DOI:10.16136/j.joel.2023.06.0303

基于transformer自适应特征向量融合的图像分类

胡 义[1],黄勃淳[2],李 凡*

(昆明理工大学 信息工程与自动化学院,云南 昆明 650500)

摘要:针对目前基于transformer的图像分类模型直接应用在小数据集上性能较差的问题,本文提出了transformer自适应特征向量融合网络,该网络在特征提取器中将不同阶段的特征进行融合,减少特征信息丢失的同时获得更多不同感受野下的信息,同时利用最大池化来去除特征中的冗余信息,从而使提取的特征更具有判别性。此外,为了充分利用图像的各级特征信息来进行分类预测,本文将网络各阶段产生的特征向量进行融合,使融合后的特征向量更具有表征能力,从而减少网络对大数据集的依赖,使网络在小数据集中也能获得很好的性能。实验表明,本文提出的算法在数据集Mini-ImageNet-100、CIFAR-100和ImageNet-1k上的TOP-1准确率分别达到了74.22%、85.86%和81.4%。在没有增加计算量的情况下,在baseline上分别提高了6.0%、3.0%和0.1%,且参数量减少了18.3%。本文代码开源在"https://github.com/xhutongxue/afvf"。

关键词:transformer;图像分类;自适应特征向量融合;卷积神经网络(CNN);模式识别

中图分类号:TP391.4 **文献标识码**:A **文章编号**:1005-0086(2023)06-0602-08

Image classification based on transformer adaptive feature vector fusion

HU Yi[1], HUANG Bochun[2], LI Fan*

(Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China)

Abstract: Aiming at the problem of poor performance that the current transformer-based image classification model is directly applied to the small data sets, this paper proposes a transformer adaptive feature vector fusion network, which fuses features at different stages in the feature extractor, reduces the loss of feature information and obtains more information under different receptive fields, and uses maximum pooling to remove redundant information of features, so that the extracted features are more discriminative. In addition, in order to make full use of the feature information at all levels of the image for classification prediction, this paper fuses the feature vectors generated at each stage of the network to make the fused feature vectors more representative. Thereby reducing the network's dependence on large data sets, so that the network can also obtain good performance in small data sets. Experiments show that the algorithm proposed in this paper achieves 74.22%, 85.86% and 81.4% of the TOP-1 accuracy on the datasets Mini-ImageNet-100, CIFAR-100 and ImageNet-1k, respectively. Without increasing the amount of computation, the baselines are improved by 6.0%, 3.0%, and 0.1%, respectively, and the amount of parameters is reduced by 18.3%. The code of this article is open source at "https://github.com/xhutong xue/afvf".

Key words: transformer; image classification; adaptive feature vector fusion; convolutional neural network (CNN); pattern recognition

* **E-mail**:478263823@qq.com

收稿日期:2022-04-27 **修订日期**:2022-06-26

基金项目:国家自然科学基金(61862036,61962030,81860318)资助项目

0 引　言

图像分类的任务是学习和判断图像中是否包含某种特定的目标内容，并依据其内容信息进行分类的过程。图像分类是最基础的计算机视觉任务，其已经在人机交互、生物医学、航空航天和公安司法等领域取得了广泛应用。

用于图像分类的传统机器学习方法有 k 最近邻(k near neighbor，KNN)[1]、支持向量机(support vector machine，SVM)[2]等，这些传统的图像分类方法存在计算量大、训练时间长、对参数调节和核函数的选取比较敏感等诸多问题。因此，基于深度学习的图像分类方法引起了研究人员的关注，并得到了迅速发展。

