计算机科学
COMPUTER SCIENCE

基于独立注意力机制的图像检索算法

张舜尧, 李华旺, 张永合, 王新宇, 丁国鹏

引用本文

张舜尧, 李华旺, 张永合, 王新宇, 丁国鹏基于独立注意力机制的图像检索算法[J]. 计算机科学, 2023, 50(6A): 220300092-6.

ZHANG Shunyao, LI Huawang, ZHANG Yonghe, WANG Xinyu, DING Guopeng. Image Retrieval Based on Independent Attention Mechanism [J]. Computer Science, 2023, 50(6A): 220300092-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220300092

基于独立注意力机制的图像检索算法

张舜尧[1,2,3]　李华旺[1,2,3]　张永合[1,3]　王新宇[1,3]　丁国鹏[1,3]
1 中国科学院微小卫星创新研究院　上海 201210
2 上海科技大学　上海 201210
3 中国科学院大学　北京 100094
(zhangshy4@shanghaitech.edu.cn)

摘　要　近年来,深度学习的方法在基于内容的图像检索领域已经占据主导地位。为了改善主干网络提取出的特征,使得网络能计算出更具区分度的图像描述,提出了一种独立于输入特征的注意力模块 ICSA(Independent Channel-wise and Spatial Attention)。该模块与其他的注意力机制的主要区别在于它的注意力权重在输入不同特征时保持一致,传统注意力模块通过对输入特征进行处理得到注意力,因此它的模型更为精简,其参数大小仅有 6.7kB,为 SENet 大小的 5.2%和 CBAM 的 2.6%,运行时间与 SENet 基本一致,为 CBAM 的 14.9%。ICSA 的注意力分为通道和空间注意力两部分,分别储存输入特征不同方向上的权重。在 Pittsburgh 数据集上进行实验,实验结果表明,对于不同的主干网络,在添加了 ICSA 模块后 Recall@1 有 0.1%～2.4%的提升。

关键词:基于内容的图像检索;注意力机制;特征增强
中图法分类号　TP391

Image Retrieval Based on Independent Attention Mechanism

ZHANG Shunyao[1,2,3],LI Huawang[1,2,3],ZHANG Yonghe[1,3],WANG Xinyu[1,3] and DING Guopeng[1,3]
1 Innovation Academy for Mircrosatellites of Chinese Academy of Sciences,Shanghai 201210,China
2 Shanghai Tech University,Shanghai 201210,China
3 University of Chinese Academy of Sciences,Beijing 100094,China

Abstract　In recent years,deep learning methods has taken a dominant position in the field of content-based image retrieval. To improve features extracted by off-the-shelf backbones and enable the network produce more discriminative image descriptors,the attention module ICSA(independent channel-wise and spatial attention),which is independent with features input into the module,is proposed. Attention weights of the proposed module keeps the same when input features change,while attention weights are usually computed with input features in other attention mechanisms,which is a main difference between ICSA and other attention modules. This feature also enables the module to be quite small(only 6.7kB,5.2% the size of SENet,2.6% of the size of CBAM) and relatively fast(similar with SENet in speed and 14.9% the time of CBAM). The attention of ICSA is divided as two parts:channel-wise and spatial attention,and they store the weights along orthogonal directions. Experiments on Pittsburgh shows that ICSA made improvement from 0.1% to 2.4% at Recall@1 when with different backbones.

Keywords　Content based image retrieval,Attention mechanism,Feature enhancement

1　引言

基于内容的图像检索(Content-Based Image Retrieval,CBIR)是一项根据查询图像的内容,在图像库中检索与其语义匹配或相似的图像的问题。CBIR 是计算机视觉与多媒体领域中经典的问题[1-2]。如今随着摄像设配的普及,图像和视频数据的数量急剧增加,为方便对如此庞大的数据进行有效管理,图像检索的技术变得尤为重要。

20 世纪 70 年代,基于文本的图像检索(Text-based Image Retrieval)曾十分流行[3]。该方法首先通过人工方式将图像注释为文本,再使用文本检索的方式来检索图像。这种方法的主要缺点如下。

