计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201100026

基于残差注意力网络的跨媒体检索方法

冯 姣 陆昶谕
南京信息工程大学电子与信息工程学院 南京 210044

摘 要 随着多媒体技术的快速发展,跨媒体检索逐渐替代传统的单媒体检索成为主流的信息检索方式。现有跨媒体检索方法复杂度高,且不能充分挖掘数据的细节特征,在映射的过程中会产生偏移,难以学习到精准的数据关联。针对上述问题,提出了一种基于残差注意力网络的跨媒体检索方法。首先,为了更好地提取不同媒体数据的关键特征,同时简化跨媒体检索模型,提出了融入注意力机制的残差神经网络。然后,提出了跨媒体检索联合损失函数,通过约束网络的映射过程,增强网络的语义辨别能力,提高网络检索精度。实验结果表明,与现有的一些方法对比,本文提出的基于残差注意力网络的跨媒体检索方法能够较好地学习到不同媒体数据之间的关联,有效地提高了跨媒体检索的精度。

关键词:跨媒体检索;注意力机制;残差神经网络;联合损失函数
中图法分类号 TP391

Cross Media Retrieval Method Based on Residual Attention Network

FENG Jiao and LU Chang-yu
College of Electronic and Information Engineering,Nanjing University of Information Science and Technology,Nanjing 210044,China

Abstract With the rapid development of multimedia technology,cross-media retrieval has gradually replaced traditional single-media retrieval as the mainstream information retrieval method. Existing cross-media retrieval methods are highly complex,and cannot fully mine the detailed characteristics of the data,which will cause deviations in the mapping process,and it is difficult to learn accurate data associations. To solve the above problems,this paper proposes a cross-media retrieval method based on residual attention network(CR-RAN). First of all,in order to better extract the key features of different media data and simplify the cross-media retrieval model,this paper proposes a residual neural network incorporating the attention mechanism. Then this paper proposes a cross-media retrieval joint loss function,which enhances the semantic discrimination ability of the network and improves the accuracy of network retrieval by constraining the mapping process of the network. Experimental results show that,compared with some existing methods,the cross-media retrieval method based on residual attention network proposed in this paper can better learn the association between different media data and effectively improve the accuracy of cross-media retrieval.

Keywords Cross media retrieval,Attention mechanism,Residual neural network,Joint loss function

1 引言

随着多媒体技术和网络技术的快速发展,海量的多媒体数据喷涌而出,关于多媒体内容理解的研究也得以迅速发展,跨媒体检索技术便是其中最新的研究热点之一。跨媒体检索,即使用者输入任意媒体类型的数据作为查询对象,检索出不同媒体类型中与之语义相关的数据[1]。如图1所示,使用者输入图像或文本作为查询对象,来检索与“狮子”相关的图像或文本。与单媒体检索相比,跨媒体检索能够为使用者提供更加灵活有效的检索方式。

然而不同媒体之间的数据分布以及表征不一致,存在“异构鸿沟”问题,难以实现不同媒体之间的语义关联,这给度量不同媒体之间的相似性带来了很大的挑战[2]。目前,解决“异构鸿沟”问题的一个主要思路是将不同媒体类型的数据统一映射到一个共同子空间,利用不同媒体数据在此空间中的表征来度量数据之间的距离,通过距离的大小来表示数据之间相关性的程度。在过去的研究中,有大量方法使用上述思想来解决跨媒体检索问题,其大致可以分为两类:基于统计分析的传统方法和基于深度学习的方法[3]。

传统方法主要是通过统计分析的方式来完成媒体特征的投影映射,例如早期的典型相关分析(Canonical Correlation Analysis,CCA)[4]。CCA作为一种常用的空间学习方法,通常用于寻找一对映射矩阵,使两种特征表示之间的相关性达到最大,是传统方法中最具代表性的一种。CCA对跨媒体检索研究有着巨大的影响,后续的大量工作都是在CCA方法的基础上进行扩展。例如,Hardoon等[5]将核函数的思想引入传统的CCA方法中,提出了核典型相关分析方法(Kernel Canonical Correlation Analysis,KCCA),增强了CCA学习非线性跨媒体关联的能力。Nikhil等[6]提出了语义关联匹配(Semantic Correlation Matching,SCM)方法,SCM将交叉关联和语义抽象引入跨媒体检索任务中,提高了检索的准确度。Zhang等[7]提出了一种弱匹配概率典型相关性分析模型,

