DOI:10.19651/j.cnki.emt.2416963

基于多层注意力和度量学习的商品识别方法*

李 婕[1] 张新月[1] 涂静敏[1] 陈记文[1] 李 礼[2,3]
(1.湖北工业大学太阳能高效利用及储能运行控制湖北省重点实验室 武汉 430068;
2.武汉大学遥感信息工程学院 武汉 430079; 3.武汉大学深圳研究院 深圳 518057)

摘 要: 针对自动售货柜场景中存在的复杂背景和商品包装高度相似导致的识别难题,提出了一种融合多尺度注意力机制和度量学习的商品识别方法。首先,基于 ResNet 层级结构引入多头自注意力,充分挖掘卷积神经网络(CNN)多尺度特征提取优势和 Transformer 全局建模能力,并设计一种新的多尺度空洞注意力,使模型关注到相似包装中商标形状和局部纹理等局部特征,以及上下文全局特征;其次设计降采样多尺度特征融合策略,有效提高算法的多尺度特征表达能力;最后采用 ArcFace 损失函数以增强模型的识别能力。为了验证所提出方法的有效性,构建了一个真实场景下的商品数据集,由自动售货柜的顶视摄像头采集。实验结果表明,该方法在 Commodity 553 数据集上的 MAP@1 准确率达到 87.4%,优于当前的主流识别方法,可实现更精确的商品识别。

关键词: 商品识别;深度学习;注意力机制;度量学习
中图分类号: TN911.73; TN919.81; TP391.41 **文献标识码**: A **国家标准学科分类代码**: 510.4050

Commodity recognition method combining multi-layer attention mechanism and metric learning

Li Jie[1] Zhang Xinyue[1] Tu Jingmin[1] Chen Jiwen[1] Li Li[2,3]
(1. Hubei Key Laboratory of Solar Energy Efficient Utilization and Energy Storage Operation Control, Hubei University of Technology, Wuhan 430068, China; 2. Institute of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China; 3. Shenzhen Institute, Wuhan University, Shenzhen 518057, China)

Abstract: Aiming at the recognition problem caused by the complex background and the high similarity of commodity packaging in the vending machine scene, a commodity recognition method combining multi-scale attention mechanism and metric learning is proposed. Firstly, based on the ResNet hierarchical structure, multi-head self-attention is introduced to fully exploit the advantages of multi-scale feature extraction of convolutional neural network (CNN) and the global modeling ability of Transformer, and a new multi-scale hollow attention is designed to make the model focus on local features such as trademark shape and local texture in similar packaging, as well as context global features. Secondly, a down-sampling multi-scale feature fusion strategy is designed to effectively improve the multi-scale feature expression ability of the algorithm. Finally, ArcFace loss function is used to enhance the recognition ability of the model. In order to verify the effectiveness of the proposed method, a commodity data set in a real scene is constructed, which is collected by the top-view camera of the vending cabinet. The experimental results show that the MAP @ 1 accuracy of this method on the Commodity 553 dataset reaches 87.4%, which is better than the current mainstream recognition methods and can achieve more accurate commodity recognition.

Keywords: commodity recognition; deep learning; attention mechanism; metric learning

0 引 言

自动售货柜采用人工智能的高效结算,极大的节省了人力成本,成为现代无人零售行业的新模式。尽管基于人工智能和机器学习在图像识别方面取得了巨大进步,但如何在商品多样、环境复杂等零售环境下提高商品识别精度

收稿日期:2024-09-23
* 基金项目:国家自然科学基金(42101440,42301515)、智能光电系统感知及应用四川省高校重点实验室开放基金(ZNGD2308)、湖北工业大学博士科研启动基金(XJ2021004501)、深圳市科技计划资助项目(JCYJ20230807090206013)、深圳市科技重大专项(KJZD20230923114611023)资助

和效率，是自动售货行业所面临的重要挑战。

早期的商品图像识别主要基于手工设计特征，如经典的 SIFT(scale-invariant feature transform，SIFT)和 SURF(speeded up robust features，SURF)特征提取算法结合支持向量机等机器学习方法进行商品识别。然而，实际应用场景中，自动售货柜摄像头在动态获取商品图片时，存在图像模糊、变形以及部分遮挡等问题，使得传统的识别方法存在较大的局限性。

随着深度学习的兴起，卷积神经网络(convolutional neural networks，CNN)通过在大规模数据集的学习训练获得鲁棒的图像特征，提高了图像识别的准确率[1]。Wei 等[2]引用多角度生成对抗网络，丰富了不同角度的商品训练样本，提高了商品识别精度。然而同一商品在不同角度呈现完全不同的特征，且商品图像种类繁多导致外观相似度高。如何提取并表达商品不同角度特征以及区分不同商品间的细粒度差异，是提高商品识别率的有效途径。

ViT(vision transformer)[3]因具有捕捉复杂特征和全局上下文信息的优势，成为图像识别等任务的研究热点[4-6]。例如：Liu 等[7]提出基于 ViT 的商品识别模型，在公开的零售商品数据集中取得了较高识别精度；Jing 等[8]设计一种渐进式的 Token 生成模块，提升识别精度的同时缓解了 Transformer 参数量大的问题。