20 世纪 90 年代，LECUN 等[3]提出 LeNet-5(Le-Cun network)网络，该网络利用卷积层、池化层、激活层来对手写数字图像进行特征提取，最后利用全连接层来得到预测值，但对于复杂问题的分类结果并不理想。因此，研究者在后续提出 AlexNet(Alex Krizhevsky network)[4]、VggNet(visual geometry group)[5]、Google-LeNet(google Le-Cun network)[6]和 ResNet(residual network)[7]等经典卷积神经网络(convolutional neural network，CNN)架构并用于更加复杂的图像分类任务，取得了更加优秀的性能。虽然 CNN 有着归纳偏置、低冗余以及具有平移不变性、平移等变性等诸多优点，使得 CNN 适合于图像处理任务，但 CNN 的缺点也较为突出，其实际感受野远小于理论感受野，不利于模型充分地利用上下文信息进行特征提取。虽然模型能够靠堆叠更深的卷积层来获得更大的感受野，但这样显然会造成模型过于臃肿、计算量急剧增加。由于 transformer[8]能够建立长期依赖关系且能并行训练，因此，研究人员将 transformer 应用于图像分类中，其中，ViT[9]是首个用于图像分类的纯 transformer 架构，虽然它获得了比许多利用先进 CNN 作为特征提取器的图像分类模型更好的性能，但它严重依赖于 JFT-300 M(共 3 亿张图像)或 ImageNet-22K(共 750 万张图像)这样规模的数据集来进行预训练。直接将 ViT 应用于规模较小的数据集 CIFAR-100(共 6 万张图像)上时，其性能比相似大小的 ResNet50 低 13.6%，这严重限制了 transformer 在图像分类中的应用。为了解决 transformer 图像分类模型对大规模数据的依赖，研究人员将 CNN 的优点直接或间接地引入到 transformer 中。其中 DeiT[10]采用知识蒸馏的方式将 CNN 作为教师网络来引导基于 transformer 的学生网络进行学习，从而间接地将 CNN 的优点引入到 transformer 网络中，使得 DeiT 仅使用 ImageNet-1k(共 133 万张图像)数据便可得到高于 ViT 的性能，极大地改善了 transformer 图像分类模型对大规模数据的依赖。但将其直接应用于更少数据量的 CIFAR-100 数据集上时，其性能并无明显提升。Swin[11]通过移动窗口分割的方式来解决不同窗口之间的信息交流问题，从而使 Swin 在 CIFAR-100 数据集上的性能得到极大提高，达到 78.03%，但仍然低于 Res-Net50 的 81.34%。Twins[12]在块嵌入(patch embedding)中使用卷积来改变特征尺寸，同时在网络中交替使用局部分组自注意力和全局子采样自注意力，从而直接将 CNN 引入到 transformer 网络中，增强了网络提取局部特征的能力。因此，使得 Twins 在数据集 CIFAR-100 上获得优于 CNN(ResNet50)的性能，但是在更少数据量的 Mini-ImageNet-100 数据集(共 3 万张图像)上性能仍低于 ResNet50。

为了解决基于 transformer 的图像分类网络直接应用在小数据集上性能较差的问题，本文提出了一种自适应特征向量融合网络，该网络一方面对 baseline 中的特征提取器进行改进使其提取出更具有判别性的特征，另一方面通过融合图像的各级特征信息来增强特征的表现能力，加快网络对目标特征的学习。这些使网络不再依赖大量的数据来进行学习，直接在小数据集上进行训练测试也能获得很好的性能。

1 提出的模型

本文提出的网络如图 1 所示，该网络由特征提取器及分类器组成。其中特征提取器由 4 个阶段组成，每个阶段包含一个改进的块嵌入(improved patch embedding)和 4 个 transformer 编码器(transformer encoder)，并且在网络前后两个阶段间增加了特征补偿模块(feature compensation)，其主要负责对图像进行特征提取。而分类器由本文提出的自适应特征向量融合模块(adaptive feature vector fusion module，AFVFM)组成，其主要负责将特征提取器提取的各级特征融合起来进行分类预测。

1.1 特征提取器

为了减少网络的计算量及参数量，便于后续添加其他模块，本文将 Twins 网络中第一、第二、第三和第四阶段的 transformer 编码器数量由原来的 2、2、10、4 分别改为 4、4、4、4，并将调整后的网络作为本文的 baseline。同时，为了提取出更具有判别性的图像特征，本文对 baseline 网络中的特征

