文章编号:1005-0523(2020)06-0124-07

基于依存关系注意力增强的跨模态检索研究

曾 辉,胡 蓉,淦修修,彭志颖,熊李艳

(华东交通大学信息工程学院,江西 南昌 330013)

摘要:随着互联网技术的极速发展,不同模态的多媒体数据呈指数增长,人们已经无法满足于原始的图片检索的单模态数据检索方式,跨模态的多媒体检索成为信息检索的一个重要研究领域。针对该任务,提出一种增加句子依存关系词组注意力机制的双分支网络结构跨模态检索方法。该方法基于CNN模型提取图像特征,基于句法结构分析获得文本的依存关系片段,构建双分支网络结构模型,并嵌入注意力机制学习各依存关系片段的权重分布,使文本的特征表示能够更偏重于关键的句子片段特征。实验结果表明该方法相比于其他方法在*P@K*检索准确率评估指标上都有较好的提高,验证了算法的有效性。

关键词:依存关系元组;句子折分;注意力机制;双分支网络

中图分类号:TP391 **文献标志码**:A

本文引用格式:曾辉,胡蓉,淦修修,等. 基于依存关系注意力增强的跨模态检索研究[J]. 华东交通大学学报,2020,37(6):124-130.

Citation format:ZENG H,HU R,GAN X X,et al. A cross-modal retrieval method based on sentence dependency attention[J]. Journal of East China Jiaotong University,2020,37(6):124-130.

DOI:10.16749/j.cnki.jecjtu.2020.06.017

在信息时代,为了便于人们获取感兴趣的视觉数据,从一个大规模数据库中快速地找到和查询图像内容相关或相似的图像,并按相关的排序返回给用户是很常见的需求。但是随着多媒体数据,特别是图像视频文字数据,呈爆炸式增长,单纯使用图像来检索图像的单模态检索方式已经难以满足当前用户的需求[1],以互联网上庞大数目的图像视频以及描述性文本等数据为基础的多模态数据研究已经成为当前检索系统领域的研究热点。

目前相关的研究大多利用数据间的共存与互补特征来分析多模态之间的语义理解与关联性表示。其中典型相关分析(canonical correlation analysis,CCA)[2]是跨模态检索研究中最常用的方法,通过寻找一个线性映射向量能够将映射至相同子空间后的两类模态数据相关系数最大化来实现跨模态检索。Tenenbaum J B 等[3]为跨模态识别提出双线性因子模型(bilinear model,BLM),能够对因子交互进行充分表达,并基于奇异值分解算法来拟合数据。Akaho S[4]就运用了改进的非线性CCA算法KCCA,即核典型相关分析算法,把核函数的思想引入CCA中。Gong Y[5]使用改进的KCCA算法,但是增加了图像和文本之外的第三类特征—语义关键词。语义相关匹配[6](semantic correlation matching,SCM)将文本语义抽象化并融合入典型相关分析学习过程中,提高了跨模态检索的准确性。Jacobs D W 等[7]提出广义多视判别分析方法(generalized multiview analysis,GMA),将特征提取转化为求解广义特征值问题,解决在不同特征空间的联合表示,获得有效的潜在子空间表示,这是CCA方法的有监督扩展。Kan M[8]在研究中基于神经网络架构而改进典型相关分析CCA算法,通过多分支的多层神经网络对图像和文本数据进行特征提取, 并将CCA的优化目标设计成为双分支网络的优化函数。Wang L[9]设计一个双分支的全连接神经网络(convolutional neural network,CNN)来对图像和文本特征信息分别进行表示学习,进而进行总体的联合嵌入空间表示,作者新颖地使用类间距离和类内距离作为优化目标。深度学习算法对这类包含位置、色彩、时序性的结构数据拥有很好的特征学习能力,其中卷积神经网络被广泛应用于图像表征的学习。Diaz-chito K[10]使用CNN和LSTM网络对图片和文本进行编码,再

收稿日期:2020-09-01

基金项目:国家自然科学基金项目(62067002)

