计算机科学
COMPUTER SCIENCE

基于共同子空间分类学习的跨媒体检索研究

韩红旗，冉亚鑫，张运良，桂婕，高雄，易梦琳

引用本文

韩红旗，冉亚鑫，张运良，桂婕，高雄，易梦琳. 基于共同子空间分类学习的跨媒体检索研究[J]. 计算机科学, 2022, 49(5): 33-42.

HAN Hong-qi, RAN Ya-xin, ZHANG Yun-liang, GUI Jie, GAO Xiong, YI Meng-lin. Study on Cross-media Information Retrieval Based on Common Subspace Classification Learning[J]. Computer Science, 2022, 49(5): 33-42.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210200157

基于共同子空间分类学习的跨媒体检索研究

韩红旗[1,2]　冉亚鑫[1,2]　张运良[1,2]　桂　婕[1]　高　雄[1,2]　易梦琳[1,2]
1 中国科学技术信息研究所　北京 100038
2 富媒体数字出版内容组织与知识服务重点实验室(国家新闻出版署)　北京 100038

摘　要　不同媒体数据间由于存在严重的异构鸿沟和语义鸿沟,而不能直接计算它们之间的语义相似度,从而影响了跨媒体检索的实现和效果。当前提出的共同子空间学习虽能实现跨媒体语义关联和检索,但多采用一般的特征提取技术,且在语义匹配时的分类效果较差,不能有效实现跨媒体数据的高层语义关联计算,影响了检索效果。对此,提出 Stacking-DSCM-WR 跨媒体关联方法,用于文档和图像之间的跨媒体检索。该方法基于词向量技术形成文档的特征表示向量,通过残差网络技术抽取图像的特征表示向量,采用深度典型相关性分析技术将不同模态的数据投影到共同子空间下,然后采用 Stacking 集成学习算法获取文本和图像在同一高层概念语义空间上的分布,使得两种不同模态的数据可以进行语义匹配、相似性计算。在 Wikipedia 和 Pascal Sentence 两个小型跨媒体数据集和一个较大规模跨媒体数据集 INRIA-Websearch 上分别开展跨媒体检索实验,证实了所提方法能够有效地抽取文本和图像的特征,实现跨媒体数据在高层语义空间上的关联和匹配,与相近跨媒体检索方法在 MAP 指标上的对比显示,该方法能够取得较好的检索效果。

关键词:跨媒体信息检索;语义关联;集成学习;词向量;残差网络
中图法分类号　TP391;G354

Study on Cross-media Information Retrieval Based on Common Subspace Classification Learning

HAN Hong-qi[1,2],RAN Ya-xin[1,2],ZHANG Yun-liang[1,2],GUI Jie[1],GAO Xiong[1,2] and YI Meng-lin[1,2]
1 Institute of Scientific and Technical Information of China,Beijing 100038,China
2 Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content,National Press and Publication Administration,Beijing 100038,China

Abstract　The semantic similarity between two different media data can not be calculated directly because of the serious heterogeneous gap and semantic gap between them,which affects the implementation and effect of cross media retrieval. Although the common space learning can achieve cross media semantic association and retrieval,the retrieval performance is not satisfied. The main reason is that it uses common feature extraction technology and general classification algorithm to implement semantic correlation and match. Aiming at this problem,the study proposes a novel cross media correlation method called Stacking-DSCM-WR for cross media retrieval between documents and images. WR means that text feature extraction is based on word-embedding technique and the image feature extraction is based on ResNet technique. DSCM means that the deep semantic correlation and match technology is exploited to project data of different modalities into a common subspace. Stacking is a kind of ensemble learning algorithm. It is employed to produce the distribution of text documents and images on the same high-level conceptual semantic space for cross-media retrieval. The experiments are carried out on two smaller cross-media datasets,Wikipedia and Pascal Sentence,and one larger cross-media dataset,INRIA-Websearch,respectively. The results show that the proposed method can effectively extract the features of text and image,and realize the correlation and match of cross media data in high-level semantic space. The comparisons with similar cross media retrieval methods show that the proposed method achieves the best retrieval effect based on MAP metric.

Keywords　Cross-media information retrieval,Semantic correlation,Ensemble learning,Word embedding,Residual networks

到稿日期:2021-02-24　返修日期:2021-07-15
基金项目:中国科学技术信息研究所重点工作项目(ZD2020-09);国家自然科学基金(71473237)
This work was supported by the ISTIC Key Work Project(ZD2020-09) and National Natural Science Foundation of China(71473237).
通信作者:韩红旗(bithhq@163.com)

1 引言

当前已进入跨媒体时代，新闻网站、电子商务网站以及近年发展起来的社交网络、图像视频分享网站等各类网络平台，将不计其数的数字图像、视频等多媒体资源推到了人们面前，文本、图像、音频、视频等复杂媒体对象的混合并存。“跨媒体”能从各自媒体的侧面表达相同的语义信息，比单一的媒体对象及其特定的模态更加全面地反映特定的内容信息[1]。这里的跨媒体数据主要指那些表达的语义内容相近，但以不同模态、不同来源、不同背景等形式出现的数据。一般来说，多媒体信息与相关联的文本是相互补充的，它们对应的高层语义往往具有很强的关联性[2]。

随着跨媒体数据总量增长和媒体类型的不断扩展，通过跨媒体检索来获取不同媒体信息已经成为用户的迫切需求[3]。然而，由于不同媒体数据具有不同的特征表示形式，即便它们采用相同维度的向量表示，也不能直接计算它们的相似性，因此跨媒体数据间存在异构鸿沟[4]。此外，跨媒体数据的底层特征与用户对数据理解的高层语义信息之间存在语义鸿沟[5]。例如，同为表达“机器人”概念，人能够理解文本、图像及视频中蕴含的概念语义、空间语义和事件语义等高层语义信息，但是计算机很难通过这些数据的底层特征有效地识别这些语义信息[6]。因此跨媒体数据间的异构鸿沟和语义鸿沟是跨媒体检索要解决的核心难题，有效理解不同模态数据之间的关联性是跨媒体检索任务的核心挑战。

基于共同子空间学习是解决跨模态语义关联的主流方法。它将不同模态数据的底层特征投影到一个共同子空间，然后在该空间学习不同模态数据的语义关联映射，从而能够计算其语义相似度[7]。基于共同子空间学习的方法可以分为基于典型相关性分析（Canonical Correlation Analysis，CCA）[8]的方法和基于深度学习的方法等。基于CCA方法的主要代表是语义关联匹配（Semantic Correlation Matching，SCM）模型[9]，它通过对不同模态数据进行最大相关性分析得到一个共同子空间。后来研究者提出了KCCA[10]，DCCA（Deep Canonical Correlation Analysis）[11]，Cluster-CCA[12]等变体方法，并在跨媒体语义关联和检索中取得了更好的效果，其中DCCA方法由于能较好地处理变量之间的非线性关系而获得了较多的研究。后续一些研究主要是解决DCCA算法占用内存、运算速度慢等问题。然而这些方法在提取数据特征时使用的大多是传统的特征提取技术，限制了其跨媒体检索性能的提高[13]。随着深度学习技术的兴起，研究者开始将其应用到跨媒体的特征表示和信息检索。例如，Wei等[13]使用卷积神经网络（Convolutional Neural Networks，CNN）技术提取图像的视觉特征，并提出深度语义匹配（Deep-SM）模型来实现跨媒体关联，取得了较好的效果。除了这些方法，也有研究者采用哈希算法学习公共汉明空间来加快检索速度，解决大规模跨媒体检索[14]，或者基于图模型的图正则化来表示跨媒体复杂的关联，实现跨媒体检索[15]。尽管这些方法取得了较大的进步，但仍不能达到令人满意的效果[3]。近几年，开始出现一些细粒度跨媒体检索[16]，主要是通过识别图像中细粒度对象与文本的关联来实现更有效的检索，但这些方法依赖于大规模的深度标注，限制了其实现和应用。