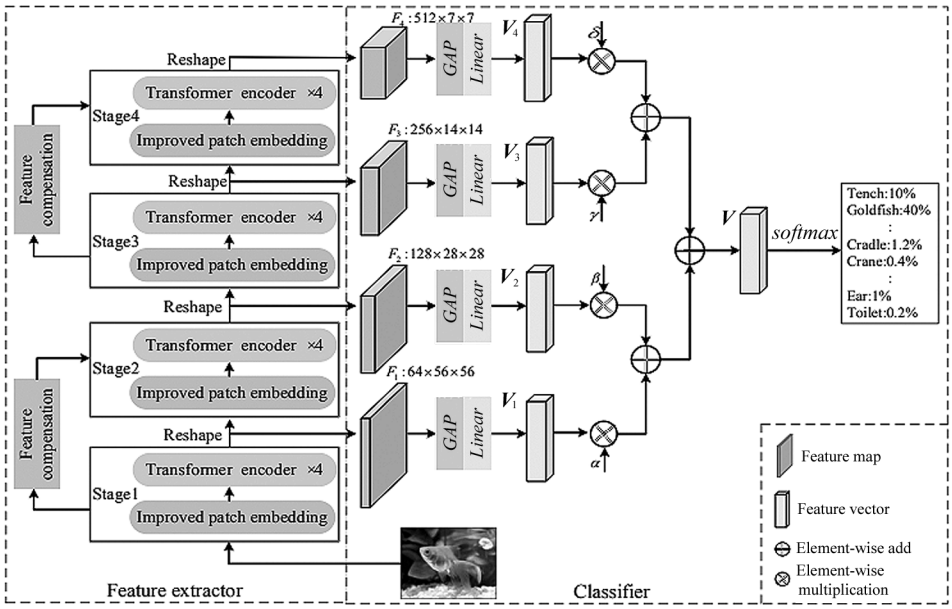

图 1　Transformer 自适应特征向量融合网络

Fig. 1　Transformer adaptive feature vector fusion network

提取器进行改进，其改进部分主要有以下两点：一是对特征提取器中的块嵌入模块进行改进，二是在特征提取器中增加特征补偿模块。

1.1.1　改进的块嵌入模块

为了在不增加过多计算量以及参数量的情况下，尽可能多地将特征图或者图像中有用的特征语义信息传入到下一阶段 transformer 编码器中，本文对 baseline 中的块嵌入模块进行改进。如图 2 所示，改进的块嵌入模块将卷积和最大池化结合，利用卷积来增强网络提取局部特征和底层特征的能力，并利用最大池化来去除特征中的冗余信息，从而使更多有利于网络识别预测的特征传入到下一阶段编码器中。

由于网络第一阶段的输入为二维图像，而其他阶段的输入为一维特征，因此，在第一阶段时，直接将图像进行 4×4 卷积和最大池化并利用 1×1 卷积来改变通道，使其卷积和最大池化后得到的特征通道数一致，便于将卷积和池化后得到的特征进行融合。而在其他阶段时，需要将输入的一维特征重塑为二维特征，然后才对其进行卷积和最大池化操作，此时的卷积为 2×2 卷积。表达式如下所示：

$$F_b = Conv(F_a) \oplus MaxPool(1\times 1Conv(F_a)) \quad (1)$$

式中，在第一阶段时，F_a 表示输入图像，在其他阶段时，F_a 表示将上一阶段输出的一维特征重塑后的二维特征，F_b 表示经过公式(1)处理后得到的二维特征。最后将融合后的二维特征重塑为一维特征输入到 transformer 编码器中。

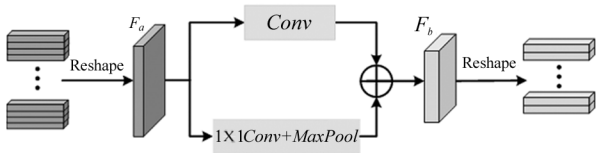

图 2　改进的块嵌入模块结构图

Fig. 2　Structure diagram of improved patch embedding model

1.1.2　特征补偿模块

为了减少特征信息的丢失以及进一步使特征提取器提取出更具有判别性的特征，本文在特征提取器中的前后两个阶段间增加了特征补偿模块。如图 3 所示，该模块由一个 2×2 卷积组成。其具体操作是将上一阶段前两个 transformer 编码器所提取特征输入到特征补偿模块中进行处理，由于 transformer 编码器的输出为一维特征，因此，需要先将其重塑为二维特征，之后对其进行卷积操作，将得到的结果重塑为一维特征并与下一阶段 transformer 编码器所输出的特征相融合，从而使融合后的特征具备网络两个阶段提取的信息。这样一方面可以减少特征信息的丢失，另一方面将两个阶段的特征融合起来增强了目标的特征信息，使提取的特征更具有判别性，更有利于网络对目标特征进行学习。同时本文在特征补偿模块中使用卷积来改变特征大小，进一步增强了网络提取局部特征的能力。

