计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190600181

面向跨模态检索的协同注意力网络模型

邓一姣　张凤荔　陈学勤　艾　擎　余苏喆
电子科技大学信息与软件工程学院　成都 610054
(1277325835@qq.com)

摘　要　随着图像、文本、声音、视频等多模态网络数据的急剧增长,人们对多样化的检索需求日益强烈,其中的跨模态检索受到广泛关注。然而,由于其存在异构性差异,在不同的数据模态之间寻找内容相似性仍然具有挑战性。现有方法大都将异构数据通过映射矩阵或深度模型投射到公共子空间,来挖掘成对的关联关系,即图像和文本的全局信息对应关系,而忽略了数据内局部的上下文信息和数据间细粒度的交互信息,无法充分挖掘跨模态关联。为此,文中提出文本-图像协同注意力网络模型(CoAN),通过选择性地关注多模态数据的关键信息部分来增强内容相似性的度量。CoAN 利用预训练的 VGGNet 模型和循环神经网络深层次地提取图像和文本的细粒度特征,利用文本-视觉注意力机制捕捉语言和视觉之间的细微交互作用;同时,该模型分别学习文本和图像的哈希表示,利用哈希方法的低存储特性和计算的高效性来提高检索速度。在实验得出,在两个广泛使用的跨模态数据集上,CoAN 的平均准确率均值(*mAP*)超过所有对比方法,文本检索图像和图像检索文本的 *mAP* 值分别达到 0.807 和 0.769。实验结果说明,CoAN 有助于检测多模态数据的关键信息区域和数据间细粒度的交互信息,充分挖掘跨模态数据的内容相似性,提高检索精度。

关键词:跨模态检索;协同注意力机制;细粒度特征提取;深度哈希;多模态数据
中图法分类号　TP391

Collaborative Attention Network Model for Cross-modal Retrieval

DENG Yi-jiao, ZHANG Feng-li, CHEN Xue-qin, AI Qing and YU Su-zhe
School of Information and Software Engineering, University of Electronic Science and Technology of China, 610054, Chengdu

Abstract　With the rapid growth of image, text, sound, video and other multi-modal network data, the demand for diversified retrieval is increasingly strong. And cross-modal retrieval has been widely concerned. However, there are heterogeneity differences among different modes. It is still a challenging to find the content similarity of heterogeneous data. Most of the existing methods project heterogeneous data into a common subspace by a mapping matrix or a deep model. In this way, a pair of correlation relation is mined, and the global information correspondence relation between image and text is obtained. However, these methods ignore the local context information and the fine-grained interaction information between the data, so the cross-modal correlation cannot be fully mined. Therefore, a text-image collaborative attention network model (CoAN) is proposed. In order to enhance the measurement of content similarity, we selectively focus on key information parts of multi-modal data. The pre-trained VGGNet model and LSTM model are used to extract the fine-grained features of image and text, and the CoAN model is used to capture the subtle interaction between text and image by using text-image attention mechanism. At the same time, this model studies the hash representation of text and image respectively. The retrieval speed is improved by using the low storage and high efficiency of hashing method. Experiments show that, on two widely used cross-modal data sets, the mean Average Precision (mAP) of CoAN model is higher than that of all other comparative methods, and the mAP value of text retrieval image and image retrieval text reaches 0.807 and 0.769. Experimental data show that CoAN model is helpful to detect key information and fine-grained interactive information of multi-modal data, and the retrieval accuracy is improved by fully mining the content similarity of cross-modal data.

Keywords　Cross-modal retrieval, Collaborative attention mechanism, Fine-grained feature extraction, Deep hash, Multi-modal data

---

到稿日期:2019-06-28　　返修日期:2019-10-08　　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金项目(61272527);四川省科技计划项目(2016GZ0063)
This work was supported by the National Natural Science Foundation of China (61272527) and Science and Technology Program of Sichuan Province, China (2016GZ0063).
通信作者:张凤荔(fzhang@uestc.edu.cn)

1 引言

随着多媒体数据的爆炸式增长，传统的单模态检索已不能满足人们在多媒体领域下的检索需求，用户迫切希望能够利用其中一种数据类型作为查询对象来检索具有相似语义的另一种数据类型的内容，例如用文本检索图像，或图像检索文本、视频等，即跨模态检索。跨模态检索技术能有效满足人们对信息检索方式多样化的需要，为用户提供更贴心的服务，如境外旅游以图搜索文字介绍、小学生看图写话等应用。同时，跨模态检索技术促进了机器学习的发展和应用，为新检索技术的发展提供了理论支撑，对跨模态检索方法的研究具有广泛的应用价值和理论意义。