(1)人工标注图像费时费力,难以被应用到庞大的图像数据集中。

(2)文本不足以描述图像的全部内容,同时图像标注成文本产生的信息损失影响了检索的质量。

2003 年开始,BoW(Bag of Words)模型被应用到图像检索以及其他的计算机视觉领域中[4-5],BoW 模型最初用于文档建模,它的主要思想是统计文档中词出现频率的直方图。由于一种具有平移、缩放、旋转不变性等特征的图像局部特征描述算子 SIFT(Scale-Invariant Feature Transform)[6] 的出现,使得图像可以表示成局部特征的集合,从而应用 BoW 模型。类似于 BoW 模型,许多方法通过图像的局部特征编码成整张图像的描述算子,如 VLAD(Vector of Locally Aggregated Descriptors)[7] 和 FV(Fisher Vector)[8]。其中 FV 能统计到比 VLAD 更高阶的特征,因此其性能略高于 VLAD,但在时间和空间上具有更高的复杂度。

自从 2012 年 Krizhevsky 等[9] 使用深度学习在 ILSVRC

通信作者:李华旺(lihw@microsate.com)

(ImageNet Large Scale Visual Recognition Challenge)2012 比赛中获得巨大成功,深度学习成为了 CBIR 以及其他计算机视觉领域的主要研究方法。近年来,基于深度学习的 CBIR 方法主要分为两个步骤:1)使用网络提取图像的特征;2)将特征编码成整张图像的描述算子。

一方面,VGG[10],ResNet[11],ViT[12] 等网络架构的提出提供了丰富的图像抽象语义特征;另一方面,对图像特征的编码或者增强方法使得图像的描述更有区分度,Babenko 等[13-16] 直接将 CNN(Convolutional Neural Network)的输出池化后作为图像的描述算子用于图像检索,该方法在仅使用其他数据集上预训练的参数就能取得不错的结果。除了池化和哈希函数之外,可以利用传统图像检索中使用的特征编码方式将神经网络提取出来的特征编码到高维空间中,包括 BoW,VLAD 以及 FV,如基于 VLAD 的 NetVLAD[17]、基于 FV 的 Deep Fisher Vector Siamese Net-work[18] 以及文献[13,19-22]提出的方法等。

在计算机视觉领域,还有一种通用的特征增强方法,即注意力机制。RAM[24] 首次将注意力机制与 RNN(Recurrent Neural Network)相结合,实现了一个端到端的网络,用于计算图像的特征。但是 RAM 的注意力机制不能用于 CNN,直到 Jaderberg 等提出了 STN[25]。STN 的注意力机制由一个子网络实现,该子网络将输入特征进行仿射变换,保留重要区域的特征并且舍弃掉不相关部分。SENet[26] 提出了一种新的通道注意力机制,它能根据输入的特征自适应地预测其中关键的通道。在 SENet 的基础上 CBAM[27] 将空间注意力机制融入其中。Vaswani 等[28] 提出的自注意力机制在自然语言处理领域取得了突破性的成就,non-local network[29] 将自注意力机制引入到计算机视觉领域,并在物体检测与视频理解中取得了不错的成绩。ViT[12] 进一步使用自注意力的网络取代传统的 CNN,并达到了与 CNN 相比拟的效果,证明了自注意力机制在计算机视觉领域中的巨大潜力。

本文在 NetVLAD 网络的基础上,提出了两种新的注意力模块 IA(Independent Attention)和 ICSA(Independent Channel-wise and Spatial Attention),通过学习整个数据集对图像中不同位置的特征以及特征的不同通道的信息的偏好,赋予特征不同的权重,这样的设计使得模块变得更加精简,便于在小数据集上优化模块参数。尤其是 ICSA 在仅增加较小的参数数量以及计算量的情况下,显著提升了算法的性能。

2 相关技术

2.1 NetVLAD

VLAD 将一张图片表示为矩阵 $\boldsymbol{V}\in\mathbb{R}^{K\times D}$,在$(j,k)$位置上的数值为:

$$\boldsymbol{V}(j,k)=\sum_i a_k(\boldsymbol{x}_i)(x_i(j)-c_k(j)) \tag{1}$$

其中,$x_i(j)$是图像第 i 个特征向量$\boldsymbol{x}_i$ 的第 j 维,$c_k(j)$是第 k 个特征聚类中心c_k 的第 j 维。$a_k(\boldsymbol{x}_i)$表示特征$\boldsymbol{x}_i$ 是否属于第 k 个特征类,具体计算方式如下:

$$a_k(\boldsymbol{x}_i)=\begin{cases}1, & \text{if } \arg\min_j \| \boldsymbol{x}_i-\boldsymbol{c}_j \| =k\\ 0, & \text{otherwise}\end{cases} \tag{2}$$

