DOI:10.19651/j.cnki.emt.2313218

注意力增强的视觉 Transformer 图像检索算法*

刘华咏 黄 聪 金汉均
(华中师范大学计算机学院 武汉 430070)

摘 要：基于深度哈希的图像检索方法往往利用卷积和池化技术去提取图像局部信息，并且需要不断加深网络层次来获得全局长依赖关系，这些方法一般具有较高的复杂度和计算量。本文提出了一种注意力增强的视觉 Transformer 图像检索算法，算法使用预训练的视觉 Transformer 作为基准模型，提升模型收敛速度，通过对骨干网络的改进和哈希函数的设计，实现了高效的图像检索。一方面，本文设计了一个注意力增强模块，来捕获输入特征图的局部显著信息和视觉细节，学习相应的权重以突出重要特征，并增强输入到 Transformer 编码器的图像特征的表征力。另一方面，为了提高图像检索的效率，设计了一种对比哈希损失函数，生成具有判别力的二进制哈希码，从而降低了内存需求与计算复杂度。在 CIFAR-10 和 NUS-WIDE 数据集上的实验结果表明，本文提出的方法，在两个不同数据集上使用不同哈希码长度的平均精度均值达到了 96.8% 和 86.8%，性能超过多种经典的深度哈希算法和其他两种基于 Transformer 架构的图像检索算法。

关键词：图像检索；视觉 Transformer；深度哈希；注意力模块
中图分类号：TP391 **文献标识码**：A **国家标准学科分类代码**：510.4050

Image retrieval method with attention-enhanced visual Transformer

Liu Huayong Huang Cong Jin Hanjun
(School of Computer Science, Central China Normal University, Wuhan 430070, China)

Abstract: The image retrieval methods based on deep hashing often use convolution and pooling techniques to extract local information from images and require deepening the network layers to obtain global long-range dependencies. These methods generally have high complexity and computational requirements. This paper proposes a vision Transformer-based image retrieval algorithm enhanced with attention, which uses a pre-trained vision Transformer as a benchmark model to improves model convergence speed and achieves efficient image retrieval through improvements to the backbone network and hash function design. On the one hand, the algorithm designs an attention enhancement module to capture local salient information and visual details of the input feature map, learns corresponding weights to highlight important features, enhances the representativeness of image features input to the Transformer encoder. On the other hand, to generate discriminative hash codes, a contrastive hash loss is designed to further ensure the accuracy of image retrieval. Experimental results on the CIFAR-10 and NUS-WIDE datasets show that the proposed method achieves an average precision of 96.8% and 86.8%, respectively, using different hash code lengths on two different datasets, outperforming various classic deep hashing algorithms and two other Transformer-based image retrieval algorithms.

Keywords: image retrieval; vision Transformer; deep hash; attention module

0 引 言

随着深度学习技术的发展和 GPU 硬件的性能提升，基于卷积神经网络（convolutional neural network，CNN）的图像检索逐渐开始成为比较主流的图像检索算法，得到了越来越多的研究人员的认可。基于 CNN 的哈希方法能够将 CNN 产生的图像特征学习生成二进制哈希码，大大地减少特征相似度计算的复杂度。Xia 等[1]率先提出了一种分两步的深度哈希检索，即第 1 步通过相似性矩阵计算哈希编码，然后第 2 步根据哈希编码和类别标签对 CNN 进行

收稿日期：2023-03-28
* 基金项目：教育部人文社会科学研究项目(21YJA870005)资助

训练的策略。然而,在这个过程中图像特征通过符号激活函数会产生梯度不适应的问题,导致检索质量降低。针对这个问题,Cao 等[2]提出的 HashNet 通过加入正则化和采用训练 tanh 激活函数使得结果不断逼近符号函数的延续方法产生更精确的哈希码。Su 等[3]提出的 GreedyHash 采用贪心算法,设计了一个哈希编码层,在正向传播中使用符号函数来保持约束,在反向传播中记录并更新梯度,避免了梯度消失的问题,提高了检索质量。Yuan 等[4]提出 CSQ,针对以往有监督的哈希方法在学习哈希函数时仅仅采用局部相似性比较,学习效率和哈希碰撞率低的问题,提出了一种基于全局相似度的中心相似度量化,对图像数据点的中心相似度进行优化,对相似数据生成内聚哈希码,对不同的数据生成分散的哈希码,显著提高了在大规模图像和视频检索任务上的性能。