跨模态检索需要同时处理不同模态的数据，这些数据在内容上具有一定的相似性，但是其底层特征是异构的，难以直接计算它们的相似度，即存在“语义鸿沟问题”。异构数据的语义鸿沟问题一直是跨模态检索难以解决的问题[1]。现有的基于深度神经网络的研究大多是为不同的模态构造一个公共子空间，挖掘异构数据的全局对应关系，而忽略了其内部具有语义辨识性的局部细粒度特征；同时，这些方法大都假设不同模态的语义相关数据具有相同的信息量，然而不同模态往往具有不平衡和互补的关系，在描述同一语义时包含的信息量是不等的。

针对不同模态描述同一语义时包含的信息量不等，难以提取局部细粒度相关特征等问题，本文提出一种基于注意机制与深度哈希的文本-图像协同注意网络模型。本文的主要贡献如下：1）利用深度神经网络提取模态特征，深层次的神经网络利于细粒度信息的提取；2）提出文本-图像协同注意网络模型，借助注意机制检测捕捉文本与图像间细微的交互作用，从而学得更好的特征表示；3）学习哈希函数将特征转化为二进制表示，加快了检索速度；4）在两个广泛使用的数据集上将本文模型与 6 种现有方法进行了实验对比，验证了本文模型的有效性和先进性。

2 相关工作

为了满足人们多样化的检索需求，国内外研究人员提出各种方法和模型来探索跨模态数据间的潜在关联。其中，成分分析方法、主题模型方法、深度学习方法、哈希方法等都得到了有效应用。

2.1 跨模态哈希方法

跨模态哈希方法[2]对所有模态数据的特征进行提取和表示，再将这些多模态特征投影到一个共同的海明空间中进行相似性度量。

Kumar 等[3]提出的跨视图哈希算法（Cross-View Hashing，CVH），将谱哈希算法（Spectral Hashing，SH）[4]从传统的单模态扩展到多模态场景，通过最小化相似样本的海明距离而最大化不相似样本的海明距离来学习哈希函数。Ding 等[5]提出了集合矩阵分解哈希算法（Collective Matrix Factorization Hashing，CMFH）。CMFH 假设不同模态的同一个样本生成相同的哈希码。Zhang 等[6]提出语义相关性最大化（Semantic Correlation Maximization，SCM）模型，在不受正交约束的情况下，利用序列学习方法（SCM-Sequential Learning，SCM-Seq）逐位学习哈希函数。为了捕获更复杂的数据结构，Lin 等[7]提出了语义保留哈希（Semantics-Preserving Hashing，SePH），将训练数据的语义亲和力转换为概率分布，通过最小化 KL 散度（Kullback-Leibler Divergence）将其与海明空间中待学习的哈希码进行近似。这些方法大多依赖由浅层架构提取的手工特性，在一定程度上限制了实例的可区分性表征，继而降低了学习哈希码的准确率。

与浅层结构相比，深度学习能更加充分地提取语义特征。Jiang 等[8]提出深度跨模态哈希（Deep Cross-Modal Hashing，DCMH），其把特征和哈希码学习集成到同一个框架，该框架是一个具有深度神经网络的端到端学习框架。Yang 等[9]提出基于成对关系的跨模态检索深度哈希（Pairwise Relationship Guided Deep Hashing，PRDH）方法，采用深度卷积神经网络模型，在端到端体系结构中同时学习每种模态的特征表示和哈希码。

2.2 Attention 机制

视觉注意机制被广泛应用于图像分类[10-11]、图像生成[12]、图像字幕[13]、视觉问答[14-15]等场景。同时，文本注意机制已成功应用于机器翻译[16]、文本生成[17]、句子摘要[18]、答疑[19-20]等任务。

近年来，学者们提出基于注意机制的方法来解决图文匹配的问题。Huang 等[21]提出选择性多模态长短记忆网络（Selective Multimodal-LSTM，sm-LSTM），该方法是一种上下文调节的注意方案，可以选择性地注意图像-文本对中的关键区域。Nam 等[22]提出双注意网络（Dual Attention Networks，DAN），通过多个步骤捕捉视觉和语言之间的精细交互作用。Zhang 等[23]提出一种具有注意机制的对抗哈希网络，通过选择性地关注多模态数据的信息性部分来增强内容相似性的度量。上述方法表明，注意机制在跨模态检索领域的应用研究仍有较大的发展空间。

3 跨的模态检索问题定义

模态是指数据的存在形式，如文本、音频、图像、视频等文件格式。有些数据的存在形式不同，但都是描述同一物体或事件。跨模态检索根据不同模态间丰富的内部联系，给定一种模态数据，检索与之语义相似的另一模态数据。