NetVLAD[17] 对 VLAD 进行了以下改进:

(1)将初始神经网络提取出的特征 $\boldsymbol{X}$ 聚类作为初始 VLAD 中心 $\boldsymbol{c}$,并在训练过程中继续微调 $\boldsymbol{c}$ 的参数。

(2)另外用 softmax 函数替代了 VLAD 模块中不可微的特征分配的操作 $a_k(\boldsymbol{x}_i)$,从而使其可以嵌入 CNN 网络完成端到端的训练。新的函数为:

$$\bar{a}_k(\boldsymbol{x}_i)=\frac{\mathrm{e}^{-\alpha\| \boldsymbol{x}_i-\boldsymbol{c}_k \|^2}}{\sum_{k'}\mathrm{e}^{-\alpha\| \boldsymbol{x}_i-\boldsymbol{c}_{k'} \|^2}} \tag{3}$$

本文使用 NetVLAD 层对主干网络提取出来的图像特征以及其被注意力模块增强后的结果进行编码,用于检索实验。

2.2 注意力机制

注意力模块通常会根据不同的输入特征产生不同的注意力,再将注意力作用于特征,可以用以下公式描述[30]:

$$Feature_{\text{Attention}}=f(g(x),x) \tag{4}$$

其中,x 表示输入特征,$g(x)$表示根据输入特征产生的表征区分输入特征不同区域重要性的注意力,$f(g(x),x)$表示将生成的注意力作用域输入特征上计算出的输出特征。

比如,对于 SENet 中的注意力模块,可以表示为:

$$\begin{aligned}&g(x)=\text{Sigmoid}(MLP(GAP(x)))\\&f(g(x),x)=g(x)x\end{aligned} \tag{5}$$

下文将沿用此处对注意力机制的定义,介绍提出的注意力模块。

3 方法介绍

本文中用于图像检索的网络分为两个部分,即特征提取和特征增强,如图 1 所示。

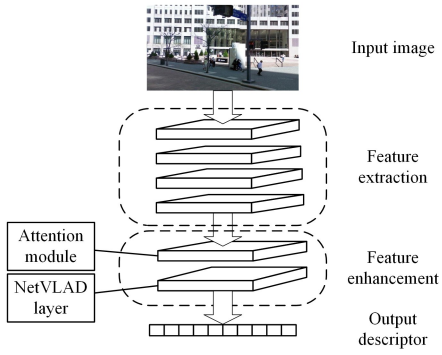

图 1 基于 NetVLAD 的图像检索网络模型

Fig. 1 CBIR network model based on NetVLAD

特征提取部分的主干网络将直接使用计算机视觉领域常见的 VGG[10]、ResNet[11] 等网络。特征增强部分分为注意力模块和 NetVLAD 层[17] 两部分,它们负责对主干网络提取出的图像特征进行增强处理,以生成更具辨别力的图像描述算子。

3.1 NetVLAD 层

NetVLAD 层是一种 VLAD 算法的近似,它模拟了 VLAD 对图像的描述方式,并且可以嵌入神经网络完成端到端的训练。如第 2.1 节中的描述,它包含了所有图像特征与特征聚类中心残差和的信息。显然,NetVLAD 的性能与聚类的数量相关,过大的聚类数量对性能的提升较少但会增加较多的计算量。为了提升 NetVLAD 编码的质量,本文通过在 NetVLAD 层之前添加一个注意力模块,来改善输入特征的分布,从而使得 NetVLAD 层能够生成更具区分度的图像编码。

3.2 IA、ICSA 注意力模块

根据第 4 节中介绍的实验,随输入特征变化的注意力 $g(x)$并不能提升如图 1 所示网络的性能,相反,不同注意力机制会造成不同程度的性能下降,该结论将在第 4 节的实验结论中详细阐述。因此,本文尝试固定注意力 $g(x)$,使之不随

