基于注意力金字塔与监督哈希的细粒度图像检索

殷梓轩,孙 涵
(南京航空航天大学 计算机科学与技术学院/人工智能学院,江苏 南京 211106)

摘 要:大规模细粒度图像检索是一项极具挑战性的任务。由于图像间具有类间距离小、类内距离大的特点,传统的深度神经网络学习到的图像特征存在高度冗余,导致检索速度慢、存储成本高昂。为解决该问题,提出了一种基于注意力金字塔与监督哈希的深度神经网络模型。在特征提取网络中,针对细粒度图像的特点,采用了双通路金字塔结构,并设计了自上而下的特征通路及自下而上的注意力通路,借此更好地融合高层与低层特征。在分类网络中,为压缩存储空间、提高检索效率,在深度哈希的基础上使用 $\tanh(x)$ 代替 $\mathrm{sign}(x)$ 作为激活函数,使学习到的哈希函数更容易达到平稳分布;同时结合量化损失与分类损失,使生成的哈希码更好地与原始输入图像的特征匹配。在 FGVC-Aircraft 及 Stanford Cars 两个标准细粒度数据集上的准确率分别达到 82.3%、83.3%,均优于其他对比算法,证明了算法的有效性。

关键词:细粒度图像检索;注意力金字塔;双通路;监督哈希;稳定分布
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2023)03-0020-07
doi:10.3969/j.issn.1673-629X.2023.03.004

Fine-grained Image Retrieval Based on Supervised Hashing with Attention Pyramid

YIN Zi-xuan, SUN Han
(School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: Large-scale fine-grained image retrieval is a challenging task. Due to the small inter-class variations and the large intra-class variations among images, features learned by traditional CNNs is highly redundant, which results in slow query speed and expensive storage cost. To address this problem, we propose a novel convolutional neural network which combines attention pyramid and supervised hashing. Specifically, in order to extract finer features, we introduce a dual pathway hierarchy structure in the feature extraction network with a top-down feature pathway and a bottom-up attention pathway, which is utilized to combine high-level semantic information and low-level detailed feature representations. Furthermore, to reduce storage cost and increase query speed, we improve deep hashing by using $\tanh(x)$ instead of $\mathrm{sign}(x)$ as the activation function to make sure that the learned hash function achieves stable distribution. At the same time, we adopt both quantization loss and classification loss to map the binary codes to the origin images better. The experimental results demonstrate that the proposed algorithm is superior to other comparison algorithms, for it achieves 82.3% and 83.3% accuracy on the FGVC-Aircraft and the Stanford Cars test set, which proves the effectiveness of the algorithm.

Key words: fine-grained image retrieval; attention pyramid; dual pathway; supervised hashing; stable distribution

0 引 言

细粒度图像检索是细粒度图像处理领域中的子方向,要求给定一张图片,算法能返回数据集中与该图片同属一个子类的图片。它与一般图像的检索不同,例如一般图像的检索要求在猫、狗、兔子中检索出来猫,细粒度图像检索则要求在波音 737-300、737-400、737-500 中找到波音 737-400。可以看出细粒度图像具有类间差异小、类内差异大的特点。因此,通常在检索的范围内,所有图像在视觉上都是相似的,但它们会在一些局部的细节处存在微小的区别。这无疑是非常具有挑战性的。

图像检索的原理,其实就是计算数据库中所有图像与待检索图像在特征空间中的距离并进行排序,距离越近,说明该图像与待检索图像相似度越高。但在

收稿日期:2022-05-11 **修回日期**:2022-09-14
基金项目:国防科技创新特区项目(XX)
作者简介:殷梓轩(1998-),男,硕士研究生,研究方向为图像处理、计算机视觉;孙 涵(1978-),男,博士,副教授,CCF 会员(33361S),研究方向为图像处理、计算机视觉。

如今大数据时代,线性搜索显然是不现实的,不仅因为时间开销大,而且存储成本也高。而深度哈希[1-2]恰恰能解决这两个弊端,因为它可以把提取到的图像特征压缩成紧凑的二进制哈希码,从而减小存储空间,同时大大提升检索速度。因此,深度哈希方法受到了越来越多的关注。