作者简介:曾辉(1973—),男,副教授,硕士生导师,研究方向为软件工程和数据挖掘,E-mail:macrohui29@sina.com

嵌入联合空间表示,并通过解码网络来保证各自的信息的重构性。Karpathy A[11]通过多目标识别算法 RCNN 对图片样本进行区域切割和文本识别,利用句法依赖树对文本句子样本进行词语切割。最后比对各片段相似度和整体相似度,通过保证类内差异小,类间差异大来进行模型学习。然而对于描述图像内容的文本数据来说,真正能够代表图像主要语义内容的只有一个或几个词,即不同的词在语义表达中的重要程度存在差异性。本文参考注意力机制在机器翻译、图文理解任务中的广泛应用模式,通过对句子依存树结构的片段化处理,添加关于依存关系元组注意力机制,使关键词组片段在文本数据表示时具有更加重要的影响作用,并基于双分支网络结构模型训练,实现跨模态检索任务。

1　模型方法

1.1　模型结构

本文所提的模型框架主要包含 3 部分,双分支网络、注意力机制嵌入和联合空间表示。如图 1 所示,双分支网络分别由图像特征输入部分和文本特征表示部分组成,每个分支都存在 3 个全连接隐含层,其中使用 relu 函数作为全连接层之间对各层输出作非线性映射;注意力机制嵌入部分通过学习基于句子依存关系拆分的不同词组的权重分布,使重要的词组在特征表示中占更多的权重;而联合嵌入表示空间将图像和文本在相同的维度空间内进行表示,从而计算相似性,并通过 hinge loss 目标函数来对整个模型进行训练,使对应的图片—句子组合具有更高的相似度。

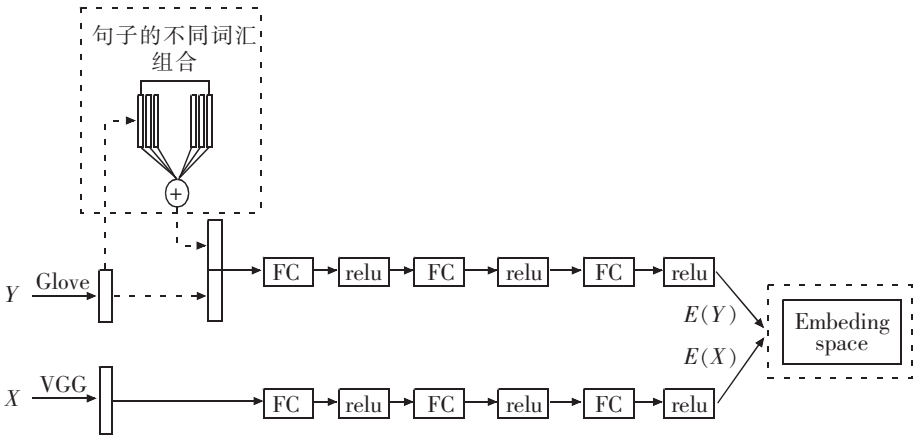

图 1　跨模态检索模型框架

Fig.1　Cross-modal retrieval model framework

1.2　注意力机制嵌入

1.2.1　图像表示

现实的图像是通过无数像素点的集合来表示,而单纯的像素点的灰度值无法有效表达图像的高层语义特征,故需要进行图像数据特征向量化。对于图像数据,本文采用迁移学习的思想,选择当前图像表征效果卓越的卷积神经网络 CNN 模型,使用基于 CNN 网络架构的 VGG 预训练模型[12],从原始图像 I 中提取图像特征 V_I,VGG 模型是由牛津大学计算机视觉组合和 Google DeepMind 公司研发的一种 16~19 层深度的卷积神经网络,并在 2014 年的机器视觉领域顶级学术竞赛(imagenet large scale visual recognition challenge, ILSVRC)中获得分类项目第二名以及图像定位比赛的第一名,目前为止被广泛运用于深度学习中提取图像特征。

如图 2 所示,VGG 模型能够有效对图像做特征表示, 而 VGG 模型需要输入图像的维度是 224×224,本文首先进行图像增强操,用于增加数据量,将图像先统一放缩成 256×256 大小,然后分别裁剪左上角、左下角、右上角、右下角和中部 5 个位置 224×224 大小的图像扩充数据,选择 VGG 模型最后一层池化层输

