DOI:10.19651/j.cnki.emt.5005407

基于注意力机制的深度哈希图像检索方法*

金汉均 曾 星
(华中师范大学 计算机学院 武汉 430070)

摘 要：传统的深度哈希图像检索方法所生成的二进制哈希码存在信息冗余,不能很好地反映图像局部语义信息。提出一种卷积神经网络同注意力模型相结合的深度哈希图像检索方法,使用VGG16网络作为图像的特征提取器,接着在模型的卷积层之后添加注意力模块,提炼出更有效的特征图,最后在模型的全连接层输出二进制哈希码作为图像的特征,从而提高图像检索任务的精确度。在CIFAR-10和NUS-WIDE数据集上的实验表明,添加注意力机制后,模型在两个数据集下使用不同位数二进制哈希码的检索精度最高达到85.3%与78.1%,均高于未使用注意力机制的情况,验证了注意力机制的有效性。

关键词：图像检索;注意力模型;卷积神经网络;哈希
中图分类号：TP391 **文献标识码**：A **国家标准学科分类代码**：510.4050

Deep hash image retrieval method based on attention model

Jin Hanjun Zeng Xing
(School of Computer Science, Central China Normal University, Wuhan 430070, China)

Abstract: The binary hash code generated by the traditional deep hash image retrieval method has information redundancy and cannot reflect the local semantic information of the image well. Proposes a deep hash image retrieval method that combines a convolutional neural network with an attention model. It uses the VGG16 network as the image feature extractor, and then adds an attention module after the convolutional layer of the model to refine the feature map, and finally output the binary hash code as the feature of the image in the fully connected layer of the model, thereby improving the accuracy of the image retrieval task. Experiments on the CIFAR-10 and NUS-WIDE datasets show that after the attention mechanism is added, the model uses different digit binary hash codes in the two datasets to achieve the highest retrieval accuracy of 85.3% and 78.1%, higher than the case where the attention mechanism is not used, verifying the effectiveness of attention model.

Keywords: image retrieval;attention model;convolutional neural network;hash

0 引 言

随着网络与多媒体技术的发展,每天都会产生大量的图像数据,如何从大量图像中找到自己想要的结果是我们需要考虑的问题,而基于内容的图像检索技术(content-based image retrieval,CBIR)则可以解决这个问题。基于内容的图像检索技术旨在分析图像内容来搜索相似图像,是目前最常使用的图像检索方法。

在早期的CBIR系统中,人们使用手工提取的特征来表示图像的内容。随着深度学习技术的发展,研究人员发现深度卷积神经网络在图像处理中拥有强大的特征提取能力[1],使用卷积神经网络已成为当前主流的特征提取方法[2]。为了提高图像检索的效率,人们又提出了可以通过学习哈希函数将图像特征映射成二进制编码的深度哈希图像检索方法。如今深度哈希图像检索方法已成为图像检索领域的研究热点,被广泛应用于图像检索任务中[3]。

2012年,Krizhevsky等[4]将深度神经网络模型AlexNet应用于图像分类并取得了优异的结果,从此基于深度学习的图像检索方法成为了新的研究方向。尽管从卷积神经网络中提取到的特征能够很好地表达图像的高层语义信息,但该特征存在维数过高、计算量大等缺点。为了解决这个问题,研究者将哈希算法与深度学习结合,形成了基于深度哈希的图像检索方法。2014年,Xia等[5]提出了CNNH方法,该方法分为两个阶段,在第一阶段学习哈希

收稿日期:2020-11-21
*基金项目:教育部人文社科规划基金(17YJA870010)项目资助

码,第二阶段训练神经网络,同时学习图像特征与哈希函数。在此之后,Lai 等[6]对 CNNH 方法进行改进,提出了 DNNH 模型,使得神经网络学习到的特征可以及时反馈给哈希编码。Lin 等[7]提出了一个端到端的神经网络模型,可以直接在神经网络的隐藏层输出图像的二进制编码,取得了较好的检索效果。Jiang 等[8]提出了深度离散哈希算法(deep supervised discrete hashing, DSDH),利用成对标签信息和分类信息学习二进制哈希码。