由于现有深度学习方法主要是通过分类算法实现底层特征到高层的语义空间映射，因此其分类的效果将会影响高层语义空间中不同模态的语义关联效果，进而影响跨媒体检索的准确率。基于此，本文提出了一种基于共同子空间分类学习的Stacking-DSCM-WR（Stacking-Deep Semantic Correlation Matching with Word Embedding and ResNet）跨媒体关联和检索方法，其目的在于取得更好的检索性能。该方法采用词向量技术形成文本特征，利用残差网络技术提取图像的语义特征，融合Stacking集成学习算法和深度语义关联匹配算法学习一个共同语义空间，实现文本和图像的语义关联和相互检索。

2 相关技术

2.1 词向量技术

分布式词向量是自然语言处理领域中的一类重要技术，其核心是对文本中的单词建模，用一个较低维的连续、低维、实值向量来表征每个单词[9]。词向量通常为100～300维，每一维度代表了一定的语义。可通过词向量之间的距离来判断词之间的语义相似度，距离越近，词表示的语义就越相似。词向量的生成方法很多，目前性能最佳的是基于深度神经网络的语言模型生成的分布式词向量，它通过无监督的机器学习方法从海量数据中自动学习词汇的语义特征，无需人工标注和复杂繁琐的特征工程。Word2vec词向量是2013年由Mikolov等[17]从海量的新闻语料中训练得到的，是目前被使用得最广泛的神经网络词向量。后来的研究者借鉴词嵌入向量的思路，提出了一些新的词向量模型，如GloVe[18]，FastText[19]，BERT[20]等。相比于经典的Word2vec词向量模型，GloVe词向量使用全局语料特征，一般情况下其效果稍优于Word2vec[18]；而FastText词向量考虑了字符级别的n-gram特征，能够表示出字级别的语义信息[19]；BERT通过预训练和微调的方式能够表示出句子级别的语义信息，研究表明，其大幅提升了文本语义特征表示的效果[20]。分布式词向量现在已被广泛应用于分类、聚类、命名实体识别、词性分析等自然语言处理任务中。

2.2 残差网络

残差网络（ResNet）是一种深度学习的卷积神经网络模型，由微软亚洲研究院的He等[21]提出。他们在研究图像识别的经典问题CIFAR-10时，发现一个56层的简单神经网络识别错误率反而高于一个20层的模型，即随着网络深度的增加，学习的效率反而下降。为了解决深层网络的“退化”问题，他们给非线性的卷积层增加直连边的方式来提高信息的传播效率。ResNet由许多堆叠的残差单元（Residual Units）组成。残差单元由多个级联的（等长）卷积层和一个跨层的直连边组成，再经过ReLU激活后得到输出。ResNet使用所谓的“跳跃链接”（Shortcut Connection）的方法，把底层的输出值每隔几层跳跃直接传递到更高层的输入，保证有效的信息不会在深层网络中被淹没。ResNet模型使用了深达152层的神经

网络，Top 5 的图像识别错误率创造了(3.57%)新低，这个数字低于一个接受良好训练的正常人的错误率(大约为 5%)，获得了 2015 年以 ImageNet 为基础的大型图像识别竞赛冠军，并且 ResNet 计算模型的复杂度还不到 2014 年获奖模型 VGG 的 19 层神经网络的 60%，且没有使用丢弃(dropout)算法。Feng 等[22]将残差网络用于跨媒体检索，取得了较好的效果。Gao 等提出了 Res2Net 网络结构，Res2Net 以更细的粒度来表示多尺度特征，并且增加了每个网络层的感受野[23]。实验结果表明，Res2Net 在多个图像数据集中取得了更好的成绩。

2.3 Stacking 集成学习方法

Stacking 是一种机器分类的集成学习方法。集成学习分类方法对多个弱分类器按照某种方式结合得到一个更好的强分类器，从而获得更好的模型泛化能力。集成学习方法一般包括 3 类，分别是 Bagging，Boosting 和 Stacking。其中 Bagging 方法通过联合多个弱模型来构建一个效果更好的强模型，从而提高分类器的预测能力；Boosting 方法通过赋予训练集权重来训练一个弱分类器并根据弱分类器，分类错误的结果更新训练样本的权重，从而将弱分类器训练为强分类器[24]；Stacking 方法一般先从初始数据集中训练出基分类器，通过一组训练好的基分类器对训练集进行预测，将预测输出的特征值作为新的训练集特征，然后基于新的训练集重新训练一个融合分类器。近几年，集成学习分类算法在各种分类任务中取得了瞩目成就[25]，已有研究者将其运用到跨媒体检索，如 Chen 提出了 Bagging-SM 方法来构建不同模态的高层语义空间，相比于传统方法，其提高了跨媒体检索的准确率[26]。已有研究证明，Stacking 通过二次训练能有效提升分类器的预测效果[27]，但目前尚没有发现有研究者将其应用到跨媒体检索领域。

3 基于 Stacking 和 DSCM 的跨媒体检索方法

3.1 跨媒体检索实现流程

为了实现不同模态数据的高层语义关联和跨媒体检索，本文提出了一种基于共同子空间分类学习的跨媒体检索方法 Stacking-DSCM-WR。该方法结合词向量和残差网络技术实现文本和图像的特征表示，融合集成学习算法学习一个共同语义空间，以实现文本和图像的语义关联，达到跨媒体检索的目的。跨媒体检索实现的流程如图 1 所示。

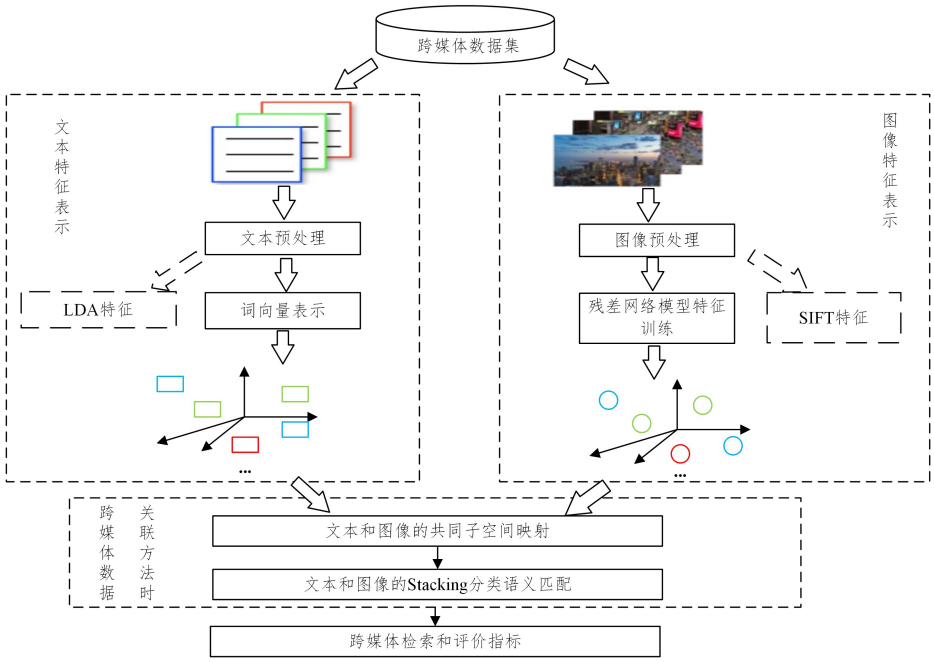


图 1 跨媒体检索实现流程(电子版为彩图)