虽然 CNN 已经在众多视觉任务上取得了良好的效果,但是仍然存在一些不足之处。CNN 在提取图像特征时通常采用固定大小的卷积核和步幅,对信息的提取通常沿着一个方向流动,这使得基于 CNN 的模型很难处理远距离依赖关系,例如在一张图片中出现的两个不同物体之间的关系。

近些年,基于自注意力机制的 Transformer[5]模型已经成为计算机视觉领域热门的研究对象,自注意力机制能够实现信息的跨位置关联,可以更好地处理远距离依赖关系。由 Transformer 模型改进的 Vision Transformer(ViT)[6]和 Swin Transformer[7]等模型在图像分类和目标检测等各种计算机视觉任务上相较于 CNN 取得了更优的结果。因此,一部分人也开始研究视觉 Transformer 在图像检索任务上的有效性。Gkelios 等[8]开创性地将 Vision Transformer 用于图像检索,提出了一种无需预训练和初始化的全局特征描述符,对比其他基于 CNN 的方法的结果,取得了具有竞争力的结果。El-Nouby 等[9]在采用 ViT 来生成图像描述符的同时利用度量学习的方法,结合对比损失和交叉熵正则项来进行图像检索,证明了基于 ViT 的方法相较于基于卷积的方法在检索任务的准确率上有显著的提升。Li 等[10]提出的 HashFormer 使用 ViT 作为骨干网络提出了一种新的平均精度损失函数,优化了检索的精度。Chen 等[11]提出的 TransHash 对 ViT 做了改动,加入双流特征学习模块去提取局部和全局特征。此外,TransHash 在哈希码的学习上采用动态构造相似度矩阵的贝叶斯学习策略,以此生成更有判别力的哈希码,从而提高检索精度。

本文提出了一种注意力增强的视觉 Transformer 图像检索算法,分别在单标签和多标签数据集上使用不同长度哈希码进行实验,验证了本文提出的算法相较于传统的基于 CNN 架构的深度哈希算法以及两种最近提出的基于 Transformer 架构的哈希算法,可以进一步提高检索的效率和准确性。

1 注意力增强的视觉 Transformer 图像检索算法

本文提出了一种基于注意力增强的视觉 Transformer 图像检索算法(attention enhanced vision transformer,AE-ViT)如图 1 所示,主要分以下 3 个模块:

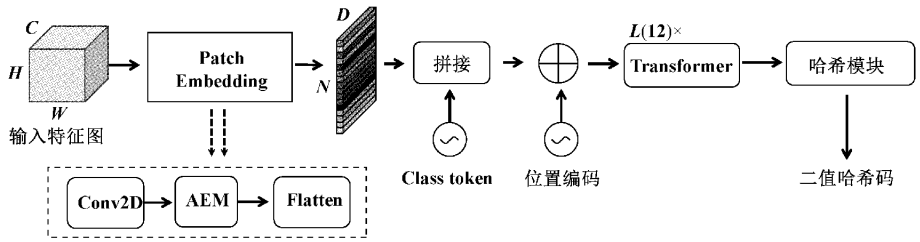


图 1 AE-ViT 模型图

1)注意力增强的 Patch Embedding 层,对于输入的图像 $\boldsymbol{I}\in\boldsymbol{R}^{H\times W\times C}$,$H$、$W$ 和 C 分别为图像的长度、宽度以及通道数。为了将三维图像转化为标准 Transformer 所需要的 $N\times D$ 大小的二维矩阵(N 为输入序列的长度,D 为输入序列中词向量的维度),AE-ViT 首先将输入图像划分为多个 $P\times P$ 大小的图像块(ViT 中称为 Patch),然后输入的序列 N 则可以被计算为:$N=H\times W/P^2$,与此同时,图像输入维度 D 可以被计算为:$D=P^2\times C$。将 $P^2\times C$ 尺寸的向量(ViT 中称为 Token)压缩为 D 大小的过程称为 Patch Embedding。为了完成最后的分类任务,ViT 参考了 BERT[12]中的分类标记,加入了一个用于分类的尺寸为 $1\times D$ 的可学习向量。此外,ViT 还加入位置编码来标记各个 Patch 的位置信息,以此来保证输入图像的位置关系。最后经过 Patch Embedding 层处理过的图像 $\boldsymbol{I}_0$ 的计算如下:

$$\boldsymbol{I}_0=[\boldsymbol{I}_{cls};\boldsymbol{I}_P^1E;\boldsymbol{I}_P^2E;\cdots;\boldsymbol{I}_P^NE]+\boldsymbol{E}_{pos} \tag{1}$$

式中:$\boldsymbol{I}_{cls}$ 是用于分类的可学习向量,$\boldsymbol{I}_P^1$ 到 $\boldsymbol{I}_P^N$ 为切分为 $P\times P$ 大小的 N 个 Patch,$\boldsymbol{E}$ 为用来做压缩的全连接层参数矩阵,大小为 $P^2\times C$,输出的大小为 D。$\boldsymbol{E}_{pos}$ 为位置编码矩阵,大小与 $\boldsymbol{E}$ 相同。与 ViT 相同,AE-ViT 拼接了大小为 $1\times D$ 的可学习的分类参数,然后按位置叠加了可学习的位置编码,位置编码的大小为 $(N+1)\times D$。最后,将得到的大小为 $(N+1)\times D$ 的二维向量经过 L(在本文实验中 L 设置为 12)次 Transformer 编码器得到大小同样为 $(N+1)\times D$ 的具有高层语义信息的图像特征。

2)Transformer 编码器，Transformer 编码器主要由归一化层、多头注意力层和全连接层构成，采用残差连接将 Encoder 模块堆叠 L 次。经过 Embedding 层处理过后的图像 Token 经过自注意力机制学习图像特征的全局长依赖关系，由此生成表征能力强的图像特征。

3)哈希模块，AE-ViT 去掉了 ViT 中的用于分类任务的分类头，加入了针对图像检索任务的哈希模块来进行哈希学习和生成二进制哈希码。ViT 为了完成分类任务，对 Transformer 编码器输出的特征只提取了第一维的分类特征来进行后续实验，而本文提出的 AE-ViT 中将所有的特征输入到哈希层，得到融合了经过卷积和 AEM 模块产生的局部显著信息和经过 ViT 的自注意力机制提取的全局语义信息的高质量二值哈希码。

本文接下来将详细介绍本文提出的注意力增强模块、用于哈希学习和生成哈希码的哈希模块以及针对图像检索设计的对比哈希损失函数。

1.1　注意力增强模块

AE-ViT 在 Embedding 层中加入注意力增强模块(attention enhanced module，AEM)，可以捕获输入特征图的局部显著信息和视觉细节，学习相应的权重以突出重要特征，以此来提升 ViT 模型对局部显著特征的处理能力并加强输入到 Transformer 编码器的图像特征质量，从而加快模型训练的收敛速度。本文提出的 AEM 如图 2 所示，其中 GAP 和 GMP 分别表示全局平均池化和全局最大池化。将输入的三维特征图记为 $\boldsymbol{F}_{in}\in\boldsymbol{R}^{H\times W\times C}$，AEM 的原理分成两步，第一步是对 $\boldsymbol{F}_{in}$ 进行注意力分布计算，经过池化处理和激活函数生成一张带有权重的注意力分布特征图。第二步是将生成的注意力特征图与 $\boldsymbol{F}_{in}$ 进行残差连接并按照矩阵位置计算输出带有权重的特征图 $\boldsymbol{F}_{out}\in\boldsymbol{R}^{H\times W\times C}$。具体的公式如下：