出作为图像表示特征向量，每张图片向量维度为4 096维。

图像特征提取过程用公式(1)表示，其中 I 表示原始图像，V_I 表示提取的图像特征

$$V_I=\mathrm{CNN}_{\mathrm{VGG}}(I) \tag{1}$$

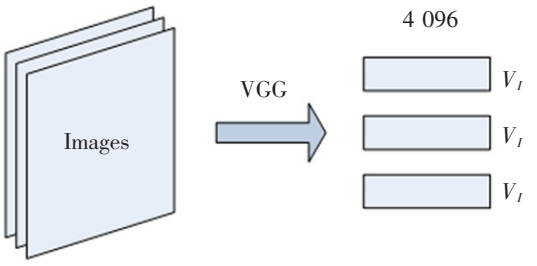

图2 VGG预训练模型提取特征

Fig.2 VGG pre-training model extraction feature

1.2.2 文本表示与句子依存结构构建

对于文本数据的表示，分为两个方面，一方面需要对句子文本自身作句子向量表示；另一方面需要将句子进行结构拆分，构造出多个子结构区域，用于注意力机制嵌入计算。文本数据集合用 U_{T} 表示。

1）句子向量表示。对于单个句子文本 $S(S\in U_{\mathrm{T}})$，由于句中副词等词语出现的频次较高，会影响句子的语义表达。本文首先对句子进行去停用词处理。句子是由多个词组成，要对句子进行表示首先需要进行词表示，本文使用基于word2vec[13]模型改进的GLOVE方法训练的词向量来对每个词作词向量表示，经GLOVE训练好的词向量表示维度分别有50维、100维、200维和300维等，能够根据不同的需求选择不同的向量维度。本文考虑模型结构和计算的复杂性，选择300维的GLOVE词向量来表示每个词 v_t，然后以整个训练集文本为集合，计算句子中每个词在整个训练集中的 *TF-IDF* 值 $W_{TF\text{-}IDF_t}(t\in S)$，并以每个词的 *TF-IDF* 值作为GLOVE词向量的权重进行加权求和，如公式(2)所示，最终得到300维的向量即代表此句子表示向量 $\boldsymbol{V}_S$。W_{TF_t} 表示对应单个句子文本中的某一个词 t 在整个训练集中的 *TF-IDF* 值。

$$\boldsymbol{V}_S=\sum_{t\in S} W_{TF_t}\cdot v_t \tag{2}$$

2）基于依存语法拆分句子。句子结构拆分的目的是将句子拆分成多个元组，每个元组由多个词组成，并且分别代表了某种词组结构关系，这样就能够将句子片段化。因此首先需要对句子作句法结构分析，并确定句子中词汇之间的依赖关系，本文选择当前广泛使用的句法解析工具斯坦福分析器(Stanford Parser)来处理英文文本，获得有效的句子依存结构关系。

句法结构普遍分成两种表现形式，经过分词和词性分析过程后，生成如图3的树结构，称为短语句法结构树。短语结构树能够表达句子的句法结构，源自传统的句子图解法，把句子分割成各个组成部分，较大的组成成分能够由较小的组成成分合并得到，由此逐级传递分解。只有叶子结点代表句子中各个词本身，而其他的中间结点是短语成分的标记，短语结构树能够表示每个词在句子中的所属成分和位置，但是无法直接处理词与词之间的依赖关系。图4是简化的图结构。依存关系树用于表达句子词与词之间的