Fig. 1 Process of cross-media retrieval

对于跨媒体数据集，首先分别对文本和图像进行预处理，实现文本和图像的特征表示。其中，文本特征表示采用词向量技术实现，图像特征表示采用 ResNet 技术实现。当然这里的文本特征也可以采用传统的 TFIDF 或 LDA 技术实现，图像特征也可以采用经典的 SIFT 技术实现(图 1 中用虚线表示)，但这些传统技术的性能已经无法和深度学习技术相比，我们将在实验部分进行证实。此时，每一个文本数据或图像数据就在各自的语义空间得到表示并与一个预定义的类别建立关联。然后，使用 DSCM 技术[11]实现文本和图像的子空间映射，采用 Stacking 集成学习算法实现文本和图像的高层分类语义匹配。最后，基于文本和图像的语义关联，实现文本查图像或图像查文本的跨媒体检索，并选择平均精度均值(Mean Average Precision，MAP)指标来评价检索性能。MAP 值与检索结果的排名情况有关，正确的检索结果在返回列表中的排名越靠前，MAP 值就越高。

3.2 跨媒体数据特征表示

3.2.1 文本特征表示方案

对于文本，为获得质量更好的文本表示，首先需要进行预处理，包括去除停用词等。基于词向量的文本表示模型由于能够提取文本深层的语义信息[28]，因此已成为当前常用的方法。

考虑到 GloVe, FastText 和 BERT 这 3 种词向量模型各有特点，因此拟分别采用它们表示文本特征，通过数据探索了解它们在跨媒体检索中的效果。在使用词向量技术获取了词向量库后，使用文本中包含词的词向量的平均向量作为文本向量表示。每一个文本特征化为向量后，还对应了一个类别标签，图 1 中不同颜色的小方块表示不同的类别。设文本集合为 $D=(d_1,d_2,\cdots,d_m)$，其中 m 为文档数量，经过词向量表示后的集合为 $X=(x_1,x_2,\cdots,x_m)$，集合中每一个文本向量的特征维度记为 n_1。

3.2.2 图像特征表示方案

图像预处理主要包括图像缩放和图像归一化等操作。由于残差网络模型(ResNet)[21]能够表示图像深层的语义信息，在图像识别竞赛中取得了超过人类识别能力的优秀效果，因此采用 ResNet 技术提取图像特征。将经过归一化后的图像数据作为输入，通过训练得到图像的特征向量。同样，每一个图像在特征化为向量后，也对应了一个类别标签，图 1 中相同颜色的方块表示文本和图像同属于一个类别。设图像集合为 $I=(i_1,i_2,\cdots,i_n)$，其中 n 为图像的数量，经过 ResNet 技术抽取特征后的向量表示集合为 $Y=(y_1,y_2,\cdots,y_n)$，其中每一个图像向量的特征维度记为 n_2。

3.3 融合 Stacking 分类学习的深度语义关联匹配方法

文本和图像数据的语义关联是实现跨媒体检索的基础。与传统的语义关联匹配方法不同的是，本文提出的 Stacking-DSCM-WR 方法考虑了文本和图像两种媒体数据的非线性关系，并在语义匹配阶段集成多种分类模型来训练文本和图像的类别语义映射，可有效提升类别语义表示能力。图 2 给出了实现文本和图像语义关联的过程。

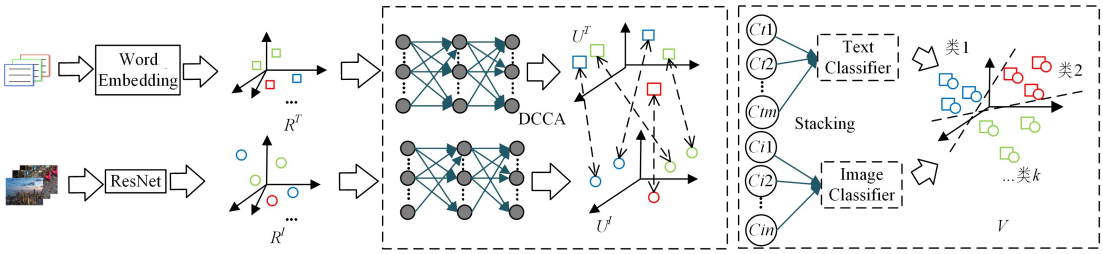


图 2 跨媒体语义关联流程图

Fig. 2 Process of cross-media semantic correlation

跨媒体语义关联的实现主要包括两个阶段，分别是共同子空间映射和语义匹配。首先文本和图像特征空间通过 DCCA 方法[11]学习得到其最大相关性子空间，然后利用 Stacking 方法训练其语义分类器，实现文本和图像在高层概念类别上的语义匹配，从而可以实现文本和图像两种不同模态数据的语义相似度计算。

3.3.1 文本和图像的共同子空间映射

假设原始的文本和图像特征空间分别为 R^{n_1} 和 R^{n_2}，其中 n_1 和 n_2 分别是文本向量和图像向量的特征数量，每个文本和图像均可在各自空间中抽象化为一个点。跨媒体检索的目标是构建一个文本和图像之间的互逆映射，使得给定一个文本 d_i，其向量表示后为文本空间中的点 $x_i\in R^{n_1}$，可以在图像空间中找到与之语义最相近的图像点 $y_q\in R^{n_2}$，返回 y_q 的一个近邻集合 $neighbor(y_q)\subset Y$ 并排序作为输出。根据图像检索文本是类似的过程。

由于文本和图像特征是异构的，因此无法直接计算 x_i 与 $neighbor(y_q)$ 的相似性，也不能进行排序。相关性分析技术的解决方案是建立它们同构的中间空间 U^T 和 U^I，并在同构空间中最大化它们的相关性来实现文本和图像之间的映射：

$$f:R^T\rightarrow U^T \tag{1}$$

$$g:R^I\rightarrow U^I \tag{2}$$

$$h:U^T\rightarrow U^I \tag{3}$$

对于一个文本查询 $x^q\in R^T$，首先通过 f 映射到共同子空间的一个点 $f(x^q)\in U^T$，然后通过 h 映射到同构共同子空间的点 $h(f(x^q))\in U^I$，最后通过 g 映射到图像语义空间中，找到对应的图像 $y=g^{-1}((f(x^q)))\in R^I$。根据图像查询文本也是相似的过程。

由于经典的典型相关性分析(CCA)方法[9]的基本原理是通过优化统计值来学习共同空间的线性投影矩阵，只能对两种线性相关的变量进行求解，无法处理变量之间的非线性关系[11]，因此本研究采用深度典型相关性分析(DCCA)方法。DCCA 方法将深度神经网络与 CCA 相结合来学习两组不同模态数据的最大相关性的非线性映射，可以获得更好的跨媒体关联效果。其实现的结构如图 3 所示。