深度哈希模型一般包含两个阶段:特征提取及特征编码。特征提取阶段其实就是要进行细粒度图像识别任务。现有的细粒度图像识别算法主要分两类[3]:带有定位-分类子网络的模型以及进行端到端特征编码的模型。但是现有的方法大部分只关注到了卷积神经网络中高层的语义信息,却忽略了低层的细节特征,该文则希望能把二者结合在一起。在特征编码阶段,深度哈希可以把图像特征映射为二进制形式的哈希码,从而在减小存储空间的同时大大提升检索速度。

具体地,该文提出了一个基于注意力金字塔与监督哈希的深度神经网络,用于解决大规模细粒度图像检索问题。它主要分为两大部分,分别是两阶段的特征提取网络以及哈希分类网络。在特征提取网络中,使用了一种双通路的金字塔结构,包含自上而下的特征通路以及自下而上的注意力通路,借此融合高层的抽象语义信息与低层的细节特征;与此同时,还在网络的两个阶段之间引入了一个中间层对输入图像进行细化。在哈希分类网络中,使用 $\tanh(x)$ 代替 $\mathrm{sign}(x)$ 作为激活函数,从而使学习到的哈希函数达到平稳分布;同时结合量化损失与分类损失,使生成的哈希码能更好地与原始输入图像的特征进行匹配。在两个标准细粒度图像数据集[4-5]上进行了大量实验,实验结果表明,提出的模型能进行高精度细粒度图像检索。同时还与当前的一些经典算法进行了实验比较,比较结果展示了该方法的优越性。

1　相关工作

1.1　细粒度图像检索

细粒度图像检索是近年来比较活跃的研究方向,其任务为:给定一个数据集,该数据集中的所有图片属于同一大类下的若干子类,要求输入一张待检索图片,系统能返回数据集中与其同属一个子类的若干张图片。现有方法可以分为两类:监督方法[6-7]和无监督方法。其中,监督方法中的代表方法是 Zheng 等人提出的度量学习方法[6]。为了更好地从背景中提取目标主体的部件特征,作者使用了一种弱监督的特征提取方式,通过由上到下的显著性分割目标轮廓。而无监督方法中的代表就是 Wei 等人提出的 SCDA 方法[8]。该方法首先从图片中定位目标主体,去除背景噪声并保留关键的深度描述子,接着再由这些深度描述子生成特征向量。

1.2　细粒度图像识别

细粒度图像检索是以识别任务为基础的,因为识别的准确性会直接影响后续分类网络的性能。现有的细粒度图像识别算法主要分为以下两类:(1)带有定位-分类子网络的模型[9-10]。如 Zhang 等人提出了一种学习部件级检测器的网络[9],Wei 等人通过分割的方法[10]对关键部位进行定位。(2)端到端特征编码模型[11-12]。这一类方法对监督信息的依赖相对较低,它们一般仅要求提供图像级别的标注。其中最具代表性的是 Lin 等人提出的双线性卷积神经网络。

上述方法均能取得良好的效果,但它们都忽略了低层的细节信息。因为细粒度图像具有类间差异小、类内差异大的特点,因此低层的细节信息,如纹理、边缘等对识别任务是非常关键的。因此,受 Ding 等人 AP-CNN[13]的启发,该文希望把高层的语义信息与低层的细节特征结合在一起。

1.3　深度哈希

深度哈希基于训练数据学习哈希函数,在还原原始数据空间距离顺序的同时生成更短的哈希码。这种紧凑的特征表示有利于存储大量样本并节省检索时间。依据是否使用标注信息,可以分为无监督方法和监督方法两类。前者如 HashGAN[14]等在学习哈希函数时不需要监督信息,但是容易造成原始数据和哈希编码之间的语义鸿沟。而后者充分利用监督信息学习哈希函数并生成哈希码,具有更高的准确性,代表方法如 KSH[15]等。

2　模型框架

2.1　网络基本架构

提出的网络分为特征提取以及特征编码两个阶段,如图 1 所示。其中特征提取是两阶段的:第一阶段为原始特征提取,以原始图片作为输入;第二阶段为细化特征提取,以第一阶段的输出作为输入。

网络接收到一张图片后,首先使用骨干网络生成特征金字塔,其信息流动方向是自上而下的。然后引入注意力机制,得到注意力金字塔,其信息流动方向是自下而上的。以上是网络的第一阶段。使用中间层对原始图像进行细化,并输入二阶段网络。网络第二阶段的工作流程与第一阶段是相同的,并最终得到两个注意力金字塔,对它们做连接操作后即可用于后续的特征编码。