输入特征 x 变化。即对于给定特征空间 $X\subseteq\mathbb{R}^{C\times H\times W}$，有 $\forall x_0$，$x_1\in X, g(x_0)=g(x_1)$。根据上述思路，可以将注意力视为一个与输入特征 x 形状相同的 3 维数组 $A\in\mathbb{R}^{C\times H\times W}$，即：

$$\forall x\in X, g(x)=A \tag{6}$$

数组中每一个数字对应输入特征中相应位置数据的权重，在网络中直接与输入特征一一相乘，即：

$$Feature_{\text{Attention}}=A\otimes x \tag{7}$$

其中，符号⊗表示哈达玛积（Hadamard product 或 Element-wise product）。由此，本文首先提出了一种注意力与输入特征不相关的注意力模块 IA（Independent Attention），它为图像特征中的数值提供了一个可学习的权重，如图 2 所示。

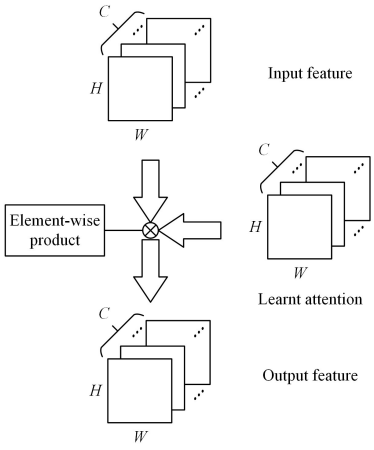

图 2　IA 注意力模块

Fig. 2　IA attention module

但为特征中每一个数据都使用一个独立权重参数会增加较多的显存占用，为进一步减少模块中的参数量，在假设特征的通道与空间位置相互独立的情况下，可以将注意力 A 拆分成通道注意力 $A_{\text{channel}}\in\mathbb{R}^{C\times 1\times 1}$ 和空间注意力 $A_{\text{spatial}}\in\mathbb{R}^{1\times H\times W}$ 两个部分，使得：

$$A=A_{\text{channel}}\otimes A_{\text{spatial}} \tag{8}$$

其中，在通道注意力 A_{channel} 和空间注意力 A_{spatial} 进行哈达玛积运算前先对其进行广播（Broadcast）操作，以保证它们的维度相同。这种通道与空间分离的注意力模块如图 3 所示。

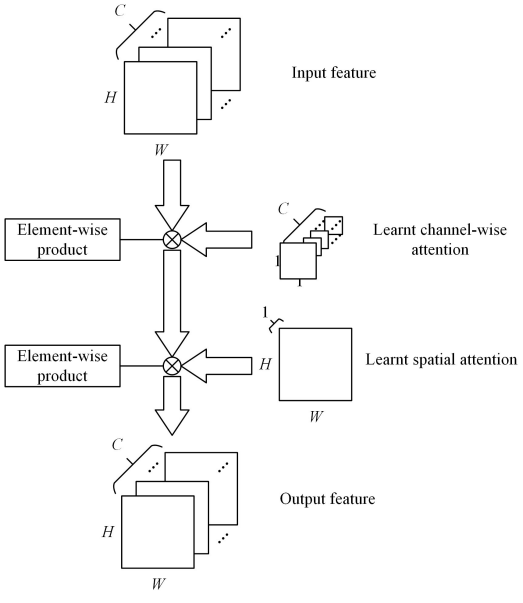

图 3　ICSA 注意力模块

Fig. 3　ICSA attention module

4　实验与分析

4.1　实验设置

本文使用的数据集是 Pittsburgh[31]，该数据集包含 250000张从 10000 张谷歌街景全景图上获取的图像作为数据库，和 24000 张同样方式获取的图像作为查询图像。实验结果指标采用前 N 个检索结果的召回率（Recall@N）。训练过程中使用的损失函数为 Triplet margin loss[31]。网络参数优化器使用的是 SGD[32]。所有实验均基于 Pytorch1.7.1 并在 Nvidia RTX3090 上运行。

4.2　实验结果与分析

4.2.1　注意力模块的比较

首先本文将第 3.2 节提出的两种注意力模块 IA、ICSA 与经典注意力模块 SENet、CBAM 以及不加注意力模块的网络进行比较。其中主干网络采用在 NetVLAD 官方实验中效果最优的 VGG16，并且只对其最后一个网络模块进行微调，其他参数保持预训练的权重不变，训练不同部分主干网络的结果将在后续实验中进行讨论。