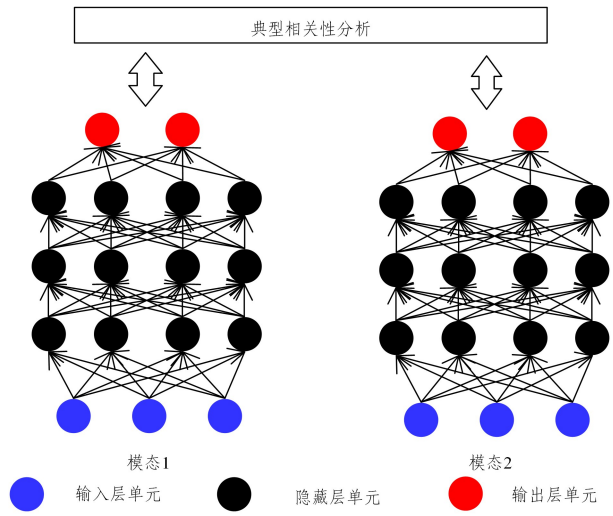


图 3 DCCA 结构示意图

Fig. 3 Network structure of DCCA

DCCA 包含两个深度网络，每个深度网络包含若干隐藏层，通过若干层非线性变换学习，使得输出层的相关性最大。对于两种模态的数据 X 和 Y，经过包含 d 层的深度网络非线性变换后，其在输出层可以分别表示为 $f(X)$和 $g(Y)$，分别对应了一组参数 w_l^v, b_l^v，其中 $l=1,2,\cdots,d$，$v=\{1,2\}$，$v=1$ 表示第 1 种模态数据，$v=2$ 表示第 2 种模态数据。DCCA 的目标是通过联合学习这两组模态的参数，使得不同模态的变量输入深度学习模型后，在输出层得到的 $f(X)$和 $g(Y)$相关性最大。设 θ_1 和 θ_2 分别代表两个深度网络的参数集合，则优化目标可以定义为：

$$(\theta_1^*, \theta_2^*) = \arg\max_{\theta_1,\theta_2} corr(f(X), g(Y)) \tag{4}$$

DCCA 基于梯度下降优化方法对参数进行优化，通过计算训练数据上相关目标函数的梯度估计来求解参数（θ_1^*，θ_2^*）。这个过程完成后，相当于实现了式(1)和式(2)的两个映射。

3.3.2 文本和图像的 Stacking 分类语义匹配算法

经过 DCCA 处理后，将文本和图像映射到共同子空间，下一步需要建立它们共同的语义空间(Semantic Space)，实现语义匹配(Semantic Matching，SM)。在语义空间上，两种模态数据的维度是一样的，每一个维度一般为一个语义概念。通过学习文本和图像的语义类别映射实现不同媒体数据的语义匹配。现有的语义匹配方法往往利用单一的分类算法对文本和图像进行分类训练，其分类效果受限于模型局限性。利用集成学习 Stacking 分类方法良好的分类效果和泛化性能学习文本和图像到共同语义空间的映射，可以有效地提升文本和图像的语义匹配效果。

(1)Stacking 分类方法的模型选择

Stacking 分类方法分为两层训练，通过结合不同学习器可以获得较好的泛化性能，并且第一层分类模型的好坏会影响第二层的融合模型的分类效果。因此，我们在第一层选择性能较好且有一定差异化的强分类模型作为训练模型，如随机森林(Random Forest，RF)、支持向量机(Support Vector Machine，SVM)、XGBoost(eXtreme Gradient Boosting，XGB)、梯度提升决策树(Gradient Boosting Decision Tree，GBDT)等。由于将初级分类器的输出类别概率作为融合分类器的输入特征将有效提升 Stacking 的分类效果，因此在上述列举的分类算法中，在其输出环节增加一个逻辑回归(Logistic Regression，LR)函数对这些算法的分类结果进行校准(具体实现时是对训练数据进行交叉验证)，从而得到分类结果的概率输出形式。在第二层训练选择逻辑回归这种简单高效的分类算法作为融合模型，因为第一层通过已经训练得到拟合效果比较好的模型，若第二层仍选择比较复杂的模型将会增加整个模型过拟合的风险。

(2)文本和图像的语义匹配

假设文本和图像共同的语义空间表示为 $V=(v_1, v_2, \cdots, v_k)$，其中 $v_1, \cdots, v_k$ 对应了分类的 k 个类别。文本和图像经过 DCCA 学习后得到了特征表示向量集合 $f(X)\in U^T$ 和 $g(Y)\in U^I$，它们作为 Stacking 分类算法的输入数据，对应于每一个文本 d 和图像 i，输出数据是它们在语义空间 V 上的每个类别的概率分布。

在 Stacking 分类模型训练时，将 $f(X)$和 $g(Y)$各分为两个集合，比例约为 7∶3。其中 70%的数据用于第一层分类模型的训练，生成基分类器后，将 30%的数据输入基分类器，基分类器的输出用于第二层分类模型的训练。第二层模型的输出是一个 k 维的概率向量，分别对应了与 k 个类别的语义相关性。这个过程完成后，相当于完成了式(3)的映射。

3.4 跨媒体检索的实现及评价

3.4.1 跨媒体检索的实现

在建立文本和图像在共同子空间的关联映射后，可以在共同的高层语义空间上实现语义匹配，从而可以根据一个文本来查询对应的图像，或者根据图像来查询文本。假设一个文本 d 和图像 i 通过特征表示、关联映射和语义匹配后，分别形成了一个 k 维的输出向量，则可以利用余弦相似度公式计算两者的相似性。

$$V_d = [d_{v_1}, d_{v_2}, \cdots, d_{v_k}] \tag{5}$$

$$V_i = [i_{v_1}, i_{v_2}, \cdots, i_{v_k}] \tag{6}$$

$$Similarity(d,i) = \frac{V_d \cdot V_i}{\| V_d \| \times \| V_i \|} \tag{7}$$

其中，$d_{v_1}, \cdots, d_{v_k}$ 和 $i_{v_1}, \cdots, i_{v_k}$ 分别表示文本 d 和图像 i 在语义空间 V 上 k 个特征类别上的分布。

当我们输入一个文档检索图像时，可以在其高层语义空间计算它与图像的相似性，获取与其语义相似的图像集合，然后将相似的图像通过式(2)映射回原图像空间，并将检索结果返回给用户。同理，当我们输入一个图像检索文档时，在高层语义空间获取与其相似的文档集合，并通过式(1)映射回原文本空间实现检索。

3.4.2 跨媒体检索评价

跨媒体检索效果评价采用被广泛使用的平均精度均值(MAP)指标，它是平均精度(Average Precision，AP)的平均值。AP 的计算公式如下：

$$AP = \frac{1}{T}\sum_{r=1}^{R} P(r)\delta(r) \tag{8}$$

其中，T 代表检索结果中与查询相关的样本个数；R 指考查的检索结果数量；$P(r)$代表 top r 的准确率，即返回的前 r 个结果中相关结果占的比例；$\delta(r)$表示第 r 个结果和查询的样本是否相关，若相关则为 1，否则为 0。AP 实际上是以准确率和召回率作为纵横轴形成的曲线下的面积。对一个有序的列表计算 AP 时，要先求出每个查询位置上的 $P(r)$，然后对所有的位置的 $P(r)$取平均值。MAP 值则是对所有查询的 AP 值取平均值。因此，检索出来的相关文档越靠前，MAP 值就越高。

