中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2023)02-0535-10

论文引用格式:He C and Wei H X. 2023. Image retrieval based on transformer and asymmetric learning strategy. Journal of Image and Graphics, 28(02):0535-0544(贺超,魏宏喜. 2023. 结合 Transformer 与非对称学习策略的图像检索. 中国图象图形学报,28(02):0535-0544)[DOI:10.11834/jig.210842]

结合 Transformer 与非对称学习策略的图像检索

贺超,魏宏喜*

1. 内蒙古大学计算机学院,呼和浩特　010010;2. 内蒙古自治区蒙古文信息处理技术重点实验室,呼和浩特　010010;
3. 蒙古文智能信息处理技术国家地方联合工程研究中心,呼和浩特　010010

摘　要:**目的**　图像检索是计算机视觉领域的一项基础任务,大多采用卷积神经网络和对称式学习策略,导致所需训练数据量大、模型训练时间长、监督信息利用不充分。针对上述问题,本文提出一种 Transformer 与非对称学习策略相结合的图像检索方法。**方法**　对于查询图像,使用 Transformer 生成图像的哈希表示,利用哈希损失学习哈希函数,使图像的哈希表示更加真实。对于待检索图像,采用非对称式学习策略,直接得到图像的哈希表示,并将哈希损失与分类损失相结合,充分利用监督信息,提高训练速度。在哈希空间通过计算汉明距离实现相似图像的快速检索。**结果**　在 CIFAR-10 和 NUS-WIDE 两个数据集上,将本文方法与主流的 5 种对称式方法和性能最优的两种非对称式方法进行比较,本文方法的 mAP(mean average precision)比当前最优方法分别提升了 5.06% 和 4.17%。**结论**　本文方法利用 Transformer 提取图像特征,并将哈希损失与分类损失相结合,在不增加训练数据量的前提下,减少了模型训练时间。所提方法性能优于当前同类方法,能够有效完成图像检索任务。

关键词:图像检索;Transformer;哈希函数;非对称式学习;哈希损失;分类损失

Image retrieval based on transformer and asymmetric learning strategy

He Chao, Wei Hongxi*

1. School of Computer Science, Inner Mongolia University, Hohhot 010010, China; 2. Provincial Key Laboratory of Mongolian Information Processing Technology, Hohhot 010010, China; 3. National and Local Joint Engineering Research Center of Mongolian Information Processing Technology, Hohhot 010010, China

Abstract: **Objective** Image retrieval is one of the essential tasks in computer vision. Most of deep learning-based image retrieval methods are implemented via learning-symmetrical strategy. Training images are profiled into a pair and then melt into convolutional neural network (CNN) for features extraction. Similar loss is utilized to learn the hash-relevant images. Consequently, a feasible performance can be obtained through this symmetric learning scheme. In recent years, to improve its performance better, a mass of the CNNs-improved are extended horizontally or deepened vertically. CNNs-based structure is complicated and time-consuming for large scale image datasets. Recently, the Transformer has been developing to the domain of computer vision and its image classification ability is improved intensively. Our Transformer-relevant research

收稿日期:2021-09-23;**修回日期**:2022-02-14;**预印本日期**:2022-02-21

*通信作者:魏宏喜　cswhx@imu.edu.cn

基金项目:内蒙古自治区自然科学基金重大项目(2019ZD14);内蒙古自治区科技计划项目(2019GG281);内蒙古自治区高等学校青年科技英才支持计划项目(NJYT-20-A05)

Supported by: Natural Science Foundation of Inner Mongolia Autonomous Region, China (2019ZD14); Project for Science and Technology of Inner Mongolia Autonomous Region, China (2019GG281); Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (NJYT-20-A05)

is melted into the task of large-scale image retrieval method because the Transformer is feasible for large scale dataset like ImageNet-21k and JFT-300M. For symmetric methods, the overall image dataset-related should be involved in the training phase. Meanwhile, query images have to be added into a pair for training, which results in the problem of time-consuming. The hash function between the training image and the query image is learnt in terms of similar calculation. The information-supervised is used for the similarity matrix only, which is insufficient to be used. A learning-asymmetric scheme is optimized for training based on some images-selected only. Furthermore, the corresponding hash function can be learnt from the hash loss. For the rest of images, their feature representations can be picked out as well. Moreover, classification constraints can be used for the query images and the corresponding classification loss can be optimized in terms of learning-altered technology. To resolve the problems of time-consuming and the insufficient information-supervised, we develop a deep supervised hash image retrieval method in terms of the integrated Transformer and learning-asymmetric strategy. **Method** For the training images, a Transformer-designed can be utilized to generate the hash representation of these images. The hash loss is used to guarantee the hash representation of images is closer to the real hash value. In the original Transformer, its input is derived of one-dimensional data. First, each image is required to be divided into multiple blocks. Then, each block is mapped into one-dimensional vector. At last, to form the one-dimensional vector, these one-dimensional vectors of all blocks of one image are concatenated together. Our Transformer is designed and composed of 1) two normalization layers, 2) a multi-head attention module, 3) a fully-connected module, and 4) a hash layer. First, the input one-dimensional vector is penetrated into the normalization layer. Next, the output of the normalization layer is fed into the multi-head attention layer of those are 16 heads. Hence, multiple local features of images can be captured. Additionally, the residual link is developed to integrate the initial one-dimensional vector with the output of the multi-head attention layer. By this way, the global features of images can be preserved better. Finally, the representation vector of each image can be obtained through the fully-connected module and the hash layer. In this study, the process for generating representation vectors mentioned above will be replicated for 24 times. For the rest of images, the classification loss is taken as a constraint, which can be used to learn the hash representation of these images in an asymmetric way. To improve the training efficiency, supervised information can be used effectively. The query images are not required to be involved in. The model can be trained via learning-alternated technology. Specifically, first, hash representation of the query images and the weights of the classification are configured and initialized randomly. Then, the parameters of the model can be optimized in terms of stochastic gradient descent algorithm. Following by an epoch of training, the weights of the classification can be balanced by the trained model. Meanwhile, the hash representation ability of the rest of images is improved gradually. In this manner, the hash code of the rest images can be obtained from a well-trained model directly, which optimizes the training efficiency. Finally, our method can realize fast retrieval of similar images through the calculation of the Hamming distance in the hash space. **Result** In our experiment, our method is compared to five categories of symmetric methods and two sorts of asymmetric methods based on two large scale image retrieval datasets. The performance of the proposed method is increased by 5.06% and 4.17% of each on the two datasets (i.e., CIFAR-10 and NUS-WIDE). The evaluation metric is based on mean average precision (mAP). Ablation experiments validate that the classification loss can make the images closer to the real hash representation. Furthermore, the hyper-parameters of the classification loss are tested as well, and the appropriate hyper-parameters are obtained. **Conclusion** To complete the image retrieval task effectively, our Transformer-based method has its potentials on image features extraction for large scale image retrieval, and the hash loss can be melted into classification loss for model training further.

