软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(4):884–895 [doi: 10.13328/j.cnki.jos.005664]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

跨媒体深层细粒度关联学习方法*

卓昀侃, 綦金玮, 彭宇新

(北京大学 计算机科学技术研究所,北京 100871)

通讯作者: 彭宇新, E-mail: pengyuxin@pku.edu.cn

摘 要: 随着互联网与多媒体技术的迅猛发展,网络数据的呈现形式由单一文本扩展到包含图像、视频、文本、音频和 3D 模型等多种媒体,使得跨媒体检索成为信息检索的新趋势.然而,“异构鸿沟”问题导致不同媒体的数据表征不一致,难以直接进行相似性度量,因此,多种媒体之间的交叉检索面临着巨大挑战.随着深度学习的兴起,利用深度神经网络模型的非线性建模能力有望突破跨媒体信息表示的壁垒,但现有基于深度学习的跨媒体检索方法一般仅考虑图像和文本两种媒体数据之间的成对关联,难以实现更多种媒体的交叉检索.针对上述问题,提出了跨媒体深层细粒度关联学习方法,支持多达 5 种媒体类型数据(图像、视频、文本、音频和 3D 模型)的交叉检索.首先,提出了跨媒体循环神经网络,通过联合建模多达 5 种媒体类型数据的细粒度信息,充分挖掘不同媒体内部的细节信息以及上下文关联.然后,提出了跨媒体联合关联损失函数,通过将分布对齐和语义对齐相结合,更加准确地挖掘媒体内和媒体间的细粒度跨媒体关联,同时利用语义类别信息增强关联学习过程的语义辨识能力,提高跨媒体检索的准确率.在两个包含 5 种媒体的跨媒体数据集 PKU XMedia 和 PKU XMediaNet 上与现有方法进行实验对比,实验结果表明了所提方法的有效性.

关键词: 跨媒体检索;5 种媒体;细粒度信息挖掘;跨媒体循环神经网络;跨媒体联合关联约束

中图法分类号: TP37

中文引用格式: 卓昀侃,綦金玮,彭宇新.跨媒体深层细粒度关联学习方法.软件学报,2019,30(4):884–895. http://www.jos.org.cn/1000-9825/5664.htm

英文引用格式: Zhuo YK, Qi JW, Peng YX. Cross-media deep fine-grained correlation learning. Ruan Jian Xue Bao/Journal of Software, 2019,30(4):884–895 (in Chinese). http://www.jos.org.cn/1000-9825/5664.htm

Cross-media Deep Fine-grained Correlation Learning

ZHUO Yun-Kan, QI Jin-Wei, PENG Yu-Xin

(Institute of Computer Science and Technology, Peking University, Beijing 100871, China)

Abstract: With the rapid development of the Internet and multimedia technology, data on the Internet is expanded from only text to image, video, text, audio, 3D model, and other media types, which makes cross-media retrieval become a new trend of information retrieval. However, the “heterogeneity gap” leads to inconsistent representations of different media types, and it is hard to measure the similarity between the data of any two kinds of media, which makes it quite challenging to realize cross-media retrieval across multiple media types. With the recent advances of deep learning, it is hopeful to break the boundaries between different media types with the strong learning ability of deep neural network. But most existing deep learning based methods mainly focus on the pairwise correlation between two media types as image and text, and it is difficult to extend them to multi-media scenario. To address the above problem, Deep Fine-grained Correlation Learning (DFCL) approach is proposed, which can support cross-media retrieval with up to five media types (image, video, text, audio, and 3D model). First, cross-media recurrent neural network is proposed to jointly model the fine-grained

* 基金项目: 国家自然科学基金(61771025, 61532005)
Foundation item: National Natural Science Foundation of China (61771025, 61532005)
本文由“多媒体数据的知识关联与理解专题”特约编辑蒋树强研究员、刘青山教授、孙立峰教授、李波教授推荐.
收稿时间: 2018-04-16; 修改时间: 2018-06-13; 采用时间: 2018-09-30

information of up to five media types, which can fully exploit the internal details and context information of different media types. Second, cross-media joint correlation loss is proposed, which combines distribution alignment and semantic alignment to exploit both intra-media and inter-media fine-grained correlation, while it can further enhance the semantic discrimination capability by semantic category information, aiming to promote the accuracy of cross-media retrieval effectively. Extensive experiments on 2 cross-media datasets are conducted, namely PKU XMedia and PKU XMediaNet datasets, which contain up to five media types. The experimental results verify the effectiveness of the proposed approach.

Key words: cross-media retrieval; quintuple-media; fine-grained information mining; cross-media recurrent neural network; cross-media joint correlation constraint

在大数据时代,互联网数据以图像、视频、文本、音频等多种媒体形式广泛存在,它们是计算机感知和认知真实世界的重要载体.由于数据总量和媒体类型的迅猛增长,多媒体信息检索[1]的相关研究得以迅速发展,其中跨媒体检索[2−4]是最新的研究热点之一.跨媒体检索是指用户通过输入任意媒体类型的查询数据,检索出所有媒体类型中的语义相关数据,如图 1 所示,用户可以输入“飞机”的相关图像作为查询来检索和飞机相关的图像、视频、文本、音频和 3D 模型.相比传统的单媒体检索,例如图像检索[5]、视频检索[6]等,跨媒体检索能够更加灵活、全面地满足用户的检索需求.然而,“异构鸿沟”问题导致不同媒体类型的数据分布和特征表示之间存在不一致性,因此难以直接度量多种媒体数据之间的相似性,使得跨媒体检索面临巨大挑战.