2.2　双通路金字塔

2.2.1　自上而下的特征金字塔

一般地,原始图像输入经过卷积神经网络的处理会得到具有高层语义信息的特征表示,但这会使低层

的许多细节特征被丢失,不利于细粒度图像的处理。为了能保留这些细节特征,受 FPN 网络[16]的启发,该文设计了一条自上而下的特征通路来提取不同尺度的特征。

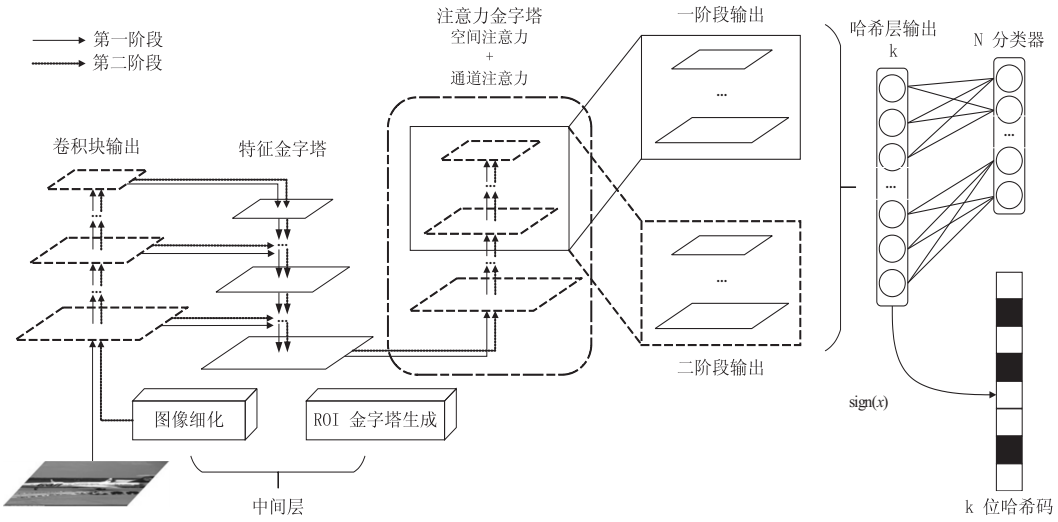

图 1 网络结构

原始的输入图像在经过若干卷积块后可以得到若干特征图,把这些特征图表示为 $\{B_1,B_2,\cdots,B_l\}$,这里 l 表示卷积块的数量。很多针对粗粒度图像处理的方法会直接把最后一块的输出 B_l 用于分类,但该文希望尽量利用好每一个 B_i 。由于越靠近低层,特征的抽象层次越低,如果要利用每一个卷积块的输出,将不可避免带来巨大的成本开销,因此选择其中的后 N 个输出并生成对应的特征金字塔。具体如图 2 所示。最终得到的特征金字塔为 $\{F_{l-N+1},F_{l-N+2},\cdots,F_l\}$ 。

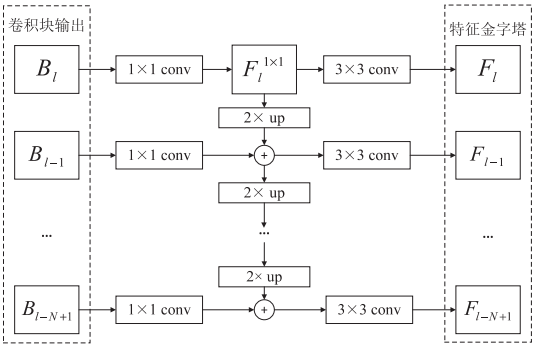

图 2 特征金字塔

2.2.2 自下而上的注意力通路

得到特征金字塔后,该文设计了一条自底向上的注意力通路,其中包含空间注意力以及通道注意力。空间注意力[17]用于定位输入图像在不同尺度上具有辨识度的区域。其形式化描述为:

$$A_i^s=\sigma(K*F_i) \tag{1}$$

式中,σ 为 sigmoid 激活函数,$*$ 代表反卷积操作,而 K 代表卷积核。通道注意力[18]用于加入通道之间的关联,并同时把低层的细节信息一层一层传递给高层。其形式化描述如下:

$$A_i^*=\sigma(W_2\cdot\mathrm{relu}(W_1\cdot\mathrm{GAP}(F_i))) \tag{2}$$

式中,σ 为 sigmoid 激活函数,$\cdot$ 代表元素间乘法,W_1 和 W_2 为两个全连接层的权重参数。GAP($\cdot$) 代表全局平均池化。为了把低层的细节信息传递给高层,从而生成一条自底向上的通路,还需要对 A_i^* 进行处理。除了最底层的 A_{l-N+1}^c 可以直接等于 A_{l-N+1}^* ,其他的 A_i^* 都需要先和 A_{i-1}^c 做加法,然后再进行二倍下采样才能得到 A_{i-1}^c 。

在得到注意力金字塔(见图 3)后,可结合空间金字塔生成一条自底向上的注意力通道。具体地,该文先把空间注意力 A_i^s 和通道注意力 A_i^c 做加法,然后与特征金字塔中的 F_i 做点乘运算,最终得到 F_i' 。该过程的形式化描述如下:

$$F_i'=F_i\cdot(A_i^s+\alpha A_i^c) \tag{3}$$

最终可以得到 $\{F'_{l-N+1},F'_{l-N+2},\cdots,F'_l\}$ 用于后续的分类。

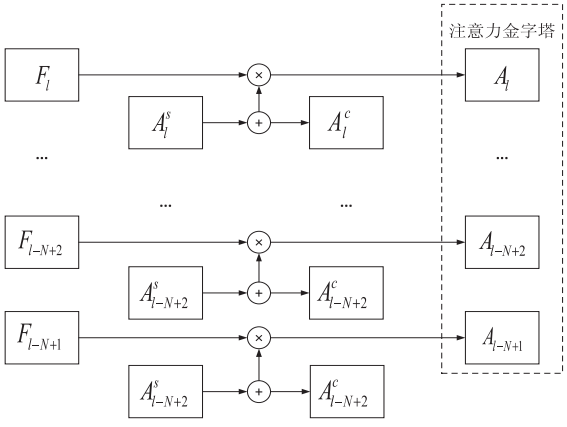

图 3 注意力金字塔

2.3 中间层

2.3.1 ROI 金字塔

该文在 RPN 网络[19]的基础上进行了改进。由于

空间注意力金字塔的值属于(0,1)区间,于是可以利用它来代替不同尺度的特征图作为锚点的评分标准,依此以一种弱监督的形式来定位部件区域。具体地,使用空间注意力作为每一个 ROI 的分数,并把小于平均分的 ROI 舍弃,剩下的作为候选。接着,使用自适应 NMS 算法[13]把多个 ROI 整理、合并到所需的 k 个。对于金字塔中的第 i 层,该文选择得分最高的 ξ_i^k 个,而且因为低层比高层拥有更多的细节信息,因此会在低层选择比高层更多的区域,即对于集合 $\{\xi_{l-N+1}^k,\xi_{l-N+2}^k,\cdots,\xi_l^k\}$ 有约束 $\xi_{l-N+1}^k > \xi_{l-N+2}^k > \cdots > \xi_l^k$。最终可以得到 ROI 金字塔 $R^{\text{nms}} = \{R_{l-N+1}^{\text{nms}},R_{l-N+2}^{\text{nms}},\cdots,R_l^{\text{nms}}\}$。

2.3.2　基于 ROI 的图像细化

为减少过拟合现象,Golnaz 等人提出了 Dropblock 算法[17],但因为丢弃区域的选择是随机的,因此缺乏针对性。该文基于 ROI 金字塔对其进行了改进:每次随机选择 ROI 金字塔中的第 i 层,并从 R_i^{nms} 中选择评分最高的 ROI 区域作为被丢弃区域。此外,还在中间层去除了背景噪声,只保留图像主体。具体地,把金字塔中所有层的所有 ROI 都进行合并,从而得到一个整体的 bounding box 以代表图像主体,最后去除多余部分即可。最终会得到 B_{l-N+1} 并把它作为网络第二阶段的输入。