Key words: image retrieval; Transformer; hash function; asymmetric learning; hash loss; classification loss

0 引 言

在大数据时代,互联网等媒介中的图像数量呈指数级增长,如何从海量图像中快速查找出所需图像是值得研究的问题。近似最近邻(approximate nearest neighbor,ANN)搜索(Andoni 和 Indyk,2008;Zhang 等,2010)是解决此问题的常用方法。作为一种广泛使用的近似最近邻搜索技术,哈希方法(Wu 等,2019;Guo 等,2017,万方 等,2021)的目的是将

原始空间上的数据点映射到汉明空间上的离散 2 进制哈希码。同时,原始空间上的相似性也保留在汉明空间上。哈希码具有存储空间占用小、计算速度快等优点,广泛应用于图像的大规模检索。

哈希方法分为与数据无关的方法和与数据相关的方法两种(刘颖 等,2020)。前者将图像随机投影在特征空间中,用 2 进制方法生成哈希码(Gionis 等,1999)。后者通过机器学习方法学习哈希函数,将图像特征映射为 2 进制代码(Lai 等,2015)。对于依赖数据的方法,维数较少的哈希码可以很好地表示图像特征并获得更好的结果,因此与数据相关的哈希方法在现实中广泛使用。与数据相关的哈希方法分为有监督的哈希方法和无监督的哈希方法。最具代表性的无监督哈希方法是迭代量化(iterative quantization,ITQ)(Gong 等,2013),它根据给定的训练样本,通过迭代投影和阈值法对投影矩阵进行优化。为了更好地利用语义标记进行特征表示学习,研究人员提出了监督学习方法。监督学习方法分为非深度监督哈希和深度监督哈希。传统的非深度监督哈希算法利用图像的颜色和纹理学习哈希函数。具有代表性的非深度监督哈希方法有核监督哈希法(supervised hashing with kernels, KSH)(Liu 等,2012)、隐因子哈希法(latent factor hashing, LFH)(Zhang 等,2014)和快速监督哈希法(fast supervised hashing,FastH)(Lin 等,2014)等。随着深度学习的发展,哈希函数的学习嵌入到深度学习框架中。

随着卷积神经网络(convolutional neural networks,CNN)在图像分类、目标检测和文字识别等方面的优异表现,许多基于 CNN 的深度监督哈希方法相继提出。基于卷积神经网络的哈希算法(convolutional neural networks based hashing,CNNH)(Xia 等,2014)是利用深度神经网络学习哈希码的早期工作之一。在 CNNH 的基础上,网中网哈希(network in network hashing,NINH)(Lai 等,2015)提出了一种利用三重排序损失保持相对相似性的深层结构。之后,一些基于成对标签的深度哈希方法应用于相关任务中。深度成对监督哈希(deep pairwise supervised hashing,DPSH)(Li 等,2016)在使用成对标签的同时学习特征表示和哈希函数。深度哈希网络(deep hashing network,DHN)(Zhu 等,2016)优化了语义相似性损失和更紧凑的哈希码量化损失。为了更好地利用标签信息,深度监督离散哈希(deep supervised discrete hash,DSDH)(Li 等,2017)综合了语义相似性损失和分类损失。对称式深度监督哈希方法目前取得了相对较好的效果,如果想进一步提升性能,可以使用更大规模的卷积神经网络。但是大规模模型的训练周期较长,并且待检索图像参与网络训练也会增加训练时间。为了解决上述问题,Jiang 和 Li(2018)提出了非对称深度监督哈希(asymmetric deep supervised hashing,ADSH),这是第 1 个以非对称方式学习训练图像和待检索图像哈希码的基于卷积神经网络实现的算法,待检索图像哈希码可直接由训练图像哈希码计算得到,这使得训练效率大幅提升。基于联合学习的深度监督哈希算法(joint learning based deep supervised hashing,JLDSH)(Gu 等,2020)以 ADSH 为基础,将分类损失与哈希损失结合起来,更加充分利用了监督信息。

一些研究人员尝试将 Transformer(Vaswani 等,2017)应用到计算机视觉领域。Transformer 在自然语言处理任务中有着十分出色的表现,在机器翻译和语音识别等领域得到了广泛应用。主流方法大都是在大型文本语料库上进行预训练,然后在较小的特定任务数据集上进行微调。Transformer 具有出色的计算效率和扩展性,使得训练规模庞大的模型成为可能。这些特性使得将 Transformer 应用于计算机视觉领域的研究变得流行起来。VIT(vision transformer)(Dosovitskiy 等,2021)在图像分类领域有着突出表现,它将图像切成小块输入到 Transformer 中,并加入图像块的位置信息和分类位,经过 Transformer 编码器可以对图像进行分类,在海量的训练数据下,达到了很高的分类正确率,性能超过现有的分类模型。图像处理 Transformer(image processing transformer,IPT)(Chen 等,2021a)可以应用于图像去噪和超分辨率等图像处理任务,在大规模数据集上训练的该模型可在上述任务中获得最先进的性能。基于 Transformer 的端到端目标检测网络(detection transformer,DETR)(Carion 等,2020)使用 2 进制匹配和 Transformer 编码器—解码器进行预测,极大简化了目标检测过程。

受 VIT 和 ADSH 的启发,本文提出了一种基于 Transformer 的非对称监督深度哈希方法(asymmetric deep hashing method based on transformer,ADSHT)。本文的主要贡献如下:1)使用 Transformer 生成图像

的哈希表示,结合非对称学习策略将 Transformer 应用到大规模图像检索任务中,在提高训练效率的同时,提升了检索性能。2)为了更加充分地利用监督信息,将哈希损失与分类损失相结合,使模型能够更好地学习哈希函数,使图像的哈希表示更加真实。3)在两个公开数据集 CIFAR-10(Krizhevsky,2009)(该数据集为单标签数据集)和 NUS-WIDE(Chua 等,2009)(该数据集为多标签数据集)上进行实验,通过与主流的对称式方法和最优的非对称式方法进行比较,验证了所提出方法的有效性。

