一种面向商品检索的多尺度度量学习方法

行阳阳[1],张索非[1],宋 越[2],吴晓富[1],周 全[1]
(1. 南京邮电大学,江苏 南京 210003;
2. 95958 部队,上海 200120)

摘 要:商品图像检索是一个典型的大规模度量学习任务,其特点在于商品零售平台需要定期上架新类型的商品,且同一类型的商品外观会不时发生变化。已有的工作表明:传统基于单一的度量学习虽然可以将商品检索模型的识别范围扩展到未知商品类别上,但是其性能仍然受限。为此,提出了一种基于多尺度监督信息的深度度量学习商品检索方法。该方法利用商品多个尺度的标签信息训练并使用协同注意力机制对不同尺度的深度特征进行有效融合,提高了深度学习模型挖掘重要信息的能力,从而有效提高了其在细粒度级别下的检索性能。在大规模商品检索数据集上的实验结果表明,该方法在 mAP 和 Rank-1 上分别为 43.0% 和 65.9%。相比于传统度量学习方法分别提升了 6.4% 和 7.8%。

关键词:度量学习;商品识别;多尺度;图像检索;特征融合
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2024)01-0065-06
doi:10.3969/j.issn.1673-629X.2024.01.010

A Multi-scale Metric Learning Approach for Product Retrieval

XING Yang-yang[1], ZHANG Suo-fei[1], SONG Yue[2], WU Xiao-fu[1], ZHOU Quan[1]
(1. Nanjing University of Posts and Telecommunications, Nanjing 210003, China;
2. 95958 Troop, Shanghai 200120, China)

Abstract: Product image retrieval is a typical large-scale metric learning task. The specificity of this task is that the commodity retail platform needs to import new types of items regularly, and the appearances of existing products also change from time to time. The previous work show that although the traditional metric learning can extend the recognition range of product retrieval to unseen product types, the performance of traditional metric learning in product retrieval is still limited, because it can only use a single scale of regulatory information to deal with such large-scale retrieval problems. Therefore, we propose a product retrieval method based on multi-scale deep metric learning. The proposed method uses the label information of multiple scales to train the model and adopts the co-attention module to integrate the deep features of different scales effectively, which can improve the ability of the model to obtain important information and effectively improve the retrieval performance of the deep learning model at the fine-grained level. The proposed method achieves 43.0% and 65.9% on mAP and Rank-1 in experiments on large-scale product retrieval dataset, improved by 6.4% and 7.8%, respectively compared with the traditional metric learning.

Key words: metric learning; product identification; multiple scale; image retrieval; feature fusion

0 引 言

随着互联网技术和电子商务的飞速发展,人们的购物方式逐渐从传统的线下购物转变为线上购物。为了充分满足客户海量、多样化的网上购物需求,人工智能零售系统需要快速、自动地从图像和视频中识别出存货单元(Stock Keeping Unit,SKU)级别的商品类别。与此同时,深度学习在计算机视觉领域取得了突破性进展,特别是在大型图像分类任务方向如 ImageNet。大规模商品识别是计算机视觉领域的一个新兴课题。但是,许多 SKU 级商品都是细粒度的,并且它们在视觉上是相似的。如何正确并快速地通过深度神经网络来进行快速识别仍面临技术挑战。

网上购物平台上的商品种类繁多。如图 1 所示,为了方便商品的管理,很多商品的子类别根据不同的用途或储存方法被划分在不同的父类别中。换句话说,一个父类别包含许多子类别。一种商品既属于某

收稿日期:2023-03-06 **修回日期**:2023-07-07
基金项目:国家自然科学基金面上项目(61876093)
作者简介:行阳阳(1999-),男,硕士研究生,研究方向为计算机视觉、图像处理;通信作者:吴晓富(1975-),男,博士,研究员,研究方向为编码与信息论、计算机视觉。