4 实验

4.1 实验数据

为了验证提出的跨媒体检索方法的效果，使用了两个在

跨媒体检索领域运用广泛的公开数据集 Wikipedia[9] 和 Pascal Sentence[29]。其中，Wikipedia 数据集由一篇篇英文文档和其对应的图片组成，文档单词数量一般在 400～1000 之间，共包含 10 个类别和 2866 对文本图像对。Wikipedia 原始数据集被随机分成了 2173 对训练集和 693 对测试集样本。Pascal Sentence 数据集包含了 20 个类别信息的 1000 个文档和图像对，每个文档由 5 句人工标注的语句组成，每个类别均包括 50 张文本和 50 个图像。实验时将随机分配 800 对作为训练数据集，200 对作为测试数据集。

Wikipedia 和 Pascal Sentence 数据集的规模较小。为了进一步说明本文提出的方法在较大规模数据集上的有效性，选择 INRIA-Websearch 数据集[30]。该数据集包含 71478 对图像文本，分为 353 个语义类别。其中实验数据集由 100 个最大的类别数据组成，共包含 14698 个图像-文本对。

4.2 数据预处理

文本数据的预处理主要是去除停用词和格式转换工作。对于 Wikipedia 文本数据集，将所需的文本数据从每个 xml 文件中包含在〈text〉和〈\text〉标签中的内容抽取出来，然后去除停用词。对于 Pascal Sentence 文本数据集，由于每个文档包括几行独立的句子，因此只需要去除其中的停用词并将其拼接为一行句子。最后将处理好的文本数据写入文件中，一行句子代表一篇文档。

对于图像数据，主要进行图像归一化处理。Wikipedia 和 Pascal Sentence 的图像数据集格式基本一致，因此采取同样的方式处理。首先将图像大小调整为 256 * 256，然后利用 Pytorch 框架的 transforms 模块对图像进行归一化，像素值范围转换为区间[－1，1]，最后将处理好的图像数据保存在文件中。

4.3 Stacking 集成学习第一层和第二层分类算法的选择

为了融合各种模型的优点，Stacking 集成学习在第一层通常使用一些分类性能较好的算法作为基本模型，如 XGB 和 GBDT 等，在第二层使用简单的分类算法作为融合模型，如 SVM 和 LR 等。在此基础上，Stacking 集成学习也会根据数据集的特点选择具体的分类模型。比如 Wikipedia 数据集中的文本内容较长、特征丰富，除了 XGB 和 GBDT 模型，另外选用具有优异特征选择能力的 RF 模型作为基本模型。而 Pascal Sentence 数据集中的文本内容较短、特征较少，而 SVM 适合处理稀疏特征，因此选用 SVM，XGB 和 GBDT 作为基本模型，融合模型均采用 LR。对于两种数据集的图像来说，其特点相差不多，均使用了与其对应的文本分类一致的模型。

4.4 文本分类实验

由于不同的词向量各有特点，为了达到更好的跨媒体检索效果，需要先了解各词向量在数据集上的分类效果。这里选择 GloVe，FastText 和 BERT 这 3 种词向量库，它们的维度(dimension)和词表大小(vocabulary size)如表 1 所列。

表 1 词向量库参数

Table 1 Specifications of word embeddings

词向量类别	维度	词表大小
GloVe	300	400000
FastText	300	400000
BERT	256	30522

基于词向量库，文档的向量利用平均词向量法表示，经过标准 4 层网络的 DCCA 变换后形成分类算法的输入特征。在 Wikipedia 和 Pascal Sentence 两个数据集上分别采用 LR，KNN，SVM，XGB，GBDT 和 Stacking 算法来比较它们在 LDA，GloVe，FastText 和 BERT 这 4 种文本表示下的分类准确率，结果如表 2 所列。从表中可以看出，基于 3 种词向量库的文本表示方法在两种数据集上的分类准确率均高于 LDA 模型，且 Stacking 分类算法整体上具有最佳的效果。采用 BERT 时，SVM 算法在 Pascal Sentence 上表现最好，可能是因为该数据集文档数量较少，SVM 采用线性核函数，所以在特征数量相对样本数量更多时会表现得更好。这说明了基于词向量模型的文本特征表示效果优于传统的 LDA 模型，提出的集成分类方法是有效的。在两种数据集上的对比显示，词向量模型在 Wikipedia 上的分类效果整体优于 Pascal Sentence 数据集。这是由于后者的每篇文档仅由几个句子组成，特征词数量较少，这使得模型在分类时难以有效利用文档特征，从而达到较好的效果。而前者文档较长，特征词丰富，因此模型能更有效地提取出文档的语义特征。

表 2 文本分类结果比较

Table 2 Results of text classification

数据集	分类算法	LDA	GloVe	FastText	BERT
Wikipedia	LR	67.4	77.3	84.3	80.2
	KNN	66.8	76.9	86.7	84.4
	SVM	66.0	77.8	87.0	84.6
	XGB	67.2	78.8	85.1	81.8
	GBDT	66.7	78.2	83.3	79.2
	Stacking	**69.4**	**80.1**	**88.2**	**85.9**
Pascal Sentence	LR	63.2	66.0	67.5	62.0
	KNN	63.0	73.5	74.0	65.0
	SVM	67.5	72.0	74.5	**72.5**
	XGB	66.8	69.5	73.0	65.5
	GBDT	65.4	71.2	73.8	67.3
	Stacking	**68.7**	**74.0**	**75.0**	72.0

对比 3 种词向量技术对应的分类准确率，可以发现 FastText 效果最好。GloVe 的表现不佳，这可能是因为 GloVe 忽略了单词内部的词序信息，而 FastText 引入了字符级别的 n-gram 特征，所以其能够较好地处理训练样本中的未登录词问题。BERT 效果差于 FastText，可能是因为 BERT 适合处理复杂的句子关系，能够提取出更丰富的句子级别的语义信息，并且在大规模文本的分类任务上更有优势。而本文研究的实验数据集样本规模较小，句子较为简单，无法发挥出 BERT 模型的优势。

4.5 图像分类实验

我们选取了 ResNet50 和 Res2Net50 两种残差网络模型进行实验。它们的模型层数(layers)、模型参数总量(params)和模型计算量(Gflops，10 亿次浮点运算数/s)的参数信息如表 3 所列。

表 3 残差网络模型的参数

Table 3 Specifications of ResNets

模型	层数	参数总量	计算量
ResNet50	50	25.56×10^6	8.19
Res2Net50	50	48.4×10^6	8.3

与文本特征处理相似，两个数据集上的图像采用残差网络抽取特征后，经过标准 DCCA 网络变换后的特征作为输入。实验比较了 LR，KNN，SVM，XGB，GBDT 和 Stacking 算法在 SIFT，ResNet50 和 Res2Net50 这 3 种图像特征表示下的分类准确率，结果如表 4 所列。从表中可以看出，基于两种残差网络模型的图像分类准确率均显著高于 SIFT，但它们之间没有表现出明显的差异。不同算法的比较显示，Stacking 算法在两个数据集上较其他算法均有明显的优势。比较 Stacking 算法在两种数据集上的表现可以发现，其在 Pascal Sentence 上的分类效果整体优于 Wikipedia，这主要是由于 Wikipedia 数据集中一些不同类别的图像描述的语义内容较为接近，还有一些同一类别下的图像存在语义内容相距较远的情况，如艺术类、历史类中均存在建筑物图像。经过统计，Wikipedia 数据集中有歧义的图片占整个数据集的 10%～15%左右，这会严重影响图像的分类效果。相比 Wikipedia，Pascal Sentence 图像数据集的类别区分性较高，很少出现有歧义的图像。