1 问题定义

将 m 个查询图像标记为 $\boldsymbol{X}=\{\boldsymbol{x}_i\}_{i=1}^{m}$,$n$ 个待检索图像表示为 $\boldsymbol{Y}=\{\boldsymbol{y}_j\}_{j=1}^{n}$,成对监督的相似信息记为 $\boldsymbol{S}=\{-1,+1\}^{m\times n}$。$S_{ij}=1$ 说明 $\boldsymbol{x}_i$ 和 $\boldsymbol{y}_j$ 是相似的,反之,$S_{ij}=-1$ 则代表不相似。深度监督哈希利用监督信息从查询图像和待检索图像中学习哈希函数,两幅相似图像中至少包含 1 个相同的标签。$\boldsymbol{R}=\{\boldsymbol{r}_j\}_{j=1}^{n}\in\{-1,+1\}^{n\times c}$ 表示待检索图像的哈希码,$\boldsymbol{Q}=\{\boldsymbol{q}_i\}_{i=1}^{m}\in\{-1,+1\}^{m\times c}$ 表示查询图像的哈希码,其中 c 表示哈希码的长度。本文方法通过从整个待检索图像中获取查询图像来学习哈希函数,在这种情况下,$\boldsymbol{X}\subseteq\boldsymbol{Y}$。

2 提出方法

2.1 模型架构

本文提出的 ADSHT 模型主要包括图像块嵌入部分、特征提取部分(Dosovitskiy 等,2021)和损失函数部分(Gu 等,2020),结构图如图 1 所示。特征提取模块是提取图像的特征,并将特征转换为哈希编码表示。损失函数部分是使图像特征更接近真实的哈希码,并保持查询图像与待检索图像的相似性。原始的 Transformer 只能处理 1 维数据。为了能够处理 2 维图像,本文将图像 $\boldsymbol{x}\in\mathbf{R}^{H\times W\times C}$ 重塑为一系列扁平的 2 维图像块 $\boldsymbol{x}_p\in\mathbf{R}^{N\times(P^2\times C)}$。$H$、$W$ 和 C 为原始图像的高度、宽度和通道数,(P,P) 是每个图像块的高度和宽度,p 为图像的序号。$N=HW/P^2$ 是 Transformer 输入的序列长度。

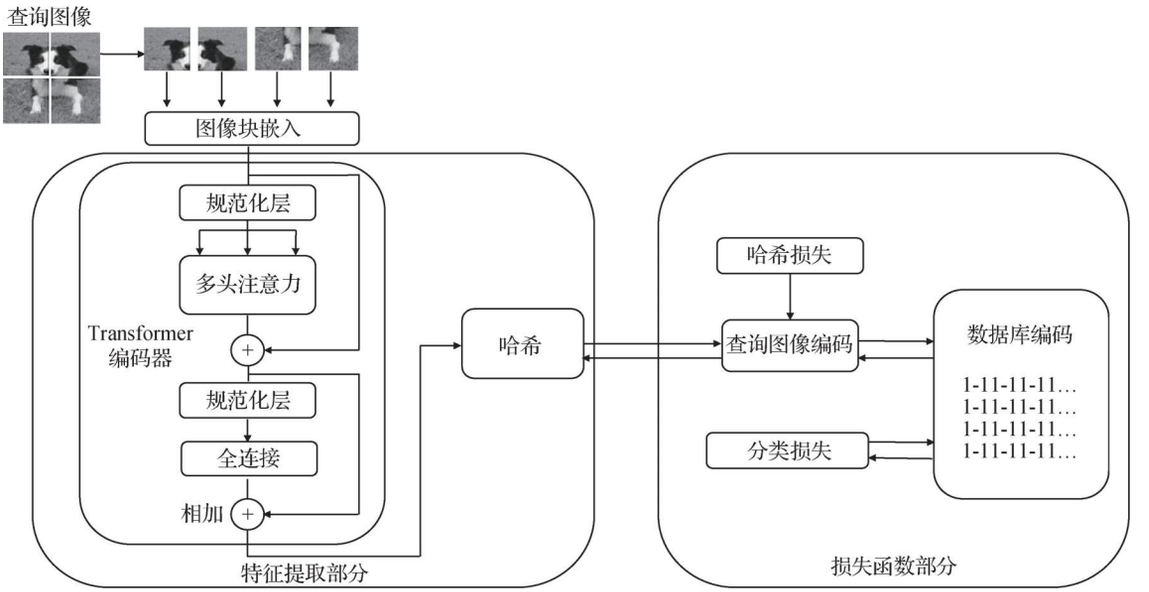


图 1 ADSHT 模型结构图

Fig. 1 Architecture of ADSHT

本文使用线性投影层 $\boldsymbol{E}\in\mathbf{R}^{(P^2\times C)\times D}$ 将每个图像块向量映射到 D 维空间。随后,拼接所有图像块向量,使之成为一个完整的图像向量,并在该图像向量第 1 位加入分类位向量 $\boldsymbol{x}_{\text{class}}$。最后,将每个图像块的位置信息 $\boldsymbol{E}_{\text{pos}}\in\mathbf{R}^{(N+1)\times D}$ 与图像块向量 $\boldsymbol{x}_p$ 相加,最终得到嵌入的图像向量 $\boldsymbol{s}_q\in\mathbf{R}^{(N+1)\times D}$。上述过程称为图像块嵌入,如图 2 所示。

本文方法的特征提取部分使用 Transformer 的编码器模块,并在该模块后加入哈希模块。编码器模块在 ImageNet1k 数据集上进行预训练,该模块由

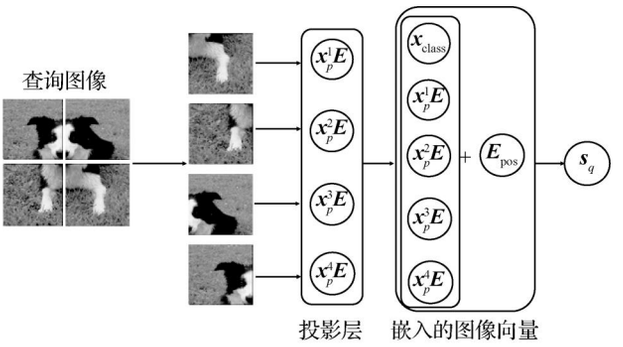


图 2　图像块嵌入

Fig. 2　Patches embedding

多层交替的多头注意力模块(multi-head self-attention,MSA)(Vaswani 等,2017)和全连接模块(fully-connected network,MLP)组成,具体为

$$\boldsymbol{s}_q^* = MSA(LN(\boldsymbol{s}_{(q-1)})) + \boldsymbol{s}_{(q-1)} \quad q = 1,2,\cdots,B \tag{1}$$

$$\boldsymbol{s}_q = MLP(LN(\boldsymbol{s}_q^*)) + \boldsymbol{s}_q^* \quad q = 1,2,\cdots,B \tag{2}$$

