情报学报　2018 年 4 月　第 37 卷　第 4 期
Journal of the China Society for Scientific and Technical Information, April 2018, 37(4): 422-435

DOI: 10.3772/j.issn.1000-0135.2018.04.008

基于表示学习的跨模态检索模型与特征抽取研究综述

李志义，黄子风，许晓绵

（华南师范大学经济与管理学院，广州　510006）

摘　要　以深度学习为代表的表示学习在语音识别、图像分析和自然语言处理领域获得了广泛关注与应用，它不仅推动了人工智能的深入研究和快速发展，而且促使企业思索新的运营与盈利模式。本文拟通过综述的形式对这些研究进行梳理，形成较为完整的综述。通过对国内外相关文献的调查和整理，从信息抽取与表示、跨模态系统建模两维度评述了基于表示学习的跨模态检索与特征抽取方面的研究成果。文章首先概括了自动编码器、稀疏编码、限制玻尔兹曼机、深度信念网络、卷积神经网络等五个经典的表示学习算法，然后从基于共享层建立各模态间的关联、表示空间中各模态间的关联、以深度学习为基础的跨模态建模算法等三方面归纳跨模态系统建模研究的现状，最后总结了跨模态检索的评价指标。研究发现：已有检索研究对于单模态信息检索较为丰富，查询和候选集的内容均属于同一模态；跨模态检索也仅限于对图像、文本两个模态对齐的语料。未来需要增加语音、视频、图像、文本等多模态数据的检索，改进深度学习算法构建多模态检索模型，实现三种或以上的跨模态检索。此外，尚需建立适合多模态检索系统的评价指标。

关键词　表示学习；跨模态检索；特征抽取；模型；综述

A Review of the Cross-Modal Retrieval Model and Feature Extraction Based on Representation Learning

Li Zhiyi, Huang Zifeng and Xu Xiaomian

(*Economic & Management College of South China Normal University, Guangzhou*　510006)

Abstract: Representation learning, particularly deep learning, has received wide attention and seen application in speech recognition, image analysis, and natural language processing fields. It not only promotes the research and development of artificial intelligence, but urges enterprises to consider new business and profit models. This paper aims to examine these studies in the form of reviews, and ultimately form a complete overview of the topic. Through the investigation and organization of relevant literature locally and internationally, this paper summarizes the research results of cross-modal retrieval and feature extraction based on representation learning from the two dimensions of information extraction and representation, and cross-modal system modeling. The main research includes summarizing five traditional representation learning algorithms, which are the autoencoder, sparse encoding, the restricted Boltzmann machine, deep belief networks, and convolutional neural networks. From the shared layer relationship between each mode, the representation space, and the correlation between each mode's in-depth learning-based

收稿日期：2017-12-03；修回日期：2018-04-11

基金项目：国家社会科学基金项目“基于表示学习的跨模态检索模型与特征抽取研究”（17BTQ062）。

作者简介：李志义，男，1968 年生，副教授，硕士生导师，主要研究方向为 Web 挖掘，E-mail: leeds@scnu.edu.cn；黄子风，男，1993 年生，在读硕士研究生，主要研究方向为信息抽取与观点挖掘；许晓绵，女，1992 年生，在读硕士研究生，主要研究方向为信息抽取与观点挖掘。

cross-modal modeling algorithm, the present state of research on modeling systems based on cross- modal modeling is summed up. Finally, the evaluation index of cross-modal retrieval is summarized. The study finds that the existing retrieval research is rich in single-modal information retrieval and that the content of queries and candidate sets belong to the same modality, whereas cross-modal retrieval is limited to two modal alignment languages of images and texts. Future research needs to see an increase of modal retrieval of audio, video, images, text, and other multimodal data, and using deeper constructing multimodal retrieval models and feature extraction algorithms to achieve three-or-greater cross-modal retrieval. In addition, an evaluation index of multimodal retrieval systems must be established.
Key words: representation learning; cross modal retrieval; feature extraction; model; review

1　引　言

表示学习（Representation Learning）旨在将研究对象的语义信息表示为稠密低维实值向量，以在低维空间中高效计算实体和关系的语义联系，并有效解决数据稀疏问题，使知识获取、融合和推理的性能显著提升。由于这种低维度的表示向量能够有效地显示出词语之间的语义关系，且更易于被应用到其他的系统当中，目前表示学习的距离模型、单层神经网络模型、双线性模型、矩阵分解和翻译等模型已广泛应用在信息抽取、知识库自动问答系统，以及多媒体信息资源的处理与识别中。特别地，以深度学习为代表的表示学习技术在音频识别、图像分析和自然语言处理领域获得了更大关注，特征表示学习（Feature Representation Learning）逐步成为机器学习的一个新兴分支。