上述基于 Transformer 方法虽然提高了商品识别精度，但这种将图像分为若干个固定大小的补丁，然后对其进行特征提取和分类的方法，会导致原始 ViT 模型对输入图像的尺寸比较敏感而且无法充分利用图像的全局信息，因此研究者们开始探索将 CNN 和 Transformer 结构进行融合，将 CNN 常见的多尺度金字塔、残差连接等结构引入 ViT 中，从而提高模型对于输入图像尺寸的适应性和特征提取能力[9]。在 ResNet 的 bottleneck 结构中引入多头自注意力机制(multi-head self-attention，MHSA)，从而替代传统的 3×3 卷积层。这种结构上的改变使得 BotNet 在实例分割和目标检测任务上取得了显著的性能提升，同时减少了参数数量，在 COCO 实例分割任务中的准确度提升了 1.2%。CSWin Transformer[10]模型采用了多层 Transformer 结构，并在每个模块之间引入了类似于残差连接结构，其通过参照并学习 ResNet 的架构实现了模型效果的优化，提高了模型的表达能力和学习效率。Mehta 等[11]在 MobileViT 中采用了类似的设计策略。他们在该模型中参照 ViT 的结构设计了 MobileViT 块，该模块用更擅长处理全局信息的 Transformer 块替代了常规卷积中的矩阵乘法运算过程。然后将该模块与倒置残差块进行交替串联堆叠，实现对 CNN 和 Transformer 的优势的融合。MobileViT 块的主要思想是将 Transformer 视为卷积，这使得它可以同时获得卷积的归纳偏置特性和 Transformer 的全局性。与 ViT 相比，MobileViT 块既不会失去图像块的顺序信息，也不会丢失图像块中像素的空间位置信息。

上述这些模型通过参照 CNN 的架构设计，对原始 ViT 的不足进行改进。基于架构设计参考所提出的 CNN-Transformer 混合模型能更高效地提取特征，在不同的任务和数据集上具备了更好的可拓展性，使得模型对光照、遮挡等复杂场景具备更好的鲁棒性。但在商品识别领域依旧存在相似商品细粒度差异特征辨别度不够的问题。因此在基于架构参考设计混合模型时，仍需要在商品识别实际研究中对相应细节进行改进。

此外，如 Triplet Loss[12](三元组损失)、Center Loss[13](中心损失)和 ArcFace Loss[14]等基于度量学习的方法关注样本之间的相似性和距离度量，在细粒度识别领域体现出显著优势。相比于传统图像识别，自动售货柜中商品识别所面临的主要挑战在商品类内差异大、类间差异小，即同一个商品由于不同的采集环境和角度导致特征变化大，而不同商品类别之间又存在外观、形状、颜色等特征相似的问题。因此，本文设计一种 CNN 和 Transformer 混合框架下的自动售货柜商品识别方法，通过构建同类商品在不同角度、背景和光照等情况下的样本数据，使网络能够学习鲁棒的商品图像特征，有效地表征样本特征；同时构建多层注意力模块，引导网络关注商标、文字等细粒度差异的同时，也关注上下文全局特征；最后将特征映射到度量空间，采用 ArcFace 损失函数以提高特征辨别力，以实现自动售货柜商品的高精度识别。

1　方　　法

1.1　商品识别模型整体网络结构

本文提出的识别网络结构如图 1 所示，主要由 4 个部分组成，分别是特征提取网络、多层自注意力模块、降采样多尺度特征融合模块以及度量学习模块。相较于 ResNet 这种常见的 CNN 多尺度金字塔结构，以及 ViT 这种利用 Transformer 架构捕捉全局信息的架构，本文将 ResNet 结构中四层多尺度特征图输入至多层自注意力模块，从而有效地增强模型对深层全局信息的建模能力以及对局部细节的表征能力，显著改善了 ResNet 结构对于局部细节特征提取不足的问题，以及 ViT 结构对于全局信息表征不充分的问题。不仅如此，针对 ViT 中的多头注意力在提取浅层信息时有信息丢失的问题，本文进一步提出了多尺度空洞注意力，通过适当增加感受野的手段，改善多头注意力在浅层的信息丢失情况。

整体网络模型以自动货柜采集到的商品图像作为输入，采用 ResNet-50 作为主干网络提取商品特征，输入 112×112 三通道的 RGB 商品图像，输出为 $C1$、$C2$、$C3$ 和 $C4$ 四个不同尺度的特征图。其中 Stage1 和 Stage2 提取出的 $C1$ 特征图和 $C2$ 特征图为浅层特征，Stage3 和 Stage4 提取出的 $C3$ 和 $C4$ 特征图表示深层特征。