式中,B 表示编码器模块的层数。在每个模块之前使用规范化层(layernorm,LN)(Wang 等,2019),在每个模块之后均使用残差连接。最后通过哈希模块 *HASH* 输出图像的哈希表示 $\boldsymbol{h} \in \mathbf{R}^c$,具体为

$$\boldsymbol{h} = HASH(LN(\boldsymbol{s}_B^0)) \tag{3}$$

本文用最后一个全连接模块输出的向量的分类位 $\boldsymbol{s}_B^0$ 作为提取到的图像特征,输入到哈希模块中获取图像的哈希表示。全连接模块包括两个具有 GELU(gaussian error linear units)激活函数的非线性的层。哈希模块包括一个具有 TANH(hyperbolic tangent)激活函数的非线性的层。

传统的对称式学习方法只利用监督信息获得相似矩阵,很少关注监督信息对图像的分类作用。本文对检索图像添加分类约束,充分利用监督信息,使待检索图像得到更真实的哈希表示。为了使损失函数最小化,在待检索图像哈希码与查询图像哈希码均不更新的情况下,对损失函数求导,计算分类损失函数参数。模型结束训练时,会输出查询图像的哈希码和分类损失函数的参数,令上述两个参数不更新,计算待检索图像的哈希码。

2.2　损失函数

为了更好地学习能够保持查询图像和待检索图像相似性的哈希码,可以优化监督信息的相似度矩阵与查询—待检索图像哈希码对的内积之间的损失。损失函数(Jiang 和 Li,2018)定义为

$$\min_{U,V} J(\boldsymbol{Q},\boldsymbol{R}) = \sum_{i=1}^{m}\sum_{j=1}^{n}(\boldsymbol{q}_i^{\mathrm{T}}\boldsymbol{r}_j - cS_{ij})^2 \quad \text{s.t. } \boldsymbol{Q} \in \{-1,+1\}^{m\times c}, \boldsymbol{R} \in \{-1,+1\}^{n\times c} \tag{4}$$

式中,$\boldsymbol{q}_i = h(\boldsymbol{x}_i)$ 很难学习,它分布是离散的,因此设置 $h(\boldsymbol{x}_i) = \mathrm{sign}(F(\boldsymbol{x}_i;\boldsymbol{\Theta}))$,$F(\boldsymbol{x}_i;\boldsymbol{\Theta}) \in \mathbf{R}$,而 $h(\boldsymbol{x}_i)$ 分布依然是离散的,很难通过反向传播来优化网络参数。其中,$F(\boldsymbol{x}_i;\boldsymbol{\Theta})$ 是 Transformer 编码器网络的输出,$\boldsymbol{\Theta}$ 是该网络的参数。因此,本文使用 $\tanh(F(\boldsymbol{x}_i;\boldsymbol{\Theta}))$ 代替 $h(\boldsymbol{x}_i)$ 来解决上述问题。$\boldsymbol{\xi} = \{1,2,\cdots,n\}$ 表示所有待检索图像的索引值,$\boldsymbol{\Omega} = \{1,2,\cdots,m\} \subseteq \boldsymbol{\xi}$ 表示所有查询图像的索引值。损失函数(Jiang 和 Li,2018)为

$$\min_{\boldsymbol{\Theta},\boldsymbol{R}} J(\boldsymbol{\Theta},\boldsymbol{R}) = \sum_{i\in\boldsymbol{\Omega}}\sum_{j\in\boldsymbol{\xi}}(\tanh(F(\boldsymbol{x}_i;\boldsymbol{\Theta}))^{\mathrm{T}}\boldsymbol{r}_j - cS_{ij})^2 \quad \text{s.t. } \boldsymbol{Q} \in \{-1,+1\}^{m\times c}, \boldsymbol{R} \in \{-1,+1\}^{n\times c} \tag{5}$$

由于 $\boldsymbol{X} \subseteq \boldsymbol{Y}$,$\boldsymbol{x}_i$ 有两种表示,分别是待检索图像中的 2 进制代码 $\boldsymbol{r}_j$ 和查询图像中的 2 进制代码 $\tanh(F(\boldsymbol{x}_i;\boldsymbol{\Theta}))$。本文添加另一个约束来减少它们之间的差异。新的损失函数(Jiang 和 Li,2018)定义为

$$\min_{\boldsymbol{\Theta},\boldsymbol{R}} J(\boldsymbol{\Theta},\boldsymbol{R}) = \sum_{i\in\boldsymbol{\Omega}}\sum_{j\in\boldsymbol{\xi}}[\tanh(F(\boldsymbol{x}_i;\boldsymbol{\Theta}))^{\mathrm{T}}\boldsymbol{r}_j - cS_{ij}]^2 + \gamma\sum_{i\in\boldsymbol{\Omega}}[\boldsymbol{r}_i - \tanh(F(\boldsymbol{y}_i;\boldsymbol{\Theta}))]^2 \quad \text{s.t. } \boldsymbol{Q} \in \{-1,+1\}^{m\times c}, \boldsymbol{R} \in \{-1,+1\}^{n\times c} \tag{6}$$

式中,γ 是超参数。为了更好地利用监督信息,本文采用简单的线性分类器对检索图像的哈希码和对应的标签信息进行建模,利用标签信息约束待检索图像的哈希码的学习,让学习到的待检索图像哈希码更接近真实的哈希码。令 $\boldsymbol{L} = \{\boldsymbol{i}_1,\boldsymbol{i}_2,\boldsymbol{i}_3,\cdots,\boldsymbol{i}_n\}$ 表示真实图像标签信息的 One-hot 编码。$\boldsymbol{i}_i \in \{0,1\}^h$,$i = 1,\cdots,n$ 是每幅图像的标签信息组成的向量,h 表示类别的数量。元素 1 表示图像包含此标签,元素 0 则相反。用 $\boldsymbol{W}$ 表示分类器的权重,将分类损失和哈希损失结合起来。最终的损失函数(Gu 等,2020)为