另一方面，目前图像、文本检索的研究主要集中在单模态检索上，查询和候选集的内容均属于同一模态。而跨模态检索（Cross-Modal Retrieval）依然局限于对两个模态对齐的语料，尤其以图像、文本模态对齐的语料为多见。未来研究的趋势将结合现有的卷积神经网络（Convolutional Neural Networks，CNN）、连续词袋模型（Continuous-Bag-Of-Words，CBOW）等在训练图像样本的优点以及在提取文本特征的优越性，设计图像、文字特征统一表示的新模型——二值特征向量（Binary Codes Feature Vector，BCFV），并构建基于表示学习的文本、图像、语音等多模态检索框架和检索系统。

从文献调研看，利用 CNKI、Web of Knowledge 数据库对 2006—2017 年国内外关于表示学习和跨模态检索领域的研究文献进行检索并统计，在表示学习研究方面，国外相关文献量在 2013 年之前以 200~300 篇的速度持续增长，自 2014 年起呈爆发式增长，并于 2017 年达到 6536 篇的高峰，大量的期刊论文研究领域涵盖了各大行业；而近年来国内表示学习相关研究才逐渐兴起，以深度学习为主题的博硕士论文开始涌现。在跨模态检索研究方面，近 10 年国外跨模态相关研究文献也呈逐渐上升趋势，逐渐出现大量与深度学习相结合的研究文献；国内相关文献上升趋势虽不及国外明显，且大部分研究仍处于探索与介绍阶段，但其研究深度也在不断深入。总体上，国内外跨模态相关文献量随时间变化的趋势略同，并呈螺旋式增长趋势。

2　现存的主要问题与不足

基于表示学习的跨模态检索系统建模主要解决两个问题[1-2]：其一，如何完成不同模态信息特征的统一映射？其二，如何在提高检索模型召回率的基础上保证检索速率？这两个问题之间是相互依赖的，由于不同模态信息具有多样性和异构性，各模态的特征抽取方法和统一表示形式便成为解决问题的关键。目前国内外相关研究仍处于起步阶段，存在的问题和不足主要表现在以下方面。

2.1　信息抽取与表示方法方面

目前表示学习模型在自动抽取特征时所得到的维度均相对较高，尤其以深度学习为基础的跨模态检索模型，在表示阶段所得到的样本特征维度通常不少于 4096，最终特征维度仍然偏高。其次，直接使用 PCA（Principal Component Analysis）等降维的手段，虽然在一定程度上能约减特征维度，但在保持必要的检索精度前提下，能够降低的维度相当有限，且缺乏高效合理、能适应大规模图像集的检索机制。如张昭旭[3]使用 AlexNet 倒数第二层提取图像的本质特征，发现 CNN 用于一般自然条件下的人脸表情识别具有相当的优势，但对高维非线性数据效果不佳；孙志军等[4]为弥补 PCA 在高维特征空间进行线性特征提取方面的不足，在预训练阶段进行无监督正则化，提出了基于深度学习的边际 Fisher 分析特征提取算法 DMFA（Deep Marginal Fisher Analysis），但其特征提取算法性能高度依赖于样本

集的规模，而实际情况下对非合作目标训练样本的获取难度较大。如何高效设定参数范围仍需进一步探究。

2.2 跨模态检索建模方面

对于跨模态检索建模方面，整体上国内外仍处于对内容的抽取及独立词语标注的阶段，跨模态检索依然仅实现图像内容与主题词语的匹配，忽略了大量基于内容的、细微且重要的图像信息。针对此类情况，王剑[1]采用CNN提取图像模态的语义特征，采用词向量的方式表示文本，用一维卷积神经网络从词向量表示中提取文本模态的语义特征，这在一定程度上弥补了基于图像内容信息的提取缺失，但在跨模态检索过程中面对大型数据集仍需较长的检索响应时间。为解决大数据处理的速率优化问题，何泳澔[2]提出了一种无需降维、利用多种图像特征快速求解的图像-标签关联学习的自动图像标注方法。该方法能够在处理大规模标注问题的同时实现在线学习过程，部分解决了目前海量图像数据结构化问题。此外，面对大数据多样的数据形式和丰富的数据内容，深入挖掘数据内在联系也变得越发困难，同时跨模态表示学习的方法和框架缺乏规范和标准，影响了数据集的组织、使用和质量。2004年起，国内外对跨模态检索的模型构建研究进入了快速发展阶段，国内外学者在模型的改进方面进行了大量研究，既有基于内容的模型[5]，也有基于关联学习的模型以及基于神经网络的模型[6-7]。然而，各种模型均有其特定的适应对象、优势与局限，如何在实际应用中结合模型与各种算法的优势，构建普适性的跨模态检索模型，这是目前跨模态检索研究中亟待解决的问题之一。

除此之外，三个模态及以上对齐的语料研究较少，两个模态对齐的语料较为多见，尤其是图像、文本模态对齐的语料较为常见。国内外的研究工作也大多集中在对两个模态的语料研究上，并发布了很多公开语料。