传统的深度哈希图像检索方法通常是设计端到端的模型,在卷积神经网络中添加哈希层,使用哈希层输出的二进制哈希码作为图像的特征。但使用卷积神经网络所提取到的特征是图像的全局特征,在图像检索任务中我们所关注的是图像中的某个具体目标,需要与目标相关的局部特征,目标之外的背景信息可能会对图像检索结果造成干扰。因此,这类基于哈希的图像检索方法所提取到的特征存在冗余。

为了解决神经网络模型中存在的信息冗余问题,研究人员在神经网络中添加了注意力机制,注意力机制通过对模型不同关注部分赋予不同的权重,学习数据中关键信息,使得模型能够做出更加准确的判断[9]。2017 年 ImageNet 图像分类大赛的冠军模型 SENet[10]中的 SE 模块就采用了注意力机制,通过将 SE 模块添加到残差网络的残差模块[11]以及 GoogLeNet 的 Inception 结构[12]中,提高了模型的性能。2018 年,Woo 等[13]提出了卷积注意力模块(convolutional block attention module, CBAM),同时考虑了通道域注意力与空间域注意力,并且相较于仅使用通道域注意力机制的 SENet,CBAM 在 ImageNet 数据集上的分类任务中取得了更好的效果。

在图像检索领域,注意力机制也有相关应用,Noh 等[14]提出了一种空间域注意力模型,将该模型添加到卷积神经网络的卷积层中,选取图像中的关键点,提高了图像检索的精度。李宗民等[15]将注意力机制应用于手绘图像检索领域,在 VGG16 网络中添加注意力模块,将提取到的 4 608 维特征向量用于手绘图像检索,取得了优秀的检索效果,但该方法提取到的特征维度过高,不利于计算。

针对深度哈希图像检索方法中存在的信息冗余问题,本文提出一种基于注意力机制的深度哈希图像检索方法,通过在卷积神经网络中添加注意力模块,训练注意力模块使模型提取到更有效的二进制编码。在 CIFAR-10 和 NUS-WIDE 数据集上的实验结果表明,加上注意力机制后一定程度上提高了模型检索精度,从而验证了注意力机制在图像检索任务中的有效性。

1 基于注意力机制的深度哈希图像检索

本文将注意力机制应用在深度哈希图像检索方法上,通过训练注意力模块,对检索任务中重要的特征赋予更高的权重,使得模型能够提取到更加有效的二进制哈希码,模型的结构如图 1 所示,在 VGG16 模型的卷积层之后添加注意力模块,然后再全连接层网络中添加用于生成二进制编码的哈希层。当图像输入到模型中时,首先经过卷积层得到一个特征图,再将特征图输入到注意力模块中,注意力模块会对特征图的通道域和空间域分别进行加权操作,得到一个新的特征图,特征图的大小在输入到注意力模块前后不发生改变,最后将新特征图展平输入到全连接层神经网络中,全连接层的倒数第二层为哈希层,可以输出二进制哈希码,该哈希码用于表示输入图像的特征,再对该特征进行相似度计算即可得到最终检索结果。