具体实验结果如表 1 所列，Recall@N 越大表明结果越好。可以看到，相比不加注意力模块的网络，本文提出的两种注意力模块的性能总体上有提升。IA 和 ICSA 的 Recall@1 分别有 5%和 1.2%的提升，Recall@5 稍微降低了 0.1%左右，Recall@10 提升了约 1.3%，总体上有明显的性能提升。另一方面，计算量和参数量都远大于 ICSA 的 SENet 和 CBAM 注意力模块的添加反而使得性能有不同程度的下降。具体的图像检索结果如图 4 所示。

表 1　不同注意力模块的性能比较

Table 1　Performance comparison of different attention modules

（单位：%）

network	R@1	R@5	R@10
VGG16	84.1	**94.6**	95.5
VGG16+IA	84.6	94.4	**96.9**
VGG16+ICSA	**85.3**	94.5	96.8
VGG16+SENet	73.5	88.0	91.9
VGG16+CBAM	74.4	89.3	93.0

当主干网络为 VGG16 时，网络中输入注意力模块特征的尺寸为 $C\times H\times W=512\times 30\times 40$，不同注意力模块的参数量与运行一次的耗时如表 2 所列。虽然 IA 的运行时间短于其他模块，但参数量却远大于 ICSA 及其他模块，本文认为如此大的参数量是 IA 性能低于 ICSA 的一个原因。而 ICSA 耗时仅略小于 IA，参数量较少，且在性能上有优势，因此本文牺牲了时间上的微小优势，在后续实验中会使用 ICSA 模块进行实验。

为了进一步分析不同注意力模块的区别，提取不同注意力模块处理得到的特征，以及不添加注意力模块时输入 NetVLAD 层的图像特征，将其根据不同空间位置进行拆分，对于每个输入图像获得了 $H\times W$ 个长度为 C 的特征向量，对测试集中所有图片的特征向量进行分析研究。将这些特征向量拼接为矩阵后进行奇异值分解（Singular Value Decomposition，SVD），其每个奇异值与最大奇异值的比值如图 5 所示，横坐标表示第 x 大的奇异值，纵坐标表示该奇异值与最大奇异值的比值。在 VGG16 和 VGG16+IA 两个网络中，超过最大奇异值 0.1 的奇异值有 3 个，而 VGG16+ICSA 中只有两个。超过最大奇异值 0.01 的奇异值个数在

VGG16，VGG16＋IA 和 VGG16＋ICSA 中分别为 35，30 和 22。因此，VGG16＋ICSA 网络中特征向量的主成分数量明显小于另外两种网络。

图 4 图像检索结果示例

Fig. 4 Example of CBIR results

表 2 不同注意力模块参数大小与运行时间的比较

Table 2 Comparison of different attention modules on size and speed

network	params size/kB	time
IA	2400.0	8.4×10^6
ICSA	6.7	11.9×10^6
SENet	128.0	10.2×10^6
CBAM	256.4	79.6×10^6

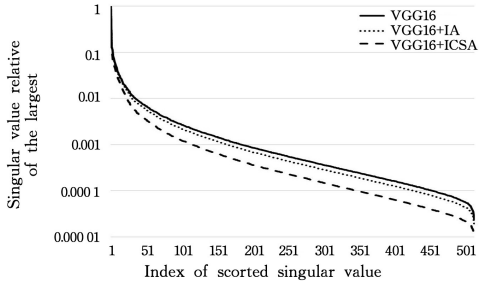

图 5 不同网络中特征向量的奇异值

Fig. 5 Singular values of feature vectors in different networks

为了更直观地观察特征向量在空间中的分布，本文使用 t-SNE[33] 将不同网络的特征向量一并投影到 2 维平面上将其可视化，VGG16，VGG16＋IA 和 VGG16＋ICSA 的投影结果分别如图 6－图 8 所示。图 6－图 8 中，由方框内所示的部分可以发现从 VGG16，VGG16＋IA 到 VGG16＋ICSA 的特征向量依次变得稀疏。