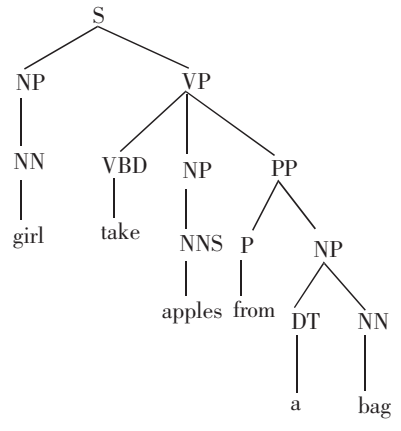

图3 短语结构树

Fig.3 Phrase tree

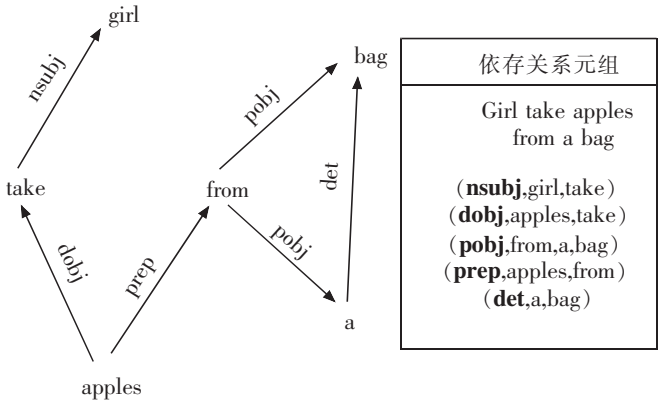

图4 依存关系结构图

Fig.4 Dependency structure chart

依存关系，依存结构中每个结点都是一个词，通过词之间的连接弧表示词之间的“主谓宾”和“定状补”语法修饰关系，同时由于总体的节点数大大减少，使结构更加简洁清晰。本文通过 Stanford Parser 生成句子依存关系树，获得如图 4 中右框中所示的依存关系元组，以 *TF-IDF* 作为词权重加权求和获得句子依存关系元组表示矩阵 $\boldsymbol{V}_Q$。

1.2.3　句子依存关系的注意力机制嵌入

不同的词汇依存关系组合对于句子语义表达的重要程度是存在差异的，因此本文在模型中加入注意力机制来学习每个词汇组合表示的权重分布[14]，以此反映出句子各片段区域的偏重性，并将其加入句子表征向量中。如公式(3)所示，首先通过一个单层神经网络对句子表示向量 $\boldsymbol{V}_S$ 和句子依存关系元组表示矩阵 $\boldsymbol{V}_Q$ 进行非线性变换来激活组合，然后通过公式(4)的 softmax 函数获得关于各依存关系元组的关注度概率分布。

$$h_A=\tanh(\boldsymbol{W}_{Q,A}\boldsymbol{V}_Q\oplus(\boldsymbol{W}_{S,A}\boldsymbol{V}_S+b_A)) \tag{3}$$

$$P_Q=Softmax(\boldsymbol{W}_P h_A+b_P)) \tag{4}$$

式中：$\boldsymbol{V}_Q\in\boldsymbol{R}^{d\times m}$，$d$ 代表每个依存关系元组向量的维度，m 是句子元组的数目，即句子经过依存语法拆分后保留的词语组合的个数；$\boldsymbol{V}_S\in\boldsymbol{R}^d$ 是 d 维句子表示向量。设定映射参数矩阵 $\boldsymbol{W}_{Q,A}$，$\boldsymbol{W}_{S,A}\in R^{k\times d}$ 以及 $\boldsymbol{W}_P\in\boldsymbol{R}^{k\times k}$，通过 softmax 函数计算后得到 $P_I\in\boldsymbol{R}^m$ 的 m 维向量，表示此句子每个依存关系组合片段的注意力概率值。由于 $\boldsymbol{W}_{Q,A}\boldsymbol{V}_Q\in\boldsymbol{R}^{k\times d}$ 是 $k\times m$ 的矩阵而 $\boldsymbol{W}_{S,A}\boldsymbol{V}_S\in\boldsymbol{R}^k$ 是 k 维向量，故将矩阵的每一列分别与该向量进行求和，得到 $h_A\in\boldsymbol{R}^{k\times m}$。$\boldsymbol{W}_{Q,A}$，$\boldsymbol{W}_{S,A}$，$\boldsymbol{W}_P$，$b_A$，$b_P$ 是通过模型学习得到的参数。P_Q 表示每个依存关系词组对于句子语义内容的重要性度量。