跨模态检索问题的目标是学习一种跨模态的相似性度量 Sim，对于给定的查询词 $\boldsymbol{x}_q \in \boldsymbol{X}$，返回最相似的另一模态的样本，如式（1）所示：

$$\boldsymbol{y} = \min_i Sim(\boldsymbol{x}_q - \boldsymbol{y}_i) \tag{1}$$

假设有 n 个训练样本，每个样本以音频、视频、图像、文本等多种形式表示，本文仅考虑文本与图像两种模态的跨模态检索问题。给定图片数据集 $\boldsymbol{X}$、文本数据集 $\boldsymbol{Y}$ 和其对应的标签集 $\boldsymbol{C}$，分别提取模态特征，学习哈希映射函数 $H(\boldsymbol{X})$ 和 $H(\boldsymbol{Y})$，将图像和文本转换至一个共同的二进制码空间，在该空间中利用相似性 Sim 度量图像和文本数据的相似性。

本文用到的所有符号和相关解释如表 1 所列。

表 1 符号及解释

Table 1 Symbols and interpretation

符号	解释
$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]\in R^{d_1*n}$	图片集 $\boldsymbol{X}$：n 个样本，每个样本的特征维度为 d_1
$\boldsymbol{Y}=[\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n]\in R^{d_2*n}$	文本集 $\boldsymbol{Y}$：n 个样本，每个样本的特征维度为 d_2
$\boldsymbol{C}=[\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_n]\in R^{1*n}$	类别标签集 $\boldsymbol{C}$：图片集 $\boldsymbol{X}$ 中第 i 个样本与文本集 $\boldsymbol{Y}$ 中第 i 个样本是一个样本对$\{\boldsymbol{x}_i,\boldsymbol{y}_i\}$，同属于一个类别 $\boldsymbol{c}_i$，其中 $i=1,2,\cdots,n$
$\boldsymbol{V}^I=\{v_1^I,v_2^I,\cdots,v_i^I,\cdots,v_n^I\}$	输入图像的特征表示，v_i^I 表示第 i 个图像区域的特征向量
$\boldsymbol{V}^T=\{v_1^T,v_2^T,\cdots,v_i^T,\cdots,v_t^T\}$	文本的语义特征表示，v_i^T 表示第 i 个单词在上下文中的语义特征
$\boldsymbol{q}_{\text{img}},\boldsymbol{q}_{\text{text}}$	图像、文本查询向量
w_i^k,b_i^k,w_t^k,b_t^k	第 k 层的网络参数
$v_{\text{img}}^k,v_{\text{text}}^k$	第 k 层带注意的图像、文本特征表示向量
$\boldsymbol{h}_i^k,\boldsymbol{h}_t^k$	第 k 层图像、文本隐藏层表示向量
Sim_i^k,Sim_t^k	第 k 层图像、文本与查询向量的相似性
α_i^k,α_t^k，	第 k 层图像、文本注意力权重
$I\to I,T\to T,I\to T,T\to I$	文本检索文本、图像检索图像、图像检索文本、文本检索图像
$F_{A\to B}$	A 检索 B 的损失函数：$A\to B$ 表示用 A 模态检索 B 模态的相关数据，A 表示待查询数据，B 表示检索数据库，$A,B\{I,T\}$
$H\{0,+1\}^c$	哈希函数

4 面向跨模态检索的协同注意力网络模型

4.1 模型流程

本文提出的跨模态检索模型由特征学习与表示、注意交互、哈希学习 3 个模块构成，总体流程如图 1 所示。

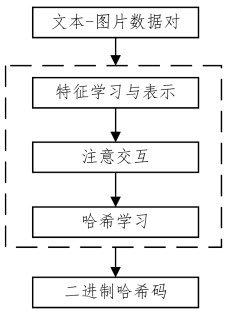


图 1 跨模态检索模型的流程

Fig. 1 Flowchart of cross-modal retrieval model

(1)基于深度学习强大的特征学习能力，本文在特征学习模块采用深度卷积神经网络(Convolutional Neural Network，CNN)提取图像特征，采用长短期记忆网络(Long-Short Term Memory，LSTM)提取文本特征。

(2)注意力感知能够检测多模态数据的关键信息区域，有助于识别不同模态数据之间的内容相似性。注意交互模块利用注意机制提取不同模态数据的交互特征和底层细节特征。

(3)哈希学习模块通过学习哈希函数将特征转化为二进制表示，保留模态内数据结构的相似性和模态间的语义相似性，提高了检索速度。

4.2 特征学习与表示

对于图像模态，本文利用 CNN 提取特征。在卷积神经网络中，利用卷积和池化操作对图像进行特征提取，并将提取的特征矩阵输入到全连接层或全局均值池化层，从而产生图像的特征向量[24]。具体来说，本文采用 ImageNet 数据集[25]上预训练的 VGGNet 模型[26]提取图像特征。VGGNet 模型探索了网络的深度与性能之间的关系，通过构建含有多个卷积子层的卷积层来实现网络深度的拓展。它反复堆叠 3 * 3 的小型卷积核和 2 * 2 的最大池化层，用较小的卷积核和多个卷积层实现了对图片特征的精细抓取。VGGNet 的结构使其能深度提取图像中精细的语义特征，以获得更好的图像模态表示。如图 2 所示，该网络模型有 5 个卷积层 *Conv*1—*Conv*5、2 个全连接层 *Fc*6 和 *Fc*7(用于图像特征表示)、1 个分类特征层 *Fc*8。