基金项目:国家自然科学基金(61501244)
This work was supported by the National Natural Science Foundation of China(61501244).
通信作者:冯姣(jiao.feng@nuist.edu.cn)

解决了数据不足的情况下出现的过拟合问题。传统方法在早期的跨媒体检索任务中取得了不错的效果，为后续的研究带来了很大的启发。但无论是 CCA 还是其扩展方法，都仅仅是最大化不同媒体数据之间的相关性，并没有在共同子空间中拉近数据之间的距离，数据映射会产生一定的偏差。

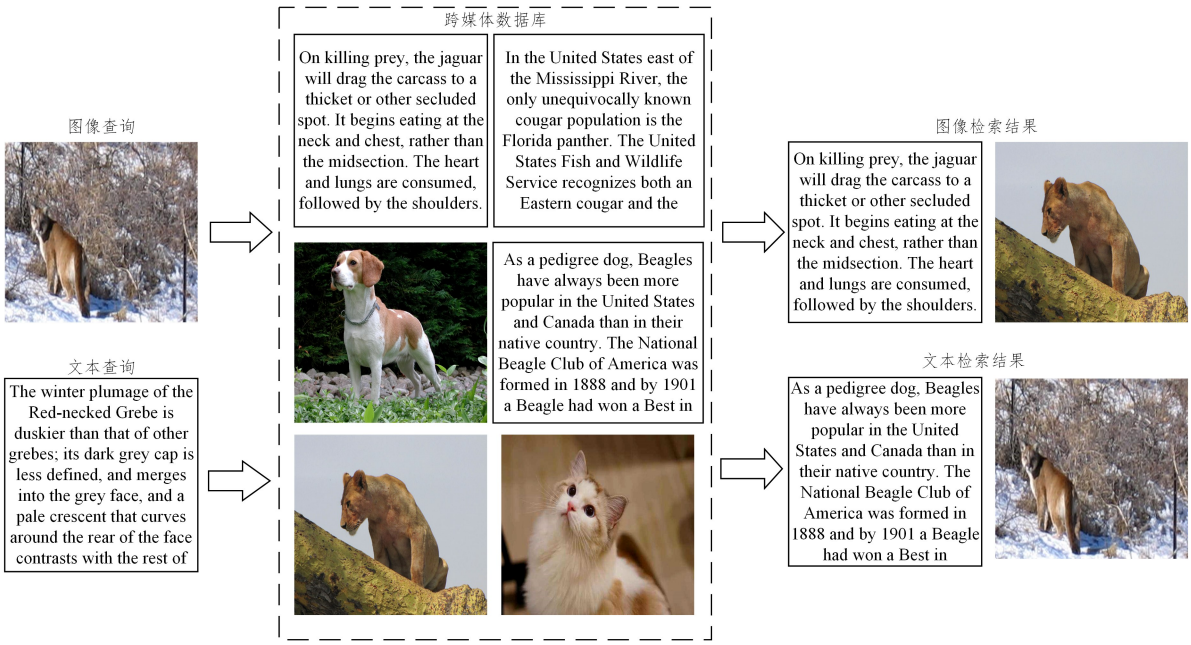

图 1　跨媒体检索示例

Fig. 1　Cross-media retrieval example

基于深度学习的方法，利用多层非线性神经网络对多媒体数据进行特征提取，相较于传统方法有效提高了对多媒体数据的映射能力。Andrew 等[8]提出了深度典型相关性分析(Deep Canonical Correlation Analysis，DCCA)方法，DCCA 将深度学习与传统的 CCA 相结合，以此来学习不同媒体数据形式间的复杂非线性投影关系。Peng 等[9]提出了跨媒体多深度网络(Cross-media Multiple Deep Network，CMDN)，通过层次学习来利用复杂的跨媒体相关性。He 等[10]提出了一个统一的深度模型，通过残差网络，在不区分处理的情况下，同时学习多种媒体数据。基于深度学习的方法充分挖掘了多媒体数据之间的语义关联，有效提高了跨媒体检索任务的精度。但现有方法往往忽略了数据内部的细粒度信息，无法有效地突出数据内部具有语义辨识性的关键区域，从而降低了模型在细粒度图像检索任务上的精度。