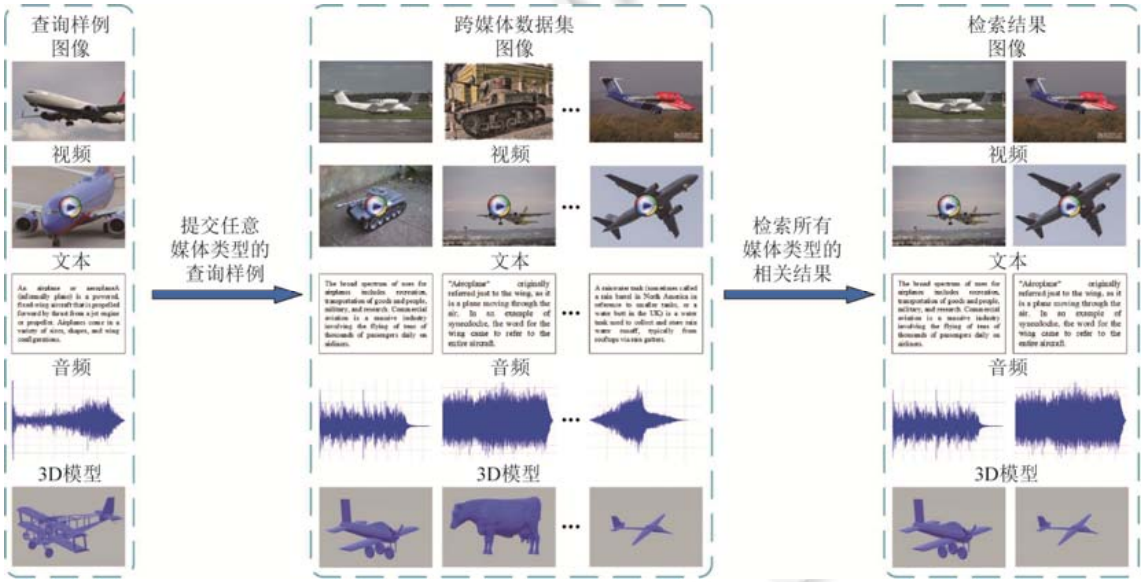

Fig.1 An example of cross-media retrieval

图 1 跨媒体检索示例

事实上,认知科学的研究表明,人类大脑能够通过多种感官信息的融合来认知外部世界[7],视觉、听觉和语言等系统能够很好地协同处理从外界接受的信息.因此,如何通过模拟人脑的认知过程,实现多媒体数据的语义互通与关联理解,是跨媒体检索需要解决的关键问题.对此,现有方法的解决思路通常是建立一个共同子空间,将不同媒体类型的异构数据映射到这个共同子空间中得到统一表征,然后通过常用的距离度量方法来直接计算不同媒体数据之间的相似性,实现跨媒体交叉检索.

根据以上思路,已有一些工作[8−10]尝试为不同媒体类型的数据学习统一表征,可以将其主要分为两类:传统方法和基于深度学习的方法.传统方法通过统计分析学习线性映射矩阵,其中,最具代表性的是典型相关分析(canonical correlation analysis,简称 CCA)[11],该方法通过最大化成对媒体数据间的关联来优化映射矩阵.另有一些工作基于典型相关分析,尝试引入其他信息提升其性能,例如语义类别信息[12]等.近年来,随着深度学习在计

算机视觉[13,14]等领域取得巨大进展,研究人员尝试通过深度网络的非线性建模能力来分析不同媒体类型数据间的复杂关联关系.Feng 等人[8]提出对应自编码器(correspondence autoencoder,简称 Corr-AE)同时对关联关系和重建信息进行建模.Peng 等人[15]提出将媒体内和媒体间的关联信息通过层次化网络的方式进行联合学习以提升检索准确率.图 2 给出跨媒体关联学习方法的框架示意.

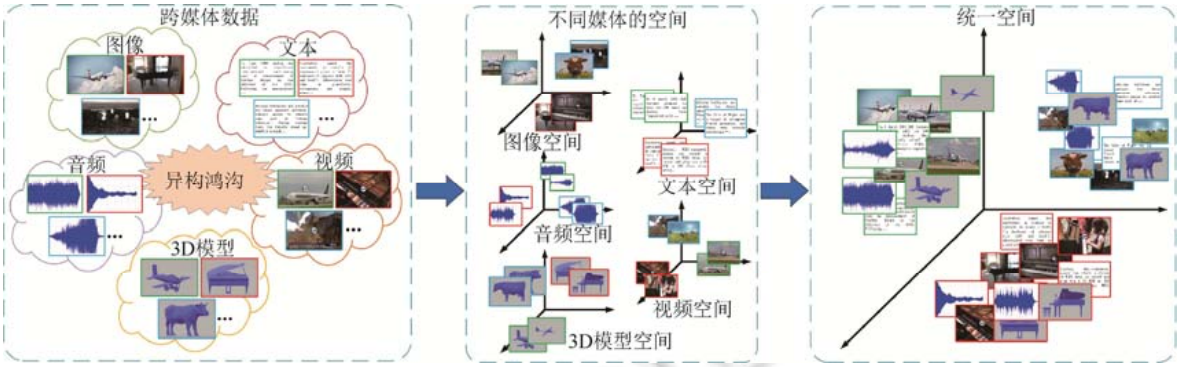

Fig.2 An illustration of the mainstream framework for cross-media correlation learning

图 2 跨媒体关联学习方法框架示意图

然而,上述方法一般仅针对图像和文本两种媒体类型的跨媒体检索任务,由于它们的泛化性能有限,很难将其扩展至更多种媒体类型的交叉检索,如典型相关分析及其变种方法[16−18]旨在分析两组变量之间的相关关系,尽管可以通过两两组合的方式来将这些方法扩充至多种媒体交叉检索的场景,但不仅无法在一个模型内解决问题,算法复杂度高,而且忽视了多种媒体关联的共存和互补性,导致关联信息有限,降低了检索的准确率.显然,在多种媒体交叉检索的场景下,挖掘不同媒体类型数据之间的语义关联更加困难.由于任意两种媒体之间都存在着异构鸿沟,而且不同媒体类型数据之间的关联关系也有各自独特的特性,现有方法很难将其同时建模在一个模型中.

事实上,描述同一语义的不同媒体类型数据存在天然的语义一致性,且数据内部蕴含着丰富的细粒度上下文信息.其中,细粒度指的是数据的局部区域或片段,上下文指的是这些区域或片段间的关联关系,如图像前景区域和背景区域之间的关系或前后视频帧之间的关系,充分利用细粒度上下文信息能够有效挖掘不同媒体数据之间的关联.例如,在多种媒体交叉检索的场景下,很可能文本的某一部分描述并未在图像中体现,但却和音频或视频的某一片段存在明显的关联.这表明,在多种媒体相互检索的任务中,不同媒体数据之间存在着丰富的语义互补关系,能够为跨媒体关联学习提供充足的线索,而且挖掘其中细粒度信息之间的语义关联尤为重要.然而,现有方法一般仅考虑了不同媒体数据的成对关联,忽略了细粒度局部上下文信息之间的语义关联.此外,现有方法一般仅使用语义类别信息来约束不同媒体数据之间的关联学习,在多种媒体的场景下,其约束能力不足以弥补多种媒体数据间的分布差异.针对上述问题,本文提出了跨媒体深层细粒度关联学习方法,同时在语义和分布两个方面挖掘多达 5 种媒体类型数据(图像、视频、文本、音频和 3D 模型)细粒度上下文信息间的关联关系.本文主要贡献如下.

(1) 提出了针对 5 种媒体的跨媒体循环神经网络,构建统一的网络结构联合建模不同媒体数据内部的细粒度信息,并进一步挖掘不同媒体数据细粒度局部区域或片段之间的上下文关系,充分学习各种媒体内独有的内在信息,为跨媒体关联学习提供更加细粒度的线索.