表 4 图像分类结果比较

Table 4 Results of image classification

数据集	分类算法	SIFT	ResNet50	Res2Net50
Wikipedia	LR	25.5	47.8	50.6
	KNN	27.4	46.2	46.3
	SVM	26.0	50.1	50.2
	XGB	28.3	49.8	51.2
	GBDT	28.0	48.1	50.6
	Stacking	**29.6**	**51.5**	**51.7**
Pascal Sentence	LR	49.3	64.0	58.5
	KNN	50.4	70.5	69.5
	SVM	48.9	64.0	65.0
	XGB	53.5	64.0	66.5
	GBDT	51.2	68.5	68.0
	Stacking	**56.2**	**71.0**	**70.5**

4.6 跨媒体检索实验

基于文本和图像分类实验的结果，文本特征表示采用 FastText 模型，图像特征采用残差网络模型进行跨媒体检索实验，并与基本原理相近的跨媒体检索方法进行了对比。对比方法分别是 SCM(2010)[9]，DCCA(2013)[11]，Deep-SM(2016)[13]，Ada-SCM (2016)[31]，Bagging-SM (2018)[25]，CHRAN (2018)[32]。其中 SCM 是传统的语义关联匹配模型；DCCA 是将深度神经网络和典型相关性分析(CCA)相结合的跨媒体数据关联模型；Deep-SM 是引入了卷积神经网络(CNN)特征提取技术的深度语义匹配模型；Ada-SCM 是融合了 Ada-Boost 集成方法的跨媒体语义关联匹配模型；Bagging-SM 是融合了 Bagging 集成方法的跨媒体语义匹配模型；CHRAN 是文献[32]提出的跨媒体层级循环注意力网络(Cross-media Hierarchical Recurrent Attention Network)模型。这些方法提供了可用的代码，为开展对比实验提供了基础。表 5 列出了采用 ResNet50 抽取图像特征时，不同检索算法在两个数据集上图像检索文本(I2T)的 MAP 值、文本检索图像(T2I)的 MAP 值和两者的平均值。

表 5 跨媒体检索结果的比较(ResNet50)

Table 5 Results of cross-media retrieval(ResNet50)

数据集	算法	I2T	T2I	平均值
Wikipedia	SCM	45.4	44.9	45.2
	DCCA	44.0	40.6	42.3
	Deep-SM	44.1	49.5	46.8
	Ada-SCM	46.2	47.8	47.0
	Bagging-SM	48.3	46.7	47.5
	CHRAN	48.7	42.8	45.8
	Stacking-DSCM-WR	**53.5**	**50.4**	**52.0**
Pascal Sentence	SCM	53.5	58.6	56.0
	DCCA	57.0	56.8	56.9
	Deep-SM	**63.2**	59.0	61.1
	Ada-SCM	58.6	60.5	59.5
	Bagging-SM	52.4	54.1	53.3
	CHRAN	59.1	56.8	58.0
	Stacking-DSCM-WR	61.2	**65.8**	**63.5**

由表 5 可见，Stacking-DSCM-WR 方法在两个数据集的跨媒体检索实验中均取得了最好的 MAP 平均值，且在 Wikipedia 数据集上不管是图像检索文本、文本检索图像，还是平均性能均优于其他方法。在 Pascal Sentence 数据集上，该方法虽然在图像检索文本实验中性能较 Deep-SM 方法稍差，但在文本检索图像、平均值上均高于其他方法。Deep-SM 在 Pascal Sentence 上进行图像检索文本时的表现比本文方法稍好，可能是因为 Pascal Sentence 中的歧义图像较少，有利于发挥深度语义匹配算法的效果。

表 6 列出了采用 Res2Net50 抽取图像特征时各个算法的跨媒体检索 MAP 结果。

表 6 跨媒体检索结果的比较(Res2Net50)

Table 6 Results of cross-media retrieval(Res2Net50)

数据集	算法	I2T	T2I	平均值
Wikipedia	SCM	46.6	48.9	47.7
	DCCA	44.5	39.9	42.2
	Deep-SM	49.8	47.2	48.5
	Ada-SCM	46.4	**50.3**	48.3
	Bagging-SM	47.6	48.5	48.1
	CHRAN	49.9	43.9	46.9
	Stacking-DSCM-WR	**52.6**	48.0	**50.3**
Pascal Sentence	SCM	58.4	55.1	56.7
	DCCA	56.8	52.9	54.9
	Deep-SM	58.3	62.1	60.2
	Ada-SCM	**60.8**	55.3	58.1
	Bagging-SM	54.7	55.6	55.2
	CHRAN	58.8	56.1	57.5
	Stacking-DSCM-WR	60.6	**66.4**	**63.5**

同样，使用 Res2Net50 作为图像特征时，Stacking-DSCM-WR 方法在两个数据集上均取得最好的 MAP 平均值。在图像检索文本上，该方法在 Wikipedia 上取得了最好的结果，但在 Pascal Sentence 上较 Ada-SCM 方法稍差，主要原因可能是 Ada-SCM 融合了 Boosting 集成方法，在语义分类时，通过优化最小化损失函数来不断降低分类错误率，从而提升语义的匹配效果；而 Stacking 是通过融合各种基模型来实现分类，因此一些效果不好的基模型将会影响其最终的语义匹配效果。在文本检索图像上，本文方法在 Wikipedia 数据集上较 Ada-SCM，SCM，Bagging-SM 稍差，但在 Pascal Sentence 上取得了最好的结果，且效果显著，这可能是因为 Pascal Sentence 中有歧义的图像较少，而 Wikipedia 中因为存在较多有歧义的图像，所以限制了该算法性能的发挥。

对比 Stacking-DSCM-WR 方法在两种残差模型下的

结果可以看出，它们在 Pascal Sentence 数据集上的效果几乎一样，而在 Wikipedia 数据集上，ResNet50 模型稍有优势。原因可能是跨媒体语义关联主要取决于图像语义类别特征，对图像的多尺度特征要求不高，因此 Res2Net50 模型在跨媒体检索任务中并未表现出优势。

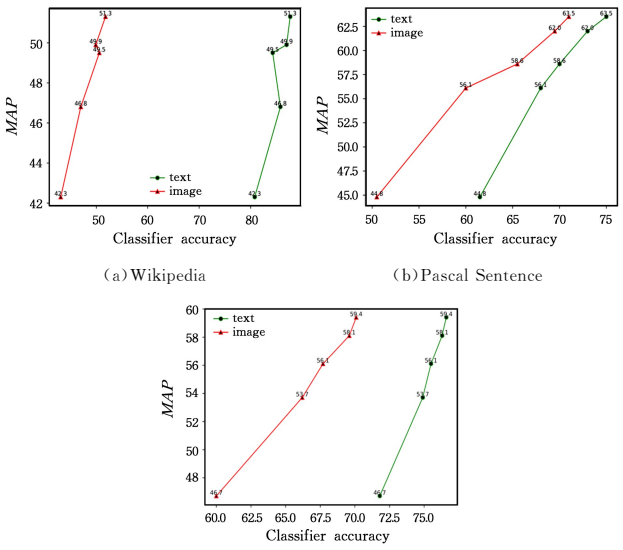
Wikipedia 和 Pascal Sentence 两个数据集虽然得到了广泛使用，但是它们的规模较小。为了说明本文方法在较大规模数据集上的效果，增加了 INRIA-Websearch 数据集的实验。之所以没有采用规模更大的数据集，是因为它们没有提供图像和文本的高层语义标签，无法采用本文及相关方法开展实验。在 INRIA-Websearch 上的 MAP 实验结果如表 7 所列。通过表 7 可以看出，Stacking-DSCM-WR 方法仍然取得了最好的效果。需要说明的是，Ada-SCM 和 Bagging-SM 两种算法在我们的设备上出现无法处理该较大规模数据的情况，长时间训练后算法没有停止，因此没有报告结果。