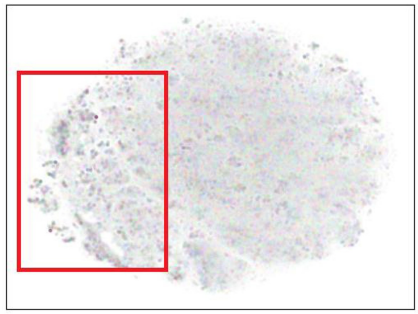

图 6 VGG16 中特征向量使用 t-SNE 可视化

Fig. 6 Visualization of feature vectors in VGG16 with t-SNE

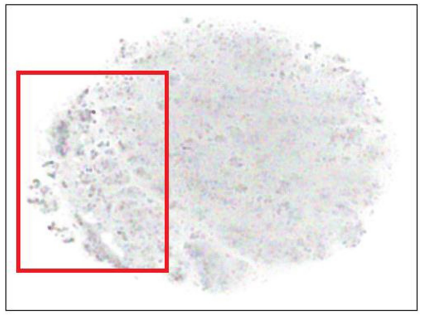

图 7 VGG16＋IA 中特征向量使用 t-SNE 可视化

Fig. 7 Visualization of feature vectors in VGG16＋IA with t-SNE

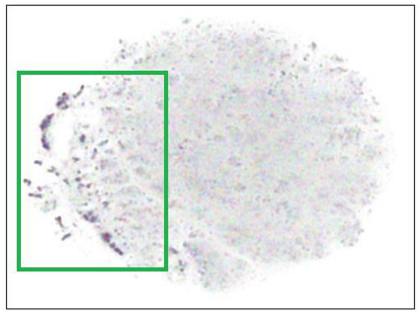

图 8 VGG16＋ICSA 中特征向量使用 t-SNE 可视化

Fig. 8 Visualization of feature vectors in VGG16＋ICSA with t-SNE

图 9 与图 10 分别为表 1 中 VGG16＋ICSA 网络训练结果的 ICSA 模块中通道注意力以及空间注意力权重的热力图，其中 512 维的通道注意力被转换成了 16×32 的矩阵以方便可视化。通道注意力权重在 0.989 到 0.994 之间浮动，不同通道之间的注意力有些许区别，而各个空间注意力保持一致，其权重均为 0.992。为了防止结果的偶然性，对本文后续中提到的 ICSA 模块训练的权重也进行了分析，均为通道注意力权重在小范围内波动，而各个空间注意力权重基本一致（波动范围小于10^{-10}）。这样的权重分布也使得 ICSA 具有更好的泛化性能。

为了验证波动较小的通道和空间注意力对网络的性能有促进作用，本文对只使用 ICSA 的通道注意力模块的情况进行了实验，如表 3 中的 VGG16＋CH 部分所示。由表 3 可以看

出，相比单纯的 VGG16 网络，VGG16＋CH 在 Recall@5 指标上基本一致，Recall@1 和 Recall@10 分别有 0.6%和 1.5%的性能提升。在此基础上加上空间注意力后（即 VGG6＋ICSA），Recall@1 进一步增加了 0.6%。因此，ICSA 模块中通道和空间注意力均对网络性能的提升有帮助。

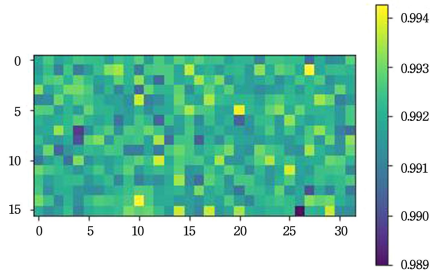

图 9　VGG16＋ICSA 中通道注意力权重可视化

Fig. 9　Visualization of reshaped channel attention weights in VGG16＋ICSA

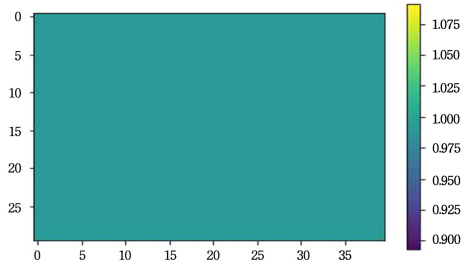

图 10　VGG16＋ICSA 中空间注意力可视化

Fig. 10　Visualization of spatial attention weights in VGG16＋ICSA

表 3　ICSA 模块不同注意力的有效性分析

Table 3　Analysis of contribution of different parts in ICSA module

network	R@1	R@5	R@10
VGG16＋ICSA	**85.3**	94.5	96.8
VGG16＋CH	84.7	94.5	**97.0**
VGG16	84.1	**94.6**	95.5