个子类别,同时也属于某个父类别。不同语义下的类别信息即商品图像多尺度标签信息。需要注意的是,商品的层级分类并不完全迎合商品的视觉相似性。不同父类别下的商品图像也可能具有相似的外观。例如,如图 1 所示,以“护手霜”为父类别的第三行第四列商品与以“洗面奶”为父类别的第四行第二列商品外观较为相似。造成这种现象的原因是商品并不是按照外观进行分类。这种现象会给商品图像的标签带来噪音并给检索出正确的商品类别带来技术挑战。

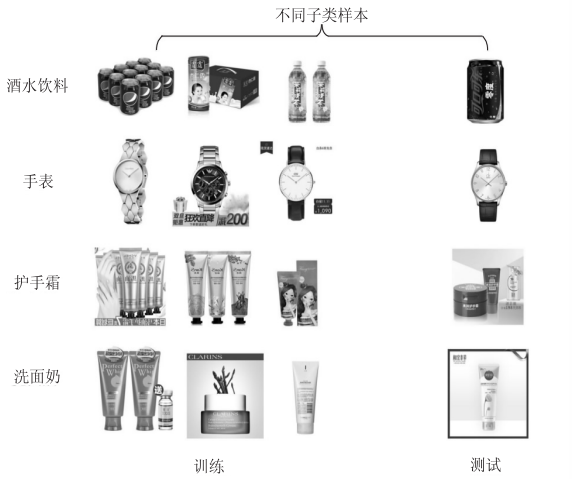

图 1 在线零售平台商品类别分类情况示意图

此外,在商品识别任务中,手动标记标签并收集所有类别商品图像的数据集总是费时且昂贵的。首先,要在网上购物平台上识别不同商品的数量可能是巨大的。对于每一类商品,需要数百张训练图像,通常从几个不同的角度拍摄。其次,商品零售平台需要定期上架新的商品类型,现有商品的外观也会不时发生变化。在实际部署过程中,更新不断增加的训练图像是一个棘手的问题。由于上述原因,如果没有这些新商品类别的训练样本,传统的图像分类模型往往无法获得令人满意的性能。相比之下,度量学习方法更适合于商品检索,因为它可以将输入图像嵌入到一个紧凑但有分辨能力的特征空间中。这种嵌入可以很容易地推广到未知类别,而不需要任何额外的训练成本。因此,将网络购物平台中的商品图像识别问题转化为大规模度量学习任务有利于问题的解决。目前,许多 SKU 级别的商品图像数据集已经进行了公开。例如,AliProducts-Challenge[1] 数据集包含近 300 万张图像,覆盖 5 万个 SKU 级商品类别;Products-10K[2] 数据集包含近 15 万张图像,覆盖 1 万个 SKU 级商品类别。

综上所述,商品检索问题可以看作是一个多尺度度量学习问题。商品类别的标签通常符合一种层级结构。许多公开可用的商品图像数据集还包含多尺度的标签信息,而不是只包含单尺度的标签信息。利用多尺度标签信息进行模型训练,可以使得网络充分挖掘商品图像不同尺度特征的关系并更可能满足不同尺度下的识别需求。

该文提出了一种充分利用商品图像的多尺度监督信息的 MSML(Multi-Scale Metric Learning)模型。在大规模商品图像检索数据集上的实验结果表明,该方法是有效的,与传统的单尺度度量学习相比,显著提高了识别的综合性能。

1 相关工作

1.1 商品识别

在过去的十年中,深度学习在计算机视觉领域取得了巨大的成功。近年来,基于深度学习的商品识别得到了广泛的研究,关于这一领域已经有了大量的工作。在文献[3]中,作者提出了一个多任务级联的卷积神经网络(MTCD-CNN)进行商品图像检测并采用分层频谱聚类进行层级的图像分类。文献[4]提出了一种对商品模型进行无标签半监督的商品识别方法。该方法基于 Self-training 训练两个目标检测模型,提高了无标签预测的准确性。文献[5]报道了通过 AlexNet 学习到的特征被用于杂货商品识别。该研究表明,深度学习方法在复杂场景下更有效。此外,文献[6]提出了一种基于 YOLOX 模型的商品检索算法。该算法采用轻量级网络 MobileNet-V2 作为主干网并使用改进的相似度检索方法进行推理,使得网络在不增加检索速度的情况下增加识别准确度。文献[7]融合商品图像的图像特征和文本特征的识别算法,利用商品的图像和文本进行多模态融合,提高了识别系统鲁棒性。文献[8]提出了一种融合金字塔池化策略并使用一种名为哈希网络的 SHN 模型提高了模型对于图像形变带来的负面影响。