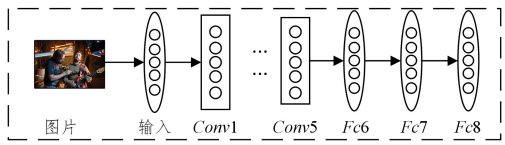


图 2 图像特征提取

Fig. 2 Image feature extraction

输入图像的大小为 $N\times N$，为了得到不同区域的特征向量，本文选取 16 层的 VGGNet 的最后一个池化层来提取图像特征。最终输入图像的特征表示为 $\boldsymbol{V}^I=\{\boldsymbol{v}_1^I,\boldsymbol{v}_2^I,\cdots,\boldsymbol{v}_i^I,\cdots,\boldsymbol{v}_n^I\}$，其中，$n$ 表示图像区域的数量，$\boldsymbol{v}_i^I$ 表示第 i 个图像区域的特征向量。

对于文本模态，本文采用 LSTM 和 CNN 提取特征，如图 3 所示。文本特征提取的流程如下：

(1) 从文本集中有放回随机采样选取 n 个样本；

(2) 通过词嵌入将文本中的单词转化为词向量；

(3) 将词向量输入 LSTM 网络模型提取语义特征；

(4) 将 LSTM 的输出作为 CNN 的输入，获得文本的语义特征，表示为 $\boldsymbol{V}^{\mathrm{T}}=\{v_1^{\mathrm{T}},v_2^{\mathrm{T}},\cdots,v_i^{\mathrm{T}},\cdots,v_t^{\mathrm{T}}\}$，其中，$t$ 表示单词的数量，vv_i^{T} 表示第 i 个单词在上下文中的语义特征。

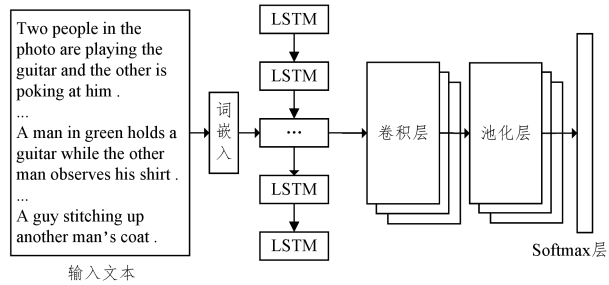


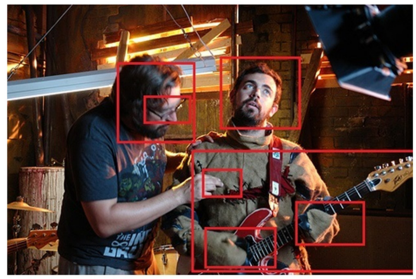
图 3 文本特征提取

Fig. 3 Text feature extraction

4.3 文本-图像协同注意网络

注意机制允许模型在任务的每个步骤中关注视觉或文本输入的关键信息部分。视觉注意机制有选择地关注图像的子区域，从而提取图像的核心特征，减少待处理的信息量；文本注意机制通常在编码器-解码器框架下寻找语义或语法的输入-输出对齐，处理长期依赖关系时有显著的效果。图像和文本通过注意机制相互指导学习，得到细粒度的关键特征，如图4中方框区域所示。

为了获得文本和图像之间细微的交互作用，本文提出文本-图像协同注意网络（Co-Attention Network，CoAN），其总体架构如图5所示。首先将文本特征向量 $\boldsymbol{V}^T$ 和图像特征图 $\boldsymbol{V}^I$ 输入网络，随机初始化图像查询向量 $\boldsymbol{q}_{\text{img}}$ 和文本查询向量 $\boldsymbol{q}_{\text{text}}$；然后使用图像特征来查询文本注意网络的关键内容，同时利用文本特征来查询图像注意网络的关键内容，文本和图像经过协同注意网络数次的交互作用后，可得到文本和图像相互指导的有注意的关键特征；最后分别将有注意的特征表示输入哈希层，通过学习哈希函数得到特征的二进制表示。文本-图像协同注意网络有效缩减了异构模态数据的语义鸿沟，从而更容易比较文本和图像的内容相似性。



A man in green holds a guitar while the other man observes his shirt .