$$\min_{\theta,\boldsymbol{R},\boldsymbol{W}} J(\theta,\boldsymbol{R},\boldsymbol{W}) = \sum_{i\in\boldsymbol{\Omega}}\sum_{j\in\boldsymbol{\xi}}(\tanh(F(\boldsymbol{x}_i;\theta))^{\mathrm{T}}\boldsymbol{r}_j - cS_{ij})^2 +$$

$$\gamma \sum_{i\in\Omega}\left[\boldsymbol{r}_i-\tanh\left(F\left(\boldsymbol{x}_i;\boldsymbol{\Theta}\right)\right)\right]^2+\mu\sum_{i\in\Omega}\left\|\boldsymbol{i}_i-\boldsymbol{r}_i\boldsymbol{W}\right\|_{\mathrm{F}}^2+\varphi\left\|\boldsymbol{W}\right\|_{\mathrm{F}}^2$$
$$\text{s.t. } \boldsymbol{Q}\in\{-1,+1\}^{m\times c},\boldsymbol{R}\in\{-1,+1\}^{n\times c} \tag{7}$$

式中，μ 和 φ 为超参数。

2.3 优化方法

2.3.1 优化参数 $\boldsymbol{\Theta}$

当 $\boldsymbol{R}$ 和 $\boldsymbol{W}$ 不改变时，令 $\boldsymbol{z}_i=F(\boldsymbol{x}_i;\boldsymbol{\Theta})$，$\boldsymbol{u}_i=\tanh(F(\boldsymbol{x}_i;\boldsymbol{\Theta}))$，$\boldsymbol{z}_i$ 的梯度计算为

$$\frac{\partial J}{\partial \boldsymbol{z}_i}=\left[2\sum_{j\in\xi}\left(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{r}_i-cS_{ij}\right)\boldsymbol{r}_i+2\gamma\left(\boldsymbol{r}_i-\boldsymbol{u}_i\right)\right]\odot\left(\boldsymbol{1}-\boldsymbol{u}_i^2\right) \tag{8}$$

式中，$\odot$ 是哈达玛积（Hadamard）（Jiang 和 Li，2018）。随后，根据链推导规则基于 $\frac{\partial J}{\partial \boldsymbol{z}_i}$ 计算 $\frac{\partial J}{\partial \boldsymbol{\Theta}}$，通过反向传播（backpropagation，BP）算法更新网络参数 $\boldsymbol{\Theta}$。

2.3.2 优化参数 $\boldsymbol{W}$

当 $\boldsymbol{\Theta}$ 和 $\boldsymbol{R}$ 不改变时，式（7）变为

$$\min_{\boldsymbol{W}} J(\boldsymbol{W})=\mu\sum_{i\in\Omega}\left\|\boldsymbol{i}_i-\boldsymbol{r}_i\boldsymbol{W}\right\|_{\mathrm{F}}^2+\varphi\left\|\boldsymbol{W}\right\|_{\mathrm{F}}^2 \tag{9}$$

这是一个具有封闭解的最小化问题，对其求解，解得 $\boldsymbol{W}$ 为

$$\boldsymbol{W}=\left(\boldsymbol{r}_i^{\mathrm{T}}\boldsymbol{r}_i+\frac{\varphi}{\mu}\boldsymbol{I}\right)^{-1}\boldsymbol{r}_i^{\mathrm{T}}\boldsymbol{L} \tag{10}$$

式中，$\boldsymbol{L}$ 代表真实图像标签信息的 One-hot 编码。

2.3.3 优化参数 $\boldsymbol{R}$

当 $\boldsymbol{\Theta}$ 和 $\boldsymbol{W}$ 不改变时，式（7）变为

$$\min_{\boldsymbol{R}} J(\boldsymbol{R})=\left\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{R}-cS_{ij}\right\|_{\mathrm{F}}^2+\gamma\left\|\boldsymbol{R}-\boldsymbol{U}\right\|_{\mathrm{F}}^2+\mu\left\|\boldsymbol{L}-\boldsymbol{R}\boldsymbol{W}\right\|_{\mathrm{F}}^2 \tag{11}$$

式中，$\boldsymbol{U}=\{\boldsymbol{u}_j\}_{j=1}^n\in\{-1,+1\}^{n\times c}$，然后，逐位更新 $\boldsymbol{R}$。也就是说，每更新 $\boldsymbol{R}$ 的一列时，其他的列是不更新的。令 $\boldsymbol{U}_{*j}$ 表示 $\boldsymbol{U}$ 的第 j 列，$\hat{\boldsymbol{U}}_j$ 表示 $\boldsymbol{U}$ 的矩阵但不包括 $\boldsymbol{U}_{*j}$。$\boldsymbol{R}_{*j}$ 表示 $\boldsymbol{R}$ 的第 j 列，$\hat{\boldsymbol{R}}_j$ 表示 $\boldsymbol{R}$ 的矩阵但不包括 $\boldsymbol{R}_{*j}$。$\boldsymbol{O}_{*j}$ 表示 $\boldsymbol{O}$ 的第 j 列，$\hat{\boldsymbol{O}}_j$ 表示 $\boldsymbol{O}$ 的矩阵但不包括 $\boldsymbol{O}_{*j}$。$\boldsymbol{W}_{*j}$ 表示 $\boldsymbol{W}$ 的第 j 列，$\hat{\boldsymbol{W}}_j$ 表示 $\boldsymbol{W}$ 的矩阵但不包括 $\boldsymbol{W}_{*j}$。这里 $\boldsymbol{O}=-2c\boldsymbol{S}^{\mathrm{T}}\boldsymbol{U}-2\mu\boldsymbol{L}\boldsymbol{W}^{\mathrm{T}}-2\gamma\overline{\boldsymbol{U}}$。其中，$\overline{\boldsymbol{U}}=\{\overline{\boldsymbol{u}}_j\}_{j=1}^n\in\{-1,+1\}^{n\times c}$，当 $j\in\Omega$ 时，$\overline{\boldsymbol{u}}_j=\overline{\boldsymbol{u}}_j$，否则 $\overline{\boldsymbol{u}}_j=0$。可以得出

$$\boldsymbol{R}_{*j}=-\operatorname{sign}\left(2\hat{\boldsymbol{R}}_j\hat{\boldsymbol{U}}_j^{\mathrm{T}}\boldsymbol{U}_{*j}+2\mu\hat{\boldsymbol{R}}_j\hat{\boldsymbol{W}}_j\boldsymbol{W}_{*j}^{\mathrm{T}}+\boldsymbol{O}_{*j}\right) \tag{12}$$