1.2 深度度量学习

深度度量学习(DML)广泛应用于计算机视觉任务,包括人脸识别、行人重识别、车辆再识别和商品识别。通常,这些任务的目标是检索与查询图像最相似的所有图像。近年来,深度度量学习取得了显著的进展。这些方法主要分为两类,即成对样本计算嵌入特征度量差异的方法和基于分类区分嵌入特征的方法。基于样本对的方法在深度嵌入的特征空间中优化样本对之间的相似性,例如,Triplet Loss[9],N-pair Loss[10],Multi-Simi Loss[11]。相比之下,基于分类的方法通过在训练集上训练各种分类模型来学习嵌入,例如 Cosface[12],ArcFace[13],NormSoftmax[14] 和 Proxy NCA[15]。最近的一项工作[16],考虑从统一的角度结合这两种方法。通过对两种损失进行加权,给出了一般的损失函数。与传统的度量学习只利用单一尺度的监督信息不同,该文提出的方法充分利用了多个语义

尺度的监督信息来对模型进行训练。

2　基于多尺度度量学习的商品检索

使用商品多个尺度的标签信息用于度量学习模型的训练即为多尺度度量学习。在多尺度度量学习中,会考虑商品图像的多个尺度。例如,该文考虑了商品标签的两个尺度,即粗粒度的组别和细粒度的类别。组标签和类标签符合层次结构,其中一个组别包含多个类别。根据实际的应用场景,不妨假设类级别的任务是开集识别任务,组级别的任务是闭集识别任务。网络设计的目标即同时完成组级分类任务和类级检索任务,并使两者尽可能没有干扰。文中模型采用了三个分支网络来满足这两个层次的识别需求。特别地,该方法利用协同注意力分支将组别级特征与类别级特征相结合,充分利用了图像标签的层次性信息。

2.1　总体方案

如图 2 所示,所提网络模型是一个以 ResNet50 作为骨干网的三个分支网络。三个分支分别是粗粒度特征提取分支、细粒度特征提取分支和融合特征提取分支。粗粒度特征提取分支和细粒度特征提取分支分别对商品图像进行粗粒度和细粒度特征的提取。融合特征提取分支则对另外两个分支提取到的特征以一定方式进行融合形成新的融合特征并最终用于商品图像的检索。

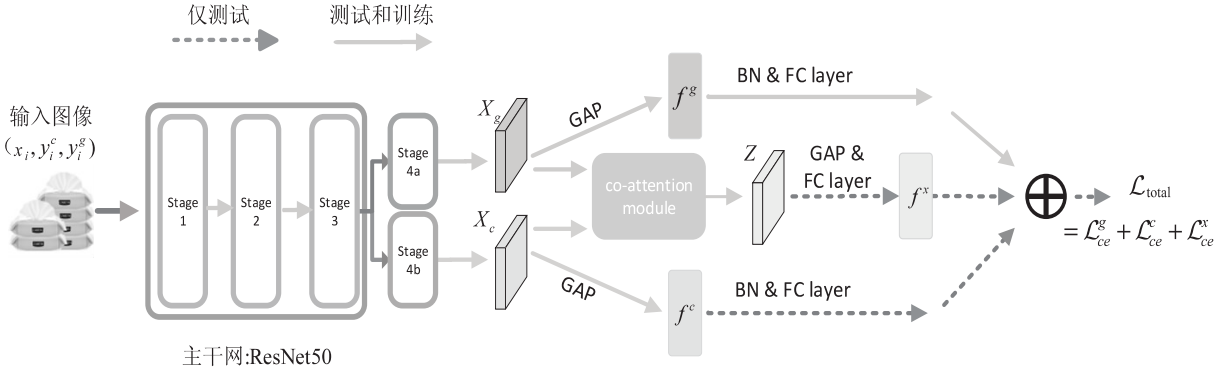

图 2　所提出的 MSML 深度神经网络模型架构

MSML 可以采用任何用于图像分类的深度网络作为主干网,例如谷歌 Inception 和 ResNet。考虑到 ResNet50 的竞争性能和相对简洁的架构,该文主要采用 ResNet50 作为主干网。在文献[17]的基础上,去掉了 ResNet 的 Stage4(包括框架中的 Stage4a 和 Stage4b)中最后一个空间向下采样操作,以增加特征图的大小。