4.2.2　不同主干网络的比较

本文对不同主干网络在使用以及不使用 ICSA 模块下的性能进行了实验对比，其结果如表 4 所列。在该实验中所有的主干网络只对最后一个模块进行微调，其他参数不参与训练。从结果中可以看到，对于所有主干网络，添加 ICSA 模块后对性能均有一定的提升。VGG16＋ICSA 在所有指标上超过了其他网络。

表 4　不同主干网络的比较

Table 4　Comparison of different backbone network

（单位：%）

network	R@1	R@5	R@10
VGG16	84.6	94.4	97.1
VGG16＋ICSA	**85.3**	**94.5**	**96.8**
ResNet18	44.6	66.0	75.1
ResNet18＋ICSA	44.7	66.1	75.3
AlexNet	53.8	72.9	79.9
AlexNet＋ICSA	54.6	73.7	80.7

4.2.3　训练部分主干网络参数研究

为了进一步优化网络性能，尝试在主干网络为 VGG16 的前提下，从它的不同网络块开始训练，结果如表 5 所列。其中，VGG16 从输入到输出共有 5 个网络模块，分别用 block1，block2，…，block5 表示。如 block1 表示从 VGG16 的最开始一个模块进行训练，none 表示只是用预训练的 VGG16 网络，不对其参数进行微调。从结果中我们可以看到，总体上来说，随着训练层数的增加，召回率有一定的提升。但训练最开始两块网络后召回率反而有些许下降，该下降可能是由于网络对训练数据集过拟合导致的。

表 5　部分训练的 VGG16 主干网络的性能比较

Table 5　Performance comparison of partial trained VGG16 backbone

Lowest trained block	VGG16			VGG16＋ICSA		
	R@1	R@5	R@10	R@1	R@5	R@10
none	80.5	91.8	95.2	76.7	89.9	93.5
block5	84.1	94.6	95.5	85.3	94.5	96.8
block4	85.1	94.4	96.1	86.8	95.0	96.8
block3	**85.5**	94.6	96.5	86.8	95.0	96.8
block2	84.5	94.6	**96.6**	**86.9**	**95.2**	**97.0**
block1	84.2	**94.7**	96.1	86.4	94.6	96.6

4.2.4　泛化能力

表 6 列出了不同网络在 TokyoTM[31] 数据集上的测试结果，测试使用的网络模型参数仅在 Pittsburgh 数据集上进行了训练，与表 1 中对应项的网络参数保持一致，未在 TokyoTM 进行进一步微调，用于比较网络在添加 ICSA 注意力模块与否时的泛化能力。在添加 ICSA 模块后，Recall@5 和 Recall@10 指标有一定下降，但 Recall@1 提升了 2.3%，说明 ICSA 模块并没有限制网络的泛化能力，反而在一定情况下有提升。

表 6　网络的泛化能力比较

Table 6　Generalization ability comparison of networks

（单位：%）

network	R@1	R@5	R@10
VGG16	0.191	**0.761**	**0.898**
VGG16＋ICSA	**0.214**	0.752	0.877

结束语　本文提出了两种新型的注意力机制 IA 和 ICSA，它们与传统的注意力机制的主要区别在于独立于输入特征。ICSA 参数量远小于传统注意力模块，并且保证了较快的运行速度。并且由于该机制结构简单，可以方便地融入各种模型中。实验结果表明，ICSA 模块结合 NetVLAD 层后在图像检索任务中对各种不同主干网络都有一定的性能提升。

参 考 文 献

[1]　LEW M S，SEBE N，DJERABA C，et al. Content-based multimedia information retrieval[J]. ACM Transactions on Multimedia Computing，Communications，and Applications，2006，2(1)：1-19.

[2]　SMEULDERS A W M，WORRING M，SANTINI S，et al. Content-based image retrieval at the end of the early years[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2000，22(12)：1349-1380.