$$F_{out1}=f_{att}\cdot F_{in} \tag{2}$$

$$F_{out2}=F_{out1}\times F_{in}+F_{in} \tag{3}$$

式中：f_{att} 表示对输入的特征图做注意力计算，$\boldsymbol{F}_{out1}$ 表示做完注意力计算之后得到的注意力分布特征图，$\boldsymbol{F}_{out2}$ 表示将得到的注意力特征图与原始输入特征图进行残差连接后得到的输出特征图。

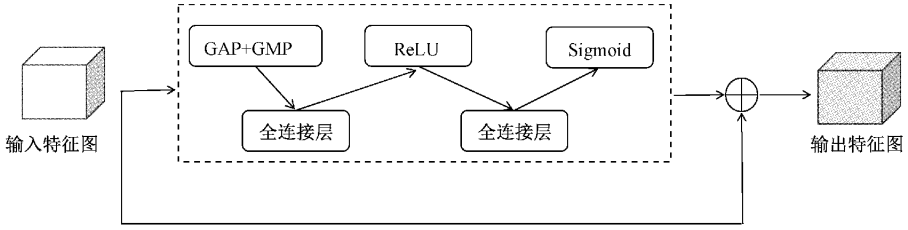


图 2　AEM 模块结构图

1.2　哈希模块

哈希模块主要包括两个全连接层、一个 Dropout 层和一个 LeakyReLU 层。首先，Transformer 编码器输出的特征经过 Dropout 层能缓解过网络拟合的情况。在本文中将 Dropout 比率设置为 0.5，使得神经网络在训练过程中随机使得一半左右的神经元暂时失效，以此来达到正则化的效果。接下来，输出的特征经过一个输出大小为 1 024 维的全连接层，然后再输入 LeakyReLU 激活函数。最后，将 1 024 维的图像特征经过最后一个全连接层转化为维度与特定哈希码长度相同的哈希特征，在模型评估阶段使用符号函数来将哈希特征转化为二进制哈希码。

此外，在激活函数的使用上，本节提出的哈希模块没有使用容易产生梯度爆炸或者梯度消失的 Sigmoid 函数和 Tanh 函数，而是使用改进的 ReLU 函数——LeakyReLU 函数，增加了网络的稀疏性并减少了计算量。

1.3　损失函数

损失函数的主要目标是使得哈希层能够学习并生成高质量的二进制哈希码，高质量指的是生成的二进制哈希码可以有效地表征图像的高级语义。受 DSH[13] 的启发，本文使用一种对比损失函数，将 RGB 空间内的相似图像在汉明空间内量化生成汉明距离较近的哈希码，反之，对不相似的图像在汉明空间内量化生成汉明距离较远的哈希码。在松弛方案上，引入正则项来使得哈希层产生的实际值逼近汉明空间的范围为 -1 或 1 的离散值，更进一步使得哈希层产生的实际值从图像的 RGB 空间映射到汉明空间时具有可微性，从而优化了网络反向传播的过程。

设 Ω 为图像空间，对于输入的图像 $\boldsymbol{I}_1,\boldsymbol{I}_2\in\Omega$，经过哈希层输出的哈希码为 $\boldsymbol{H}_1,\boldsymbol{H}_2$。设置一个标记符号 S，如果 $\boldsymbol{I}_1,\boldsymbol{I}_2$ 具有共同的标签，S 置为 1，否则将 S 置为 0，损失函数计算的公式如下：

$$L(\boldsymbol{H}_1,\boldsymbol{H}_2,S)=\frac{1}{2}(1-S)\cdot D_h(\boldsymbol{H}_1,\boldsymbol{H}_2)+\frac{1}{2}S\cdot\max(m-D_h(\boldsymbol{H}_1,\boldsymbol{H}_2),0) \tag{4}$$