为了更好地提取粗粒度和细粒度级别的特征,MSML 使用了 Stage4a 和 Stage4b 从骨干网将两个级别所优化的特征空间分开。Stage4a 和 Stage4b 都是从原始 ResNet 中的 Stage4 复制而来,并在 Stage3 后面并行连接。Stage4a 和 Stage4b 在结构上是相同的,但是在网络训练过程中网络参数的更新是不同的。从 Stage4a 派生的分支用于粗粒度级别的特征提取并在经过全连接层用于商品组别的分类,从 Stage4b 派生出的分支用于细粒度级别特征的提取。对 Stage4a 和 Stage4b 的输出进行 GAP(Global Average Pooling)运算,可以得到两个 2 048 维的特征向量。不同粒度的特征提取使用不同的分支网络可以缓解粗粒度特征和细粒度特征在一个特征空间提取所造成的相互干扰。此外,利用协同注意模块将粗粒度和细粒度特征相结合,充分利用了图像标签的层次信息。

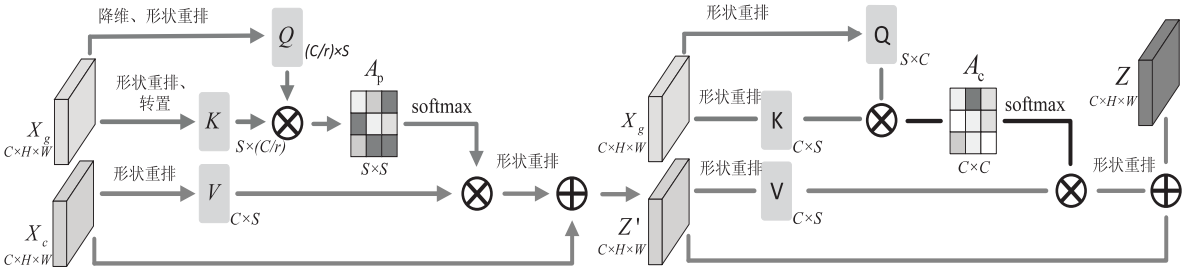

图 3　协同注意力模块结构示意图

2.2　融合特征提取分支

融合特征提取分支使用协同注意力机制将粗粒度和细粒度特征进行融合形成新的融合特征。注意力模块已被证明是多种机器学习场景下的有效机制[18],并广泛应用于自然语言处理(NLP)、图像处理(CV)、语音信号识别等各类机器学习任务中。注意力机制利用特征之间的相关性,迫使网络更加关注有用的信息。

在商品图像识别的情况下,对于同一个样本,网络在粗粒度级别和细粒度级别学习的特征应该是不同的。为了更好地使网络挖掘到图像的重要特征,该文

使用一种协同注意机制,将粗粒度级别的特征引入细粒度级别的特征中,如图 3 所示。这两个协同注意模块由一个空间注意力模块(PAM)和一个通道注意模块(CAM)组成。设 Stage4a 和 Stage4b 的输出特征 $\boldsymbol{X}_c, \boldsymbol{X}_g \in \mathbb{R}^{C\times(H\times W)}$ 为协同注意模块的输入特征,其中 C, H, W 分别为特征图的通道数、高度和宽度。PAM 将每个位置的特征 $\boldsymbol{X}_g$ 映射重塑到两个低维子空间上,这些子空间是由核大小为 1×1 的二维卷积实现的。经过重塑降维操作最终得到查询 $\boldsymbol{Q} \in \mathbb{R}^{C/r\times S}$,键 $\boldsymbol{K} \in \mathbb{R}^{C/r\times S}$。其中 $S = H \times W$ 为特征图的空间大小,r 为控制子空间维数的超参数。该文将遵从注意力机制一般的实验配置,简单的将其设置为 8。与自注意模块不同的是,协同注意中的值 $\boldsymbol{V}$ 并不是来自与键相同的特征图,而是来自另一个特征图。在所提出的模型中,直接将 Stage4b 的输出 $\boldsymbol{X}_c$ 经过大小重排作为值 $\boldsymbol{V} \in \mathbb{R}^{C\times S}$。那么 $\boldsymbol{X}_c$ 和 $\boldsymbol{X}_g$ 的经过空间注意力的输出可以由查询 $\boldsymbol{Q}$、键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$ 计算为:

$$\boldsymbol{Z}' = \text{attention}_p(\boldsymbol{X}_c, \boldsymbol{X}_g) = \boldsymbol{V}\sigma(\boldsymbol{A}_p) = \boldsymbol{V}\sigma(\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{K}) \tag{1}$$

其中,$\sigma(\cdot)$ 为 Softmax 函数,$\boldsymbol{Z}' \in \mathbb{R}^{C\times(H\times W)}$ 为 PAM 的输出,$\boldsymbol{A}_p$ 为位置权重矩阵,可以度量 $\boldsymbol{X}_g$ 不同位置特征之间的相关性。与 PAM 类似,同样使用了通道注意力模块(CAM),从通道角度赋予网络关注关键信息的能力。CAM 直接从 PAM 的输出 $\boldsymbol{Z}'$ 中获取键 $\boldsymbol{K}$ 和查询 $\boldsymbol{Q}$。最终融合特征分支可表示为:

$$\boldsymbol{Z} = \text{attention}_c(\boldsymbol{Z}', \boldsymbol{X}_g) = (\sigma(\boldsymbol{A}_c)\boldsymbol{Z})^{\mathrm{T}} = (\sigma(\boldsymbol{X}_g\boldsymbol{X}_g^{\mathrm{T}})\boldsymbol{Z}')^{\mathrm{T}} \tag{2}$$

2.3 损失函数

在 MSML 中,最终的总损失是三个单独损失的加权和,即:

$$\mathcal{L}_{\text{total}} = k_c\mathcal{L}_{ce}^{c}(\sigma(\boldsymbol{W}^c f_i^c), y_i^c) + k_g\mathcal{L}_{ce}^{g}(\sigma(\boldsymbol{W}^g f_i^g), y_i^g) + k_x\mathcal{L}_{ce}^{x}(\sigma(\boldsymbol{W}^x f_i^x), y_i^c) \tag{3}$$

其中,i 是训练样本 x_i 的指标。$\mathcal{L}_{ce}(\cdot)$ 为交叉熵损失函数。$\boldsymbol{W}^c$,$\boldsymbol{W}^g$ 和 $\boldsymbol{W}^x$ 是 f_i^c,f_i^g 和 f_i^x 之后 FC 层的权重参数。y_i^c 和 y_i^g 是训练样本 x_i 的组标签和类标签。k_c,k_g 和 k_x 是三个权重超参数,用于平衡三个损失的权重。经实验验证,所提模型最优配置为 1∶1∶1。粗粒度特征提取分支使用商品图像的组别标签进行训练,细粒度特征提取分支和融合特征提取分支使用类别标签进行训练。如图 2 所示,对特征图 $\boldsymbol{X}_c, \boldsymbol{X}_g, \boldsymbol{Z} \in \mathbb{R}^{C\times H\times W}$ 进行 GAP 运算,可以得到特征向量 f^c,f^g 和 f^x。将特征向量 f^c,f^g 和 f^x 分别经过全连接层后使用交叉熵损失函数进行训练。其中 Stage4b 输出的特征只用来作为融合特征分支的输入并不用作最后的图像检索。

在训练阶段,ResNet 的 Stage1,Stage2 和 Stage3 会同时被这三个损失函数进行优化,网络的其余部分仅由相应的损失函数进行优化。基于以上分析,任何分支的优化都可以通过影响公共部分 Stage1,Stage2 和 Stage3 来影响其他分支的优化。