(2) 提出了基于分布对齐和语义对齐的跨媒体联合关联损失函数.一方面,通过分布对齐弥补不同媒体类型数据之间的分布差异;另一方面,通过语义对齐增强关联学习过程中的语义辨识能力.使分布对齐与语义对齐相互促进,实现对不同媒体数据的语义一致性表达,更好地在 5 种媒体条件下实现细粒度跨媒体关联分析与挖掘,提升跨媒体检索的准确率.

为了验证方法的有效性,本文在两个包含 5 种媒体(图像、视频、文本、音频和 3D 模型)的跨媒体数据集 PKU XMedia 和 PKU XMediaNet 上与现有方法进行实验对比,结果表明,本文方法有效地提高了跨媒体检索的准确率.

1 相关工作

1.1 针对两种媒体的跨媒体检索方法

现有方法往往旨在解决两种媒体类型数据之间的异构鸿沟问题,通常是针对图像和文本,将其映射至统一空间得到跨媒体统一表征.其中,传统方法通过优化特定统计量来学习线性映射矩阵.典型相关分析(canonical correlation analysis,简称 CCA)[11]是第一个被广泛使用的跨媒体模型,该方法通过最大化不同媒体类型成对数据之间的关联来优化模型.一些后续工作基于典型相关分析进行了扩展,例如,Hardoon 等人[17]提出核典型相关分析(kernel canonical correlation analysis,简称 KCCA),利用核函数实现非线性典型相关分析.此外,Li 等人[18]提出了跨媒体因子分析(cross-modal factor analysis,简称 CFA)算法,通过最小化成对数据之间的 Frobenius 范数来优化跨媒体模型.

近年来,深度网络在图像识别[19,20]、视频分类[21]等领域显示出强大的学习能力.受此启发,一些工作尝试使用深度网络来学习统一表征以实现跨媒体检索.Andrew 等人[22]提出深度典型相关分析(deep canonical correlation analysis,简称 DCCA)方法,通过两个子网络的输出关联来优化模型.Feng 等人[8]构建对应自编码器(correspondence autoencoder,简称 Corr-AE),通过中间层来链接两路子网络,同时对关联关系和重建信息进行建模.Wei 等人[23]提出的深度语义匹配(deep semantic match,简称 Deep-SM)模型使用卷积神经网络来建模图像数据,从而进一步挖掘语义关联信息.Peng 等人[15]提出了跨媒体多网络结构(cross-media multiple deep network,简称 CMDN)模型,将媒体内和媒体间的关联信息通过层次化网络的方式进行联合学习以提升检索准确率.他们在此基础上进一步提出了跨模态关联学习(cross-modal correlation learning,简称 CCL)方法[24],通过多任务学习的方式挖掘不同媒体类型数据的粗细粒度信息.Huang 等人[25]提出了基于混合迁移网络的跨媒体统一表征(cross-modal hybrid transfer network,简称 CHTN)方法,实现了从单媒体源域到跨媒体目标域的知识迁移.此外,对抗式学习也被应用在跨媒体检索中[26].

1.2 针对多种媒体的跨媒体检索方法

目前仅有很少的工作针对多于两种媒体的交叉检索任务,其中,Zhai 等人[27]尝试构建图模型来学习映射矩阵,首先将 5 种媒体同时在传统框架中建模,并进一步提出了联合表示学习(joint representation learning,简称 JRL)方法[10],加入语义信息和半监督规约来构建统一空间.此外,Peng 等人[28]提出构建统一的跨媒体关联超图,同时利用了不同媒体的细粒度信息并结合半监督规约来学习跨媒体统一表征.然而,由于以上方法均使用传统框架学习线性映射,难以充分挖掘多达 5 种媒体类型数据之间的关联关系.而某些基于深度学习的方法,如深度语义匹配模型,尽管可以通过增加子网络的方式将其扩展至多种媒体,但其仅考虑了数据内部的语义类别信息,难以挖掘多种媒体之间复杂且多样的关联关系.

本文旨在弥补上述缺陷,联合建模多达 5 种媒体类型数据的细粒度上下文信息,同时实现不同媒体数据类型数据之间的语义对齐和分布对齐,从而提升 5 种媒体交叉检索的准确率.

2 本文方法

本文方法的网络结构如图 3 所示.首先,构建针对 5 种媒体数据的跨媒体循环神经网络,通过将不同媒体类型数据的局部区域或片段序列输入到循环神经网络中建模数据内部的细粒度上下文信息.然后,在循环神经网络之上设计跨媒体联合关联损失函数,通过语义对齐和分布对齐相结合的方式,联合优化异构数据到统一空间的映射,从而学习更加精确的细粒度跨媒体关联.

首先介绍本文的形式化定义,其中,$D=\{D^I,D^T,D^A,D^V,D^M\}$为包含 5 种媒体类型的跨媒体数据集,$\{x^i,x^t,x^a,x^v,$

$x^m\}\in D$ 分别代表数据集中图像、文本、音频、视频和 3D 模型数据.此外,定义 $l\in\{i,t,v,a,m\}$表示任意一种媒体类型,这样,$\{x^l,y^l\}\in D$ 分别代表数据集中的任意媒体类型的数据及其类别标签.跨媒体检索旨在给定任意一种媒体类型的数据,返回与其语义相关的所有媒体类型的检索结果.