式（12）用来更新 $\boldsymbol{R}$。最后可以得到整个数据库对应的哈希码。

2.4 学习算法

ADSHT 学习算法的具体步骤如下：

输入：n 个数据点 $\boldsymbol{Y}=\{\boldsymbol{y}_j\}_{j=1}^n$，监督相似矩阵 $\boldsymbol{S}\in\{-1,+1\}^{n\times n}$，分类损失参数 $\boldsymbol{W}\in\mathbf{R}^{c\times h}$，哈希码的长度 c。

输出：待检索图像的哈希码 $\boldsymbol{R}$，神经网络参数 $\boldsymbol{\Theta}$。

1）初始化 $\boldsymbol{\Theta}$、$\boldsymbol{R}$ 和 $\boldsymbol{W}$，批处理大小 M，迭代次数 T_{out} 和 T_{in}。

2）FOR $i\rightarrow T_{\mathrm{out}}$ DO；

（1）从待检索图像中随机选取查询图像 $\boldsymbol{X}=\boldsymbol{Y}^{\Omega}$，并获取查询图像的相似矩阵 $\boldsymbol{S}^{\Omega}$。

（2）FOR $j\rightarrow T_{\mathrm{in}}$ DO；

FOR $s=1,2,\cdots,m/M$ DO；

①从 $\boldsymbol{X}$ 中随机选取 $\boldsymbol{M}$ 幅图像，构建一次批处理。

②通过前向传播计算批处理中各幅图像的 $\boldsymbol{z}_i$ 和 $\boldsymbol{u}_i$。

③计算批处理的哈希损失，并根据式（8）计算梯度。

④根据反向传播算法更新网络参数 $\boldsymbol{\Theta}$，并根据式（10）更新 $\boldsymbol{W}$。

3）FOR $k=1,2,\cdots,c$ DO；

根据式（12）更新 $\boldsymbol{R}_{*j}$。

3 实　验

3.1 数据集

CIFAR-10 是单标签数据集，包含 60 000 幅图像，50 000 幅为查询图像，10 000 幅为测试图像，共 10 个类别，每个类别包含 6 000 幅图像，每幅图像的大小为 32×32 像素。NUS-WIDE 是大规模图像检索任务常用的多标签数据集，包含 269 648 幅图像，共 81 个类别。

3.2 评价指标

实验采用全局平均精度（mean average precision，mAP）作为评价指标。mAP 是图像检索中最重要的衡量模型检索性能的指标，是一组查询中每个查询的平均精度（AP）相加所取的平均值，定义为

$$AP = \frac{1}{R}\sum_{c}^{N}\frac{R_c}{c}\times rel_c \tag{13}$$

式中，N 表示数据集的大小，R 是数据集中相关图像的数量，c 是进行检索以后，数据集中返回图像的数量，R_c 是前 c 中返回相关图像的数量，如果排名在第 c 位置的图像是相关的，则 rel_c 为 1，否则为 0（Zheng 等，2020）。

3.3　实现细节

深度监督哈希算法大多基于对称式方法，少部分基于非对称式方法，目前非对称式方法都是基于 CNN 实现的。实验通过与目前最优的对称、非对称深度监督哈希算法对比，验证本文方法的性能。

在 CIFAR-10 数据集上，选择 Transhash（Chen 等，2021b）、ADSH（Jiang 和 Li，2018）、DBDH（deep balanced discrete hashing）（Zheng 等，2020）、JLDSH（Gu 等，2020）、DFH（deep Fisher hashing）（Li 等，2019）、DSDH（Li 等，2017）和 DSHSD（deep supervised hashing based on stabe distribution）（Wu 等，2019）进行对比。实验中，遵循 Jiang 和 Li（2018）的方法，随机选取 1 000 幅图像（每个类别 100 幅）用做测试集，其余 59 000 幅图像用做检索集，并从检索集中随机抽取 5 000 幅图像（每个类别 500 幅）作为训练集。

在 NUS-WIDE 数据集上，选择 Transhash（Chen 等，2021b）、ADSH（Jiang 和 Li，2018）、DBDH（Zheng 等，2020）、DFH（Li 等，2019）、DSDH（Li 等，2017）和 DSHSD（Wu 等，2019）进行对比。与 Li 等人（2017）的方法类似，共选择 195 969 幅图像，包含 21 个常见类别，每个类别至少有 5 000 幅图像。实验随机选取 2 100 幅（每类 100 幅图像）作为测试集，其余图像作为检索集，并从检索集中随机抽取 10 500 幅（每类 500 幅图像）图像作为训练集。特别说明，mAP 结果是基于 NUS-WIDE 数据集返回的 Top-5K 样本计算的。

实验参数参考 Jiang 和 Li（2018）使用的参数，为了避免过拟合，权重衰减参数设为 10^{-5}。批处理数量为 128，学习率的调试范围为 $[10^{-6},10^{-2}]$，超参数 γ 为 200，T_{out} 为 60，T_{in} 为 3。本文将 $\boldsymbol{S}$ 中元素 −1 的权重设为元素 1 的数量与元素 −1 的数量之比。

3.4　实验结果

ADSHT-R 和 ADSHT 模型的提取特征模块分别为 ResNet50 网络和 Transformer 编码器网络。实验使用 Tesla P40 GPU，两个模型都在 ImageNet1k 数据集上进行预训练，设置训练轮数为 60，批处理大小为 32，都使用随机梯度下降方法，超参数 γ 为 200。两个模型在 32 位哈希码上的训练效率对比如表 1 所示。可以看出，在相同训练轮数下，ADSHT 的参数量大约是 ADSHT-R 的 13 倍，但训练时间少于 ADSHT-R，表明 ADSHT 的训练效率更高。

表 1　不同网络下模型的效率对比

Table 1　Efficiency comparison of models under different networks

模型	参数量/M	训练时间/h	
		CIFAR-10	NUS-WIDE
ADSHT-R	23.6	11.1	23.8
ADSHT	305.5	2.6	7.5

本文方法结合了分类损失函数和哈希损失函数，分类损失函数是式（7）的最后两项，超参数分别为 μ 和 φ。若 μ 为 0，则最小化分类权重为

$$\min_{\boldsymbol{W}} J(\boldsymbol{W}) = \varphi\|\boldsymbol{W}\|_{\text{F}}^2 \tag{14}$$