3 实 验

在本节中,设置了一些对比神经网络模型与文中模型进行比较,并使用多个基于分类的损失函数来证明所提 MSML 模型的有效性。此外,文献[17]商品识别的冠军方案模型和文献[19]基于自注意力模块(S-A based)的图像检索方法将应用于商品检索任务,并与文中模型进行了性能比较。

3.1 数据集

实验所采用的数据集是来自公开数据集 Products-10k[2],为便于实验,进行了重新分割,最终形成新的商品检索数据集 MSML-Product。其中,源商品数据集 Products-10k 是一个基于商品识别应用场景的开放数据集。Product-10k 包含在线零售平台频繁购买的 10 000 种商品,涵盖时尚、3C、食品、保健及家居等所有品类。所有 SKU 级别的商品都被组织到一个层次结构中,总共有近 19 万张图片。在实际应用场景中,图像数量的分布并不均衡。所有图像都由生产专家团队手动检查和标记。

对 Product-10k 的所有图像进行重新分布,使数据集符合一般图像检索数据集的形式。将处理后的数据集 MSML-Product 分为三组:训练集、查询集和待查询集。查询集和待查询集用于测试模型,训练集用于训练模型。

MSML-Product 中的每个图像都有两个标签:类别和组别。这两个标签满足一个层次结构,即一个组别包含多个类别。对于类别而言,MSML-Product 保持开集设置,即训练集和测试类别没有交集。测试集中的类别对于网络来说是全新的。对于组别而言,MSML-Product 保持闭集设置,即测试的所有类别均已在训练集出现过。MSML-Product 数据集的详细信息如表 1 所示。

表 1 数据集类别分布和样本数量

File	images	classes	groups
train	100 504	4 952	350
query	27 204	4 756	350
gallery	69 479	4 756	350

3.2 评价指标

实验基于图像检索常用的三种评价指标,即累积匹配特征(Cumulative Matching Characteristics,CMC)、平均精度均值(mean Average Precision,mAP)和准确

率。CMC 表示在前 k 排序列表中存在真匹配的概率(如 Rank-1 表示第一位匹配正确的概率)。准确率(Precision)考虑在被模型判断为真的样例中,实际为真的样例比例。本次实验考虑将模型返回的前 10 个最相似的样本去计算准确率,并记为 Prec-10。相比之下,mAP 同时考虑了检索结果的精度和查全率。当一个查询有多个正确匹配时(这是常见的情况),mAP 强调识别所有正确匹配的能力,特别是那些困难的样本。

3.3　实验细节

该模型的训练主要采用典型的度量学习方法并重点参考了行人重识别领域的相关技术。在训练网络之前,先从 ImageNet 加载预训练的骨干网络,用于权重参数初始化。需要注意的是图 2 中主干网 Stage4a 和 Stage4b 参数初始化相同。它们都使用预训练模型的 Stage4 作为初始化。训练中采用标准的图像增强方法,包括随机水平翻转、随机裁剪、随机擦除。每张图像大小调整为 224×224 像素。训练方面使用了 Adam 优化器,其初始学习率为 3.5e-5,并在 30 和 50 个 epoch 时将学习率缩小 0.1 倍,直到收敛。实验在 Intel E5-2680CPU 2.4 GHz 的硬件环境下进行,4 张 NVIDIA Tesla P100 GPU。该模型每个批量包含 16 个细粒度类别在内的 256 个训练样本。对于损失函数权重参数 k_c , k_g 和 k_x ,所提模型中先固定其中两个权重参数,每隔 0.2 对另外一个权重进行每次增大或缩小 0.2,直到取最优值。经测试,权重参数最佳配置为1∶1∶1。所有的实验其损失的权重均设置为 1∶1∶1。