针对上述问题，本文提出了一种基于残差注意力网络的跨媒体检索方法。首先提出了融入注意力机制的跨媒体残差神经网络，通过同一网络并行挖掘多种媒体数据的细粒度特征，并充分学习不同媒体数据之间的关联。同时，本文提出了基于分类损失函数和岛屿损失函数(Island loss)的跨媒体联合损失函数，通过分类损失函数，最大化不同类别在统一空间的分布距离；通过岛屿损失函数，减小语义相同的数据在统一空间的映射距离。为了验证方法的有效性，本文在两个广泛使用的跨媒体数据集上与现有方法进行对比。实验结果表明，本文方法有效提高了跨媒体检索的精度。

2　CR-RAN

为了提高跨媒体检索精度，本文提出了一种基于残差注意力网络的跨媒体检索方法，其网络结构如图 2 所示。首先，对图像和文本两种媒体数据进行预处理，实现对不同媒体数据的统一表征。然后，构建针对这两种媒体数据的残差注意力神经网络，挖掘数据的细粒度特征，并充分学习数据之间的关联。最后，设计跨媒体联合损失函数，通过结合分类损失函数和岛屿损失函数的方法，联合优化多媒体数据到统一空间的映射，从而学习到更加精准的跨媒体关联关系。

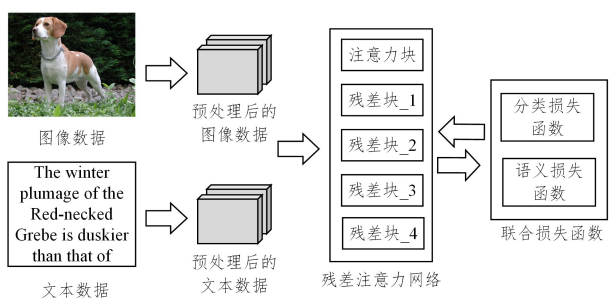

图 2　残差注意力网络框架示意图

Fig. 2　Schematic diagram of residual attention network framework

2.1　数据预处理

图像和文本两种媒体数据，在输入网络之前均需进行一系列的预处理。具体地，对于图像数据，本文使用双线性插值法将图像的大小统一调整为 $m^i \times n^i$ 像素，其中 m^i 和 n^i 分别表示图像的高和宽。对于文本数据，其处理流程如图 3 所示。首先，需要对原始文本进行清洗，其中，清洗主要包括缩写替换、标点删除、停用词删除以及词形还原等过程。对于清洗后的文本，通过使用 GloVe 词嵌入将文本中的每一个单词转化为词向量，以此获得文本向量，维度为 $m^t \times d$，其中 m^t 表示文本中所有单词的数量，d 表示词向量的维数。最后，为了对文本向量进行初步的特征提取，同时使得文本向量的维度符合残差网络的输入格式，对文本向量进行一系列的卷积处理，最终得到输出 X_t，具体设置将在 3.3 节中详细描述。

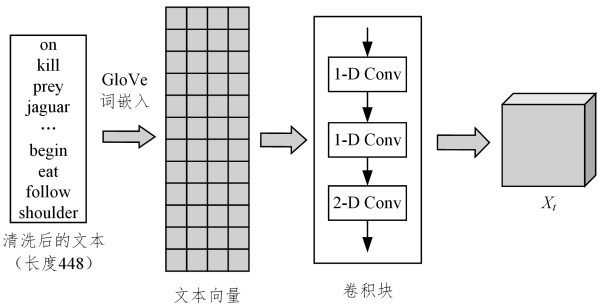

图 3 文本数据处理流程

Fig. 3 Text data processing flow

2.2 残差注意力网络

为了充分学习预处理后的图像数据以及文本数据，本文构建了用于跨媒体检索的残差注意力网络，其模型结构如图 4 所示。

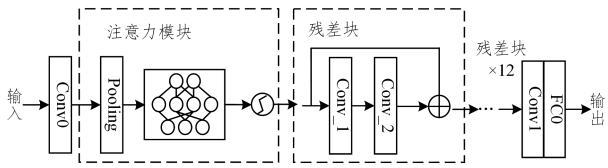

图 4 残差注意力网络模型图

Fig. 4 Residual attention network model diagram

2.2.1 残差网络

残差网络在一定程度上解决了因网络层数增加所导致的网络退化问题，且其特殊的残差块结构使得神经网络的搭建更加灵活、更容易优化[11-12]。如图 4 所示，残差网络由一系列的残差块构成。其中，每个残差块都由恒等映射和残差映射组成，其结构如图 5 所示。