$$L=\sum_{i=1}^{N}L(\boldsymbol{H}_1^i,\boldsymbol{H}_2^i,S_i) \tag{5}$$

式(4)为损失函数的计算公式，其中 D_h 为汉明距离计算函数。对于 $S=1$ 的情况，损失函数针对的是两张相似的图像的汉明距离，促进网络拉进相似图像的哈希码的汉明距离，生成更相似的哈希码。对于 $S=0$ 的情况，损失函数针对的是不相似图像的汉明距离，促进网络拉远不相似图像的汉明距离，生成差异更大的哈希码。m 为边缘阈值

参数,用来过滤计算出的汉明距离大于 m 的情况。式(5)表示总的损失函数,其中 $\boldsymbol{H}_1^i$ 表示第 i 对图像中第一张图像生成的哈希码,$\boldsymbol{H}_2^i$ 表示第 i 对图像中第二张图像生成的哈希码,S_i 表示第 i 对图像的相似情况。

式(5)虽然是一种高效的损失策略,但是却忽略神经网络直接输出的实际值需要通过阈值化函数(sigmoid 和 tanh 等)或者符号函数(sign)来产生二进制哈希码,这会使得网络的训练减慢甚至抑制网络收敛。因此,通过引入了一个正则化项来逼近汉明空间所需的离散值(+1/−1),并且使用欧氏距离来替换式(4)中汉明距离,以此来增加网络的可微性,提高网络训练的速度。

$$L_r = \sum_{i=1}^{N}\left\{\frac{1}{2}(1-S_i)\,\|\boldsymbol{H}_1^i-\boldsymbol{H}_2^i\|_2^2+\frac{1}{2}S_i\cdot\max(m-\|\boldsymbol{H}_1^i-\boldsymbol{H}_2^i\|_2^2,0)+\alpha(\|\boldsymbol{H}_1^i-sign(\boldsymbol{H}_1^i)\|_p^p+\|\boldsymbol{H}_2^i-sign(\boldsymbol{H}_2^i)\|_p^p)\right\} \quad (6)$$

式(6)为加入了正则项的损失函数,其中 $\|\cdot\|_2^2$ 和 $\|\cdot\|_p^p$ 分别表示向量的 L2 正则化和逐项正则化,α 为一个超参数,表示正则化的权重,本文参考 DSH,α 将设置为 0.1,其他变量的含义与式(4)相同。

2 实验及分析

2.1 实验数据集

本文实验使用 CIFAR-10 和 NUS-WIDE 这两个公共基准数据集来进行实验与性能评估。CIFAR-10 有 10 个种类,每个种类 6 000 张,总共 60 000 张图像。对每个种类抽取 500 张作为训练集,对每个种类随机抽样 100 张作为测试集,剩下作为图像检索数据库。

其中 NUS-WIDE 总共有 81 个类别,269 684 张图片包含 5 018 种不同的标签。实验使用包含了 21 个最常见类别的 NUS-WIDE 数据集(简称 NUS-WIDE_21),每个类别随机采样 100 张图像作为测试集,训练集从剩下图像每个种类采样 500 张图像。

2.2 评价指标

本文实验使用图像检索的常用指标平均精度(mean average precision,mAP)作为整个实验的主要评价指标,使用查准率(Precision)与查全率(Recall)曲线作为衡量检索模型性能的可视化指标。

2.3 实验设置

在实验软硬件环境上,本文使用 PyTorch 深度学习框架版本为 1.11.0,在一台配置为 RTX 3080 的服务器上进行实验,CUDA 版本为 11.3。在以上环境下,将在 ImageNet 数据集上做分类任务的 ViT 模型作为预训练模型,基于迁移学习的原理,加载预训练模型的参数,提高骨干网络的特征提取能力。此外,本文实验使用 CIFAR−10 数据集和 NUS-WIDE_21 数据集,使用 Adam 优化器,学习率设置为 10^{-5},每次在数据集上完成 150 个 epoch,每 30 个 epoch 进行一次测试和结果记录。