[3]　CHANG S K，HSU A. Image information systems：where do we go from here? [J]. IEEE transactions on Knowledge and Data Engineering，1992，4(5)：431-442.

[4]　SIVIC J，ZISSERMAN A. Video Google：A text retrieval approach to object matching in videos [C]// IEEE International Conference on Computer Vision. IEEE Computer Society，2003：1470-1470.

[5]　FEI-FEI L，PERONA P. A bayesian hierarchical model for

learning natural scene categories [C]// 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, 2005: 524-531.
[6] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.
[7] JÉGOU H, DOUZE M, SCHMID C, et al. Aggregating local descriptors into a compact image representation [C]// 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010: 3304-3311.
[8] PERRONNIN F, SÁNCHEZ J, MENSINK T. Improving the fisher kernel for large-scale image classification [C]// European Conference on Computer Vision. Springer. 2010: 143-156.
[9] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 60(6): 84-90.
[10] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv: 1409. 1556, 2014.
[11] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.
[12] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv: 2010. 11929, 2020.
[13] BABENKO A, SLESAREV A, CHIGORIN A, et al. Neural codes for image retrieval [C]// European Conference on Computer Vision. Springer. 2014: 584-599.
[14] LAI H, PAN Y, LIU Y, et al. Simultaneous feature learning and hash coding with deep neural networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3270-3278.
[15] NOROUZI M, FLEET D J, SALAKHUTDINOV R R. Hamming distance metric learning[J]. Advances in Neural Information Processing Systems, 2012, 25: 1061-1069.
[16] ZHANG R, LIN L, ZHANG R, et al. Bit-scalable deep hashing withregularized similarity learning for image retrieval and person re-identification[J]. IEEE Transactions on Image Processing, 2015, 24(12): 4766-4779.
[17] ARANDJELOVIC R, GRONAT P, TORII A, et al. NetVLAD: CNN architecture for weakly supervised place recognition [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 5297-5307.
[18] ONG E J, HUSAIN S, BOBER M. Siamese network of deep fisher-vector descriptors for image retrieval[J]. arXiv: 1702. 00338, 2017.
[19] RADENOVIĆ F, TOLIAS G, CHUM O. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples [C]// European Conference on Computer Vision. Springer. 2016: 3-20.
[20] BROWN A, XIE W, KALOGEITON V, et al. Smooth-ap: Smoothing the path towards large-scale image retrieval [C]// European Conference on Computer Vision. Springer, 2020: 677-694.
[21] BABENKO A, LEMPITSKY V. Aggregating local deep features for image retrieval [C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 1269-1277.
[22] KALANTIDIS Y, MELLINA C, OSINDERO S. Cross-dimensional weighting for aggregated deep convolutional features [C]// European Conference on Computer Vision. Springer, 2016: 685-701.
[23] ITTI L, KOCH C, NIEBUR E. A model of saliency-based visual attention for rapid scene analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(11): 1254-1259.
[24] MNIH V, HEESS N, GRAVES A. Recurrent models of visual attention[J]. Advances in Neural Information Processing Systems, 2014, 27: 2204-2212.
[25] JADERBERG M, SIMONYAN K, ZISSERMAN A. Spatial transformer networks[J]. Advances in Neural Information Processing Systems, 2015, 28.
[26] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7132-7141.
[27] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module [C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 3-19.
[28] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 2: 6000-6010.
[29] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7794-7803.
[30] GUO M H, XU T X, LIU J J, et al. Attention Mechanisms in Computer Vision: A Survey[J]. arXiv: 2111. 07624, 2021.
[31] BALNTAS V, RIBA E, PONSA D, et al. Learning local feature descriptors with triplets and shallow convolutional neural networks [C]// Bmvc. 2016.
[32] SUTSKEVER I, MARTENS J, DAHL G, et al. On the importance of initialization and momentum in deep learning [C]// International Conference on Machine Learning. PMLR. 2013: 1139-1147.
[33] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(11): 2579-2605.

ZHANG Shunyao, born in 1996, postgraduate. His main research interests include content based image retrieval and pose estimation.

LI Huawang, born in 1973, Ph. D, professor, Ph. D supervisor. His main research interests include digital signal processing and computer science.