基于注意力分布概率 P_Q，通过计算句子所有依存关系片段的加权和，获得嵌入了重要依存关系元组注意力增强机制的表示向量 $\widetilde{\boldsymbol{V}}_Q$，如公式(5)所示

$$\widetilde{\boldsymbol{V}}_Q=\sum_{q\in Q}P_q\boldsymbol{V}_q \tag{5}$$

最后将向量 $\widetilde{\boldsymbol{V}}_Q$ 与句子表示向量 $\boldsymbol{V}_S$ 进行首尾拼接融合，可得到添加了注意力机制的文本表示向量。

1.3　目标函数

图像、文本数据分别用 X，Y 表示，对于给定的训练集图像样本 x_i，设定文本样本 y_i 和 y_i^-，它们分别表示与图像 x_i 正确匹配的文本和不正确对应的文本，$\boldsymbol{E}(x_i)$，$\boldsymbol{E}(y_i)$，$\boldsymbol{E}(y_i^-)$分别代表各自的嵌入空间的最终输出向量。跨模态检索的目标是期望 $\boldsymbol{E}(x_i)$与正确文本 $\boldsymbol{E}(y_i)$之间的相似度比 $\boldsymbol{E}(x_i)$和 $\boldsymbol{E}(y_i^-)$之间的相似度更高，如公式(6)所示

$$\mathrm{Sim}(\boldsymbol{E}(x_i),\boldsymbol{E}(y_i))-m>\mathrm{Sim}(\boldsymbol{E}(x_i),\boldsymbol{E}(y_i^-)) \tag{6}$$

其中 m 是阈值参数，表示两者相似度期望的差值，设 m=0.3。Sim(·)表示两者相似性计算函数，本文采用余弦函数作为相似度计算方式，通过计算向量之间的余弦夹角体现相似程度。如公式(7)所示

$$\mathrm{Sim}(\boldsymbol{E}(x_i),\boldsymbol{E}(y_i))=\frac{(\boldsymbol{E}(x_i)\cdot\boldsymbol{E}(y_i)^{\mathrm{T}}}{\|(\boldsymbol{E}(x_i)\|\cdot\|\boldsymbol{E}(y_i)\|} \tag{7}$$

同样，对于给定的文本数据 y_i 来检索相关图像，如公式(8)所示

$$\mathrm{Sim}(\boldsymbol{E}(y_i),\boldsymbol{E}(x_i))-m>\mathrm{Sim}(\boldsymbol{E}(y_i),\boldsymbol{E}(x_i^-)) \tag{8}$$

对于图像检索文本过程，三元组$\{x_i,y_i,y_i^-\}$损失函数 $L_{(x,y_i,y_i^-)}$定义如下

$$L_{(x,y_i,y_i^-)}=\max[0,m+\mathrm{Sim}(E(x_i),E(y_i^-))-\mathrm{Sim}(E(x_i),E(y_i))] \tag{9}$$

对于文本检索图像过程，损失函数计算方式定义

$$L_{(y_i,x,x_i^-)}=\max[0,m+\mathrm{Sim}(E(y_i),E(x_i^-))-\mathrm{Sim}(E(y_i),E(x_i))] \tag{10}$$

因此，使用 hinge loss 函数方法定义图像分支 x_i 与文本分支 y_i 联合损失，如公式(11)所示

$$L_{X_i,Y_i}=L_{(x,y_i,y_i^-)}+L_{(y_i,x,x_i^-)} \tag{11}$$

故定义模型的总损失 $L_{X,Y}$ 如下

$$L_{X,Y}=\frac{1}{N}\sum_{i}^{N}L_{X_i,Y_i} \tag{12}$$

其中 N 是测试集样本数。

2 实验结果与分析

2.1 数据集

本文选择公共数据集 Flickr8K 和 Flickr30K 用来对所提算法进行实验评估，Flickr8K 和 Flickr30K 是图片分享网站 Flickr 上筛选出的专门用于图像研究的公共数据集，分别包含 8 000 和 30 000 幅图像，每幅图像使用标 5 个独立的描述句子标注，数据集样本的内容如图 5 所示。