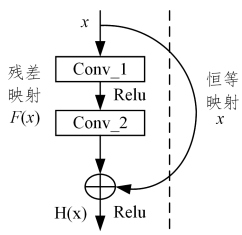

图 5 残差块结构图

Fig. 5 Residual block structure diagram

根据其结构，残差块可表示为：

$$H(x)=F(x)+x \tag{1}$$

其中，$H(x)$是残差结构的输出，$F(x)$是残差部分，x 是残差结构的输入。通过这个特殊的结构，底层的数据可以直接作为后面某一层的输入，有效解决了模型因层数增加而造成的退化问题。

搭建完成后的残差网络不需要对预处理后的图像以及文本进行区分处理，只需将两个不同数据张量在通道维度上进行拼接便可同时输入残差网络中。残差网络的最终输出便是图像和文本在统一空间的映射表征，其中一部分为图像数据的统一映射表征，另一部分为文本数据的统一映射表征。

2.2.2 注意力机制

在原有的残差网络基础上，本文将注意力机制[13]融入其中，提高了网络对关键信息的捕捉能力。注意力机制可以实现对关键信息的关注，极大提高信息处理的效率和准确率。本文所采用的注意力机制为通道注意力机制，即注意力关注通道域。参考残差网络中残差块的思想，本文将通道注意力机制设计为一种可堆叠的结构，可以很好地嵌入到网络模型中，方便网络模型的训练与优化。

通道注意力模块的具体结构如图 6 所示。通道注意力模块首先对输入同时使用最大值池化（MaxPool）和均值池化（AvgPool），以此来对空间信息进行聚合。其中，均值池化对输入数据中的每一个点都有反馈，能够综合考虑数据的整体特征。而最大值池化只会对输入数据中响应最大的地方进行反馈，能够识别到数据中相对重要的内容。

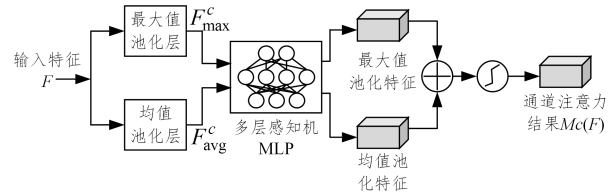

图 6 通道注意力模块结构

Fig. 6 Channel attention module structure

然后，通道注意力模块会将聚合的空间信息送入一个共享的多层感知机（Multilayer Perceptron，MLP），对空间信息进行分析。最后，使用按元素求和的方法合并感知机输出的最大值池化特征和均值池化特征，合并后的特征通过 sigmoid 函数进行输出，得到最终的通道注意力特征。其具体计算过程如下：

$$\begin{aligned} M_C(F)&=\sigma(MLP(AvgPool(F))+MLP(MaxPool(F)))\\ &=\sigma(W_1(W_0(F^c_{avg}))+W_1(W_0(F^c_{max}))) \end{aligned} \tag{2}$$

其中，F 表示输入特征，W_0 和 W_1 表示多层感知机中的两层参数，F^c_{avg}表示经过均值池化后的空间信息，F^c_{max}表示经过最大值池化后的空间信息。

2.3 联合损失函数

如图 2 所示，残差注意力网络在提取到不同媒体数据的特征之后，需要将其映射到统一空间中。在训练的过程中，需要计算残差注意力网络所提取到的数据特征与其所对应的标签之间的余弦距离[14]。

为了学习到更为精准的映射关系，本文设计了一种混合损失函数，其定义为：

$$L=L_S+\lambda L_I \tag{3}$$

其中，L_S 表示分类损失函数，L_I 表示语义损失函数，λ 用于调节两项损失函数的比例。

2.3.1 分类损失函数

分类损失函数，其主要目的在于保证网络的类间判别能力。本文采用两组交叉熵损失函数相结合的方式作为分类损失函数，具体定义如下：

$$L_S=\frac{1}{NI}\sum_{k=1}^{N_I} l(x_k^I,y_k^I)+\frac{1}{N_T}\sum_{k=1}^{N_T} l(x_k^T,y_k^T) \tag{4}$$

其中，$l(x_k^I,y_k^I)$表示图像的交叉熵损失函数，N_I 表示训练集中图像数据的个数，x_k^I 表示第 k 个图片的残差注意力网络输出特征，y_k^I 表示第 k 个图像数据的标签。$l(x_k^T,y_k^T)$表示文本的交叉熵损失函数，N_T 表示训练集中文本数据的个数，x_k^T 表示第 k 个文本的残差注意力网络输出特征，y_k^T 表示第 k 个文本数据的标签。