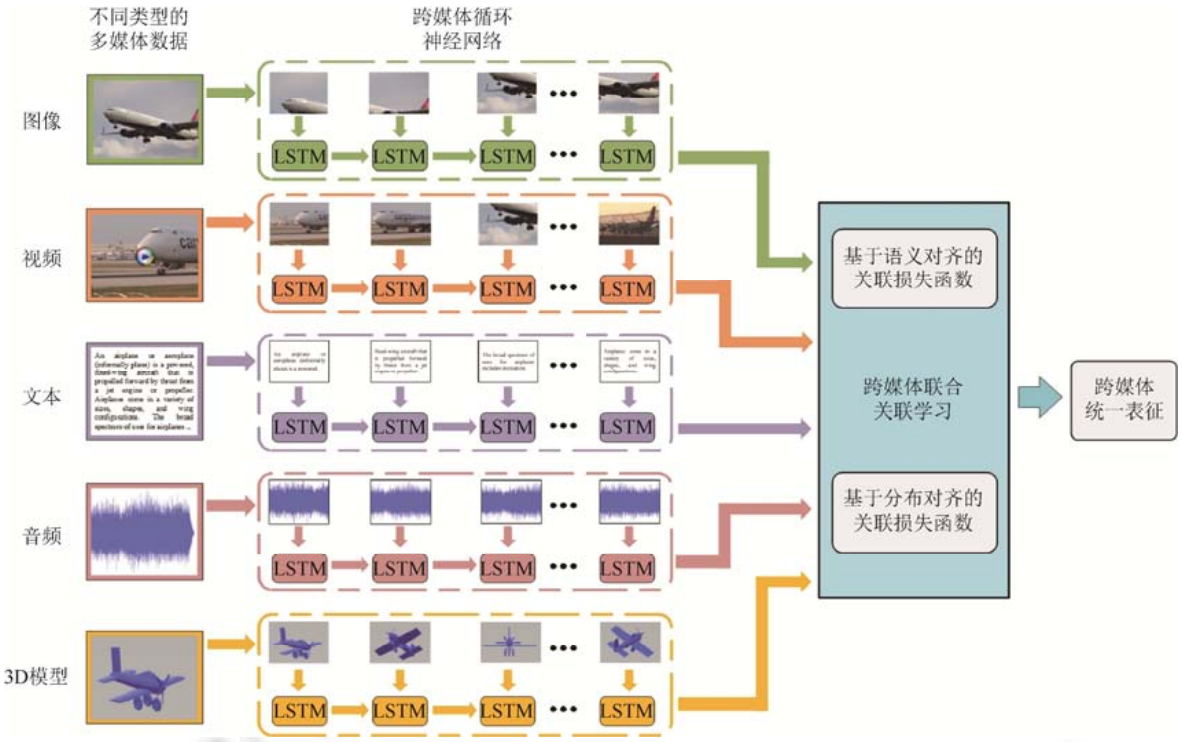

Fig.3 An overview of our proposed FGCL approach

图 3 本文方法整体框架示意图

2.1 跨媒体循环神经网络

为了充分利用多种媒体类型数据中丰富的细粒度上下文信息,本文构建了多路循环神经网络,将每种媒体类型数据的局部区域或片段的序列输入到循环神经网络来学习细粒度特征表示.对不同媒体类型数据分别进行分割并获取细粒度特征序列的具体策略将在第 2.3 节中详细加以介绍.

上述得到的每种媒体类型数据局部区域或片段的特征序列蕴含了丰富的细粒度信息,进一步将其输入到循环神经网络中来充分挖掘不同媒体类型数据内部的细粒度上下文信息.本文采用了长短时记忆(long short term memory,简称 LSTM)网络[29],LSTM 网络作为一种特殊的循环神经网络,能够利用记忆单元(cell)及门限(gate)的更新有效地学习序列数据中的长期依赖,并充分保存历史时间步中的信息.本文将上述每种媒体类型数据的特征按照序列逐步输入到 LSTM 网络中,并根据如下公式逐步更新网络:

$$\begin{Bmatrix} i_t \\ f_t \\ o_t \end{Bmatrix} = \sigma\left(\begin{Bmatrix} W_i \\ W_f \\ W_o \end{Bmatrix} x_t + \begin{Bmatrix} U_i \\ U_f \\ U_o \end{Bmatrix} h_{t-1} + \begin{Bmatrix} b_i \\ b_f \\ b \end{Bmatrix}\right) \tag{1}$$

$$u_t = \tanh(W_u x_t + U_u h_{t-1} + b_u) \tag{2}$$

$$c_t = c_{t-1} \odot f_t + u_t \odot i_t \tag{3}$$

$$h_t = o_t \odot \tanh(c_t) \tag{4}$$

其中,x 表示输入序列,i,f,o 和 c 分别表示输入门、遗忘门、输出门和记忆单元,$\odot$ 表示元素相乘,而σ表示 Sigmoid

激活函数,W 和 U 为循环神经网络中待学习的参数.将输出序列通过全连接层就可以得到每种媒体数据固定维数的序列特征 $h_{seq}^l=\{h_1^l,\ldots,h_j^l\}$, 随后将序列特征 h_{seq}^l 取平均得到 $h^l=1/j\sum_{i=1}^{j}h_i^l$, 其中,$j$ 为序列长度.这样,每个任意媒体类型数据的特征 h^l 都包含了丰富的细粒度上下文信息,为进一步挖掘跨媒体细粒度关联关系提供了重要线索.

2.2 跨媒体联合关联学习

在得到包含细粒度上下文信息的不同媒体特征之后,如何更好地将其映射至统一空间中成为解决 5 种媒体类型数据间交叉检索的关键问题.具体地,本文在上述循环神经网络顶层提出了基于分布对齐和语义对齐的跨媒体联合关联损失函数,通过弥补不同媒体类型数据之间的分布差异,同时充分利用了数据的语义类别信息增强关联学习过程中的语义辨识能力,能够更好地在 5 种媒体的条件下实现细粒度跨媒体关联的分析与挖掘.

首先,我们设计了基于语义对齐的关联损失函数.将第 2.1 节得到的不同媒体类型的数据表征 h^l 通过全连接网络(fully-connected network)映射到统一的语义空间中,并采用如下损失函数来约束不同媒体类型数据之间的语义关联:

$$L_{SA}=l_{sm}(h^l,y^l)+l_{trip}(h^l,y_c^l,\widehat{y_c^l}) \tag{5}$$

$$l_{sm}(h^l,y^l)=-\sum_{q=1}^{n}1\{y^l=q\}\log[\hat{p}(h^l,q)] \tag{6}$$

$$l_{trip}(h^l,y_c^l,\widehat{y_c^l})=\max(0,\alpha+f(h^l,y_c^l)-f(h^l,\widehat{y_c^l})) \tag{7}$$

其中,$l_{sm}(h^l,y^l)$ 为交叉熵损失函数项,y^l 为 h^l 的语义类别标签,共有 n 个类别.当 y^l=q 时,1{y^l=q}值为 1,否则,其值为 0, $\hat{p}(h^l,q)$ 表示预测该样本属于第 q 个类别的概率.

对于 $l_{trip}(h^l,y_c^l,\widehat{y_c^l})$, 我们首先将每个语义类别对应的语义标签通过 Word2Vec[30]模型提取特征,将其视作该类别的特征向量,得到 n 个类别的特征向量 $\{y_1,\ldots,y_n\}$, y_c^l 表示该样本对应类别的特征向量,而 $\widehat{y_c^l}$ 表示不匹配类别的特征向量,f 表示两个向量之间的点乘代表两个向量之间的相似度,α为固定的边界参数.

通过三元组的形式,约束属于相同语义类别的不同媒体类型数据,使其距离其对应类别的特征向量尽可能地近,同时距离其他类别的特征向量尽可能地远.由于类别标签通过 Word2Vec 模型来映射,其映射后的特征向量本身带有语义信息,通过将不同媒体数据映射到其类别向量周围,使得不同媒体数据映射后的统一表征保留其对应类别的语义信息,同时保证它们的语义一致性.因此,通过基于语义对齐的关联损失函数,能够有效地增强统一表征的语义辨识能力,促进细粒度的跨媒体关联挖掘.