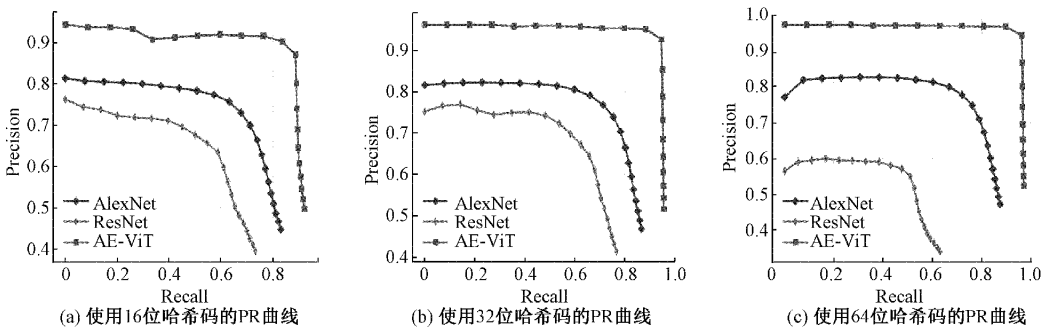
2.4 对比实验

对比实验分为以下两个部分:

1)使用经典的深度 CNN——AlexNet、ResNet 作为骨干网络分别进行实验,证明了本文提出的 AE-ViT 在图像检索任务上的有效性以及 Transformer 架构相对于纯 CNN 架构在图像检索任务上的优势。

2)将本文提出的 AE-ViT 与多种经典的深度哈希算法以及两种基于 Transformer 的图像检索算法进行对比,验证本文提出模型的性能优越性。

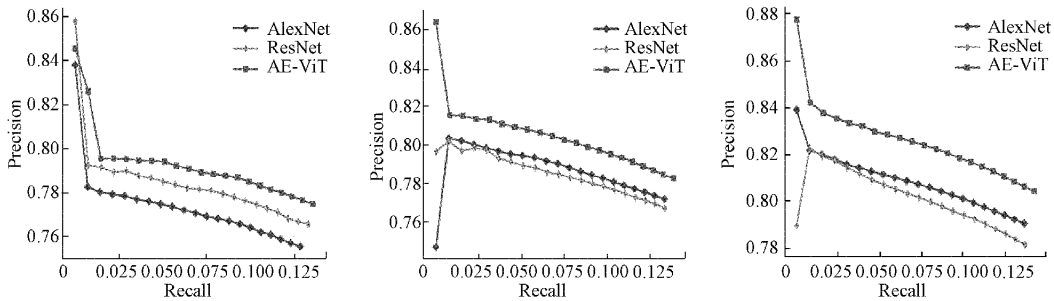
本文第 1 部分对比实验的 PR 曲线如图 3 和 4 所示,其中图 3(a)~(c)分别为在 CIFAR-10 数据集上分别在长度为 16、32 以及 64 位长度哈希码下的实验 PR 曲线图,而图 4(a)~(c)分别为在 NUS-WIDE_21 数据集上分别在长度为 16、32 以及 64 位长度哈希码下的实验 PR 曲线图,PR 曲线图越接近右上角表明检索效果越好,从曲线图中能够大致说明,本文提出的 AE-ViT 与两种经典 CNN 相比,不管是在单标签数据集在多标签数据集上,在各个哈希码长度下均取得了更好的性能。



(a) 使用16位哈希码的PR曲线 (b) 使用32位哈希码的PR曲线 (c) 使用64位哈希码的PR曲线

图 3 在不同哈希码长度下 3 种网络在 CIFAR−10 数据集上的 PR 曲线图

如表 1 所示,实验对比了本文所提出的 AE-ViT 在 CIFAR-10 和 NUS-WIDE_21 数据集上在 16、32 和 64 位哈希码长度下分别与经典的 CNN——AlexNet 和 ResNet 作为特征提取网络在图像检索上的 mAP 值对比结果。总