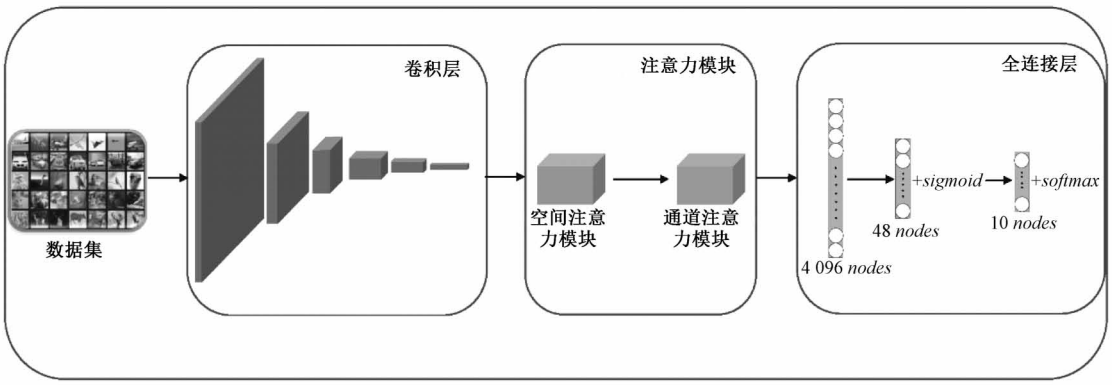

图 1 模型结构

1.1 图像特征提取

本文所使用的特征提取网络为 VGG16[16],在 VGG16 的卷积层之后添加了注意力模块,在全连接层中添加了哈希层,VGG16 的卷积层结构如图 2 所示,卷积层均采用 3×3 大小的卷积核[17]。输入图像尺寸为 224×224,当图像经过模型的卷积层之后得到 7×7×512 的特征图,将该特征图作为注意力模块的输入。

本文所使用的注意力模块为文献[13]中提出的卷积注意力模块,该模块分为通道注意力模块和空间注意力模块,可以添加到卷积层的任意位置。通道注意力模块可以表示为:

$$F_{Out}^{C}=\sigma(MLP(AvgPool(F))+MLP(MaxPool(F)))\odot F \tag{1}$$

其中,F 表示输入的特征图,F_{Out}^{C} 表示输出特征图,σ

输入224×224 RGB图像
conv3-64 conv3-64
最大池化
conv3-128 conv3-128
最大池化
conv3-256 conv3-256 conv3-256
最大池化
conv3-512 conv3-512 conv3-512
最大池化
conv3-512 conv3-512 conv3-512
最大池化
输出7×7×512特征图

图 2　VGG16 网络卷积层结构

表示 *sigmoid* 激活函数，⊙表示在通道维度上同位元素逐个相乘，*MLP* 代表多层感知机，*AvgPool* 和 *MaxPool* 分别表示全局平均池化和全局最大池化，分别如式(2)、(3)所示。

$$AvgPool(F_C)=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}F_c(i,j) \tag{2}$$

$$MaxPool(F_c)=Max(F_c(i,j)) \tag{3}$$

当图像的特征图输入到全局平均池化层和全局最大池化层后，特征图的通道数不变，在空间维度进行降维，特征图的尺寸变为，即将特征图的每个通道变成一个实数，再将这组实数输入到 *MLP* 中学习相应的权重，*MLP* 表示如下：

$$M(F)=\boldsymbol{W}_1(\delta(\boldsymbol{W}_2(F))) \tag{4}$$

式中：$\boldsymbol{W}_1$ 与 $\boldsymbol{W}_2$ 表示感知机的权重矩阵，$\boldsymbol{W}_2\in R^{\frac{C}{r}\times C}$，$\boldsymbol{W}_1\in R^{C\times\frac{C}{r}}$，$C$ 的值为输入特征图的通道数，$\frac{C}{r}$ 表示感知机中隐层节点个数，δ 为 ReLU 激活函数。在本文中，通道注意力模块的输入特征图大小为 7×7×512，512 为特征图的通道数，该特征图经过池化、感知机等操作后可以得到每一个通道的权重，该权重大小为 1×1×512，再将权重与原输入特征图逐位相乘得到新的特征图，新特征图大小与输入相同。

空间注意力机制表示为：

$$F_{Out}^{S}=\sigma(\mathrm{Conv}([\mathrm{mean}(F);\max(F)]))\odot F \tag{5}$$

其中，mean 与 max 操作分别表示将输入特征图 F 按照通道维度求平均值与求最大值，Conv 表示卷积操作，作经过这两个步骤后特征图的通道数均从 512 降维到 1，将两个通道数为 1 的特征图在通道维度上连接，再经过卷积操作后得到空间注意力权重，最后使用 *sigmoid* 激活函数将权值约束在 0～1 之间。模块输出的空间注意力权重大小为 7×7×1。将空间注意力权重与输入特征图在空间维度上对同位元素逐位相乘得到空间注意力特征图，其大小同样与输入特征图相同。

综上，卷积模块输出的特征图通过注意力模块加权后得到一个大小相同的新特征图，在输入到全连接层神经网络之前对新特征图进行展平操作。

全连接层神经网络的倒数第二层为 48 个结点的哈希层，用于输出图像特征，本文使用 *sigmoid* 激活函数将哈希层中的值约束在(0,1)区间，再通过一个阈值函数得到二值向量。本文使用的阈值函数如下：

$$H_i=\begin{cases}1, & F(i)\geqslant 0.5\\ 0, & \text{其他}\end{cases}\quad i\in[i,48] \tag{6}$$