2.3.3　细化特征提取阶段

特征提取的第二阶段与第一阶段用的是同一个网络结构,所不同的是网络的第一阶段需要经过 l 个卷积块 $(1,2,\cdots,l-N+1,l-N+2,\cdots,l)$,而网络的第二阶段只需要经过 N 个卷积块 $(l-N+1,l-N+2,\cdots,l)$。最终,记第一阶段得到的结果为 $\{F_{l-N+1}^{'},F_{l-N+2}^{'},\cdots,F_l^{'}\}$,记第二阶段得到的结果为 $\{F_{l-N+1}^{''},F_{l-N+2}^{''},\cdots,F_l^{''}\}$,对它们做连接运算:

$$F^{\text{output}} = \text{concat}(F_{l-N+1}^{'},\cdots,F_l^{'},F_{l-N+1}^{''},\cdots,F_l^{''}) \tag{4}$$

并把 F^{output} 用于后续的分类。

2.4　深度哈希

2.4.1　稳定分布与量化损失

目前许多哈希方法都使用量化正则器(如 $\|x_i - \text{sign}(x_i)\|_2^2$)对图像特征进行正规化从而得到离散的哈希码,但这样的二值化过程会导致两幅图像在汉明空间与欧几里得空间中不一致,同时也使量化损失的优化问题成为 NP 完全问题。因为像 $\|x_i - \text{sign}(x_i)\|_2^2$ 的传统量化正则器在最小化量化损失的时候会改变图像的特征分布,因此它是不稳定的,无法使量化损失达到严格意义上的最小化,这导致生成的哈希码与原本图像的特征不能很好地匹配。

受 DSHSD[20]的启发,该文引入了 p 稳定分布。记输入图像从欧几里德空间到汉明空间上的损失为量化损失,当量化损失较小时,可以认为哈希函数是一个 p 稳定分布,否则是不稳定的。因此,为了使量化损失尽可能小,选择 $\tanh(x)$ 作为激活函数,使用 $\|\tanh(x_i) - \tanh(x_j)\|_p$ 代替 $\|x_i - x_j\|_p$ 来保证分布一致性。$\tanh(x)$ 不仅梯度容易计算,而且对于反向传播算法来说是友好的:

$$h_i = \tanh(F_i^{\text{output}}) \tag{5}$$

于是量化损失可以写成:

$$L_d(H,S) = \sum_{s_{ij}\in S}(L_{ij}^a + L_{ij}^b) \tag{6}$$

$$L_{ij}^a = \frac{1}{2}(1 - s_{ij})\|h_i - h_j\|_2^2 \tag{7}$$

$$L_{ij}^b = \frac{1}{2}s_{ij}\max(0,m - \|h_i - h_j\|_2^2) \tag{8}$$

式中,S 为相似性标签的集合,m 为阈值参数且 $m > 0$,$\|\cdot\|_2$ 表示 L2 范数并用于衡量哈希码之间的距离。

2.4.2　分类损失

除了量化损失以外,为了使哈希码能保留图像特征中的分类信息,还引入了分类损失。不同于以前的一些方法,为了不引入额外的变量,该文并没有对 h_i 施加二值化约束,这也符合奥卡姆剃刀原理。于是分类损失的形式化表示如下:

$$L_c(Y,H) = L(Y,W^{\text{T}}H) \tag{9}$$

式中,$H = \{h_i\}_{i=1}^{\text{T}}$,$Y$ 是图像标签,W 是哈希全连接层的权重。对于单标签分类问题,采用 softmax 优化方法。对于多标签分类问题,使用交叉熵损失。

最后,结合公式 6 与公式 9 就可以得到总的损失函数:

$$L(Y,S,H) = L_c + \varepsilon L_d \tag{10}$$

式中,ε 是控制量化损失的权重参数。

3　实验结果与分析

在 FGVC-Aircraft 以及 Stanford Cars 两个标准细粒度数据集上进行了实验验证。其中 FGVC-Aircraft 数据集包含了 10 200 张飞机的图片,共有 102 种不同的机型,其中训练集有 6 667 张图片,测试集有 3 333 张图片。Stanford Cars 数据集包含了 16 185 张汽车的图片,共有 196 种不同型号的汽车,其中训练集有 8 144张图片,而测试集有 8 041 张图片。该文使用 mAP 作为度量指标。

3.1　实现细节

该文使用 ResNet-50 作为骨干网络,并在 ImageNet 上进行预训练。选择 ResNet-50 的后 3 个残差块输出的特征图建立特征金字塔,没有选择 conv1 和 conv2 主要是因为它们的抽象层次较低,若参与特征金字塔的构建将带来较大计算开销。所有图片在输