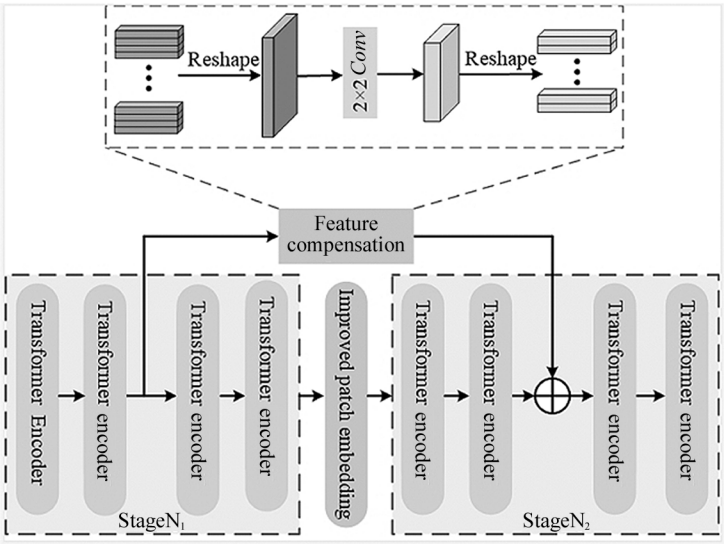

图 3　特征补偿模块结构图

Fig. 3　Structure diagram of feature compensation model

1.2　分类器(自适应特征向量融合模块)

以往的图像分类模型通常直接用全连接层或者全局平均池化(global average pooling,*GAP*)作为分类器,在特征提取器进行特征提取后,将网络最后一层的特征即图像的高级特征送入到分类器中进行分类预测,而忽略图像的低级以及中级特征。然而,在 transformer 网络进行特征提取的过程中,往往会丢失一些有用的特征信息,特别是当目标物体在图像中占比较小时,丢失的特征信息会更多。并且特征提取器往往不能精准地对图像中的小目标物体进行特征提取,反而提取了许多背景信息,使得网络最后一层特征中包含许多干扰信息。此时仅利用最后一层特征来进行分类,会使网络对目标特征的学习变得更为困难,需要大量的数据来进行学习。因此,当数据量不充分时,仅仅利用图像的高级特征来进行分类预测往往会导致性能较低。为此,本文受特征金字塔 FPN 及多特征融合表示思想[13-15]的启发,提出了 AFVFM。

为了充分利用图像的各级特征信息来进行图像分类,本文提出了 AFVFM。如图 1 分类器部分所示,在该模块中,本文将网络各阶段产生的特征图进行 *GAP* ,从而为特征的每一个通道赋予特殊的意义。同时将得到的特征向量进行线性映射,一方面可以增强网络的拟合能力,另一方可以将 4 个不同尺寸的特征向量转换到同一尺寸便于融合。最后将 4 个相同尺寸的特征向量进行融合并利用可学习参数来使网络自主学习各层特征向量的权重,从而使融合后的特征向量具备图像的各级特征信息,这样更有利于网络对目标特征进行学习,以减少网络对大量数据的依赖。

同时,由于网络低层提取的特征中虽然包含着有利于提升判别性的特征,但也包含许多干扰信息,而网络高层提取的特征中干扰信息相对较少,语义信息相对丰富。因此,在融合网络各层提取的特征时,既希望融合后的特征中含有网络低层提取的特征,又不希望引入太多干扰信息,而导致对网络的学习带来负面影响。基于此,本文在对特征向量融合权重 α、β、γ 和 δ 进行初始化时,将 α、β、γ 和 δ 的初始值分别设置为 0.1、0.2、0.3 和 0.4,即网络从低层到高层提取的特征在融合后的特征中所占比重逐层增大,从而使融合后的特征既引入了网络各层提取的特征信息,又避免了大量干扰信息的引入。

具体地,将网络 4 个阶段产生的一维特征,重塑为二维特征 F_1、F_2、F_3 和 F_4。接下来对 F_1、F_2、F_3 和 F_4 依次进行 *GAP* 和线性映射,得到特征向量 $\boldsymbol{V}_1$、$\boldsymbol{V}_2$、$\boldsymbol{V}_3$ 和 $\boldsymbol{V}_4$,其尺寸均为 N(N 为类别数,CIFAR-100 数据集,$N=100$,ImageNet-1k 数据集,$N=1\,000$)。随后将 4 个特征向量分别与 4 个可学习参数 α、β、γ 和 δ 相乘相加得到最终预测向量 $\boldsymbol{V}$。最后将向量 $\boldsymbol{V}$ 送入 $softmax$ 层完成分类预测。如式(2)—(4)所示:

$$\boldsymbol{V}_k = Linear(GAP(F_k)),\ k \in \{1,2,3,4\}, \tag{2}$$

$$\boldsymbol{V} = \alpha \times \boldsymbol{V}_1 \oplus \beta \times \boldsymbol{V}_2 \oplus \gamma \times \boldsymbol{V}_3 \oplus \delta \times \boldsymbol{V}_4, \tag{3}$$

$$p(v_{ij}) = softmax(v) = \frac{e^{v_{ij}}}{\sum_{j=0}^{N} e^{v_{ij}}}, \tag{4}$$

式中，F_k 为二维特征，$\boldsymbol{V}_1$、$\boldsymbol{V}_2$、$\boldsymbol{V}_3$ 和 $\boldsymbol{V}_4$ 分别表示网络4个阶段产生的特征向量。α、β、γ 和 δ 表示可学习参数，初始值分别为0.1、0.2、0.3和0.4，v_{ij} 表示第 i 张样本所得到的预测向量的第 j 个值，e为常数，$p(v_{ij})$ 表示网络预测第 i 张样本属于第 j 类的概率。

本文利用交叉熵损失对预测向量进行约束，进而迫使网络提取更具有判别性的特征。利用软标签和预测向量按照式(5)进行损失计算，

$$Loss = -\frac{1}{B}\times\sum_{i=0}^{B}\sum_{j=0}^{N}(q_{ij}\times\log(p(v_{ij}))), \quad (5)$$

式中，q_{ij} 表示第 i 张图像的类别标签经过标签平滑后所生成软标签的第 j 个值，B 表示一个批次的样本数量，N 表示类别总数，也是预测向量的长度。

2 实验与分析

本节将详细介绍所使用的数据集、实验细节和实验结果，以此来证明本文算法的有效性。

2.1 实验细节

实验中本文使用随机裁剪、随机水平翻转、标签平滑正则化、混合和随机擦除来进行数据增强，并在训练过程中使用AdamW优化器以0.9的动量、90的批量和 5×10^{-2} 的权重衰减来优化模型。初始学习率设置为0.002并使用余弦学习率衰减策略来对学习率进行衰减。所有模型都在一片3090上训练300代。同时，图像输入尺寸均为224×224。在对比实验中，计算量以及参数量是在分类数为1000的情况下计算机得来。

2.2 数据集及评价指标

本文将在数据集Mini-ImageNet-100、CIFAR-100和ImageNet-1k上进行实验。

2.2.1 数据集

(1) Mini-ImageNet-100

本文从ImageNet-1k数据集前100个类中选取3万张图像组成一个新的数据集，并将其命名为Mini-ImageNet-100。该数据集分为100个类，每个类均由ImageNet-1k数据集中对应类别的前250张训练图像和50张测试图像组成。

(2) CIFAR-100

CIFAR-100数据集由6万张32×32的彩色图像组成。共分为100个类，每个类有500张训练图像和100张测试图像。

(3) ImageNet-1k

ImageNet-1k数据集由来自1000个类别的128万张训练图像和5万张验证图像组成。

2.2.2 评价指标

本文采用♯Param (M)、FLOPs(G)和Top-1 *Acc* (%)3种评价指标。其中♯Param和FLOPs分别表示网络的6参数量和计算量，Top-1 *Acc* 表示网络在预测某张图片时，预测概率最大的类别与实际类别相符的准确率，其中准确率(*Acc*)按照式(6)计算：

$$Acc = \frac{n_a}{n}\times100\%, \quad (6)$$

式中，n 表示测试样本总数量，n_a 表示在所有测试样本中网络预测概率最大的类别与实际类别相符的样本数量。

2.3 对比实验

本文在数据集Mini-ImageNet-100、CIFAR-100和ImageNet-1k上与一些算法进行比较以此证明本文算法的优越性。这些算法包括基于CNN的算法[7,16-18]和基于transformer的算法[9-12,19]，并将对比结果记录在表1中。

表1 不同方法在数据集Mini-ImageNet-100、CIFAR-100和ImageNet-1k上的比较结果

Tab. 1 Comparison results of different methods on Mini-ImageNet-100, CIFAR-100 and ImageNet-1k data sets