3.4　实验结果

3.4.1　所提模型的消融实验结果

为便于比较,实验共设置了三种对照模型,以突出文中模型各个模块的有效性。首先采用基线网络作为第一个对照模型,然后在基线网络的基础上,在模型的基础上逐个添加一些模块,以构建其他模型。原始基线模型采用 ResNet50 骨干网将原始输入图像映射到特征空间。类别和组别共享一个 Stage4 提取特征。然后在 ResNet50 后直接连接两个 FC 层,并使用两个交叉熵损失函数进行优化。

表 2　与对照模型的实验结果比较　　%

Method	class scale (group 标签未知)			class scale (group 标签已知)			group scale
	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	Top-1
ResNet50+long FC	36.6	58.1	76.1	54.2	68.2	84.8	57.0
+BN	40.6	63.7	79.1	56.2	71.4	86.1	57.8
+Stage 4 replication	42.1	64.8	80.0	58.1	72.6	87.1	57.3
Proposed Method	43.0	65.9	80.5	57.9	72.8	87.1	57.9

与基线模型相比,第二个对照模型在 GAP 操作后只增加了一个 BN 层。第三个对照模型与第二个对照模型相比增加了 Stage4 复制操作。该模型通过使用 Stage4a 和 Stage4b 分离粗粒度和细粒度特征空间。所提模型可以通过在第二个对照模型中添加协同注意模块来获得。实验将所提方法与所有对照方法进行了比较。为了充分证明所提模型的有效性,在已知组标签的情况下也进行了类别检索的相关实验。在表 2 中列出了这些模型在 Softmax 损失函数优化下的实验结果。可以看出,在损失函数相同的情况下,所提方法的综合性能最好。从第一个基线模型到所提模型,组别标签未知时,mAP 从 36.6% 上升到 43.0%,组别标签已知时,mAP 从 54.2% 上升到 57.9%。

对比对照模型的实验结果,可以得出结论,Stage4 的复制分离了粗粒度和细粒度级别的特征空间,缓解了不同尺度之间的冲突。此外,协同关注模块融合了类级和组级的层级关系信息,提高了模型的性能。

3.4.2　与相关文献方法对比实验结果

除了与设计的对照模型比较外,实验还包括与文献[18]中最先进的(SOTA)模型与文献[17]中基于自注意力机制的检索模型的比较。SOTA 模型是为了解决商品分类问题而设计的。该 SOTA 模型同样采用 ResNet 作为骨干网,并使用一种特殊的池化操作——广义均值(GeM)对 ResNet 输出的特征图进行池化。在池化之后,使用两个 BN 和 FC 层分别对组和类进行分类。文献[17]使用将主干网提取到的局部特征经过自注意力模块得到局部融合特征并加在原来的局部特征之后得到最终融合特征进行图像的检索。

为了进一步验证所提方法的有效性,实验在每个模型上分别使用了不同的基于分类的损失函数,即 Softmax 损失,Cosface 和 Arcface。从表 3 可以看出,所提模型与文献[18]中 SOTA 模型和文献[17]中基于自注意力模块的方法相比性能是最好的。在测试阶段,SOTA 模型最后一个 FC 层将被移除用于图像检索,与所提方法进行比较。实验结果表明,在使用 Softmax 损失函数进行训练时,所提模型和 SOTA 模型的性能都是最好的。在使用 Softmax 损失函数的情况下,无论组别标签是否已知,所提模型总是表现

最佳。

表 3 与 SOTA 模型的实验结果比较 %

Method	Loss	group 标签	mAP	Rank -1	Rank -5	Prec -10
SOTA method[18]	Softmax	未知	42.2	65.2	80.5	47.1
		已知	56.8	72.3	86.8	65.3
SOTA method[18]	Focal	未知	40.7	63.3	78.6	43.2
		已知	56.7	72.0	86.0	65.1
SOTA method[18]	Softmax+ Focal	未知	41.5	64.5	77.3	45.1
		已知	56.2	71.6	84.2	60.5
Proposed method	Arcface	未知	33.1	57.0	73.3	40.1
		已知	53.0	69.1	80.3	63.3
Proposed method	Cosface	未知	35.6	60.5	72.3	40.1
		已知	53.8	71.0	75.3	62.3
S-A based method[17]	Cosface	未知	41.6	65.1	79.8	44.1
		已知	55.3	72.3	85.1	65.8
S-A based Method[17]	Softmax	未知	42.6	65.2	79.7	44.2
		已知	55.8	72.0	85.0	65.7
Proposed method	Softmax	未知	43.0	65.9	80.5	53.1
		已知	57.9	72.8	87.0	67.6