图4 文本-图像相互注意的关键特征

Fig. 4 Text-image collaborative attention key features

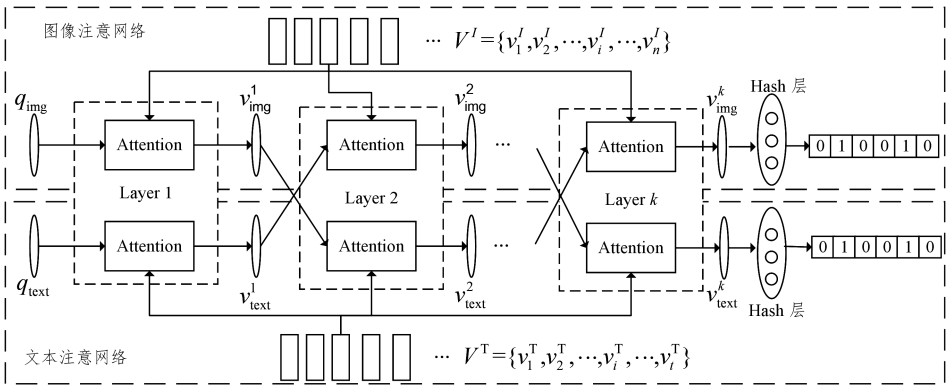


图5 文本-图像协同注意网络架构

Fig. 5 Text-image collaboration attention network architecture

在视觉注意网络的第1层注意网络单元，网络的输入为图像特征映射 v_i^I 和图像查询向量 $\boldsymbol{q}_{\text{img}}$，激活函数为 $Tanh$ 函数，使用单层感知机获得图像的隐藏层表示 $\boldsymbol{h}_i^1$：

$$\boldsymbol{h}_i^1=Tanh(w_i^1\boldsymbol{v}_i^I+b_i^1) \tag{2}$$

计算图像隐藏层表示向量 $\boldsymbol{h}_i^1$ 与图像查询向量 $\boldsymbol{q}_{\text{img}}$ 的Cosine相似性：

$$Sim_i^1=\frac{\boldsymbol{h}_i^1\boldsymbol{q}_{img}}{\|\boldsymbol{h}_i^1\|\cdot\|\boldsymbol{q}_{\text{img}}\|} \tag{3}$$

使用 $Softmax$ 函数对相似性得分进行归一化处理，得到视觉注意权重系数 α_i^1；再对图像特征映射进行注意加权求和，得到带注意的图像特征表示向量 $\boldsymbol{v}_{\text{img}}^1$。

$$\alpha_i^1=Softmax(Sim_i^1) \tag{4}$$

$$\boldsymbol{v}_{\text{img}}^1=\sum_{i=1}^{L}\alpha_i^1\boldsymbol{v}_i^I \tag{5}$$

本文使用文本特征寻找图像的关键特征映射（Text-Guided-Image，TGI）。首先，连接带注意的文本特征表示向量 $\boldsymbol{v}_{\text{text}}^1$ 和图像特征映射 $\boldsymbol{v}_i^I$，将 $[\boldsymbol{v}_i^I,\boldsymbol{v}_{\text{text}}^1]$ 作为单层感知机的输入，获得图像的隐藏表示 $\boldsymbol{h}_i^2$。其次，计算注意文本特征表示向量 $\boldsymbol{v}_{\text{text}}^1$ 和图像的隐藏表示向量的相似性，使用 $Softmax$ 函数对其进行归一化处理，最后对图像特征映射进行注意加权求和，得到带注意的图像特征表示向量。

$$\boldsymbol{h}_i^2=Tanh(w_i^2[\boldsymbol{v}_i^I,\boldsymbol{v}_{\text{text}}^1]+b_i^2) \tag{6}$$

$$\boldsymbol{v}_{img}^2=\sum_{i=1}^{L}Softmax\left(\frac{\boldsymbol{h}_i^2\boldsymbol{v}_{\text{text}}^1}{\|\boldsymbol{h}_i^2\|\cdot\|\boldsymbol{v}_{\text{text}}^1\|}\right)\boldsymbol{v}_i^I \tag{7}$$

在第 k 层注意网络单元，合并图像特征映射 $\boldsymbol{v}_i^I$ 和第 $(k-1)$ 层带注意的文本特征表示向量 $\boldsymbol{v}_{\text{img}}^{k-1}$，得到带注意的图像特征表示向量 $\boldsymbol{v}_{\text{img}}^k$。

$$\boldsymbol{v}_{\text{img}}^k=TGI([\boldsymbol{v}_i^I,\boldsymbol{v}_{\text{text}}^{k-1}]) \tag{8}$$

同时，本文使用图像特征寻找文本的关键特征（Image-Guided-Text，IGT）。同理，得到带注意的文本特征表示向量 $\boldsymbol{v}_{\text{text}}^k$ 为：

$$\boldsymbol{v}_{\text{text}}^k=IGT([\boldsymbol{v}_i^{\text{T}},\boldsymbol{v}_{\text{img}}^{k-1}]) \tag{9}$$