Method type	Network	♯Param /M	FLOPs /G	Mini-ImageNet-100 Top-1 *Acc* /%	CIFAR-100 Top-1 *Acc* /%	ImageNet-1K Top-1 *Acc* /%
Convolutional networks	ResNeXt-50(32×4d)[16]	25.0	4.3	71.42	81.77	77.6
	ResNet50[7]	25.6	4.1	70.16	81.34	76.1
	ResNet50D[17]	25.0	4.3	71.46	81.82	77.2
	RegNetY-4G[18]	21.0	4.0	69.68	82.87	80.0
Transformer networks	ViT-S/16[9]	22.1	4.2	51.96	71.62	71.6
	DeiT-S/16[10]	22.1	4.6	51.56	71.74	79.9
	Swin-Ti[11]	29.0	4.5	66.20	78.03	81.3
	ConViT-S[19]	27.0	5.4	61.48	79.28	81.3
	Twins-PCPVT-S[12]	24.1	3.7	68.82	82.77	81.2
	Twins-SVT-S[12]	24.6	2.8	68.24	82.87	81.3
	Ours	20.1	2.8	74.22	85.86	81.4

从表 1 中可以看出，本文算法在数据集 Mini-Im-ageNet-100、CIFAR-100 和 ImageNet-1k 上的 Top-1 准确率分别达到了 74.22%、84.71% 和 81.4%，均高于其他算法，并且所用参数量最少。同时可以看出，随着数据量的减少，本文算法的优势更为明显。证明了本文算法可以进一步提高 transformer 网络在小数据集中的性能，降低 transformer 在图像分类中的使用成本。

2.4　消融及可视化实验

为了证明算法中每一个模块的有效性，本文在数据集(CIFAR-100)上进行了一系列消融实验。在 baseline 上依次将改进的块嵌入、特征补偿和 AFVFM 加入训练，并将模型性能记录在表 2 中。从表 2 中可以看出，当在 baseline 中加入改进的块嵌入模块后，模型性能提升了 0.68%，说明，本文在改进的块嵌入模块中加入的最大池化可以进一步去除特征中的冗余信息，减少特征中的干扰信息，使特征更具有判别性。当继续在模型中加入特征补偿模块后，模型的性能又提高了 0.9%，这是因为特征补偿模块将不同阶段的特征融合在一起，减少特征信息丢失的同时增强了特征的表现能力。使得改进后的特征提取器提取出更具有判别性的特征。当在模型中加入自适应特征向量融合后，模型的性能继续提升了 1.21%，这是因为虽然改进后的特征提取器可以更好地提取特征，但是网络最后一层的特征中仍然包含部分干扰信息以及随着网络的加深，仍有部分特征丢失，特别是对于图像中的小目标物体进行提取时，此时最后一层的特征中包含的干扰信息会更多，有用的信息会更少。因此，将不同阶段提取的特征生成的向量融合起来有利于提高预测向量的表征能力，并且利用可学习参数来控制各阶段特征向量的占比权重，可以使网络通过损失函数来进行自主调节。

为了更加直观地展现改进后特征提取器的有效性，本文根据Grad-CAM[20]策略对改进前后特征提取器所提取的特征进行可视化处理，并将结果在图 4 中进行展示。从图 4 中可以看出。当目标物体在图像中所占比例较大时，改进前的特征提取器虽然也能提取到目标特征，但是，提取的特征中包含许多干扰信息，不利于网络对目标进行学习和识别。而改进后的特征提取器可以准确地对目标特征进行提取，仅包含少许干扰信息。当目标物体在图像中所占比例较小时，改进前的特征提取器并不能很好地提取到目标特征，而改进后的特征提取器可以精准地对图像中的小目标物体进行特征提取，减少特征中的干扰信息，这进一步说明了改进后的特征提取器优于 baseline 中的特征提取器。特别是对图像中的小目标物体进行提取时，改进后的特征提取器所提取的特征更具有判别性，优势更为明显。