进一步地,我们设计了基于分布对齐的关联损失函数.具体地,我们采用最大均值差异(maximum mean discrepancy,简称 MMD)[31]损失函数来优化不同媒体类型数据之间的分布差异.最大均值差异被广泛使用在迁移学习和域自适应中,是衡量两个数据分布差异的重要标准.其基本原理是针对两个不同分布的样本,通过寻找在样本空间上的连续函数,使不同分布的样本在该函数上函数值均值的差值最大,从而得到最大均值差异 MMD.通过最小化 MMD 损失,可以减小不同分布之间的差异,达到对齐分布的效果.基于上述思想,我们定义了如下基于分布对齐的关联损失函数:

$$L_{DA}=\sum_{i,j}g_{mmd}(h^i,h^j) \tag{8}$$

其中,i,j 表示任意两种不同的媒体类型.而任意两种媒体类型数据之间的 MMD 损失函数定义如下:

$$g_{mmd}(h^i,h^j)=\left\|E_I[\phi(h^i)]-E_J[\phi(h^j)]\right\|_H^2 \tag{9}$$

其中,MMD 损失函数是在再生希尔伯特空间(reproducing kernel Hilbert space,简称 RKHS)的平方形式.通过最小化上式,可以减小 h^i 和 h^j 之间的分布差异,达到不同媒体类型之间的分布对齐.综上,基于语义对齐和分布对齐的跨媒体联合关联损失函数定义如下:

$$L=L_{SA}+L_{DA} \tag{10}$$

通过最小化上述损失函数,不仅可以增强跨媒体统一表征的语义辨识能力,在统一空间中将不同媒体类型

的数据约束至其语义中心,同时可以减小 5 种媒体之间的数据分布差异,从而有效学习不同媒体类型数据细粒度上下文信息之间的关联关系,提高跨媒体检索的准确率.

2.3 实现细节

本文提出的网络在Torch框架上得以实现.具体地,对于每个图像样本x^i,将其缩放后输入VGG-19卷积神经网络[32],通过最后一个池化层(pool5)来提取出49个不同区域的局部特征,每个特征维数为512维,然后按照人眼观察的顺序组成序列.对于每个文本样本 x^t,首先按照段落或语句将其切分成片段,然后利用文本卷积神经网络[33]对每个片段提取 300 维特征,最后按照文本片段本身顺序组成序列.对于每个音频样本 x^a,按照固定时间间隔将其分割成片段,对每个片段分别提取 128 维 Mel 频率倒谱系数特征(mel frequency cepstrum coefficient,简称MFCC)形成序列.对于视频,对每一个视频帧提取 VGG-19 网络[32]全连接层(fc7)的 4 096 维图像特征,然后按照其原本时间顺序组成序列.对于 3D 模型,我们采用 47 个不同角度来观察 3D 模型数据,然后使用光场描述子(light field)[34]对每一个角度提取 100 维特征,再依照文献[28]将其组成序列.总的来说,针对特征选择,本文旨在探究跨媒体关联学习问题,特征选择并非本文重点,且本文的模型可以支持多种输入特征.针对序列选择,对于带有内在序列性质的媒体类型,如文本、音频和视频,我们按照其天然顺序将区域片段组成序列.对于序列性质不明显的媒体类型,如图像和 3D 模型,我们按照固定顺序组成序列,且其细粒度数据之间的顺序对关联学习的最终结果影响不大.使用上述固定切分方式不仅能够有效地保留某些媒体数据的细粒度单元,也降低了模型的复杂度.此外,在实验过程中,我们将跨媒体循环神经网络的输出,即统一表征的维数设置为 300 维,语义对齐关联损失函数(见公式(7))中的边界参数α设置为 1,网络训练的学习率固定为 1e-4.

本文模型训练过程需要 25 个 epoch,时间复杂度和其他基于深度网络的跨媒体检索方法相当,并且由于算法充分挖掘了跨媒体细粒度数据之间的上下文关系,泛化能力较强,输入特征可以直接使用预训练的深度网络或是传统特征而不需要进行微调,这也缩短了算法的运行时间.空间复杂度上,一方面循环神经网络的自身性质决定了不同时刻输入循环神经网络的数据经过同一个神经元,大大节省了参数量.另一方面,较低的统一空间维度(300 维)也减少了模型的空间复杂度.

3 实 验

本文在两个具有挑战性的跨媒体数据集PKU XMedia和PKU XMediaNet上进行了多种媒体的交叉检索实验,两个数据集均包含多达 5 种媒体类型(图像、文本、音频、视频和 3D 模型)的数据.为了更加全面地验证本文提出方法的有效性,我们进行了两大类的实验对比,包括 5 种媒体的交叉检索和 2 种媒体(图像和文本)的相互检索,与 12 种现有方法进行了对比.此外,本文还进一步通过基线实验以验证本文方法各个部分的效果.

3.1 数据集介绍

下面简要介绍本文使用的两个包含 5 种媒体类型的跨媒体数据集,每个数据集均划分为训练集、验证集和测试集 3 个部分,具体划分方式见表 1 和表 2.

数据集网址为http://www.icst.pku.edu.cn/mipl/XMedia.

PKU XMedia 数据集[2]是第一个包含 5 种媒体类型的跨媒体数据集.数据集共有 20 个常见的语义类别,比如自行车、钢琴、昆虫等,数据来源包括维基百科(Wikipedia)、Flickr、YouTube 等.

Table 1 The dataset partition on PKU XMedia

表 1 PKU XMedia 数据集的划分方式

媒体类型	训练集	测试集	验证集	总数
图像	4 000	800	200	5 000
视频	400	80	20	500
文本	4 000	800	200	5 000
音频	800	160	40	1 000
3D 模型	400	80	20	500
总数	9 600	1 920	480	12 000

Table 2 The dataset partition on PKU XMediaNet

表 2 PKU XMediaNet 数据集的划分方式

媒体类型	训练集	测试集	验证集	总数
图像	32 000	4 000	4 000	40 000
视频	8 000	1 000	1 000	10 000
文本	32 000	4 000	4 000	40 000
音频	8 000	1 000	1 000	10 000
3D 模型	1 600	200	200	2 000
总数	81 600	10 200	10 200	102 000

PKU XMediaNet 数据集[2]是目前国际上最大的包含 5 种媒体类型的跨媒体数据集,共包含超过 10 万个数据样本,其规模是 XMedia 的 10 倍.共包含了 200 个常见类别,主要分为动物和人造物两大类.图 4 展示了该数据集的部分样例.数据来源包括 Wikipedia、Flickr、YouTube、Freesound、Yobi3D 等.