图 5 数据集表示

Fig.5 Dataset representation

2.2 评估方法

本文采用排序问题评价指标 $P@K$ 对算法性能进行评估，计算方式如下

$$P@K=\frac{1}{N}\sum_{i}^{N}\delta(k) \tag{13}$$

$P@K$ 用于判断检索排在前 k 位的相关性，N 是测试集样本数，若在检索结果中排序前 k 个中有正确对应的样本，则 $\delta(k)=1$ 表示检索正确。

2.3 实验设置

在模型训练过程时通过交叉验证方式选取最优实验参数设置，选择随机梯度下降方法(stochastic gradient descent，SGD)优化目标函数，学习率设置为 0.01，隐含网络层激活函数选择 relu 函数。嵌入表示空间维度为 300，即将每类模态数据最终使用 300 维的向量进行特征表示，从而计算相似度。每层隐含层之后的 dropout 方法的节点丢弃率设为 0.5，以分批度训练的方式训练样本，每个批度数据数目设为 100，每个单分支网络的 3 层全连接网络层的神经元节点数设置为 1 000*1 000*1 000。

2.4 实验结果分析

本文分别使用 4 种算法与所提算法 VSDA 进行实验对比，其中“CCA”是典型相关分析算法，是一种经典的无监督关联学习方法，通过最大化不同模态间在投影子空间的相关性实现匹配。“SCM”是语义关联匹配方法，基于 CCA 方法学习子空间映射，再基于多标签的逻辑回归方法学习不同模态间的语义匹配。“KCCA”是一种将核函数思想引入 CCA 进行高维映射的方法。“DCCA”是基于典型相关分析理论的深度网络改进模型。结果如表 1 所示。

表 1 各算法实验结果

Tab.1 Experimental results of various algorithms

模型	Flickr8K 数据集						Flickr30K 数据集					
	Text to Image			Image to Text			Text to Image			Image to Text		
	P@1	*P*@5	*P*@10	*P*@1	*P*@5	*P*@10	*P*@1	*P*@5	*P*@10	*P*@1	*P*@5	*P*@10
CCA	15.4	22.6	24.3	12.4	17.6	22.1	10.9	24.3	28.8	9.2	18.9	23.7
SCM	12.6	19.7	26.2	14.5	16.3	20.6	11.6	21.5	26.3	12.9	18.7	19.5
DCCA	18.9	28.4	39.2	18.9	25.4	30.8	17.6	26.4	29.7	14.6	17.4	24.9
KCCA	21.8	30.5	40.6	21.6	33.6	39.2	23.5	38.5	45.2	22.3	34.3	42.8
VSDA	23.2	34.3	42.6	20.7	35.7	43.9	25.6	39.3	48.9	23.3	36.2	46.7

表 1 是各类算法在 Flicker8K 和 Flicker30K 数据集下的 *P@K* 指标的实验结果，分别对给定文本的情况下检索与文本内容匹配的图像，和给定图像的情况下查找与图像内容匹配的相关文本。整体来说，本文所提的 VSDA 算法相较于其他对比方法无论是 *P*@1、*P*@5 还是 *P*@10 的检索准确率都有一定程度的提高，尤其是相比于 CCA、SCM、KCCA 算法，在 *P*@5 和 *P*@10 指标上效果提升显著，可能是神经网络模型在异构数据之间具有复杂关系情况下对比于传统理论模型拥有更强大的学习能力。其次，检索准确率普遍非常低，但是对于检索任务来说，其目的类似于推荐系统形式是为用户提供合理的检索结果集合，对于 Top1 即正确对应的样本一定要排序在首位的需求并不是必须，故 *P*@1 的结果偏低是可以接受的。随着数据量的增加，由 Flicker8K 的 8 000 张图片到 Flicker30K 的 30 000 张图片，本文的 VSDA 算法和 DCCA 方法的 Text 检索 Image、Image 检索 Text 的检索准确率都有可观的提高，然而其他方法的检索准确率普遍都有相对应的降低，说明 VSDA 算法在大数据集情形下同样可以获得好的性能。对比 VSDA 算法和 DCCA 方法，Flicker8K 数据集和 Flicker30K 数据集的各项指标都高于 DCCA 方法，而 DCCA 方法同样是使用双分支网络结构，说明本文所提的 VSDA 算法针对句子依存关系注意力机制的嵌入改进是具有可行性的。