同时，为了直观地展示本文算法对 CIFAR-100 数据集中各类别的影响，本文将网络对测试图像的预测值按类别进行求和，并将其结果作为一个新的评价指标 P_S，P_S 表示网络对该类目标的识别能力，P_S 值越大表示网络对该类目标的识别能力越强，越容易识别出该类目标。各类别 P_S 值按照式(7)进行计算：

$$P_{S_i}=\sum_{j=0}^{100}P_{ij}\ ,\ i\in\{0,1\cdots98,99\}\ , \tag{7}$$

式中，P_{S_i} 表示测试集中所有属于类图像的预测值总和，P_{ij} 表示 i 类第 j 张图像的预测值。

按上述方法在图 5 和图 6 中分别展示出 baseline 与本文网络在 CIFAR-100 上的 P_S -Category 曲线(横坐标表示类别，纵坐标表示对应类别的 P_S 值)以及本文网络分别使用 *GAP* 和 AFVFM 作为分类器的 P_S -Category 曲线。从图 5 中可以看出，本文提出的网络可以提高网络对 CIFAR-100 数据集中大部分类别的预测精度，经统计可以提高 66%类别的预测精度，且平均精度提高了 3.4%。从图 6 中可以看出将本文提出的 AFVFM 作为分类器可以提高网络对CIFAR-100数据集中许多类别的预测精度，经统计有 50%的类别精度得到提升，其平均提升精度为2.2%。同时，从图5和图6中均可以看出，本文

表 2　模型在数据集(CIFAR-100)上的消融实验结果

Tab. 2　Experimental results of model ablation on data set (CIFAR-100)

Method	# Param/M	FLOPs/G	Top-1 *Acc* /%
Baseline	19.09	2.50	83.07
Baseline + Improved patch embedding	19.32	2.58	83.75
Baseline + Improved patch embedding + Feature compensation	19.83	2.63	84.65
Baseline + Improved patch embedding + Feature compensation + Adaptive feature vector fusion	20.06	2.81	85.86

图 4　特征提取器改进前后提取的特征可视化结果对比：(a1)(a2) 原图；(b1)(b2) 改进前提取的特征可视化结果；(c1)(c2) 改进后提取的特征可视化结果

Fig. 4　Comparison of feature visualization results extracted before and after feature extractor improvement: (a1)(a2) Original image; (b1) (b2) Feature visualization results extracted before improvement; (c1)(c2) Improved visualization of extracted features

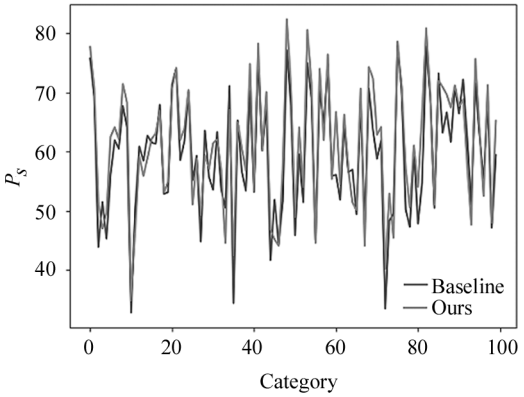

图 5　Baseline 与本文网络在 CIFAR-100 上的 P_S-Category 曲线

Fig. 5　P_S-Category curve of baseline and this network on CIFAR-100

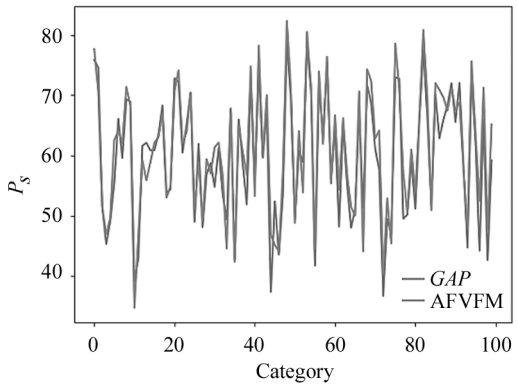

图 6　本文网络分别使用两种分类器(*GAP* 和 AFVFM)在 CIFAR-100 上的 P_S-Category 曲线

Fig. 6　P_S-Category curve of the two classifiers (*GAP* and AFVFM) on CIFAR-100 respectively using this network

网络及 AFVFM 不仅对 CIFAR-100 数据集中容易识别的类别有提升效果，而且对一些网络难以识别的类别同样有提升效果，这进一步展示了本文算法及 AFVFM 的有效性。

3　结　论

针对目前基于 transformer 的图像分类模型直接应用在小数据集上性能较差的问题，本文在 transformer 图像分类算法 Twins 的基础上对其特征提取器及分类器进行改进，提出了 transformer 自适应特征向量融合算法。在 3 种不同数据量的数据集(Mini-ImageNet-100、CIFAR-100 和 Image-Net-1k)上对该算法进行训练和测试，并与 baseline 及其他算法进行对比，结果表明，本文算法在 3 种数据集上的 TOP-1 准确率均高于其他算法，并且随着数据量的减少，本文与其他算法的差距越明显。同时，本文也做了大量的可视化实验，直观地展示了本文算法及各模块的有效性和优越性。