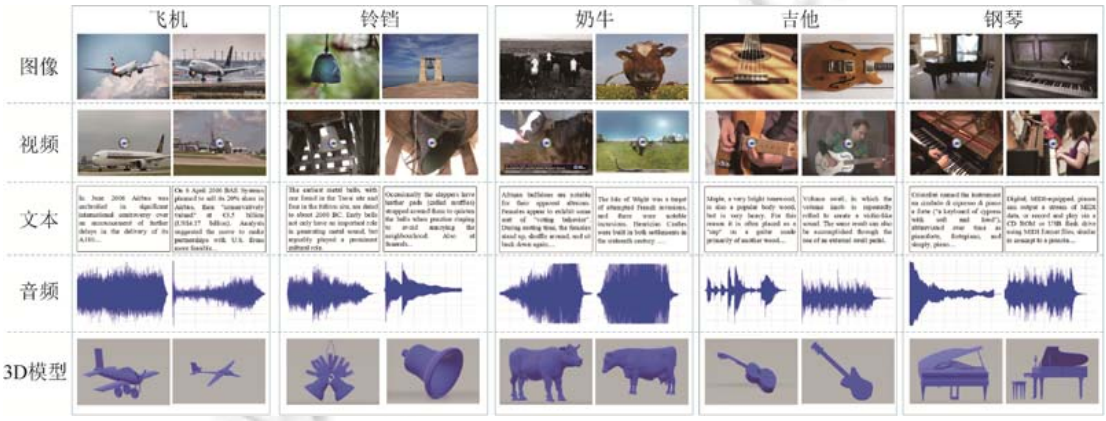

Fig.4 Quintuple-media examples from PKU XMediaNet dataset

图 4 来自 PKU XMediaNet 数据集的 5 种不同媒体类型数据示意图

3.2 评价指标和对比方法

不同媒体数据之间的相似度可以通过计算跨媒体统一表征之间的距离来得到,本文采用余弦距离来计算相似度,从而对检索结果进行排序.为了全面验证本文方法的有效性,我们分别设置了 5 种媒体交叉检索和 2 种媒体相互检索的实验.

3.2.1 5 种媒体交叉检索

5 种媒体交叉检索是指将任意一种媒体类型的查询样例作为输入,检索所有 5 种媒体类型数据中与之语义相关的结果.举例来说,将图像作为查询样例输入,检索测试集中图像、文本、音频、视频和 3D 模型的样本,表示为图像检索全部(Image→All).以其余 4 种媒体类型作为查询的检索可以表示为:文本检索全部(Text→All)、音频检索全部(Audio→All)、视频检索全部(Video→All)和 3D 模型检索全部(3D→All).

本文采用平均准确率均值(mean average precision,简称 MAP)作为评价指标,该指标能够同时兼顾返回结果的排序以及准确率,在信息检索领域被广泛使用.具体地,首先计算查询样本所有返回结果的平均准确率(average precision,简称 AP),然后计算所有查询的 AP 结果的平均值得到最终的 MAP 值.

本文方法与 3 种支持 5 种媒体场景或可以扩展至 5 种媒体场景的现有方法进行了实验对比,分别是 JRL[10]、S^2UPG[28]和 Deep-SM[23],其中,前两种是直接支持 5 种媒体的交叉检索的传统方法,而 Deep-SM[23]是基于深度学习的方法,其本身仅针对两种媒体相互检索,但可以通过扩充另外 3 路子网络的方式来支持 5 种媒体的交叉检索.为了更加公平地与现有方法进行比较,所有方法在 5 种媒体上都使用了与本文相同的深度网络或描述子来提取输入特征.具体地,对于图像,我们采用在 ImageNet 数据集上预训练,并在目标数据集上微调的 VGG-19 卷积

神经网络[32]提取 4 096 维全连接层特征(fc7).对于文本,我们依照文献[33]中的方式通过文本卷积神经网络对其提取 300 维的特征.对于音频,我们对音频帧分别提取 Mel 频率倒谱系数特征(mel frequency cepstrum coefficient,简称 MFCC),然后取平均获得 128 维 MFCC 特征.对于视频,我们通过平均每一个视频帧的 VGG-19 网络全连接层特征(fc7)得到 4 096 维特征.对于 3D 模型,我们将 47 个角度的光场描述子特征(light field)[34]拼接得到 4 700 维特征.

3.2.2 两种媒体相互检索

由于现有方法往往仅针对两种媒体的跨媒体检索任务,且以图像和文本相互检索为主,为了更全面地与现有方法进行实验比较,本文也进行了图像和文本相互检索的实验,包括两个检索任务:图像检索文本(Image→Text)和文本检索图像(Text→Image).实验结果评估同样采用了第 3.2.1 节中提到的 MAP 指标,这里需要说明的是,本文中的 MAP 值通过计算每个样例返回的所有检索结果得到,与 Corr-AE[8]以及 ACMR[26]中仅使用前 50 个返回结果的计算方式不同.图像文本相互检索的实验对比了 12 种现有方法,包括 6 种传统跨媒体检索方法:CCA[11]、CFA[18]、KCCA[17]、JRL[10]、S^2UPG[28]和 LGCFL[9],以及 6 种基于深度学习的跨媒体检索方法:Corr-AE[8]、DCCA[22]、Deep-SM[23]、CMDN[15]、CCL[24]和 ACMR[26].为了实验的公平对比,如第 3.2.1 节中所述,所有对比方法的图像和文本都使用了相同的输入特征.本文代码已经发布在https://github.com/PKU-ICST-MIPL,对比方法 JRL[10]、S^2UPG[28]、CMDN[15]和 CCL[24]的发布代码也在此目录下.

3.3 与现有方法的实验结果对比

3.3.1 5 种媒体交叉检索

5 种媒体交叉检索的实验结果见表 3 和表 4.从对比结果可以看出,本文提出的方法在两个数据集上均超过了所有对比方法,跨媒体检索的准确率有比较明显的提升.以 PKU XMediaNet 数据集为例,平均检索准确率从 0.303 提升到 0.366.对比方法中,基于深度网络的 Deep-SM 方法未能超过另外两种基于传统框架的方法 JRL 和 S^2UPG,因为其只考虑了粗粒度的全局语义信息,没有考虑不同媒体数据之间的分布差异.而本文方法充分挖掘了不同媒体数据内部的细粒度上下文信息,同时结合语义对齐和分布对齐来优化不同媒体数据到统一空间的映射,更好地克服了 5 种媒体之间的异构鸿沟问题.