该模块由多头自注意力和多尺度空洞注意力组成，深层特征经过多头自注意力以关注大尺度特征，多尺度空洞

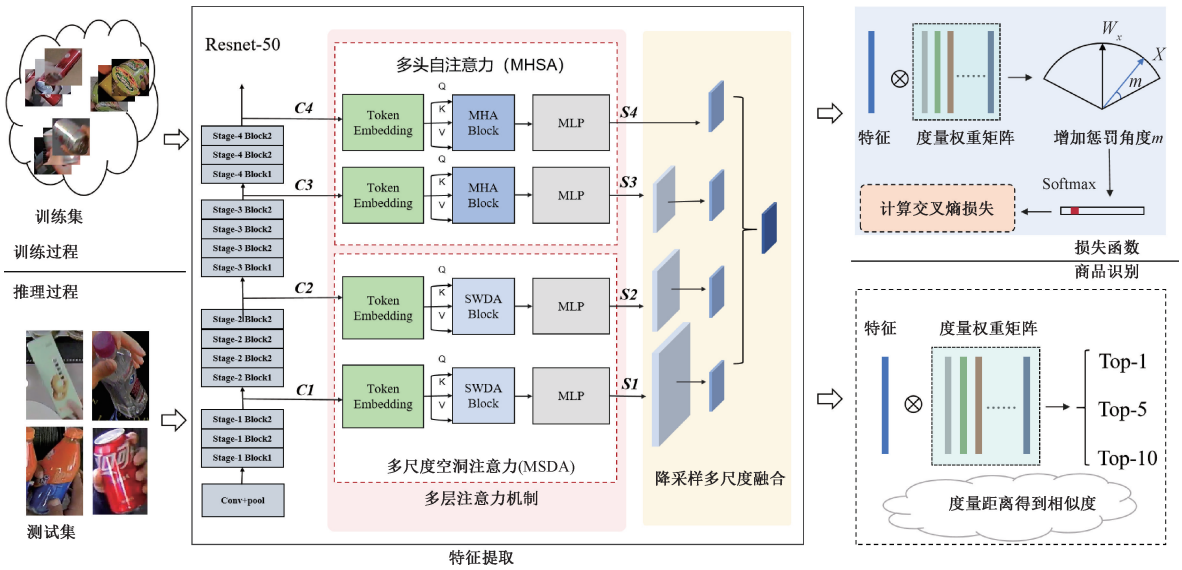

图 1　整体网络结构

Fig. 1　Overall network structure

注意力结合浅层特征捕捉局部细节特征。这种多层注意力构建方法能有效增强模型对深层全局信息的建模能力以及对局部细节的表征能力。注意力模块的输出特征图 $S1$、$S2$、$S3$ 和 $S4$,经过多尺度特征融合策略进行特征融合,以提高算法对商品在不同角度下的尺度变化适应能力。最后,不同于以往的分类模型所使用的交叉熵损失函数,本文将特征映射到度量空间,采用 ArcFace 损失函数增强网络判别能力,达到减小类内距离,增大类间距离的目的。

推理过程由特征提取和度量空间权重矩阵组成,通过训练好的网络模型进行特征提取,将测试图像转换成特征向量。再与训练过程中保存下来的度量权重矩阵计算度量距离,得到类别概率。

1.2　多层自注意力模块

在商品识别任务中,相似包装商品虽然在全局特征上呈现出相似性,但在局部细节上会存在差异,因此,强调模型在局部细节的提取能力变得至关重要。本文设计的多层自注意力结构,即在 ResNet 网络浅层采用多尺度空洞注意力(multi-scale dilated attention, MSDA),通过在低层执行动态的扩张操作以充分地捕获不同尺度细粒度局部特征。而在 ResNet 深层引入多头自注意力机制,使模型关注全局信息,增强模型的特征表达能力。

1)MSDA Block 多尺度空洞注意力机制

浅层网络通常局限于较小的感受野,导致其难以捕捉到远距离特征之间的依赖和交互,容易造成细节丢失,为此,本文在浅层中搭建一种能充分捕捉浅层信息的注意力机制,即 MSDA。MSDA 通过多尺度处理来扩大感受野,有助于捕捉到不同层次的语义信息,减小细粒度信息丢失。具体结构如图 2 所示,首先将浅层特征层分成多个特征块,加上类别标签和位置编码信息使模型学习全局上下文信息,再将整合后的特征块送入滑窗空洞注意力模块[15](sliding window dilated attention, SWDA)进行学习,采用残差结构和 Dropout 层来消除网络堆叠带来的梯度消失、爆炸和网络退化等问题,最后将输出送入 MLP 得到浅层特征。

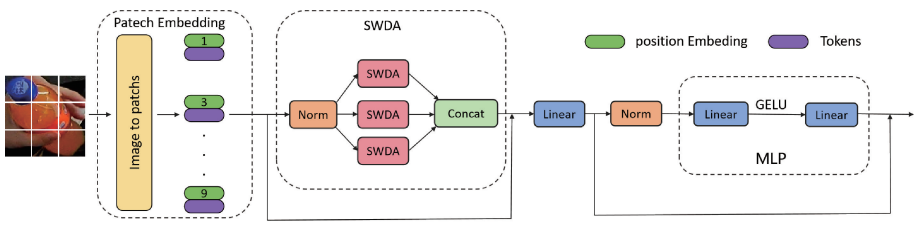

图 2　MSDA Block 网络结构图

Fig. 2　MSDA Block network structure diagram

如图 3 所示,SWDA 在不降低分辨率的情况下增加感受野,并通过设置不同的空洞率 r 来提取不同尺度下的特征图信息,从而在降低计算复杂度的同时有效扩大感受野,使模型在更大范围内聚合信息,有效改善了浅层信息丢失的情况。