表 7 INRIA-Websearch 数据集跨媒体检索结果比较

Table 7 Results of cross-media retrieval (INRIA-Websearch dataset)

模型	算法	I2-T	T2-I	平均值
ResNet50	SCM	48.7	47.9	48.3
	DCCA	47.9	46.0	46.9
	Deep-SM	51.7	50.0	50.8
	CHRAN	54.3	58.2	56.3
	Stacking-DSCM-WR	**60.3**	**58.8**	**59.5**
Res2Net50	SCM	49.8	49.2	49.5
	DCCA	47.8	46.6	47.2
	Deep-SM	52.1	50.5	51.3
	CHRAN	57.4	57.2	57.3
	Stacking-DSCM-WR	**58.0**	**57.4**	**57.7**

为了说明 Stacking-DSCM-WR 方法中文本和图像分类准确率对最终检索效果的影响，分别将 Wikipedia，Pascal Sentence 和 INRIA-Websearch 的训练数据分为 5 个等份，逐次选择 20%，40%，60%，80%，100%的文本或图像数据进行训练，并采用测试集测试，实验结果如图 4 所示。



(a) Wikipedia (b) Pascal Sentence

(c) INRIA-Websearch

图 4 不同分类准确率下的检索效果(电子版为彩图)

Fig. 4 MAP under different classification accuracy

其中，标签为“text”的绿色折线表示文本分类准确率变化时的最终检索效果，而标签为“image”的红色折线则表示图像分类准确率变化时的最终检索效果。可以看出，对于 3 个数据集，随着文本分类准确率、图像分类准确率的提高，跨媒体检索效果基本得到同步提高，这说明分类效果的改善的确有助于跨媒体语义关联和检索性能的改善。

为了进一步说明 Stacking-DSCM-WR 方法的效果，将本文方法在 ResNet50 模型下的实验结果与 2019 年、2020 年新提出的、有代表性的跨媒体检索方法报告的实验结果进行了对比，它们的 MAP 如表 8 所列。其中，SRDMH(2019)[33] 采用有监督鲁棒离散多模哈希算法；DCLMM(2019)[34] 采用基于距离保持相关学习和多模流形正则化的半监督方法；TQSL(2019)[35] 基于任务依赖和查询依赖实现跨媒体检索；GIN(2019)[36] 采用基于图卷积神经网络(GCN)对文本建模并引入基于知识图谱的先验信息，此外，L[36] 对 GIN 进一步进行改进并提出了 Eo-GCN(2019)，它采用 Elmo 和 GCN 实现语义特征表示；MMSES(2020)[37] 基于模态依赖的嵌入语义增强空间方法；MMTES(2020)[37] 基于图像特征的嵌入联合降维空间方法。这些研究中均提供了 Wikipedia 上的实验结果，但只有 4 个算法提供了 Pascal Sentence 上的实验结果，3 个算法提供了 INRIA-Websearch 上的实验结果。

表 8 本文方法与 2019－2020 年相近研究结果的比较

Table 8 Comparisons of proposed method with similar results in 2019 and 2020

数据集	算法	I2T	T2I	平均值
Wikipedia	SRDMH	32.5	69.4	50.9
	DCLMM	52.7	46.7	49.7
	TQSL	46.3	41.5	43.9
	GIN	45.3	76.7	61.0
	Eo-GCN	42.3	**84.0**	**63.2**
	MMSES	43.8	39.5	41.7
	MMTES	44.7	38.9	41.8
	Stacking-DSCM-WR	**53.5**	50.4	52.0
Pascal Sentence	TQSL	50.5	50.2	50.4
	GIN	31.7	45.2	38.4
	MMSES	48.5	49.0	48.9
	MMTES	49.1	48.8	48.9
	Stacking-DSCM-WR	**61.2**	**65.8**	**63.5**
INRIA-Websearch	TQSL	54.1	55.0	54.6
	MMSES	53.6	56.8	55.2
	MMTES	53.9	56.2	55.1
	Stacking-DSCM-WR	**60.3**	**58.8**	**59.5**

从表 8 中可以看出，本文提出的方法与近几年的一些新方法相比仍然有优势，且表现出较好的稳定性。在 3 个数据集上，本文方法在图像检索文本上均取得了最好的效果，且在 Pascal Sentence 和 INRIA-Websearch 数据集上，图像检索文本 MAP、文本检索图像 MAP 和两者平均值均取得最好的结果。GIN 方法虽然在 Wikipedia 数据集上的文本检索图像 MAP 和 MAP 平均值上取得了较好的效果，但在 Pascal Sentence 数据集上表现平平，不如本文方法稳定。需要说明的是，GIN 及其改进方法 Eo-GCN 由于在文本特征表示时采用 GCN 且引入先验信息，Eo-GCN 还采用了基于 Elmo 的词语隐层向量特征表示，因此其在文本检索图像上取得了显著

优于本文方法的结果，这一方面说明特征表示对于跨媒体检索效果提升的重要性，另一方面为未来进一步改进本文算法提供了思路。

结束语 针对当前跨媒体检索效果不佳的问题，本文提出了 Stacking-DSCM-WR 跨媒体检索方法。该方法通过词向量技术形成文档的特征表示向量，通过残差网络技术抽取图像的特征表示向量，采用深度典型相关性分析将不同模态的数据投影到共同子空间下，然后采用 Stacking 集成学习算法获取文本和图像在同一高层概念语义空间上的分布，使得两种不同模态的数据在语义上可以匹配、计算相似性。并在 Wikipedia 和 Pascal Sentence 两种跨媒体数据集上分别开展了文本分类实验、图像分类实验、跨媒体检索实验。文本分类实验证实了词向量技术比传统的 LDA 文本特征提取技术效果好，而且提出的 Stacking 分类算法在不同词向量库、不同数据集下较单一分类算法均取得了较好的效果。图像分类实验证实了残差网络技术能够比传统的 SIFT 技术更好地抽取图像特征，且同样证实了提出的 Stacking 方法较其他单一算法在图像分类上有更好的性能。采用 FastText 词向量抽取文本特征、残差网络抽取图像特征，进而使本文提出的 Stacking-DSCM-WR 方法在与 SCM，DCCA 等 6 种相近方法对比中表现出最佳的性能。实验结果说明，本文提出的 Stacking-DSCM-WR 跨媒体检索方法能有效地建立文本和图像两种模态数据的关联，实现良好的跨媒体检索效果。

本文提出的方法虽然取得了较好的效果，但仍然存在一些限制，需要进一步进行优化。从数据集的角度来说，本文该方法受制于实验数据集必须具备高层语义标签的限制，本文仅在两个小规模数据集和一个较大规模数据集上进行了跨媒体检索实验，并不能说明它在其他数据集或者更大型数据集上一定是有效的，未来需要在更多的数据集上加以验证。从算法设计的角度来说，所提方法仍有进一步改进的空间。已经有研究使用神经网络模型建立不同模态的共同嵌入空间，试图寻找一种端到端的语义关联模式。而本文提出的方法对经典的语义关联匹配方法提出改进，将特征表示模块和语义映射模块分开进行，因此未来需要对比本文方法与端到端方法的效果，探究更有效的跨媒体数据语义关联模式。从跨媒体数据类型的角度来说，本文用到的跨媒体数据类型只包括文本和图像，类似音频、视频、3D 图像等媒体数据均不涉及，因此在未来的工作中，需考虑更多的跨媒体数据类型，探究多模态数据关联和检索方法。

参 考 文 献

[1] ZHAO Y，WEI S K，WANG S H. Knowledge representation in cross media era：perception，relevance and consistency representation [J]. Communications of the CCF，2014，10(7)：8-13.