入网络时大小均会被调整为 448×448。该文并没有使用数据集提供的 bounding box 标注信息,因为在实际问题中,bounding box 的获取成本较高,过于依赖它会使方法存在较大局限性。

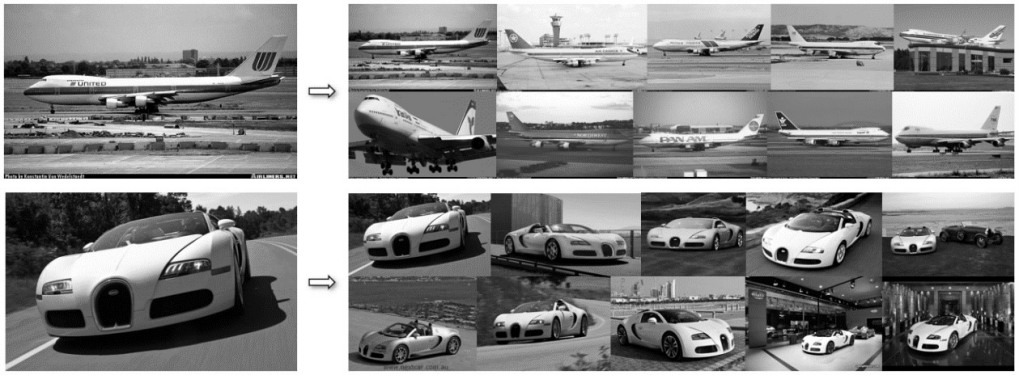

图 4 波音 747 与布加迪威龙 16.4 图片检索示例

自适应 NMS 算法中的阈值 t_1 和 t_2 分别设置为 0.05 和 0.9。公式 8 中的阈值 m 设置为 $2k$(k 为哈希码位数且 $k=\{16,32,64\}$)。公式 3 中的权重参数 α 设置为 0.5,公式 10 中的 ε 设置为 0.1。在模型训练的过程中,该文使用的优化算法为 Adam 算法,其中学习率设置为 10^{-5} ,β_1 设置为 0.9,β_2 设置为 0.999。

表 1 文中方法与其他 SOTA 方法的 mAP 对比 %

方法	FGVC-Aircraft			Stanford Cars		
	16 bit	32 bit	64 bit	16 bit	32 bit	64 bit
DHN(AAAI2016)[21]	2.9	9.2	45.1	0.6	0.7	4.3
IDHN(TMM2019)[22]	3.0	31.0	42.0	0.7	2.0	5.6
HashNet(ICCV2017)[23]	5.0	19.7	30.1	1.9	3.3	34.6
DCH(CVPR2018)[24]	56.4	68.4	64.4	55.0	60.1	62.2
Ours	75.0	79.3	82.3	64.0	78.6	83.3

3.2 对比实验

文中方法与其他 4 种 State-Of-The-Art 方法在 2 个数据集、3 种哈希码长度下的对比结果如表 1 所示。从表中可以看出,文中方法无论在 16 位哈希码、32 位哈希码以及 64 位哈希码中都能达到最高的准确率,这验证了该方法的准确性。具体地,在 FGVC-Aircraft 数据集中,文中方法在三种长度哈希码下的准确率比第二名的 DCH 分别提升了 33.0%、15.9% 以及 27.8%。在 Stanford Cars 数据集中,文中方法在三种长度哈希码下的准确率比 DCH 分别提升了 16.4%、30.8% 以及 33.9%。这得益于在哈希层中使用了 $\tanh(x)$ 作为激活函数,可以保证哈希变换过程中的分布一致性,使哈希码可以更好地保留图像的特征信息。

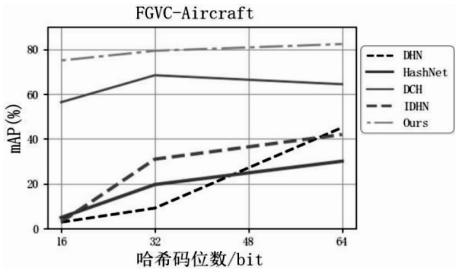

图 5 mAP 随哈希码位数的变化

图 5 展示了 5 种方法在 FGVC-Aircraft 数据集上的 mAP 随哈希码位数的变化。可以很直观地观察到,文中方法在 16 位、32 位以及 64 位长度的哈希码上效果都大幅优于其他四种方法,同时,哈希码长度的改变并不会显著影响文中方法的准确度。