式中：i 表示哈希层中某一个结点，48 为哈希层结点个数，$F(i)$ 表示哈希层中第 i 个结点的值，H_i 表示该结点的最终输出值。这样，当一幅图像通过模型之后就可以生成相应的 48 位二进制哈希码，本文使用这串二进制哈希码表示输入图像的特征。

1.2　相似性度量

对特征的相似性度量是影响图像检索结果的关键因素，本文使用上述特征提取方法提取数据库中每一张图像的二进制哈希码，构建图像特征数据库，再以同样的方法提取待查询图像的特征。检索方法如下：首先计算查询图像哈希码与数据库图像哈希码间的汉明距离，汉明距离的公式可表示为式(7)。

$$d(x,y)=\sum x\oplus y \tag{7}$$

其中，⊕为异或操作，x 和 y 表示两张图像的二进制哈希码。计算完汉明距离后可以得到与查询图像汉明距离相近的 k 张图像，再对 k 张图像进行更加细致的检索。计算查询图像与 k 张图像的欧氏距离，并根据图像的欧氏距离大小对 k 张图像重新排序，得到最终检索结果。欧氏距离如式(8)所示。

$$d_{(\boldsymbol{X},\boldsymbol{Y})}=\sqrt{\sum_{i=1}^{N}(x_i-y_i)^2} \tag{8}$$

其中，$\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 表示两张图像的特征向量，x_i 与 y_i 表示两个特征向量中第 i 个结点的值。

2　实验及分析

本实验所选用的数据集为 CIFAR-10 数据集和 NUS-WIDE 数据集，CIFAR-10 数据集一共包含 60 000 张图像，分为 10 个类别，其中训练集包含 50 000 张图像，测试集包含 10 000 张图像，每张图像的大小为 32×32。NUS-WIDE 数据集[18]是一个多标签数据集，一共包含近 27 万张图像，为了保证实验的公平性，本文参照文献[6]的设置选取数据集中 21 个常见类别标签，每个类别大约包含 5 000 张图像。最后将本文提出方法的最终检索结果与其他经典的哈希算法分别在不同位数二进制哈希码的情况下进行比较。同时为了验证注意力机制的有效性，本文还对比了未添加注意力机制的模型检索效果。本文所使用的评价指标是在图像检索领域中常用的平均精度值(mean average precision，mAP)。

本实验一共训练了两个模型,分别是添加了哈希层的VGG16网络,记为VGGH,以及添加了哈希层和注意力模块的VGG16网络,记为AM-VGGH。训练时两个模型均使用了迁移学习的方法,将在ImageNet数据集上预训练好的VGG16网络的卷积层参数加载到VGGH与AM-VGGH模型的卷积层上,再将两个模型放在数据集上训练,迭代相同次数,微调注意力模块与全连接层的参数。

本文对比了几种经典的哈希算法在不同位数二进制哈希码下的检索精度,其中包括传统的哈希算法和深度哈希算法。传统的哈希算法选择了核哈希[19](KSH)与局部敏感哈希(LSH),深度哈希算法选择CNNH[5],CNNH+[5],DNNH[6]以及DPSH[20]作为对比。

在CIFAR-10与NUS-WIDE数据集上的实验结果分别如表1和2所示,从表中数据可以看出使用深度哈希算法的检索结果均优于传统哈希算法,检索性能有较大提升,这也验证了深度哈希算法的有效性,使用深度学习方法确实可以提取到更好的图像特征。而本文所提出的方法与其他经典深度哈希方法相比,在大多数情况下检索精度更高,说明了本文所提出的模型性能更好。

表1 CIFAR-10上不同方法mAP值

方法	12位	24位	32位	48位
AM-VGGH	0.797	0.844	0.845	0.853
VGGH	0.776	0.825	0.834	0.841
DPSH	0.713	0.727	0.744	0.757
DNNH	0.552	0.566	0.558	0.581
CNNH+	0.465	0.521	0.521	0.532
CNNH	0.439	0.511	0.509	0.522
KSH	0.303	0.337	0.346	0.356
LSH	0.121	0.126	0.120	0.120