(a) 使用16位哈希码的PR曲线　(b) 使用32位哈希码的PR曲线　(c) 使用64位哈希码的PR曲线

图 4　在不同哈希码长度下 3 种网络在 NUS-WIDE 数据集上的 PR 曲线图

体而言，实验结果表明在其他条件相同时，本文提出的 AE-ViT 网络对检索性能的提升整体上均高于 AlexNet 和 ResNet。具体而言，AE-ViT 在 16、32 和 64 位哈希码长度下分别达到了 89.7%、95.0%和 96.5%的 mAP，相较于最差的 ResNet 网络的 mAP 分别有 25.7%、25.7% 和 42.2%的较大差距。在 NUS-WIDE_21 数据集上，也都是本文提出 AE-ViT 网络性能较好，在 16、32 和 64 位哈希码长度下，分别达到了 79.6%、81.3%和 83.6%，与各个长度下最差的结果相比，性能相差了 1.7%、1.4%和 2.1%。在本文的第 2 部分对比实验结果如表 2 所示，在两个数据集上和不同哈希码长度下，本文所提出的基于注意力增强的 Transformer 图像检索方法性能均好于基于 CNN 的深度哈希方法，具体来说 AE-ViT 在 CIFAR-10 数据集上分别在 16、32 和 64 位哈希码长度下取得了 96.2%、96.8%和 96.6%的最佳结果，在 NUS-WIDE_21 上取得了 82.4%、86.6 和 85.5%的最佳结果。更进一步，本文提出的 AE-ViT 与两种最近的提出的基于 Transformer 架构的哈希检索方法（HashFormer 与 TransHash）相比，依然在各个哈希码长度和两个数据集下都获得了大幅度的性能提升，验证了 AE-ViT 的性能优势。

表 1　不同骨干网络下不同编码长度的 mAP 结果 %

骨干网络	CIFAR-10			NUS-WIDE_21		
	16 位	32 位	64 位	16 位	32 位	64 位
AE-ViT	**96.2**	**96.8**	**96.6**	**82.4**	**86.8**	**85.5**
AlexNet	74.1	77.9	79.2	77.9	79.9	81.7
ResNet	64.0	69.3	54.3	79.1	80.9	81.5

表 2　AE-ViT 与其他方法在不同数据集以及不同哈希码长度下的 mAP　%

不同检索模型	CIFAR-10			NUS-WIDE_21		
	16 位	32 位	64 位	16 位	32 位	64 位
AE-ViT（本文方法）	**96.2**	**96.8**	**96.6**	**82.4**	**86.8**	**85.5**
DSH	61.5	68.1	69.1	63.4	65.1	68.6
DHN[14]	65.4	67.1	67.4	64.7	67.3	70.3
HashNet	51.5	62.8	68.3	68.2	69.5	73.4
DCH[15]	66.8	69.4	67.8	70.4	71.8	70.6
IDHN[16]	54.2	57.0	59.7	70.0	71.5	72.6
DPN[17]	82.5	83.8	82.9	—	—	—
Hash Former	91.2	91.7	92.4	73.2	74.2	76.0
Trans Hash	90.1	91.1	91.2	73.6	73.9	74.9

3　结　　论

本文提出了一种基于注意力增强的视觉 Transformer 图像检索算法——AE-ViT。在 AE-ViT 中设计了一个注意力增强模块（AEM），用于提取输入特征图的局部显著信息并分配相应的权重以突出重要特征。通过设计的 AEM 加强了输入到 Transformer 编码器的图像特征，提高了图像检索的精度。在实验阶段，本文在不同的哈希码长度下，分别在单标签数据集和多标签数据集上，首先将 AE-ViT 与经典的深度 CNN——AlexNet 和 ResNet 进行对比，将其分别作为骨干网络在不同的数据集和不同长度的哈希码进行对比实验，然后还对比了多种经典的基于 CNN 的深度哈希检索方法和两种基于 Transformer 的哈希检索方法。通过实验验证了本文提出的 AE-ViT 在图像检索任务上的有效性，并表明了基于 Transformer 架构的检索网络相对基于纯 CNN 架构的网络在图像检索任务上的性能优势。在未来的研究中，可以进一步对 ViT 模型和哈希损失进行改进或在更广泛的数据集上进行实验，以取得更好的检索效果。