总体上,5 种方法的 mAP 基本会随着哈希码位数的增加而增加,这很容易解释,因为哈希码的位数越高,它的特征表示能力也越强。但也有例外,如 DCH 在 FGVC-Aircraft 数据集中,当哈希码长度由 32 位变成 64 位时,mAP 反而从 68.4% 下降到了 64.4%。

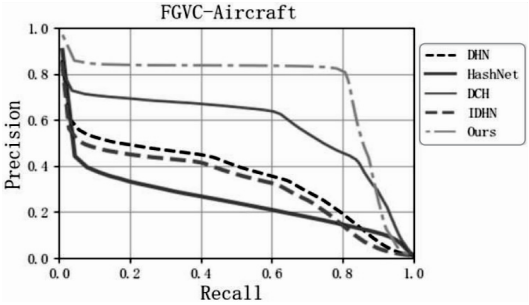

图 6 PR 曲线对比

图 6 展示了 5 种方法在 FGVC-Aircraft 数据集上的 PR 曲线。可以直观地看出,文中方法的 PR 曲线所围成的面积是最大的,这证实了该方法的有效性。

3.3　消融实验

在特征提取网络中,该文结合了特征金字塔与注意力金字塔。为了分别验证它们的作用效果,使用 32 位哈希码设计了如下消融实验。

在表 2 中,Baseline 即 ResNet-50,FP 表示只使用特征金字塔,FP+AP 表示既使用特征金字塔,也使用注意力金字塔。可以看出,特征金字塔能在 Baseline 的基础上较大地提升模型准确率,而注意力金字塔能让准确率有进一步的提升。

表 2　不同特征提取网络下的 mAP 对比　%

方法	FGVC-Aircraft	Stanford Cars
Baseline	73.8	72.0
FP	77.6	76.3
FP+AP	79.3	78.6

接下来,验证了中间层中使用到的 Dropblock 算法以及背景噪声去除算法的作用效果,消融实验同样是基于 32 位哈希码设计的。由表 3 可知,Dropblock 算法迫使网络学习不同区域的特征,从而能较大提升网络性能。而背景噪声去除算法作为一种辅助手段,虽然提升效果不如 Dropblock 算法,但也可以在一定程度上提升网络性能。

表 3　不同特征提取网络下的 mAP 对比　%

Dropblock	背景去除	FGVC-Aircraft	Stanford Cars
×	×	74.5	74.1
√	×	78.4	76.9
×	√	75.2	74.7
√	√	79.3	78.6

最后,还验证了哈希算法中量化损失与分类损失的作用。其中,Variant-Q 的损失函数中只有量化损失,而 Variant-C 只有分类损失,Variant-S 中仍然使用 $\| x_i - \mathrm{sign}(x_i) \|_2^2$ 作为量化正则器。

表 4　各哈希方法的对比

方法	FGVC-Aircraft	Stanford Cars
Variant-Q	4.5	6.8
Variant-C	78.0	77.2
Variant-S	15.7	13.2
Ours	79.3	78.6

由表 4 可以看出,在分类损失与量化损失中,分类损失起主要作用,同时也可以看出使用 $\tanh(x)$ 作为激活函数的巨大优势。

4　结束语

细粒度图像检索是一个在日常生活中有着广泛应用、但却充满挑战性的任务。该文提出了一种基于注意力金字塔和监督哈希的深度神经网络。通过两阶段的特征提取网络构建双通路金字塔,把高层语义信息与低层细节特征充分结合,同时引入了一个中间层对图像进行细化;在哈希编码阶段结合了量化损失与分类损失,并使用 $\tanh(x)$ 代替 $\mathrm{sign}(x)$ 作为激活函数,使学习到的哈希函数达到平稳分布的要求,减少了反向传播过程中量化损失的偏差,从而使生成的哈希码更好地与原本图像的特征进行匹配。为了验证模型的有效性与准确性,在两个标准的细粒度数据集上进行了大量实验。通过在 16 位、32 位以及 64 位三种长度的哈希码上与四种 SOTA 方法进行对比,证实了该模型的有效性,同时通过一系列消融实验验证了不同算法组件的作用效果。

参考文献:

[1] LI W J,WANG S,KANG W C. Feature learning based deep supervised hashing with pairwise labels[J]. arXiv:1511.03855,2015.

