognition [J]. ArXiv Preprint, 2021, ArXiv: 2103.07976.

[26] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization [C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 618-626.

作者简介

张书恒,硕士研究生,主要研究方向为细粒度图像识别。
E-mail:zhangshuheng@njust.edu.cn

李军(通信作者),教授,主要研究方向为自动检测理论与技术,工业领域中的图像处理、图像识别技术等。
E-mail:ljun1008@njust.edu.cn

张礼轩,硕士研究生,主要研究方向为智能人群计数。

王子文,硕士研究生,主要研究方向为交通标志识别技术。