[2] WEI Y C. Semantic classification and retrieval for cross-media Data [D]. Beijing：Beijing Jiaotong University，2016.

[3] PENG Y X，ZHU W W，ZHAO Y，et al. Cross-media analysis and reasoning：advances and directions[J]. Frontiers of Information Technology & Electronic Engineering，2017，18(1)：44-57.

[4] HUANG X，PENG Y X. Deep cross-media knowledge transfer [C]// 31th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway：IEEE，2018：8837-8846.

[5] ZHANG B. Research on multimodal multimedia retrieval method based on neural network [D]. Jinan：Shandong Normal University，2018.

[6] XIE Y X，LUAN X D，WU L D. Multimedia Data Semantic Gap Analysis [J]. Journal of Wuhan University of Technology(Information & Management Engineering)，2011，33(6)：859-863.

[7] PENG Y X，HUANG X，ZHAO Y，et al. An overview of cross-media retrieval：concepts，methodologies，benchmarks，and challenges[J]. IEEE Transactions on Circuits and Systems for Video Technology，2018，28(9)：2372-2385.

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

[9] RASIWASIA N，PEREIRA J C，COVIELLO E，et al. A new approach to cross-modal multimedia retrieval[C]// International Conference on Multimedia. New York：ACM，2010：251-260.

[10] HWANG S J，GRANMAN K. Learning the relative importance of objects from tagged images for retrieval and cross-modal search[J]. International Journal of Computer Vision，2012，100(2)：134-153.

[11] ANDREW G，ARORA R，BILMES J，et al. Deep canonical correlation analysis[C]// International Conference on International Conference on Machine Learning. CambridgeMA：Microtome Publishing，2013，28(3)：1247-1255.

[12] RASIWASIA N，MAHAJAN D，MAHADEVAN V，et al. Cluster canonical correlation analysis[C]// Proceedings of Machine Learning Research. Reykjavik：PMLR，2014：823-831.

[13] WEI Y，ZHAO Y，LU C，et al. Cross-modal retrieval with CNN visual features：a new baseline[J]. IEEE Transactions on Cybernetics，2017，47(2)：449-460.

[14] KUMAR S，UDUPA R. Learning hash functions for cross-view similarity search[C]// International Joint Conference on Artificial Intelligence. Barcelona：IJCAI，2011：1360-1365.

[15] ZHAI X H，PENG Y X，XIAO J G. Heterogeneous metric learning with joint graph regularization for cross-media retrieval [C]// Web Information Systems Engineering. Heidelberg：Springer，2013：1198-1204.

[16] MESSINA N，AMATO G，ESULI A，et al. Fine-grained Visual Textual Alignment for Cross-Modal Retrieval using Transformer Encoders[J]. ACM Transactions on Multimedia Computing，Communications，and Applications，2021，17(4)：1-23.

[17] MIKOLOV T，CHEN K，CORRADO G，et al. Efficient Estimation of Word Representations in Vector Space[J]. arXiv：1301.3781，2013.

[18] PENNINGTON J，SOCHER R，MANNING C. Glove：Global Vectors for Word Representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP). Doha，Qatar，2014：1532-1543.

[19] JOULIN A，GRAVE E，BOJANOWSKI P，et al. Bag of tricks for efficient text classification [C]// Proceedings of the 15th Conference of the European Chapter of the Association for Com-

putational Linguistics: Volume 2, Short Papers. Stroudsburg: ACL, 2017: 427-431.

[20] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [J]. Eprint Arxiv, 2019(5): 1-16.

[21] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [J]. Computer Vision and Pattern Recognition, 2015, 19(1): 51-59.

[22] FENG J, LU C Y. Cross Media Retrieval Method Based on Residual Attention Network[J]. Computer Science, 2021, 48(6A): 122-126.

[23] GAO S H, CHENG M M, ZHAO K, et al. Res2Net: a new multi-scale backbone architecture [EB/OL]. (2019-09-01) [2020-06-08]. https://arxiv.org/pdf/1904.01169.pdf.

[24] CAI Y, ZHU X F, SUN Z L, et al. Semi-supervised and Ensemble Learning: A Review [J]. Computer Science, 2017, 44(Z1): 7-13.

[25] SCHWENKER F. Ensemble methods: foundations and algorithms[J]. IEEE Computational Intelligence Magazine, 2013, 8(1): 77-79.

[26] CHEN X. Research on cross modal multimedia retrieval method based on semantic matching [D]. Jinan: Shandong Normal University, 2018.

[27] WU D P, ZHANG Z L, CAO T T. Research on Stability Classifier Combination Algorithm Based on Stacking Strategy [J]. Journal of Chinese Computer Systems, 2019, 40(5): 135-139.

[28] ZHAI W J, YAN Y, ZHANG B W, et al. A Model for Text Representation and Classification Based on Hybrid Deep Belief Networks [J]. Technology Intelligence Engineering, 2016, 2(5): 30-40.

[29] RASHTCHIAN C, YOUNG P, HODOSH M, et al. Collecting image annotations using amazon's mechanical turk[C]// Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Stroudsburg: ACL, 2010: 139-147.

[30] KRAPAC J, ALLAN M, VERBEEK J J, et al. Improving web image search results using query-relative classifiers[C]// Computer Vision & Pattern Recognition. IEEE, 2010.

[31] LIU Y. Cross-modal multimedia information retrieval with CCA and Adaboost [D]. Chongqing: Southwest University, 2016.

[32] JI 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.

[33] LI C X, YAN T K, LUO X, et al. Supervised Robust Discrete Multimodal Hashing for Cross-Media Retrieval[J]. IEEE Transactions on Multimedia, 2019, 21(11): 2863-2877.

[34] WANG T, ZHANG H, LI B, et al. Semisupervised Cross-Media Retrieval by Distance-Preserving Correlation Learning and Multi-modal Manifold Regularization[C]// Pacific Rim International Conference on Artificial Intelligence. Cham: Springer, 2019.

[35] WANG L. Research on Cross Media Retrieval Algorithm based on discriminative common subspace [D]. Jinan: Shandong Normal University, 2019.

[36] LU Y H. Semantic Modeling of Textual Relationship in Cross-Media Information Retrieval [D]. Beijing: University of Chinese Academy of Sciences, 2019.

[37] ZHENG S X. Research on Cross Media Retrieval Algorithm based on Embedded Spatial Representation [D]. Jinan: Shandong Normal University, 2020.

HAN Hong-qi, born in 1971, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include data mining, cross-media retrieval and knowledge engineer.

(责任编辑:李亚辉)