[2] LIU H,WANG R,SHAN S,et al. Deep supervised hashing for fast image retrieval[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas:IEEE,2016:2064-2072.

[3] WEI X S,WU J,CUI Q. Deep learning for fine-grained image analysis:a survey[J]. arXiv:1907.03069,2019.

[4] MAJI S,RAHTU E,KANNALA J,et al. Fine-grained visual classification of aircraft[J]. arXiv:1306.5151,2013.

[5] KRAUSE J,STARK M,DENG J,et al. 3D object representa-

tions for fine - grained categorization [C]//Proceedings of the IEEE international conference on computer vision workshops. Sydney: IEEE, 2013: 554-561.

[6] ZHENG X, JI R, SUN X. Centralized ranking loss with weakly supervised localization for fine - grained object retrieval [C]//International joint conferences on artificial intelligence organization. [s. l.]: Morgan Kaufmann, 2018: 1226-1233.

[7] HUANG Z, DUAN X, ZHAO B. Interpretable attention guided network for fine - grained visual classification [J]. arXiv: 2103. 04701, 2021.

[8] WEI X, LUO J, WU J. Selective convolutional descriptor aggregation for fine-grained image retrieval [J]. IEEE Transactions on Image Processing, 2017, 26(6): 2868-2881.

[9] ZHANG N, DONAHUE J, GIRSHICK R. Part-based R-CNNs for fine - grained category detection [C]//European conference on computer vision. Zurich: Springer, 2014: 834-849.

[10] WEI X S, XIE C W, WU J. Mask-CNN: localizing parts and selecting descriptors for fine-grained bird species categorization [J]. Pattern Recognition, 2018, 76: 704-714.

[11] LIN T Y, ROYCHOWDHURY A, MAJI S. Bilinear CNN models for fine-grained visual recognition [C]//Proceedings of the IEEE international conference on computer vision. Santiago: IEEE, 2015: 1449-1457.

[12] DUBEY A, GUPTA O, RASKAR R. Maximum-entropy fine-grained classification [J]. arXiv: 1809. 05934, 2018.

[13] DING Y, MA Z, WEN S. AP-CNN: weakly supervised attention pyramid convolutional neural network for fine - grained visual classification [J]. IEEE Transactions on Image Processing, 2021, 30: 2826-2836.

[14] DIZAJI K G, ZHENG F, SADOUGHI N. Unsupervised deep generative adversarial hashing network [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City: IEEE, 2018: 3664-3673.

[15] LIU W, WANG J, JI R. Supervised hashing with kernels [C]//2012 IEEE conference on computer vision and pattern recognition. Providence: IEEE, 2012: 2074-2081.

[16] LIN T Y, DOLLAR P, GIRSHICK R. Feature pyramid networks for object detection [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu: IEEE, 2017: 2117-2125.

[17] JADERBERG M, SIMONYAN K, ZISSERMAN A. Spatial transformer networks [J]. arXiv: 1506. 02025, 2016.

[18] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City: IEEE, 2018: 7132-7141.

[19] REN S, HE K, GIRSHICK R. Faster R-CNN: towards real-time object detection with region proposal networks [J]. arXiv: 1506. 01497, 2016.

[20] WU L, LING H, LI P. Deep supervised hashing based on stable distribution [J]. IEEE Access, 2019, 7: 36489-36499.

[21] ZHU H, LONG M, WANG J. Deep hashing network for efficient similarity retrieval [C]//Proceedings of the AAAI conference on artificial intelligence. Phoenix: AAAI, 2016: 2415-2421.

[22] ZHANG Z, ZOU Q, LIN Y. Improved deep hashing with soft pairwise similarity for multi-label image retrieval [J]. arXiv: 1803. 02987, 2019.

[23] CAO Z, LONG M, WANG J. HashNet: deep learning to hash by continuation [C]//Proceedings of the IEEE international conference on computer vision. Venice: IEEE, 2017: 5608-5617.

[24] CAO Y, LONG M, LIU B. Deep cauchy hashing for hamming space retrieval [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City: IEEE, 2018: 1229-1237.