2)MHSA Block 多头自注意力机制

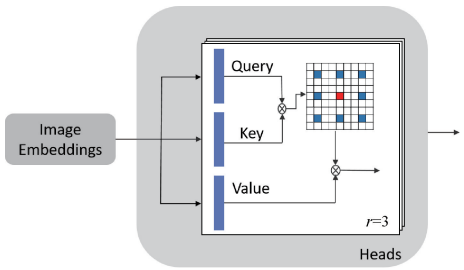

图 3　滑窗空洞注意力

Fig. 3　Sliding window empty attention

VIT 模型采用多头自注意力机制 MHSA 捕捉图像特征之间的依赖关系，并利用上下文信息进行全局理解，提高了模型对全局信息的处理能力。MHSA 网络结构图与 MSDA 结构同理，将深层特征层分成多个特征块，加上类别标签和位置编码信息使模型学习全局上下文信息，区别在于，将 SWDA 替换为多头注意力(multi-head attention，MHA)，将整合后的特征块送入 MHA 进行全局特征的学习，采用残差结构和 Dropout 层来消除网络堆叠带来的梯度消失、爆炸和网络退化等问题，最后将输出送入 MLP 得到深层特征。

1.3　多尺度特征融合策略

为了提高网络对商品尺度的适应能力，本文采用多尺度特征融合策略，即将多层注意力模块输出的 4 个尺度特征进行降采样，再进行相加融合。具体操作为将多层注意力模块输出的 $S1$、$S2$ 和 $S3$ 特征层进行降采样，最后和 $S4$ 层特征进行融合，最大限度保留浅层局部特征的基础上，降低深层特征相关性，使模型有效地关注不同尺度特征。融合后的特征再进行批量 Dropout 层、全链接层和归一化处理，生成 512 维特征向量，并送入后续模块进行分类。

1.4　基于 ArcFace 损失函数的度量空间权重

ArcFace 将特征向量映射到单位超球面，并且在该超球面上最大化类间距离，同时最小化类内距离。这种利用角度信息进行相似性度量的方式，能够约束出紧凑且具有判别性的特征，在细粒度图像识别领域表现出色[16]。本文采用 ArcFace 损失函数，使得同一类别商品特征在角度特征空间中的距离更加接近，而增加不同类别的商品特征的距离，提高了相似特征的可分性。

首先，假设有 N 种商品类别，为每个商品种类预设置一个度量权重 $\boldsymbol{W}\in\boldsymbol{R}^{N\times 512}$，利用反余弦(arc-cosine) 函数计算每一张训练图像经过主干网络提取出的特征向量 $\boldsymbol{X}_i\in\boldsymbol{R}^{512\times 1}$ 与度量权重 $\boldsymbol{W}$ 之间的角度 $\theta_{yi}=\arccos(\boldsymbol{W}^{\mathrm{T}}\cdot\boldsymbol{X}_i)$，在目标角度 θ_{yi} 中加入角度边距 m(additive angular margin)，增强特征向量的判别性；然后通过余弦(cosine) 函数获得目标得分 $\cos\theta_j$，以及超球体的半径 s 重新缩放所有目标得分 $\cos\theta_j$，最后采用交叉熵损失函数计算损失 Loss。具体可表示为：

$$L=-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\mathrm{e}^{s\cdot\cos(\theta_{yi}+m)}}{\mathrm{e}^{s\cdot\cos(\theta_{yi}+m)}+\sum_{j=1,j\neq y_i}^{n}\mathrm{e}^{s\cdot\cos\theta_j}} \tag{1}$$

在人脸识别领域，每种人脸类别取 M 张不同角度的图像，用特征提取网络将 M 个训练图像分别提取成一个 512 维特征向量以形成每一类商品的度量向量。大小表示为 $N\times M\times 512$，这种用多角度一维特征代替图像特征进行识别的方法，大幅度提高了人脸识别效率。

本文商品识别方法考虑到一个商品类别使用多个模板特征进行匹配识别时效率较低，如果将每个类别在不同角度的图像拟合成一个特征，此时模版大小将为 $N\times 1\times 512$。因此在 Arcface 损失函数中预设一个度量权重矩阵，并通过训练拟合度量权重矩阵，这样每一类商品的模版库仅有一个度量权重特征，可以极大的缩短商品识别的时间。

2　实验结果与分析

2.1　实验准备

1)数据准备与实验环境

本文采用自制数据集 Commodity 553 进行模型的训练和验证，该数据集包含了 553 种商品，每一类商品的图像信息是由自动售货柜顶端摄像头在不同光线、不同背景和不同角度下获取得到。每类商品包含 300 张图像，共 17 万张图片，包括面包，饮料，饼干等常见的零食商品。