表2 NUS-WIDE上不同方法mAP值

方法	12位	24位	32位	48位
AM-VGGH	0.753	0.765	0.769	0.781
VGGH	0.732	0.745	0.752	0.756
DPSH	0.747	0.751	0.763	0.776
DNNH	0.674	0.697	0.713	0.715
CNNH+	0.617	0.663	0.657	0.608
CNNH	0.611	0.618	0.625	0.608
KSH	0.556	0.572	0.581	0.588
LSH	0.403	0.421	0.426	0.441

相比于没有使用注意力机制的VGGH,使用了注意力机制的AM-VGGH模型在不同位数二进制哈希码的情况下检索结果均有不同程度的提高,这个结果证明了在其它条件相同的情况下,使用注意力机制的神经网络可以提取到更有效的图像特征,从而提高图像检索的精确度。

3 结 论

本文将注意力机制应用在深度哈希图像检索算法上,提出一种基于注意力机制的深度哈希图像检索方法,通过对注意力模块的训练,使得注意力模块能够学习到特征图不同区域的重要性并通过加权的方式对重要的区域赋予较高的权重,通过在不同数据集上图像检索任务的结果表明,添加了注意力机制的深度哈希算法可以生成更加有效的二进制哈希码,提高检索任务的准确率。在未来的研究中,可以对注意力模块进行改进,从而实现更优秀的图像检索效果。

参考文献

[1] 郭鹏,肖秦琨,赵一丹.基于深度图像的手势识别研究[J].国外电子测量技术,2019,38(10):6-12.

[2] 朱阳光,刘瑞敏,黄琼桃.基于深度神经网络的弱监督信息细粒度图像识别[J].电子测量与仪器学报,2020,34(2):115-122.

[3] 刘颖,程美,王富平,等.深度哈希图像检索方法综述[J].中国图象图形学报,2020,25(7):1296-1317.

[4] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet classification with deep convolutional neural networks[C]. International Conference on Neural Information Processing Systems, 2012:1097-1105.

[5] XIA R, PAN Y, LAI H, et al. Supervised hashing for image retrieval via image representation learning[C]. AAAI Conference on Artificial Intelligence, 2014: 2156-2162.

[6] LAI H, PAN Y, LIU Y, et al. Simultaneous feature learning and hash coding with deep neural networks[C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR),2015:3270-3278.

[7] LIN K, YANG H, HSIAO J, et al. Deep learning of binary hash codes for fast image retrieval, 2015[C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops(CVPRW), 2015:27-35.

[8] JIANG Q, CUI X, LI W. Deep supervised discrete hashing[J]. IEEE Transactions on Image Processing, 2018,27(12):5996-6009.

[9] 朱张莉,饶元,吴渊,等.注意力机制在深度学习中的研究进展[J].中文信息学报,2019,33(6):1-11.

[10] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 42(8): 2011-2023.

[11] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2016:770-778.

[12] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2015:1-9.

[13] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]. European Conference on Computer Vision, 2018:3-19.

[14] NOH H, ARAUJO A, SIM J, et al. Large-scale image retrieval with attentive deep local features[C]. 2017 IEEE International Conference on Computer Vision(ICCV), 2017:3476-3485.

[15] 李宗民，李思远，刘玉杰，等. 基于注意力模型的手绘图像检索方法[J]. 计算机科学，2020,47(11):199-204.

[16] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. ArXiv Preprint, 2014:1409-1556.

[17] 金汉均，吴静. 基于深度柯西哈希的图像检索研究[J]. 电子测量技术，2020,43(9):104-108.

[18] ZHAO F, HUANG Y, WANG L, et al. Deep semantic ranking based hashing for multi-label image retrieval[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015:1556-1564.

[19] CHANG S. Supervised hashing with kernels [C]. Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012:2074-2081.

[20] LI W, WANG S, KANG W. Feature learning based deep supervised hashing with pairwise labels [J]. ArXiv E-prints, 2015:1511-3855.

作者简介

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

曾星(通信作者)，硕士研究生，主要研究方向为图像处理与深度学习。

E-mail:zengxing@mails.ccnu.edu.cn