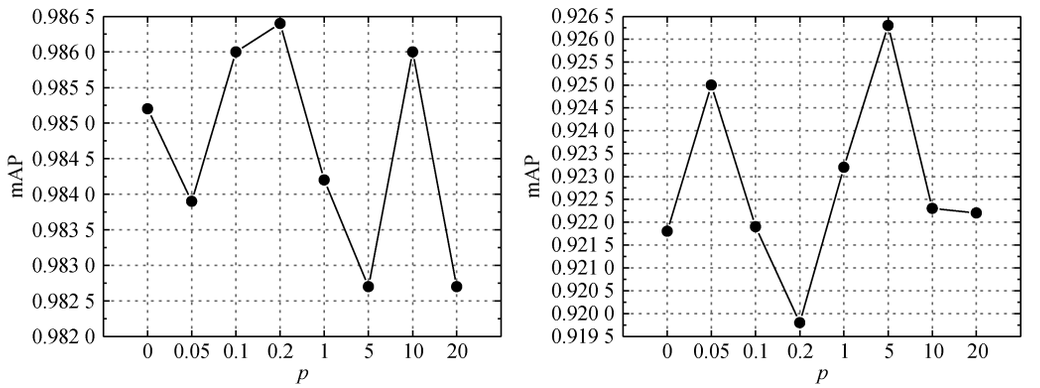
这样求得的分类权重为 0 或任意值。若 φ 为 0，则最小化分类权重为

$$\min_{\boldsymbol{W}} J(\boldsymbol{W}) = \mu\sum_{i\in\boldsymbol{\Omega}}\|\boldsymbol{i}_i - \boldsymbol{r}_i\boldsymbol{W}\|_{\text{F}}^2 \tag{15}$$

超参数 μ 不会影响分类权重，在求解当中会被消掉。因此两项要一起存在，本文令 $p = \varphi/\mu$，对不同的 p 值（0，0.05，0.1，0.2，1，5，10，20）进行实验，得到的结果如图 3 所示，本文取 $\varphi = 1$。

从图 3 可以看出，在 CIFAR-10 数据集上，模型在 $p = 0.2$ 时，$\mu = 5$，得到最好的性能。在 NUS-WIDE 数据集上，在 $p = 5$ 时，$\mu = 0.2$，模型获得最好的性能。

不同方法的检索性能如表 2 所示。ADSHT 为仅使用哈希损失函数的模型，ADSHT-F 代表结合分类损失函数的模型。从 24 位、32 位和 48 位 3 种不同的哈希位数与其他方法对比，可以发现 ADSHT-F 在 CIFAR-10 和 NUS-WIDE 数据集上的 mAP 值都高于其他方法。尤其在 CIFAR-10 数据集上，与表中 mAP 值最高的 JLDSH 相比，ADSHT-F 在 24 位时提升高达 5.06%。在 NUS-WIDE 数据集上，ADSHT-F 与 ADSH 相比，在 24 位时提升了 4.17%。ADSHT-F 模型相对于 ADSHT 模型，性能也有一定提升，说明



(a) CIFAR-10数据集　　(b) NUS-WIDE数据集

图 3　两个数据集的超参数取值

Fig. 3　Hyper parameter values of two datasets ((a) CIFAR-10 dataset; (b) NUS-WIDE dataset)

表 2　不同方法的 mAP 对比

Table 2　Comparison of mAP of different methods

方法	CIFAR-10 数据集			NUS-WIDE 数据集		
	24 bit	32 bit	48 bit	24 bit	32 bit	48 bit
DSDH(Li 等,2017)	0.786 0	0.801 0	0.820 0	0.808 0	0.820 0	0.829 0
DFH(Li 等,2019)	0.825 0	0.831 0	0.844 0	0.823 0	0.833 0	0.842 0
DSHSD(Wu 等,2019)	0.813 0	0.821 0	0.827 0	0.826 0	0.831 0	0.835 0
DBDH(Zheng 等,2020)	0.735 6	0.742 7	0.748 3	–	–	–
Transhash(Chen 等,2021b)	–	0.910 8	0.914 1	–	0.739 3	0.753 2
JLDSH(Gu 等,2020)	0.935 0	0.937 5	0.944 4	–	–	–
ADSH(Jiang 和 Li,2018)	0.928 0	0.931 0	0.939 0	0.878 4	0.895 1	0.905 5
ADSHT-R(本文)	0.745 1	0.685 5	0.711 1	0.830 7	0.841 6	0.831 4
ADSHT(本文)	0.983 7	0.985 2	0.986 6	0.914 7	0.921 8	0.926 1
ADSHT-F(本文)	**0.985 6**	**0.986 4**	**0.988 6**	**0.920 1**	**0.926 3**	**0.934 1**

注:加粗字体表示各列最优结果。“–”表示该项结果为空。

监督信息得到充分利用,分类损失会产生一定贡献。从表 2 可以看出,随着哈希表示位数的增加,模型性能会逐渐提升,因为位数增加可以更好地表示图像特征。表中前 5 行的 DSDH(Li 等,2017)、DFH(Li 等,2019)、DSHSD(Wu 等,2019)、DBDH(Zheng 等,2020)和 Transhash(Chen 等,2021b)模型均使用对称式方法,第 6—10 行的 JLDSH(Gu 等,2020)、ADSH(Jiang 和 Li,2018)、ADSHT-R、ADSHT 和 ADSHT-F 模型均使用非对称式方法,可以看出除了 ADSHT-R 模型因未完全收敛导致性能偏低以外,非对称式方法的性能明显优于对称式方法。将 ADSHT 与 ADSHT-R 对比,可以看出,ADSHT 模型性能远高于 ADSHT-R,并且从实验结果中可以观察到,训练轮数设置为 60 的情况下,ADSHT-R 模型并没有完全收敛,因此其性能偏低。这也说明 ADSHT 模型的收敛速度快于 ADSHT-R 模型。

4　结　论

本文提出了一种 Transformer 与非对称学习策略相结合的图像检索方法。一方面,利用哈希损失学习哈希函数,并使用 Transformer 生成查询图像的哈希表示,使查询图像的哈希表示更加真实;另一方面,采用非对称式学习策略,根据查询图像的哈希表

示与分类损失函数直接计算得到待检索图像的哈希表示。在此过程中,将哈希损失与分类损失相结合,充分利用监督信息,提高训练效率。与目前最优的对称、非对称深度监督哈希算法的对比实验表明,本文方法在 CIFAR-10 和 NUS-WIDE 数据集上均获得了最优性能,验证了本文方法的有效性。但是本文方法使用的 Transformer 模型未关注图像局部信息,且模型参数量较大,解决这些问题是后续工作的重点。