参考文献

[1] XIA R, PAN Y, LAI H, et al. Supervised hashing for image retrieval via image representation learning: Proceedings of the AAAI conference on artificial intelligence [C]. 2014. DOI: https://doi.org/10.1609/aaai.v28i1.8952.

[2] CAO Z, LONG M, WANG J, et al. Hashnet: Deep learning to hash by continuation[C]. Proceedings of the IEEE international conference on computer vision,

2017. DOI: https://doi.org/10.48550/arXiv.1702.00758.

[3] SU S, ZHANG C, HAN K, et al. Greedy hash: Towards fast optimization for accurate hash coding in CNN [C]. Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018: 806-815. DOI: 10.5555/3326943.3327018.

[4] YUAN L, WANG T, ZHANG X, et al. Central similarity quantization for efficient image and video retrieval [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 3083-3092. DOI: 10.1109/CVPR42600.2020.00315.

[5] HAN K, WANG Y, CHEN H, et al. A survey on vision transformer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(1): 87-110.

[6] 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. DOI: https://doi.org/10.48550/arXiv.2010.11929.

[7] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV), 2021: 10012-10022. DOI: 10.1109/ICCV48922.2021.00986.

[8] GKELIOS S, BOUTALIS Y, CHATZICHRISTOFIS S A. Investigating the vision transformer model for image retrieval tasks [C]. 2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS), IEEE, 2021: 367-373. DOI: 10.48550/arXiv.2101.03771.

[9] EL-NOUBY A, NEVEROVA N, LAPTEV I, et al. Training vision transformers for image retrieval[J]. ArXiv Preprint, 2021, ArXiv: 2102.05644. DOI: https://doi.org/10.48550/arXiv.2102.05644.

[10] LI T, ZHANG Z, PEI L, et al. HashFormer: Vision transformer based deep hashing for image retrieval[J]. IEEE Signal Processing Letters, 2022,29:827-831.

[11] CHEN Y, ZHANG S, LIU F, et al. Transhash: Transformer-based hamming hashing for efficient image retrieval [C]. Proceedings of the 2022 International Conference on Multimedia Retrieval, 2022: 127-136. DOI: https://doi.org/10.48550/arXiv.2105.01823.

[12] DEVLIN J, CHANG M, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. ArXiv Preprint, 2018, ArXiv: 1810.04805. DOI: https://doi.org/10.48550/arXiv.1810.04805.

[13] 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, 2016: 2064-2072. DOI: 10.1109/CVPR.2016.227.

[14] ZHU H, LONG M, WANG J, et al. Deep hashing network for efficient similarity retrieval [C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2016, 30(1). DOI: https://doi.org/10.1609/aaai.v30i1.10235.

[15] CAO Y, LONG M, LIU B, et al. Deep cauchy hashing for hamming space retrieval[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 1229-1237. DOI: 10.1109/CVPR.2018.00134.

[16] ZHANG Z, ZOU Q, LIN Y, et al. Improved deep hashing with soft pairwise similarity for multi-label image retrieval [J]. IEEE Transactions on Multimedia, 2019,22(2):540-553.

[17] FAN L, NG K W, JU C, et al. Deep polarized network for supervised learning of accurate binary hashing codes [C]. IJCAI, 2020: 825-831. DOI: https://doi.org/10.24963/ijcai.2020/115.

作者简介

刘华咏,博士,副教授,主要研究方向为多媒体检索。

黄聪(通信作者),硕士研究生,主要研究方向为计算机视觉与深度学习。

金汉均,博士,教授,主要研究方向为图像处理,语义理解,计算机视觉分析以及深度学习。