图 4 展示了数据集中 6 个不同种类的商品图像，其中每一类商品在不同角度下呈现出不同的图像特征，如商品 1 和 2 都是泡面类，此类商品在不同角度呈现出的商品特征差异较大，容易混淆；商品 3 和 4 都是罐状饮料，常因为人手遮挡罐身的情况导致其难以识别；商品 5 和 6 的瓶身长度、外形、颜色等极其相似。另外，顾客从自动售货柜拿出商品的动态过程中，会存在图像模糊、光线太暗以及遮挡严重等问题；商品 7、8、9 是面包类零售商品，种类繁多，包装相似。Commodity 553 数据集均为真实场景下的自动售货柜商品，且包含商品识别中的常见干扰等，因此使用该数据集进行模型的训练具有较强的代表性。

本文实验配置为 Intel Core i9-10900，NVIDIA 3090 显卡。训练批次大小(Batch Size)设置为 64，训练进程次数 Epoch 为 300，选择 Adam 作为优化器，学习率为 0.01。训练前使用填充式数据处理将图片尺寸统一填充为(112×112)，并随机打乱读取数据的顺序，以增加训练的多样性。

2)算法评价指标

本文采用识别正确率 $MAP@n\ Accuracy$ 对模型进行评价。具体表示如下：

$$MAP@n\ Accuracy=\frac{TP_n}{TP_n+FP_n} \tag{2}$$

其中，TPn 为前 n 个识别结果中包含正确标签的样本数，$TPn+FPn$ 为总样本数。MAP@1 表示第一个商品是

图 4　自制数据集 Commodity 553

Fig. 4　Self-made dataset Commodity 553

正确标签的概率,同理,MAP@5 是指识别出的前 5 个商品中包含正确标签的概率,MAP@10 表示识别出的前 10 个商品中包含正确标签的概率。

自动售货柜中商品识别任务要求能准确识别出正确商品,而识别准确度 MAP@1 最满足商品识别任务的需求,因此本文采用第一识别目标准确率(MAP@1)作为商品识别模型的最佳评价指标。

2.2　不同主干网络识别精度对比

为了验证本文所采用主干网络结构的优越性,首先将主干网络更换为经典 CNN 网络 VGGNet[17]、MobileFaceNet[18]、基于 ResNet-18 改进的 IResNet-18[19],以及 CNN-Transformer 混合架构 Next-ViT[20] 和 BoTNet。其中,Next-ViT 和 BoTNet 为 CNN 和 Transformer 混合网络。

由表 1 的评价结果可以看出,相较于 VGGNet 和 Mobilefacenet,本文所使用的模型在 MAP@1 上分别提高了 1.94% 和 3.13%,并高于 IResNet-18 网络 3.9%。此外,本文的在 MAP@1 识别率相较于近年来 CNN-Transformer 混合架构 BoTNet 和 Next-ViT 也分别提高了 2% 和 1%,表明本文所使用的 CNN 和注意力融合机制的优越性。另外,本文网络 FLOPs 和参数量 Params 相较于 BoTNet 和 Next-ViT,分别降低了 2.38 G 和 11.861 M 以及 0.579 G 和 36.336 M。

表 1　不同 backbone 下的识别精度对比

Table 1　Comparison of recognition accuracy under different backbones

主干网络	MAP@1/%	MAP@5/%	MAP@10/%	计算量/G	参数量/M
VGGNet	85.75	94.25	95.41	3.860	40.930
Mobilefacenet	84.28	93.82	94.83	**0.237**	**1.201**
IResNet-18	83.43	94.43	95.56	2.634	24.025
BoTNet	85.47	94.52	95.42	3.927	19.844
Next-ViT	86.42	**94.83**	95.62	2.126	44.319
本文	**87.41**	94.57	**95.75**	1.547	7.983

2.3　消融实验

为了进一步验证本文所提出网络结构的有效性,对本文网络进行消融实验。网络包含 3 个模块,即 ResNet 基层、浅层特征注意力模块、深层特征注意力模块。在自制数据集上以 ResNet 作为基准结构,每轮消融学习屏蔽一个或多个模块,保留其他部分不变。

根据表 2 的结果显示,在没有引入注意力模块的情况下,MAP@1 可达 80.71%。随着深层注意力的引入,相较于未引入注意力模块,其精度略有改善,MAP@1 提高了 1.26%,MAP@5 和 MAP@10 分别提高了 1.9% 和 2.69%。而浅层注意力的引入则带来了显著的精度提升,相较于原始网络,MAP@1、MAP@5 和 MAP@10 分别提高了 2.83%、2.46% 和 3.11%。单独引入浅层注意力与单独引入深层注意力相比,MAP@1 提高了 1.57%,表明浅