2.3.2 语义损失函数

为了提高网络的映射能力，获得更好的跨媒体检索性能，

网络需要在确保不同类别可分离的情况下，减少类别内部的差异以及映射偏移。为此，本文采用岛屿损失函数[15]作为语义损失函数。岛屿损失函数分别计算样本与其类别中心的余弦距离以及不同类别中心之间的距离，以此减小数据在映射过程中的类内距离，同时扩大类间差异。其定义如下：

$$L_I=\frac{1}{2}\sum_{k=1}^{n}\|x_k-c_{y_k}\|_2^2+\lambda_1\sum_{c_j\in N}\sum_{c_k\in N}\left(\frac{c_k\cdot c_j}{\|c_k\|_2\|c_j\|_2}+1\right) \tag{5}$$

不同于分类损失函数，在语义损失函数中不区分数据的类型，无论是图片数据还是文本数据，都统一对待。因此，在语义损失函数的定义中，n 表示训练集中所有媒体类型数据的个数，x_k 表示第 k 个训练数据的特征，c_{y_k} 表示 y_k 类别中心特征，λ_1 表示调节因子，N 表示类别标签的集合，c_k 和 c_j 分别表示第 k 个和第 j 个类别中心特征。在训练过程中，类别中心特征根据每批次的数据计算更新。

3　实验结果

3.1　数据集

本文在两个被广泛使用的跨媒体数据集上进行了实验：Wikipedia 数据集以及 Pascal Sentence 数据集。

Wikipedia 数据集[6]在跨媒体检索任务的评估中有着广泛的应用，它是从维基百科中的精选文章挑选所得。该数据集共有 2 866 个图像/文本对，分别来自 10 个不同的语义类别，例如生物、体育、历史等类别。文献[16]中，将其分为训练集、测试集以及验证集 3 个部分。其中，训练集包含 2 173 个图像/文本对，测试集包含 462 个图像/文本对，验证集包含 231 个图像/文本对。

Pascal Sentence 数据集[17]是从 2008PASCAL 开发工具包中提取。该数据集共有 1 000 个图像/文本对，分别来自 10 个不同的语义类别，其中每个图像所对应的文本由 5 句人工标注的语句组成。文献[16]中，将数据集随机分为训练集、测试集以及验证集 3 个部分。其中，训练集包含 800 个图像/文本对，测试集和验证集均包含 100 个图像/文本对。

3.2　实验设置

在预处理阶段，本文统一将图像转化为 448 * 448 像素大小，即 m^i 和 n^i 都设为 448。同时，本文使用 RGB 空间表示图像，最终预处理后的图像数据维度为 3 * 448 * 448。

对于文本数据，本文使用 100 维的 GloVe 词嵌入进行文本表示。同时，通过对数据集中的文本长度进行统计，发现只有极少量的文本长度会超过 448。因此，本文设定一个文本的最大单词数为 448。如果文本长度小于 448，则对文本向量填充 0；如果文本长度大于 448，则对文本向量进行截断处理，即 m^t 设为 448，d 设为 100。在得到一个维度为 448 * 100 的文本向量之后，本文通过 2 个大小为 3 * 3 的一维卷积核以及 1 个大小为 3 * 3 的二维卷积核对文本向量进行初步特征提取，同时使得文本向量的维度与处理后的图像数据维度相一致，便于输入模型之中。最终得到一个维度为 3 * 448 * 448 的文本向量。

3.3　实验方法及标准

为了验证本文方法的有效性，实验中分别进行了以下两项跨媒体检索任务。

(1)图像检索文本任务，即给定任意一个测试集中的图像样本作为查询对象，计算测试集中所有文本样本与查询图像样本之间的余弦距离，并根据余弦距离对文本样本进行排序返回。

(2)文本检索图像任务则与图像检索文本任务相反，以测试集中任意一个文本样本作为查询对象，对测试集中所有图像样本进行排序返回。

对于检索任务的准确度，主要通过检索结果的精度以及排序进行评价，因此本文采用平均准确率均值(Mean Average Precision，MAP)作为评价指标。MAP 作为一种可以公平且全面地评价检索结果优劣的评价指标，在信息检索领域有着广泛的使用。