Table 3 Results of cross-media retrieval with five media types on PKU XMedia dataset

表 3 PKU XMedia 数据集上的 5 种媒体交叉检索结果

对比方法	Image→All	Text→All	Audio→All	Video→All	3D→All	平均
本文方法	**0.870**	**0.878**	**0.583**	**0.648**	**0.654**	**0.727**
S^2UPG[28]	0.868	0.861	0.323	0.623	0.565	0.648
JRL[10]	0.843	0.828	0.249	0.519	0.295	0.547
Deep-SM[23]	0.767	0.806	0.364	0.492	0.396	0.565

Table 4 Results of cross-media retrieval with five media types on PKU XMediaNet dataset

表 4 PKU XMediaNet 数据集上的 5 种媒体交叉检索结果

对比方法	Image→All	Text→All	Audio→All	Video→All	3D→All	平均
本文方法	**0.520**	**0.581**	**0.138**	**0.343**	**0.248**	**0.366**
S^2UPG[28]	0.510	0.510	0.050	0.282	0.165	0.303
JRL[10]	0.480	0.453	0.042	0.258	0.105	0.268
Deep-SM[23]	0.314	0.345	0.043	0.148	0.069	0.184

3.3.2 两种媒体相互检索

图像文本相互检索的实验结果见表 5 和表 6,本文提出的方法在两个数据集上同样超过了 12 种对比方法,表明本文方法在两种媒体相互检索的场景下同样具有很好的效果.对比方法中,传统方法和基于深度学习的方法的检索准确率并没有很大的差异,一些传统方法甚至超过了部分基于深度学习的方法,例如 JRL[10]、S^2UPG[28]和 LGCFL[22].另一方面,CCL[24]方法采用多任务学习的方式同时考虑粗细粒度的信息,在对比方法中取得了最好的结果.而本文方法不仅充分挖掘了数据内部的细粒度信息,还考虑到了它们之间的上下文关系,有效地学习了两种媒体类型数据之间的关联关系.

Table 5 Results of cross-media retrieval between image and text on PKU XMedia dataset

表 5 PKU XMedia 数据集上的两种媒体相互检索结果

对比方法	Image→Text	Text→Image	平均
本文方法	**0.926**	**0.922**	**0.924**
CCL[24]	0.915	0.914	0.915
S^2UPG[28]	0.916	0.906	0.911
CMDN[15]	0.911	0.905	0.908
JRL[10]	0.902	0.888	0.895
ACMR[26]	0.886	0.884	0.885
Corr-AE[8]	0.872	0.874	0.873
Deep-SM[23]	0.856	0.846	0.851
LGCFL[9]	0.830	0.844	0.837
CFA[18]	0.735	0.790	0.763
DCCA[22]	0.629	0.642	0.636
KCCA[17]	0.710	0.623	0.667
CCA[11]	0.516	0.523	0.520

Table 6 Results of cross-media retrieval between image and text on PKU XMediaNet dataset

表 6 PKU XMediaNet 数据集上的两种媒体相互检索结果

对比方法	Image→Text	Text→Image	平均
本文方法	**0.607**	**0.628**	**0.618**
CCL[24]	0.537	0.528	0.533
S^2UPG[28]	0.591	0.589	0.590
CMDN[15]	0.485	0.516	0.501
JRL[10]	0.488	0.405	0.447
ACMR[26]	0.536	0.519	0.528
Corr-AE[8]	0.469	0.507	0.488
Deep-SM[23]	0.399	0.342	0.371
LGCFL[9]	0.441	0.509	0.475
CFA[18]	0.252	0.400	0.326
DCCA[22]	0.425	0.433	0.429
KCCA[17]	0.252	0.270	0.261
CCA[11]	0.212	0.217	0.215

3.4 基线实验结果分析

为了验证本文方法各个部分的效果,我们进一步进行了基线实验的对比,其中,“无三元组损失”表示去掉语义对齐关联损失函数(见公式(5))中的三元组损失函数(见公式(7))部分,“无 MMD 损失”表示去掉分布对齐关联损失函数(见公式(8)),“基线方法”表示同时去掉上述两个部分,仅使用语义类别信息(见公式(6))来约束不同媒体类型数据到统一空间的映射.从表 7 和表 8 可以看出,仅使用语义类别约束的平均检索准确率也同样高于 3 种对比方法的结果,表明充分利用数据内部的细粒度上下文信息能够更有效地建模不同媒体类型数据之间的关联关系,而三元组损失函数和分布对齐损失函数能够使模型在拥有语义辨识能力的同时,有效地将不同媒体类型数据的分布在统一空间内对齐,进一步提高了跨媒体检索的准确率.

Table 7 Baseline experiments on PKU XMedia dataset

表 7 PKU XMedia 数据集上的基线实验结果

对比方法	Image→All	Text→All	Audio→All	Video→All	3D→All	平均
本文方法	**0.870**	**0.878**	**0.583**	**0.648**	**0.654**	**0.727**
无三元组损失	0.864	0.868	0.565	0.643	0.638	0.716
无 MMD 损失	0.865	0.860	0.553	0.639	0.629	0.709
基线方法	0.856	0.853	0.532	0.611	0.606	0.691

Table 8 Baseline experiments on PKU XMediaNet dataset

表 8 PKU XMediaNet 数据集上的基线实验结果

对比方法	Image→All	Text→All	Audio→All	Video→All	3D→All	平均
本文方法	**0.520**	**0.581**	**0.138**	**0.343**	**0.248**	**0.366**
无三元组损失	0.513	0.567	0.117	0.333	0.237	0.353
无 MMD 损失	0.506	0.557	0.104	0.325	0.211	0.340
基线方法	0.499	0.546	0.092	0.314	0.203	0.334

4 结 论

本文提出了跨媒体深层细粒度关联学习方法,首先提出跨媒体循环神经网络以充分挖掘多达 5 种媒体类型数据的细粒度上下文信息,然后设计了跨媒体联合关联损失函数,将分布对齐和语义对齐相结合,在准确挖掘媒体内和媒体间细粒度关联的同时,利用语义类别信息增强关联学习过程中的语义辨识能力,有效提升了跨媒体检索的准确率.通过在两个包含多达 5 种媒体类型(图像、视频、文本、音频和 3D 模型)的跨媒体数据集 PKU XMedia 和 PKU XMediaNet 上与现有方法进行实验对比,表明了本文方法在多种媒体交叉检索任务的有效性.

下一步工作将尝试扩展现有框架,在不同尺度上挖掘跨媒体数据之间的关联关系,同时充分利用无标注数据并结合外部知识库以进一步提升跨媒体检索的准确率.

References:

[1] Lew MS, Sebe N, Djeraba C, Jain R. Content-based multimedia information retrieval: State of the art and challenges. ACM Trans. Multimedia Computing, Communication, and Applications (TOMMCCAP), 2006,2(1):1–19.

[2] Peng YX, Huang X, Zhao YZ. An overview of crossmedia retrieval: Concepts, methodologies, benchmarks and challenges. IEEE Trans. on Circuits and Systems for Video Technology (TCSVT), 2018,28(5):2372–2385.

[3] Zhuang Y, Zhuang YT, Wu F. An integrated indexing structure for large-scale cross-media retrieval. Ruan Jian Xue Bao/Journal of Software, 2008,19(10):2667–2680 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/2667.htm [doi: 10.3724/SP.J.1001.2008.02667]

[4] Wu F, Zhuang YT. Cross media analysis and retrieval on the Web: Theory and algorithm. Journal of Computer-Aided Design & Computer Graphics, 2010,22(1):1–9 (in Chinese with Emglish absrtact).