在哈希层中，$Tanh$ 激活函数使得每个神经元的输出在-1到1之间，阈值为0的 $Sign$ 函数再将其转换成二进制编码。编码值为1，代表神经元的输出大于或等于0；编码值为0，代表输出小于0。图像和文本的哈希函数如式(10)所示：

$$H^{I/T}=Sign(Tanh(\boldsymbol{v}_{\text{img/text}}^k w^{I/T}+b^{I/T})) \tag{10}$$

4.4 目标函数

跨模态检索损失函数的目标是既保留模态内的相似性，又保留异构模态间的语义相似性。跨模态检索损失函数如式(11)所示：

$$F=\min(F_{I\to I}+F_{T\to T}+F_{I\to T}+F_{T\to I}) \tag{11}$$

A 检索 B 的损失函数的定义为：

$$F_{A\to B}=\sum_{(i,j,k)}\max\{0,\varepsilon+\|H_i^A-H_j^B\|-\|H_i^A-H_k^B\|\} \tag{12}$$

其中，(i,j,k)为三元组，表示最小边距。$\| H_i^A - H_j^B \|$表示当前查询模态与正样本的欧氏距离，$\| H_i^A - H_k^B \|$表示当前模态与负样本的欧氏距离。假设 A 表示图像，B 表示文本，$F_{A\to B}$是三元组排序损失[27]，表示图像 i 与文本 j 的相似性大于图像 i 与文本 k 的相似性。

本文跨模态检索算法的流程如算法 1 所示。

算法 1 跨模态检索算法

输入：图片集 $\mathbf{X}$，文本集 $\mathbf{Y}$

输出：图片、文本的最优哈希编码

1. 初始化参数：均值 $\mu=0$，标准差 $\delta=0.01$，批量值 batch=64，总 epoch=60，学习率 $\eta=0.05$，随机初始化图像查询向量 $\mathbf{q}_{\text{img}}$ 和文本查询向量 $\mathbf{q}_{\text{text}}$；
2. 从图片集 $\mathbf{X}$ 和文本集 $\mathbf{Y}$ 中随机选择一对样本$\{\mathbf{x}_i,\mathbf{y}_i\}$，通过特征学习与表示分别获取图像、文本的特征表示 $\mathbf{V}^{\mathrm{I}}$ 和 $\mathbf{V}^{\mathrm{T}}$；
3. for i=1 to k do
   3.1. 分别计算图像、文本特征与查询向量的相似性 $\mathrm{Sim}_{\mathrm{i}}^{\mathrm{k}}$ 和 $\mathrm{Sim}_{\mathrm{t}}^{\mathrm{k}}$；
   3.2. 更新图像、文本的注意力权重 $\alpha_{\mathrm{i}}^{\mathrm{k}}$ 和 $\alpha_{\mathrm{t}}^{\mathrm{k}}$；
   3.3. 计算带注意的图像、文本的特征表示向量 $\mathrm{v}_{\mathrm{img}}^{\mathrm{k}}$ 和 $\mathrm{v}_{\mathrm{text}}^{\mathrm{k}}$；
4. 计算图像、文本的哈希值 $\mathrm{H}^{\mathrm{I/T}}$。

5 实验结果与分析

5.1 数据集与评价指标

(1)MIR-Flickr25K 数据集[28]

MIR-Flickr25K 数据集由 Flickr 网站收集的 25015 张图片组成。每张图像都与几个文本标记相关联，这些文字描述来自用户上传图片时对其添加的单词注释，平均每张图片约有 9 个注释。每张图像及其对应的多个文字注释构成了一个图像-文本对。该数据集共包含 24 个人工标注的类别标签，本文选取注释单词出现次数在整个数据集上不少于 20 的图像-文本对作为实验数据。

(2)NUS-WIDE 数据集[29]

NUS-WIDE 数据集是一个 Web 图像数据集，包含图像及相关标签，共 269648 幅图像。每张图像与 81 个语义标记相关联。本文评估了 21 个最常用标签的 195834 个图像-文本对的性能。

(3)评价指标

本文采用平均准确率均值（mean Average Precision，mAP）[30]作为评价指标。给定一个查询 $\boldsymbol{q}$，mAP 的计算如式(13)所示：

$$mAP=\frac{1}{|\boldsymbol{Q}|}\sum_{i=1}^{|Q|}AP(\boldsymbol{q}_i) \tag{13}$$

其中，$|\boldsymbol{Q}|$代表查询数据集 $\boldsymbol{Q}$ 的大小，AP 代表均值准确度(Average Precision)。

$$AP(\boldsymbol{q})=\frac{1}{L}\sum_{r=1}^{n}P_q(r)\delta(r) \tag{14}$$

其中，L 代表查询点 $\boldsymbol{q}$ 在数据中的真实近邻数；n 代表数据总量；$P_q(r)$代表前 r 个被检索到的实体的精度；$\delta(r)$为指示函数，$\delta(r)=1$ 表示第 r 个文档与被检索文档是相关的，$\delta(r)=0$ 则表示其不相关。