层注意力的引入增加了局部细节特征的提取优势。而将浅层注意力和深层注意力结合，较于初始网络，MAP@1、MAP@5 和 MAP@10 提高了 6.7%、5.49% 和 5.59%。上述消融实验数据表明，本文网络中的各模块在提升识别性能方面均起到积极作用，其相互协同作用可以进一步提升网络精度。

表 2　消融学习

Table 2　Ablation learning

浅层注意力	深层注意力	MAP@1/%	MAP@5/%	MAP@10/%
×	×	80.71	89.08	90.03
×	√	81.97	90.98	92.72
√	×	83.54	91.54	93.14
√	√	87.41	94.57	95.62

2.4　方法对比

为了验证本文识别方法的有效性，将本文的识别方法与几种当前主流的识别方法进行对比分析，如表 3 所示。首先，为了对比本文方法与主流 CNN 特征提取方法的性能，本文选取了 Res-Arc[21] 作对比试验，Res-Arc 采用 ResNet-18 作为主干网络，与 ArcFace 损失函数结合进行识别，该方法使用经典 CNN 网络进行特征提取，难以表征局部特征差异；其次，为了验证本文模型性能优于主流 Transformer 模型，选取 ViT-Arc[22] 作对比试验，该方法将 ViT 和 ArcFace 损失函数相结合完成分类识别任务，ViT 通过多头自注意力机制，使模型关注到了局部特征，但对输入图像的尺寸比较敏感而且无法充分利用图像的全局信息；本文的网络结构结合上述两种方法的优点，通过融合 CNN 的金字塔结构和 ViT 的多头自注意力机制，使模型既能充分的提取图像的全局信息，又能关注到局部特征。此外，Res-SupCon[23] 通过自监督对比学习，自监督对比学习与本文算法中的度量学习都是通过学习样本之间的距离或相似度，提高了分类识别的精度，因此选取该方法与本文的算法进行对比；ResNext-CE[24] 与本文的模型都是基于 ResNet 设计了一种新的模型，因此选取该方法作对比试验，其中 ResNeXt-CE 在网络中使用多个并行的卷积分支来捕获不同特征子空间，使得网络能够更好地学习和表示复杂的特征，从而提高了分类性能，但没有结合 Transformer 的优点，难以表征局部特征；ResNeSt-CE[25] 利用 ResNet 网络与注意力模块的特征交互提高了识别准确性，在识别领域表现优良，与本文的算法相似，设计了 CNN-Transformer 混合模型以提高网络的特征提取能力，但 ResNeSt-CE 区分局部细粒度特征的能力依然不充分，而本文的模型在金字塔浅层中使用进一步改善后的多尺度空洞注意力，进一步提高了模型关注局部细粒度特征的能力。

表 3　不同识别方法对比

Table 3　Comparison of different identification methods

方法	MAP@1/%	MAP@5/%	MAP@10/%	FLOPs/G	Params/M
Res-Arc[21]	81.81	94.32	94.92	1.728	11.433
ViT-Arc[22]	78.34	93.56	95.14	1.884	50.051
Res-SupCon[23]	85.35	94.26	95.32	6.834	11.497
ResNeXt-CE[24]	83.67	94.34	95.78	4.287	24.113
ResNeSt-CE[25]	86.23	94.53	95.43	**1.405**	26.484
本文	**87.41**	94.57	95.62	1.547	**7.983**

从表 3 可以看出，ResNeSt-CE 由于在 ResNet-50 加入了注意力机制，在本文构建的商品数据集上的表现优于 Res-Arc 这类 CNN 算法，MAP@1 提高了 4.42%，表明注意力机制的加入增强了网络对商品特征的提取。此外，ResNeSt-CE 相较于 ViT-Arc 使用纯 Transformer 算法的方式，其 MAP@1 提高了 7.89%。ViT-Arc 作为 ViT 模型的算法，识别率较低，这是由于 ViT 是基于纯 Transformer 的网络模型，需要大规模数据和训练时间以获得最优模型，说明了 CNN 和 Transformer 融合机制相较于纯 Transformer 可以在有限数据下获得更优的商品识别模型。而本文的识别方法 MAP@1 相较于 Res-Arc、ViT-Arc、Res-SupCon 和 ResNeXt-CE 网络分别提高了 5.6%、9.07%、2.06% 和 3.74%，体现了本文的网络模型优于 CNN 网络模型和纯 Transformer 模型。通过对比本文方法和 ResNeSt-CE，可以看出本文的 MAP@1 相较于 ResNeSt-CE 提高了 1.18%，优于现有的 CNN 和 Transformer 融合框架。

图 5 是几种实验方法的可视化结果展示。从识别结果中随机抽取几个商品样本，展示出不同方法的前 5 个识别结果，并用红色矩形框标记了识别正确的商品。如图 5 上半部分所示，可以看出浅色罐装饮料受到模糊和遮挡的影响较大，此时 Res-Arc、ViT-Arc、Res-SupCon 和 ResNeXt-CE 这 4 种方法无法正确识别。而 ResNeSt-CE 因为结合了注意力机制，更加关注关键特征，可以在 Top-5 中识别，但 Top-1 识别不准确，可以看出该网络在 Top-1 识别出的饮料罐与待识别图片相似度高。而采用本文提