[5] Hu Y, Xie X. Coherent phrase model for efficient image near-duplicate retrieval. IEEE Trans. on Multimedia (TMM), 2009,11(8):1434–1445.

[6] Peng YX, Ngo CW. Clip-based similarity measure for query-dependent clip retrieval and video summarization. IEEE Trans. on Circuits and Systems for Video Technology (TCSVT), 2006,16(5):612–627.

[7] McGurk H, MacDonald J. Hearing lips and seeing voices. Nature, 1976,264(5588):746–748.

[8] Feng F, Wang X, Li R. Cross-modal retrieval with correspondence autoencoder. In: Proc. of the ACM Int'l Conf. on Multimedia (ACM-MM). 2014. 7–16.

[9] Kang C, Xiang S, Liao S, Xu C, Pan C. Learning consistent feature representation for cross-modal multimedia retrieval. IEEE Trans. on Multimedia (TMM), 2015,17(3):370–381.

[10] Zhai XH, Peng YX, Xiao J. Learning cross-media joint representation with sparse and semi-supervised regularization. IEEE Trans. on Circuits and Systems for Video Technology (TCSVT), 2014,24(6):965–978.

[11] Hotelling H. Relations between two sets of variates. Biometrika, 1936, 321–377.

[12] Ranjan V, Rasiwasia N, Jawahar CV. Multi-label cross-modal retrieval. In: Proc. of the IEEE Int'l Conf. on Computer Vision (ICCV). 2015. 4094–4102.

[13] Ding MY, Niu YL, Lu ZW, Wen JR. Deep learning for parameter recognition in commodity images. Ruan Jian Xue Bao/Journal of Software, 2018,29(4):1039–1048 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5408.htm [doi: 10.13328/j.cnki.jos.005408]

[14] Bai Z, Huang L, Chen JN, Pan X, Chen SY. Optimization of deep convolutional neural network for large scale image classification. Ruan Jian Xue Bao/Journal of Software, 2018,29(4):1029–1038 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5404.htm [doi: 10.13328/j.cnki.jos.005404]

[15] Peng YX, Huang X, Qi JW. Cross-media shared representation by hierarchical learning with multiple deep networks. In: Proc. of the Int'l Joint Conf. on Artificial Intelligence (IJCAI). 2016. 3846–3853.

[16] Yan F, Mikolajczyk K. Deep correlation for matching images and text. In: Proc. of the Conf. on Computer Vision and Pattern Recognition (CVPR). 2015. 3441–3450.

[17] Hardoon DR, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: An overview with application to learning methods. Neural Computation, 2004,16(12):2639–2664.

[18] Li D, Dimitrova N, Li M, Sethi IK. Multimedia content processing through cross-modal association. In: Proc. of the ACM Int'l Conf. on Multimedia (ACM-MM). 2003. 604–611.

[19] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS). 2012. 1106–1114.

[20] He KM, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proc. of the Conf. on Computer Vision and Pattern Recognition (CVPR). 2016. 770–778.
[21] Wu Z, Jiang Y, Wang X, Ye H, Xue X. Multi-stream multiclass fusion of deep networks for video classification. In: Proc. of the ACM Int'l Conf. on Multimedia (ACM-MM). 2016. 791–800.
[22] Andrew G, Arora R, Bilmes J. Deep canonical correlation analysis. In: Proc. of the Int'l Conf. on Machine Learning (ICML). 2013. 1247–1255.
[23] Wei Y, Zhao Y, Lu C, Wei S, Liu L, Zhu Z, Yan S. Cross-modal retrieval with CNN visual features: A new baseline. IEEE Trans. on Cybernetics (TCYB), 2017,47(2):449–460.
[24] Peng YX, Qi JW, Huang X, Yuan YX. CCL: Cross-modal correlation learning with multi-grained fusion by hierarchical network. IEEE Trans. on Multimedia (TMM), 2017.
[25] Huang X, Peng YX, Yuan MK. Cross-modal common representation learning by hybrid transfer network. In: Proc. of the Int'l Joint Conf. on Artificial Intelligence (IJCAI). 2017. 1893–1900.
[26] Wang BK, Yang Y, Xu X, Hanjalic A, Shen HT. Adversarial cross-modal retrieval. In: Proc. of the ACM Conf. on Multimedia (ACM-MM). 2017. 154–162.
[27] Zhai XH, Peng YX, Xiao J. Heterogeneous metric learning with joint graph regularization for cross-media retrieval. In: Proc. of the AAAI Conf. on Artificial Intelligence (AAAI). 2013. 1198–1204.
[28] Peng YX, Zhai XH, Zhao YZ, Huang X. Semi-supervised crossmedia feature learning with unified patch graph regularization. IEEE Trans. on Circuits and Systems for Video Technology (TCSVT), 2016,26(3):583–596.
[29] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997,9(8):1735–1780.
[30] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems (NIPS). 2013. 3111–3119.
[31] Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A. A kernel two-sample test. Journal of Machine Learning Research (JMLR), 2012,13(1):723–773.
[32] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proc. of the Int'l Conf. on Learning Representations (ICLR). 2014.
[33] Kim Y. Convolutional neural networks for sentence classification. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP). 2014. 1746–1751.
[34] Chen D, Tian X, Shen Y, Ouhyoung M. On visual similarity based 3D model retrieval. Computer Graphics Forum, 2003,22(3): 223–232.

附中文参考文献:

[3] 庄毅,庄越挺,吴飞.一种支持海量跨媒体检索的集成索引结构.软件学报,2008,19(10):2667–2680. http://www.jos.org.cn/1000-9825/2667.htm [doi: 10.3724/SP.J.1001.2008.02667]
[4] 吴飞,庄越挺.互联网跨媒体分析与检索:理论与算法.计算机辅助设计与图形学学报,2010,22(1):1–9.
[13] 丁明宇,牛玉磊,卢志武,文继荣.基于深度学习的图片中商品参数识别方法.软件学报,2018,29(4):1039–1048. http://www.jos.org.cn/1000-9825/5408.htm [doi: 10.13328/j.cnki.jos.005408]
[14] 白琮,黄玲,陈佳楠,潘翔,陈胜勇.面向大规模图像分类的深度卷积神经网络优化.软件学报,2018,29(4):1029–1038. http://www.jos.org.cn/1000-9825/5404.htm [doi: 10.13328/j.cnki.jos.005404]

卓昀侃(1995－),男,福建宁德人,学士,主要研究领域为跨媒体分析与检索.

綦金玮(1994－),男,学士,主要研究领域为跨媒体分析与检索.

彭宇新(1974－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为跨媒体分析与推理,图像视频理解与检索,计算机视觉.