在未来的研究中，仍要进一步提高 transformer 模型在小数据集上的分类性能，并采用更少的计算量和参数量，降低 transformer 在图像分类应用中的成本，使 transformer 模型可以应用到更多的场景中。

参考文献：

[1] ALTMAN N S. An introduction to kernel and nearest-neighbor nonparametric regression[J]. The American Statistician, 1992, 46(3): 175-185.

[2] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[3] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[4] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [C]//Advances in Neural Information Processing Systems, December 3-6, 2012, Lake Tahoe, Nevada, USA. Massachusetts: Neur IPS, 2012, 25: 1097-1105.

[5] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2014-09-04) [2022-04-27]. https://arxiv.org/abs/1409.1556.

[6] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//IEEE Conference on Computer Vision and Pattern Recognition, June 8-10, 2015, Boston, MA, USA. New York: IEEE, 2015: 1-9.

[7] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]//IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems, December 4-9, 2017, Long Beach, California, USA. Massachusetts: Neur IPS, 2017: 5998-6008.

[9] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16 × 16 words: Transformers for image recognition at scale [EB/OL]. (2020-10-22) [2022-04-27]. https://arxiv.org/abs/2010.11929.

[10] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention [C]//International Conference on Machine Learning (PMLR), July 18-24, 2021, Virtual. New York: PMLR, 2021: 10347-10357.

[11] LIU Z, LIN Y, CAO Y, et al. Swin transformer: hierar-chical vision transformer using shifted windows [EB/OL]. (2021-03-25) [2022-04-27]. https://arxiv.org/abs/2103.14030.

[12] CHU X, TIAN Z, WANG Y, et al. Twins: revisiting the design of spatial attention in vision transformers[EB/OL]. (2021-04-28) [2022-04-27]. https://arxiv.org/abs/2104.13840.

[13] XIAO J S, RAO T Y, JIA Q, et al. A Laplacian pyramid Image fusion algorithm based on graph cutting [J]. Journal of Optoelectronics · Laser, 2014, 25(7): 1416-1424.
肖进胜，饶天宇，贾茜，等. 基于图切割的拉普拉斯金字塔图像融合算法[J]. 光电子 · 激光，2014，25(7)：1416-1424.

[14] WU X Y, WEN X B, XU H X, et al. SAR image classification based on multi-feature nonlocal dynamic kernel sparse representation based on tensor projection [J]. Journal of Optoelectronics · Laser, 2021, 32(7): 742-752.
吴效莹，温显斌，徐海霞，等. 基于张量投影的多特征非局部动态核稀疏表示的 SAR 图像分类[J]. 光电子 · 激光，2021，32(7)：742-752.

[15] YANG H, WU X T, HE B G, et al. Image fusion method based on multi-scale guided filtering[J]. Journal of Optoelectronics · Laser, 2015, 26(1): 170-176.
杨航，吴笑天，贺柏根，等. 基于多尺度导引滤波的图像融合方法[J]. 光电子 · 激光，2015，26(1)：170-176.

[16] XIE S, GIRSHICK R, DOLLáR P, et al. Aggregated residual transformations for deep neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 1492-1500.

[17] HE T, ZHANG Z, ZHANG H, et al. Bag of tricks for image classification with convolutional neural networks [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, July 16-21, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 558-567.

[18] RADOSAVOVIC I, KOSARAJU R P, GIRSHICK R, et al. Designing network design spaces[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 14-19, 2020, Seattle, WA, USA. New York: IEEE, 2020: 10428-10436.

[19] D'ASCOLI S, TOUVRON H, LEAVITT M, et al. Convit: improving vision transformers with soft convolutional inductive biases [EB/OL]. (2021-03-19) [2022-04-27]. https://arxiv.org/abs/2103.10697.

[20] SELVARAJU R, COGSWELL M, DAS A, et al. Grad-cam: visual explanations from deep networks via gradient based localization[C]//IEEE International Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 618-626.

作者简介：

李　凡　(1986－)，男，工学博士，副教授，硕士生导师，主要从事计算机视觉、图像处理等方面的研究.