5.2 结果对比与分析

实验比较了几种基于浅层结构的方法（CVH[3]，CMFH[5]，SCM[6]，SePH[7]）和两种基于深层结构的方法（DCMH[8]和 PRDH[9]）。公平起见，对于图像模态，使用在 ImageNet 数据集上预训练的 VGGNet-16 网络模型提取所有基于浅层结构的深层特征；对于文本模态，所有参数都用均值为 0、标准差为 0.01 的高斯函数随机初始化。本文通过随机梯度下降来训练网络，batch 值为 64，总 epoch 为 60，学习率为 0.05，每 20 次迭代后学习率变为当前值的 1/10。分别设置文本-图像协同注意网络的层数 k 为 1～6，实验表明当 k 为 4 时模型表现最好。

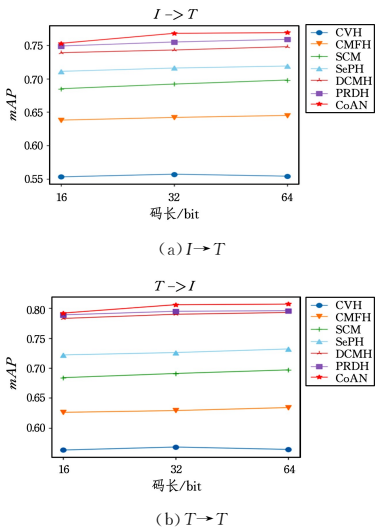
实验主要实现了编码长度为 16bit，32bit 和 64bit 时利用各种模型方法进行图文相互检索的 mAP 值，如表 2 所列。可以看出，基于深层结构（DCMH，PRDH，CoAN）的实验结果明显优于浅层方法（CVH，CMFH，SCM，SePH）的结果，这是因为深度学习具有强大的特征学习能力，可以更有效地捕捉不同模态数据间的非线性相关性，进而提取到更细粒度的特征。从整体来看，NUS-WIDE 数据集实例较多，内容复杂，比较具有挑战性，各方法在其上的检索效果与在 MIR-Flickr25K 数据集上的检索效果仍有较大差距。

表 2 本文算法与其他跨模态检索算法的 mAP 对比

Table 2 mAP comparison of different methods

任务	方法	MIR-Flickr25K			NUS-WIDE		
		16 bit	32 bit	64 bit	16 bit	32 bit	64 bit
I→T	CVH[3]	0.553	0.557	0.554	0.425	0.417	0.421
	CMFH[5]	0.638	0.642	0.645	0.490	0.505	0.509
	SCM[6]	0.685	0.692	0.698	0.521	0.529	0.535
	SePH[7]	0.711	0.716	0.719	0.564	0.569	0.571
	DCMH[8]	0.739	0.743	0.748	0.590	0.603	0.609
	PRDH[9]	0.749	0.755	0.759	0.611	0.629	0.627
	CoAN	**0.753**	**0.768**	**0.769**	**0.638**	**0.635**	**0.645**
T→I	CVH	0.563	0.568	0.564	0.455	0.467	0.463
	CMFH	0.626	0.629	0.634	0.503	0.519	0.522
	SCM	0.684	0.691	0.697	0.534	0.541	0.548
	SePH	0.722	0.726	0.732	0.598	0.603	0.611
	DCMH	0.783	0.790	0.793	0.639	0.651	0.657
	PRDH	0.789	0.795	0.796	0.653	0.692	0.672
	CoAN	**0.792**	**0.806**	**0.807**	**0.679**	**0.698**	**0.704**

MIR-Flickr25K 数据集上的实验结果如图 6 所示，可以看出 CoAN 的 mAP 值超过其余 6 种方法。CoAN 联合利用视觉-文本注意机制捕捉视觉和语言之间的细微交互作用，能更准确地捕捉两种模式之间的相关性，其性能始终优于其他方法。



(a) $I\to T$

(b) $T\to T$

图 6 MIR-Flickr25K 数据集上的 mAP 对比

Fig. 6 mAP comparison on MIR-Flickr25K dataset

结束语 本文利用深度学习在特征学习方面的优越性，提出了基于注意机制与深度哈希的跨模态检索模型，共同学习视觉和文本注意模型，通过多次交互得到图像和文本的关键特征。实验证明，本文提出的模型能够有效解决现有跨模态检索算法粒度粗、精度低等问题，为跨模态检索新技术提供了参考。然而，本文局限于图像与文本两个模态之间的检索问题，未来将会把方法扩展应用到音频、视频等较为复杂的多媒体数据上。

参 考 文 献

[1] OU W H, LIU B, ZHOU Y H, et al. Research review of cross-modal retrieval [J]. Journal of Guizhou normal university: natural science edition, 2018, 36(2): 114-120.