3　结论

本文针对跨模态多媒体检索领域中图像与文本之间的互检任务，考量句子文本的不同片段对于整体语义表达具有不同的偏重性，通过基于依存关系结构的句子拆解，嵌入注意力机制从而学习各片段元组的权重分布，并设计双分支神经网络模型实现跨模态检索。实验结果表明该方法的检索准确率相比于其他算法有着显著的提高。未来的研究工作将从以下两个方面去考虑：① 模型学习主要采用的是全连接网络，可以考虑替换成其他更优越的模型架构，可能获得更好的效果；② 可以考虑图像和文本之间是否存在其他联系。

参考文献：

[1] 李志义，黄子风，许晓绵. 基于表示学习的跨模态检索模型与特征抽取研究综述[J]. 情报学报，2018，37(4)：422-435.

[2] 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.

[3] TENENBAUM J B，FREEMAN W T. Separating style and content with bilinear models[J]. Neural Computation，2000，12(6)：1247-1283.

[4] AKAHO S. A kernel method for canonical correlation analysis[J]. International Meeting of Psychometric Society (IMPS) 2006，36(9)：347-356.

[5] GONG Y，KE Q，ISARD M，et al. A multi-view embedding space for modeling internet images，tags，and their semantics[J]. International Journal of Computer Vision，2014，106(2)：210-233.

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

[7] JACOBS D W，DAUME H，KUMAR A，et al. Generalized multiview analysis：a discriminative latent space[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society，2012：2160-2167.

[8] KAN M，SHAN S，CHEN X. Multi-view deep network for cross-view classification[C]//Computer Vision and Pattern Recognition. IEEE，2016：4847-4855.

[9] WANG L，LI Y，LAZEBNIK S. Learning deep structure-preserving image-text embeddings[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)，IEEE，2016：5005-5013.

[10] DIAZ-CHITO K，DEL RINCON J M，HEMANDEZ-SABATE A. Decremental generalized discriminative common vectors applied to images classification[J]. Knowledge-Based Systems，2017，131：46-57.

[11] KARPATHY A，JOULIN A，FEI-FEI L. Deep fragment embeddings for bidirectional image entence mapping[J]. Advances in Neural Information Processing Systems，2014，3：1889-1897.

[12] YANG Z,HE X,GAO J,et al. Stacked attention networks for image question answering[C]//Computer Vision and Pattern Recognition. IEEE,2016:21-29.

[13] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient estimation of word representations in vector space[J]. Computer Science,2013,34:226-240.

[14] XU K,BA J,KIROS R,et al. Show attend and tell:neural image caption ceneration with visual attention[J]. Computer Science,2015,15:2048-2057.

A Cross-modal Retrieval Method Based on Sentence Dependency Attention

Zeng Hui,Hu Rong,Gan Xiuxiu,Peng Zhiying,Xiong Liyan

(East China Jiaotong University School of Information Engineering,Information Engineering,Nanchang 330013,China)

Abstract:With the rapid development of Internet technology, multimedia data of different view have grown exponentially, and people have been unable to satisfy the original single-modal data retrieval methods such as image retrieval. Cross-modal retrieval has became more and more important in information retrieval field. Aiming at this task, a cross-modal retrieval method for double-branch network structure by increase the attention mechanism of sentence-dependent phrases is proposed. The paper appies the CNN model to extract image features, and obtains the dependency segments of text based on syntactic structure analysis, and designs the original double-branch network structure model which embeds the attention mechanism to learn the weight distribution of each dependent segment, so that the feature representation of the text can be more focused on key sentence segment features. The experimental results show that the proposed method has better performance in the retrieval accuracy evaluation than other methods, and verify the effectiveness of the algorithm.

Key words: dependency phrase; sentence split; attention mechanism; dual branch network