出的方法可以正确的在 Top-1 中识别出,说明本文提出的识别网络具备细粒度区区分能力。如图 5 下半部分所示,可以看出同色系瓶装饮料难以区分,ResNeXt-CE 和 ResNeSt-CE 在 Top-1 识别上无法准确识别,而通过本文提出的方法在 Top-1 识别中,在图 5 下半部分中有遮挡、模糊和相似产品中依然能正确识别。上述实验结果表明,本文算法在区分相似产品上的优势。

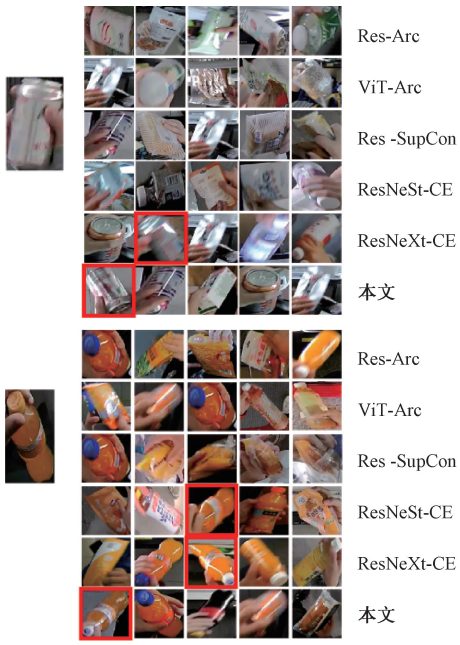

图 5　识别结果可视化展示

Fig. 5　Visual display of recognition results

2.5　度量权重矩阵效率对比

为了提高商品识别的效率,本文提出了一种度量权重商品识别方法。如表 4 所示,传统的方法是从训练数据集中每类选取 M 张图片(本实验中 m 取 10),在经过特征提取网络提取的一维特征后,N 类待识别商品构建成度量矩阵长度为 $N\times M\times 512$。而本文所提出的识别方法,是将事先拟合好的度量权重矩阵 W 作为度量矩阵用于识别,长度为 $N\times 1\times 512$,相比于传统方法可以使得识别效率提高 M 倍。表 4 共选取 3 270 张商品图像作为测试集进行识别效率测试。

表 4　识别效率对比

Table 4　Comparison of recognition efficiency

方法	MAP@1/ Time	MAP@5/ Time	MAP@10/ Time
Traditional Approaches	78.0%/ 0.088 6 s	85.1%/ 0.089 6 s	86.7%/ 0.090 4 s
本文	78.8%/ 0.007 9 s	87.5%/ 0.008 7 s	89.7%/ 0.009 2 s

通过对照表 4 的数据对比发现,传统方法平均每张图像推理时间 0.089 5 s,而本文所提出的方法平均每张图像推理时间 0.008 6 s,在时间效率上提高了 10 倍。证明本文所提出的方法在商品识别中的有效性。

3　结　　论

本文提出了一种 CNN 和 Transformer 融合框架下基于多层注意力和度量学习的商品识别方法,该方法以 ResNet 作为特征提取的骨干网络,在浅层设计了一种多尺度扩张注意力模块,有效的提取局部纹理和线条等浅层语义信息,从而更好的识别相似包装的产品;在深层采用多头自注意力模块,增强了模型对全局信息的处理能力,进一步提高了识别精度。考虑到传统模版识别方案效率低的问题,本文设计了一种新的商品识别模版设计方式,显著地提高了商品识别的效率。实验结果表明,本文方法在 Commodity 553 数据集上 MAP@1 可达 87.41%,模型 Paras 参数量为 7.983 M,FLOPs 计算量为 1.547 G,具有重要的现实意义。本文后续将探索更多特征提取和度量学习方法,以进一步提高视觉识别准确性,并加速其在实际应用中的落地。

参考文献

[1]　李奇,常立娜,武岩,等. 基于深层图卷积的 EEG 情绪识别方法研究[J]. 电子测量技术,2024,47(4):18-22.
LI Q, CHANG L N, WU Y, et al. Research on EEG emotion recognition method based on deep graph convolution[J]. Electronic Measurement Technology, 2024,47(4): 18-22.

[2]　WEI Y CH, XU SH X, KANG B, et al. Generating training images with different angles by GAN for improving grocery product image recognition [J]. Neurocomputing, 2022, 488: 694-705.

[3]　DOSOVITSKIY A, BEYER L, KOLESNIKOY A, et al. An image is worth 16×16 words: Transformers for image recognition at scale [J]. ArXiv preprint arXiv: 2010.11929, 2020.

[4]　李冰锋,刘帅,杨艺. 基于改进的 Transformer 细粒度图像识别算法研究[J]. 电子测量技术,2024,47(2):114-120.
LI B F, LIU SH, YANG Y. Research on fine-grained image recognition algorithm based on improved Transformer[J]. Electronic Measurement Technology, 2024,47(2): 114-120.