[2] FAN H, CHEN H H. Research progress of cross-modal retrieval based on hash method [J]. Data communication, 2018, 184(3): 43-49.

[3] KUMAR S, UDUPA R. Learning Hash Functions for Cross-View Similarity Search [C] // Proceedings International Joint Conference on Artificial Intelligence. 2011: 1360-1365.

[4] WEISS Y, TORRALBA A, FERGUS R. Spectral hashing[C] // International Conference on Neural Information Processing Systems. 2008.

[5] DING G, GUO Y, ZHOU J. Collective Matrix Factorization Hashing for Multimodal Data[C] // 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2014.

[6] ZHANG D, LI W J. Large-scale supervised multimodal hashing with semantic correlation maximization [C] // Twenty-eighth AAAI Conference on Artificial Intelligence. AAAI Press, 2014.

[7] LIN Z, DING G, HU M, et al. Semantics-preserving hashing for cross-view retrieval[C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015.

[8] JIANG Q Y, LI W J. Deep Cross-Modal Hashing[C] // IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2017.

[9] YANG E, DENG C, LIU W, et al. Pairwise Relationship Guided Deep Hashing for Cross-Modal Retrieval [C] // Thirty-First AAAI Conference on Artificial Intelligence. AAAI, 2017.

[10] MNIH V, HEESS N, GRAVES A, et al. Recurrent Models of Visual Attention[J]. arXiv: 1406.6247, 2014.

[11] STOLLENGA M, MASCI J, GOMEZ F, et al. Deep Networks with Internal Selective Attention through Feedback Connections [J]. Advances in Neural Information Processing Systems, 2014, 4(2): 3545-3553.

[12] GREGOR K, DANIHELKA I, GRAVES A, et al. DRAW: A Recurrent Neural Network For Image Generation [J]. arXiv: 1502.04623, 2015.

[13] XU K, BA J, KIROS R, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention[J]. arXiv: 1502.03044, 2015.

[14] YANG Z, HE X, GAO J, et al. Stacked Attention Networks for Image Question Answering[J]. arXiv: 1511.02274, 2015.

[15] SHIH K J, SINGH S, HOIEM D. Where To Look: Focus Regions for Visual Question Answering[J]. arXiv: 1511.07394, 2015.

[16] BAHDANAU D, CHO K, BENGIO Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. arXiv: 1409.0473, 2014.

[17] LI J W, LUONG M T, JURAFSKY D. A hierarchical neural autoencoder for paragraphs and documents [J]. arXiv: 1506.01057, 2015.

[18] RUSH A M, CHOPRA S, WESTON J. A Neural Attention Model for Abstractive Sentence Summarization[J]. arXiv: 1509.00685, 2015.

[19] KUMAR A, IRSOY O, SU J, et al. Ask Me Anything: Dynamic Memory Networks for Natural Language Processing[J]. arXiv: 1506.07285, 2015.

[20] XIONG C, MERITY S, SOCHER R. Dynamic Memory Networks for Visual and Textual Question Answering[J]. arXiv: 1603.01417, 2016.

[21] HUANG Y, WANG W, WANG L. Instance-aware Image and Sentence Matching with Selective Multimodal LSTM[J]. arXiv: 1611.05588, 2016.

[22] NAM H, HA J W, KIM J. Dual Attention Networks for Multimodal Reasoning and Matching[J]. arXiv: 1611.00471, 2016.

[23] ZHANG X, LAI H, FENG J. Attention-Aware Deep Adversarial Hashing for Cross-Modal Retrieval[M] // Computer Vision – ECCV 2018. Cham: Springer, 2018.

[24] LIU J W, DING X H, LUO X L. Review of multimodal deep learning [J]. Computer Application Research, 2019, 37(6).

[25] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet Large Scale Visual Recognition Challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[26] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. arXiv: 1409.1556, 2014.

[27] LAI H, PAN Y, LIU Y, et al. Simultaneous feature learning and hash coding with deep neural networks[J]. arXiv: 1504.03410, 2015.

[28] HUISKES M J, THOMEE B, LEW M S. New trends and ideas in visual concept detection the MIR Flickr retrieval evaluation initiative[C] // International Conference on Multimedia Information Retrieval. ACM, 2010.

[29] CHUA T S, TANG J, HONG R, et al. Nus-wide: a real-world web image database from national university of Singapore[C] // International Conference on Multimedia Information Retrieval. ACM, 2009.

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

DENG Yi-jiao, born in 1995, postgraduate, is a member of China Computer Federation. Her main research interests include machine learning and data mining.

ZHANG Feng-li, born in 1963, Ph. D, professor, Ph.D supervisor, is a member of China Computer Federation. Her main research interests include network security and network engineering, cloud computing and big data and machine learning.