此外,在检测速度上,所提方法的检索速度与文献[18]与文献[17]相差不大。所提方法的检索速度为 5.4e-4s/张,文献[18]和文献[17]的检索速度分别为 4.0e-4s/张和 5.3e-4s/张。可以看到,所提方法在没有显著增加检索时间的基础上显著提高了性能。

4 结束语

利用多尺度度量学习的方法,解决了大规模商品识别中使用有限类别的图像识别新增类别商品图像的问题。重点提出了一种充分利用商品图像多尺度信息的 MSML 模型。在大规模商品图像检索数据集上的实验结果表明,该方法是有效的,显著提高了商品识别的综合性能。

参考文献:

[1] CHENG L, ZHOU X, ZHAO L, et al. Weakly supervised learning with side information for noisy labeled images [C]//European conference on computer vision (ECCV). Seattle: IEEE, 2020: 306-321.

[2] BAI Y, CHEN Y, YU W, et al. Products-10K: a large-scale product recognition dataset[J]. arXiv: 2008.10545, 2020.

[3] ZOU Xiaofeng, LI Kenli, CHEN Cen. Multi-task cascade deep convolutional neural networks for large-scale commodity recognition [J]. Neural Computing and Applications, 2020, 32: 5633-5647.

[4] 刘文豪,姜胜明. 基于无标签半监督学习的商品识别方法[J]. 计算机应用与软件,2022,39(7):167-173.

[5] KRIZHEVSKY A, SUTSKEVER I, HINTON G. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[6] 张才裕. 基于图像相似度检索的商品识别系统研究[D]. 武汉:华中师范大学,2022.

[7] 张宏杰. 融合多模态数据的商品图像检索方法研究[D]. 哈尔滨:哈尔滨商业大学,2022.

[8] 贺周雨,冯旭鹏,刘利军,等. 基于 SHN 模型的商品图像检索方法[J]. 计算机工程与科学,2019,41(11):1991-1999.

[9] SCHROFF F, KALENICHENKO D, PHILBIN J. Facenet: a unified embedding for face recognition and clustering[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. Boston: IEEE, 2015: 815-823.

[10] SOHN K. Improved deep metric learning with multi-class n-pair loss objective[C]//Neural information processing systems. Barcelo: NeurIPS, 2016: 1857-1865.

[11] WANG X, HAN X, HUANG W, et al. Multi-similarity loss with general pair weighting for deep metric learning[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach: IEEE. 2019: 5022-5030.

[12] WANG H, WANG Y, ZHOU Z, et al. Cosface: large margin cosine loss for deep face recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City: IEEE, 2018: 5265-5274.

[13] DENG J, GUO J, XUE N, et al. Arcface: additive angular margin loss for deep face recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach: IEEE, 2019: 4690-4699.

[14] ZHAI A, WU H Y. Classification is a strong baseline for deep metric learning[J]. arXiv: 1811.12649, 2018.

[15] MOVSHOVITZ-ATTIAS Y, TOSHEV A, LEUNG T K, et al. No fuss distance metric learning using proxies[C]//Proceedings of the IEEE international conference on computer vision. Hawaii: IEEE, 2017: 360-368.

[16] SUN Y, CHENG C, ZHANG Y, et al. Circle loss: a unified perspective of pair similarity optimization[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle: IEEE, 2020: 6398-6407.

[17] LUO H, GU Y, LIAO X, et al. Bag of tricks and a strong baseline for deep person re-identification[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. Long Beach: IEEE, 2019.

[18] 高广尚. 深度学习推荐模型中的注意力机制研究综述[J]. 计算机工程与应用,2022,58(9):9-18.

[19] 秦姣华,黄家华,向旭宇,等. 基于卷积神经网络和注意力机制的图像检索[J]. 电讯技术,2021,61(3):304-310.