[5]　朱学岩,陈锋军,郑一力,等. 融合双线性网络和注意力机制的油橄榄品种识别[J]. 农业工程学报,2023,39(10):183-192.
ZHU X Y, CHEN F J, ZHENG Y L, et al. Olive variety identification based on bilinear network and attention mechanism [J]. Acta Agricultural Engineering Science,2023,39(10):183-192.

[6] 陈莹,匡澄. 基于 CNN 和 TransFormer 多尺度学习行人重识别方法[J]. 电子与信息学报, 2023, 45(6): 2256-2263.
CHEN Y, KUANG CH. Multi-scale learning person re-identification method based on CNN and TransFormer [J]. Journal of Electronics and Informatics, 2023,45(6): 2256-2263.

[7] LIU SH, WANG X Y, ZHU CH ZH, et al. GRVT: Toward effective grocery recognition via vision transformer [C]. Computer Graphics International Conference, Cham: Springer Nature Switzerland, 2023: 266-277.

[8] JING H R, LIN G J, ZHANG H J, et al. A face recognition algorithm based on improved resnet[J]. Frontiers in Computing and Intelligent Systems, 2022, 1(1): 22-25.

[9] SRINIVAS A, LIN T Y, PARMAR N, et al. Bottleneck transformers for visual recognition [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 16519-16529.

[10] DONG X Y, BAO J M, CHEN D D, et al. CSWin transformer: A general vision transformer backbone with cross-shaped windows [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 12124-12134.

[11] MEHTA S, RASTEFARI M. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer [J]. ArXiv preprint arXiv: 2110.02178, 2021.

[12] 张红颖, 田鹏华. 结合残差网络与多级分块结构的步态识别方法[J]. 电子测量与仪器学报, 2022, 36(6): 66-72.
ZHANG H Y, TIAN P H. Gait recognition method combining residual network and multi-level block structure[J]. Journal of Electronic Measurement and Instruments, 2022,36(6):66-72.

[13] 杜闯, 何赟泽, 邓海平, 等. 基于百度飞桨的面向黑暗环境人员行为检测与身份识别[J]. 电子测量与仪器学报, 2023, 37(8): 21-29.
DU CH, HE Y Z, DENG H P, et al. Human behavior detection and identification in dark environment based on baidu flying oar[J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(8):21-29.

[14] DENG J K, GUO J, YANG J, et al. ArcFace: Additive angular margin loss for deep face recognition[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 4690-4699.

[15] JIAO J Y, TANG Y M, LIN K Y, et al. Dilateformer: Multi-scale dilated transformer for visual recognition [J]. IEEE Transactions on Multimedia, 2023, 25: 8906-8919.

[16] XU B B, WANG W SH, GUO L F, et al. CattleFaceNet: A cattle face identification approach based on retinaface and arcface loss[J]. Computers and Electronics in Agriculture, 2022, 193: 106675.

[17] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. ArXiv preprint arXiv:1409.1556, 2014.

[18] CHEN SH, LIU Y, GAO X, et al. MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices[C]. Bio-metric Recognition: 13th Chinese Conference, CCBR 2018, Springer International Publishing, 2018: 428-438.

[19] HUA H, LIU M P, LI Y Q, et al. An ensemble framework for short-term load forecasting based on parallel CNN and GRU with improved ResNet[J]. Electric Power Systems Research, 2023, 216: 109057.

[20] LI J SH, XIA X, LI W, et al. Next-ViT: Next generation vision transformer for efficient deployment in realistic industrial scenarios[J]. ArXiv preprint arXiv: 2207.05501,2022.

[21] JING H R, LIN G J, ZHANG H J, et al. A face recognition algorithm based on improved resnet[J]. Frontiers in Computing and Intelligent Systems, 2022, 1(1): 22-25.

[22] MANESCO J R R, MARANA A N. Combining arcface and visual transformer mechanisms for biometric periocular recognition [J]. IEEE Latin America Transactions, 2023, 21(7): 814-820.

[23] KHOSLA P, TETERWAK P, WANG C, et al. Supervised contrastive learning [J]. Advances in Neural Information Processing Systems, 2020, 33: 18661-18673.

[24] XIE S N, GIRSHICK R, DOLLAR P, et al. Aggregated residual transformations for deep neural networks[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1492-1500.

[25] ZHANG H, WU CH R, ZHANG ZH Y, et al. Resnest: Split-attention networks [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 2736-2746.

作者简介

李婕,副教授,博士,主要研究方向为计算机视觉。
E-mail:jielonline@hbut.edu.cn

张新月,硕士,主要研究方向为计算机视觉、图像识别。
E-mail:xinyue000711@163.com

涂静敏,讲师,博士,主要研究方向为三维点云处理。
E-mail:jingmin.tu@hbut.edu.cn

陈记文,硕士,主要研究方向为计算机视觉。
E-mail:1595872508@qq.com

李礼(通信作者),副研究员,博士,主要研究方向为计算机视觉。
E-mail:jielonline@hbut.edu.cn