在不同的检索任务实验中，测试集中所有相应的样本分别被单独作为查询对象进行检索，并通过检索的排序结果计算得到每一个样本的平均精度(Average Precision，AP)。最后将所有样本的 AP 值进行平均，即可得到相应检索任务的 MAP 值。其中，AP 计算如下：

$$AP=\frac{1}{R}\sum_{k=1}^{m}\frac{R_k}{k}\times rel_k \tag{6}$$

其中，R 表示测试集中相关样本的个数，m 表示测试集大小，R_k 表示前 k 个结果中相关样本的个数，rel_k 在第 k 个结果是相关时为 1，否则为 0。

3.4　实验结果与分析

(1)为了对比不同层数的残差网络在跨媒体检索任务中的效果，本文对 34 层、50 层以及 101 层的残差网络进行了实验对比。在本实验中，只使用最基本的网络结构和交叉熵损失函数，其结果如表 1 所列。可以看出，与 34 层和 101 层的残差网络相比，50 层的残差网络在 Wikipedia 数据集和 Pascal Sentence 数据集上均取得了更高的检索精度。因此，本文选择 50 层的残差网络作为模型的基础网络。

表 1　不同层数的残差网络效果的对比结果

Table 1　Comparison results of residual network effects with different layers

数据集	残差网络层数	图像检索文本	文本检索图像	平均
Wikipedia	34	0.509	0.489	0.499
	50	0.514	0.492	0.503
	101	0.508	0.487	0.497
Pascal Sentence	34	0.415	0.418	0.416
	50	0.425	0.435	0.430
	101	0.413	0.416	0.414

(2)实验对比了现有的 3 种经典的传统跨媒体检索方法：CCA[4]，SCM[6]以及 CDL[18]；同时对比了 3 种基于深度学习的跨媒体检索方法：Corr-AE[14]，CMDN[9]以及 CP-VGG[19]。

本文方法以及对比方法在 Wikipedia 数据集和 Pascal Sentence 数据集的实验结果分别如表 2 和表 3 所列。

表 2　Wikipedia 数据集上跨媒体检索的 MAP 结果

Table 2　MAP results of cross-media retrieval on the Wikipedia dataset

方法	图像检索文本	文本检索图像	平均
CCA[4]	0.332	0.317	0.324
SCM[6]	0.375	0.393	0.384
CDL[18]	0.282	0.223	0.253
Corr-AE[14]	0.326	0.361	0.344
CMDN[9]	0.393	0.325	0.359
CP-VGG[19]	0.498	0.393	0.446
CR-RAN(本文)	0.514	0.492	0.503

表 3 Pascal Sentence 数据集上跨媒体检索的 MAP 结果

Table 3 MAP result of cross-media retrieval on Pascal Sentence dataset

方法	图像检索文本	文本检索图像	平均
CCA[4]	0.380	0.372	0.376
SCM[6]	0.407	0.394	0.400
CDL[18]	0.384	0.306	0.345
Corr-AE[14]	0.290	0.279	0.285
CMDN[9]	0.334	0.333	0.333
CP-VGG[19]	0.311	0.442	0.376
CR-RAN(本文)	0.425	0.435	0.430

从对比结果可以看出，本文方法在两个数据集上均取得了较高的 MAP 值，提高了跨媒体检索准确度，相比其他方法，MAP 值均有所提升。

(3)为了验证注意力机制和联合损失函数的效果，本文进行了一组基线实验的对比，其结果如表 4 所列。其中，“基线实验”表示在模型中同时去掉注意力机制以及联合损失函数，只使用交叉熵损失函数。“无注意力”表示去掉模型中的通道注意力机制。“无联合损失”表示在模型中不使用联合损失函数。从结果可以看出，单独使用通道注意力机制或联合损失函数时，能够有效地提高跨媒体检索的准确率。此外，在同时使用注意力机制和联合损失函数的情况下，跨媒体检索模型的准确率有着明显的提升，这说明两者能够相互促进，进一步提高跨媒体检索模型的关联学习能力。

表 4 Wikipedia 数据集和 Pascal Sentence 数据集上基线实验的 MAP 结果

Table 4 MAP results of baseline experiments on Wikipedia dataset and Pascal Sentence dataset