参考文献(References)

Andoni A and Indyk P. 2008. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. Communications of the ACM, 51(1): 117-122 [DOI: 10.1145/1327452.1327494]

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S. 2020. End-to-end object detection with transformers//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 213-229 [DOI: 10.1007/978-3-030-58452-8_13]

Chen H T, Wang Y H, Guo T Y, Xu C, Deng Y P, Liu Z H, Ma S W, Xu C J, Xu C and Gao W. 2021a. Pre-trained image processing transformer//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 12294-12305 [DOI: 10.1109/CVPR46437.2021.01212]

Chen Y B, Zhang S, Liu F X, Chang Z G, Ye M and Qi Z W. 2021b. Transhash: transformer-based hamming hashing for efficient image retrieval [EB/OL]. [2021-05-05]. https://arxiv.org/pdf/2105.01823.pdf

Chua T S, Tang J H, Hong R C, Li H J, Luo Z P and Zheng Y T. 2009. NUS-WIDE: a real-world web image database from national university of Singapore//Proceedings of 2009 ACM International Conference on Image and Video Retrieval. Santorini, Fira Greece: Association for Computing Machinery: #48 [DOI: 10.1145/1646396.1646452]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16 × 16 words: transformers for image recognition at scale [EB/OL]. [2021-06-03]. https://arxiv.org/pdf/2010.11929.pdf

Gionis A, Indyk P and Motwani R. 1999. Similarity search in high dimensions via hashing//Proceedings of the 25th International Conference on Very Large Data Bases. San Francisco, USA: Morgan Kaufmann Publishers Inc.: 518-529

Gong Y C, Lazebnik S, Gordo A and Perronnin F. 2013. Iterative quantization: a procrustean approach to learning binary codes for large-scale image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12): 2916-2929 [DOI: 10.1109/TPAMI.2012.193]

Gu G H, Liu J T, Li Z Y, Huo W H and Zhao Y. 2020. Joint learning based deep supervised hashing for large-scale image retrieval. Neurocomputing, 385: 348-357 [DOI: 10.1016/j.neucom.2019.12.096]

Guo Y C, Ding G G, Liu L, Han J G and Shao L. 2017. Learning to hash with optimized anchor embedding for scalable retrieval. IEEE Transactions on Image Processing, 26(3): 1344-1354 [DOI: 10.1109/TIP.2017.2652730]

Jiang Q Y and Li W J. 2018. Asymmetric deep supervised hashing//Proceedings of the 32nd AAAI Conference on Artificial Intelligence and the 13th Innovative Applications of Artificial Intelligence Conference and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence. New Orleans, USA: AAAI: 3342-3349

Krizhevsky A. 2009. Learning Multiple Layers of Features from Tiny Images. Toronto, Canada: University of Toronto

Lai H J, Pan Y, Liu Y and Yan S C. 2015. Simultaneous feature learning and hash coding with deep neural networks//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3270-3278 [DOI: 10.1109/CVPR.2015.7298947]

Li Q, Sun Z, He R and Tan T N. 2017. Deep supervised discrete hashing//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 2479-2488

Li W J, Wang S and Kang W C. 2016. Feature learning based deep supervised hashing with pairwise labels//Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York, USA: AAAI Press: 1711-1717

Li Y Q, Pei W J, Zha Y F and van Gemert J. 2019. Push for quantization: deep fisher hashing [EB/OL]. [2021-08-31]. https://arxiv.org/pdf/1909.00206.pdf

Lin G S, Shen C H, Shi Q F, Van den Hengel A and Suter D. 2014. Fast supervised hashing with decision trees for high-dimensional data//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 1971-1978 [DOI: 10.1109/CVPR.2014.253]

Liu W, Wang J, Ji R R, Jiang Y G and Chang S F. 2012. Supervised hashing with kernels//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE: 2074-2081 [DOI: 10.1109/CVPR.2012.6247912]

Liu Y, Cheng M, Wang F P, Li D X, Liu W, Fan J L. 2020. Deep Hashing image retrieval methods, 25(7): 1296-1317(刘颖,程美,王富平,李大湘,刘伟,范九伦. 2020. 深度哈希图像检索方法综述. 中国图象图形学报,25(7): 1296-1317) [DOI: 10.11834/jig.190518]

Wan F, Qiang H P, Lei G B. 2021. Self-supervised deep discrete hashing for image retrieval. Journal of Image and Graphics, 26(11):

2659-2669(万方,强浩鹏,雷光波. 2021. 自监督深度离散哈希图像检索. 中国图象图形学报, 26(11): 2659-2669)[DOI: 10.11834/jig.200212]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 5000-6010

Wang Q, Li B, Xiao T, Zhu J B, Li C L, Wong D F and Chao L S. 2019. Learning deep transformer models for machine translation//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics: 1810-1822 [DOI: 10.18653/v1/P19-1176]

Wu L, Ling H F, Li P, Chen J Z, Fang Y and Zhou F H. 2019. Deep supervised hashing based on stable distribution. IEEE Access, 7: 36489-36499 [DOI: 10.1109/ACCESS.2019.2900489]

Xia R K, Pan Y, Lai H J, Liu C and Yan S C. 2014. Supervised hashing for image retrieval via image representation learning//Proceedings of the 28th AAAI Conference on Artificial Intelligence. Québec City, Canada: AAAI Press: 2156-2162

Zhang D, Wang J, Cai D and Lu J S. 2010. Self-taught hashing for fast similarity search//Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. Geneva, Switzerland: Association for Computing Machinery: 18-25 [DOI: 10.1145/1835449.1835455]

Zhang P C, Wei Z, Li W J and Guo M Y. 2014. Supervised hashing with latent factor models//Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. Gold Coast, Australia: Association for Computing Machinery: 173-182 [DOI: 10.1145/2600428.2609600]

Zheng X T, Zhang Y C and Lu X Q. 2020. Deep balanced discrete hashing for image retrieval. Neurocomputing, 403: 224-236 [DOI: 10.1016/j.neucom.2020.04.037]

Zhu H, Long M S, Wang J M and Cao Y. 2016. Deep hashing network for efficient similarity retrieval//Proceedings of the 13th AAAI Conference on Artificial Intelligence. Phoenix, USA: AAAI Press: 2415-2421

作者简介

贺超,男,硕士研究生,主要研究方向为图像检索。
E-mail: 3089840275@qq.com
魏宏喜,通信作者,男,教授,主要研究方向为机器学习和多媒体信息处理。E-mail:cswhx@imu.edu.cn