数据集	方法	图像检索文本	文本检索图像	平均
Wikipedia	RACMR	0.514	0.492	0.503
	无注意力	0.509	0.489	0.499
	无联合损失	0.508	0.487	0.497
	基线实验	0.499	0.488	0.493
Pascal Sentence	RACMR	0.425	0.435	0.430
	无注意力	0.415	0.418	0.416
	无联合损失	0.413	0.416	0.414
	基线实验	0.340	0.376	0.358

结束语 本文提出了一个基于残差注意力网络的跨媒体检索方法。该方法首先通过残差注意力网络对预处理后的跨媒体数据提取特征并进行统一映射，然后利用联合损失函数约束网络的映射过程，使得网络学习到更加精准的跨媒体关联关系。该方法将注意力机制以及联合损失函数融入其中，增强了网络对关键特征的捕捉能力，更好地学习跨媒体数据之间的关联。实验结果表明，该方法具有较高的跨媒体检索准确度。下一步工作将尝试扩展现有模型，使得模型支持更多类型的媒体数据。同时，继续尝试在不同尺度上挖掘跨媒体数据之间的深层关联。

参 考 文 献

[1] QI J W, PENG Y X, YUAN Y X. Cross-media retrieval with hierarchical recurrent attention network[J]. Journal of Image and Graphics, 2018, 23(11): 1751-1758.

[2] PENG Y X, QI J W, HUANG X. Current Research and Prospects on Multimedia Content Understanding[J]. Journal of Computer Research and Development, 2019, 56(1): 183-208.

[3] ZHUO Y K, QI J W, PENG Y X. Cross-media deep fine-grained correlation learning[J]. Ruan Jian Xue Bao/Journal of Software, 2019, 30(4): 884-895.

[4] HOTELLING H. Relation between two sets of variates[J]. Biometrika, 1936, 28(3/4): 321-377.

[5] HARDOON D R, SZEDMAK S, SHAWE-TAYLOR J. Canonical Correlation Analysis: An Overview with Application to Learning Methods[J]. Neural Computation, 2004, 16(12): 2639-2664.

[6] RASIWASIA N, PEREIRA J C, COVIELLO E, et al. A New Approach to Cross-Modal Multimedia Retrieval [C]// International Conference on Multimedia. 2010: 251-260.

[7] ZHANG B, HAO J, MA G, et al. Automatic image annotation based on semi-paired probabilistic canonical correlation anlysis [J]. Ruan Jian Xue Bao/Journal of Software, 2017, 28(2): 292-309.

[8] ANDREW G, ARORA R, BILMES J, et al. Deep Canonical Correlation Analysis[C]// ICML. 2013.

[9] PENG Y X, HUANG X, QI J W. Cross-media shared representation by hierarchical learning with multiple deep networks [C]// IJCAI. 2016.

[10] HE X, PENG Y, XIE L. A New Benchmark and Approach for Fine-grained Cross-media Retrieval[C] // FGcross Net _ ACMMM 2019. 2019.

[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] ZHU W, WANG T Q, CHEN Y F, et al. Object-level Edge Detection Algorithm Based on Multi-scale Residual Network[J]. Computer Science, 2020, 47(6): 144-150.

[13] ZHANG Y, LI K, LI K, et al. Image Super-Resolution Using Very Deep Residual Channel Attention Networks[J]. arXiv: 1807.02758, 2018.

[14] LIU S, BAI L, YU T Y, et al. Cross-media Semantic Similarity Measurement Using Bi-directional Learning Ranking[J]. Computer Science, 2017, 44(S1): 84-87, 118.

[15] CAI J, MENG Z B, KHAN A S, et al. Island loss for learning discriminative features in facial expression recognition[C]// Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 302-309.

[16] FENG F X, WANG X J, LI R F. Cross-modal retrieval with correspondence utoencoder[C]// Proceedings of the 22nd ACM International Conference on Multimedia. Orlando, Florida, USA: ACM, 2014: 7-16.

[17] RASHTCHIAN C, YOUNG P, HODOSH M, et al. Collecting image annotations using Amazon's Mechanical Turk[C]// Proceeding of NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Los Angeles, California: ACM, 2010: 139-147.

[18] LIU Y, YU Z L, FU Q. Cross-media retravel method fusing with coupled dictionary learning and image regularization [J]. Computer Engineering, 2019, 45(6): 230-236.

[19] SUN Z Y. Research on Cross-media Retrival Method Based on Compression Convolutional Neural Networks[D]. Wuahn: Central China Normal University, 2020.

FENG Jiao, Ph. D, associate professor. Her main research interests include signal processing and deep learning.