3 国内外研究现状

3.1 表示学习与特征抽取研究进展

近年来，不断有学者对表示学习在自然语言和多媒体处理领域的应用进行研究与突破，尤其在自然语言处理领域，涌现出了大量关于单词[8]、短语[9-10]、句子[11-13]、文档[10]、社会网络[14-15]等不同对象的表示学习研究。表示学习能够充分利用自然语言对象间的语义相似度信息，弥补短文本在有效表示、数据稀疏干扰等方面的不足。在多媒体处理领域，尤其以深度学习为代表，在图像与视频分析[16]、计算机视觉[17]、语音识别、多媒体等[18]诸多领域取得了巨大成功。从模型层次的角度可将表示学习分为浅层特征学习（Shallow Learning）和深度特征学习（Deep Learning）两个阶段。

1986年，David[19]提出了经典的浅层特征学习算法——神经网络反向传播算法（Error Back-Propagation, BP算法），成为表示学习研究正式进入浅层特征学习阶段的标志。它利用人工神经网络将大量训练数据中的特征以统计学的方法进行获取与预测，更适合于学习和存储大量输入-输出模式的映射关系。BP神经网络模型虽被称作多层感知器，但实际上只支持1个隐藏层（Hidden Layer），其拓扑结构包括输入层（Input Layer）、隐藏层(Hidden Layer)和输出层（Output Layer）。

BP神经网络模型的输入层与隐藏层、隐藏层与输出层的节点间均为全连接，其主要目的是反复修正权值和阈值，使得误差函数值达到最小。根据Widrow-Hoff学习规则，通过改变神经元之间的连接权值来减少系统实际输出和期望输出的误差。设输出层输出的所有结果 $\sum y_j$ 为 d_j，则误差 E 可表示为：

$$E(w,b)=\frac{1}{2}\sum_{j=0}^{n-1}(d_j-y_j)^2 \tag{1}$$

20世纪90年代后，多种浅层机器学习模型相继出现，包括Cortes等[20]提出的支持向量机（SVM）模型，Greene[21]提出的最大熵（LR）以及Boosting模型等，在内容推荐、分类、网页搜索等方面取得了显著效果。这些模型的共同特点是不含隐藏层或仅有一层隐藏层节点，在学习过程中容易因隐藏层数量不足而造成学习效率低、收敛速度慢的问题[20-22]。其次，模型的学习信号在反向传播的过程中会逐渐减弱，学习过程对神经网络的设计要求较高，而隐藏层节点数量的确定问题目前仍缺乏较全面的指导理论[23]。由于浅层学习算法存在的局限，Hinton等[24]提出了基于深度信任网络（Deep Belief Network）的无监督贪婪逐层训练算法。其主要观点是：①充分利用多隐层人工神经网络的特征学习能力，将学习得到的特征用于数据分类处理与可视化；②通过无监督“逐层初始化”（Layer-Wise Pre-Training）以应对深度神经网络在训练上所带来的困难。随后，Ngiam等[7]进一步提出了多层自动编码器(Multilayer

Auto-encoder），为解决深层网络结构的相关优化问题带来了突破。自此，表示学习研究进入深度特征学习阶段。

与浅层特征学习相比，深度学习凸显的特点有：①强调模型结构的深度，通常有 5 层以上，甚至多达 10 余层的隐藏层节点[25]，利用多层隐藏层对海量数据进行充分的处理，得到更有针对性的特征，从而提升学习分类或预测的准确性。②突出特征学习的重要性，通过逐层特征变换，将样本在原空间的特征表示到一个新的特征空间，从而简化分类与预测的过程[7]。③能够通过深层非线性网络结构的学习实现复杂函数逼近，完成输入数据的分布式表示，具有从样本数据集中学习数据集本质特征的能力。④能够自动学习数据的另一种表示方法，并将其作为特征加入原有问题的特征集合中，从而进一步提高学习效果。

正因为如此，深度学习很快炙手可热，成果不断涌现。2009 年，Bengio[22]在研究中指出，用特定的方法设定训练样本的初始分布和排列顺序可以产生更好的训练结果；Glorot 等[26]等探讨了隐层非线性映射关系的选择和网络的深度相互影响的问题；2012 年，Bengio 等[27]进一步描述了用于有效训练的大型深度结构神经网络的超参数的影响因素。此类以深度学习为代表的算法在各种单模态数据处理上的成功为其应用到复杂多模态数据处理奠定了基础。吴海燕[28]提出了同时进行特征学习和有监督的分类学习的联合框架以及半监督自动编码器模型；2016 年，Andreas 等[29]提出了基于模块神经网络（NMN）的视觉问答框架，利用模块化神经网络对图像进行成分切分，实现动态识别图像内容及颜色；朱陶等[30]在 *K*-means 聚类获取训练样本虚拟标签和卷积核学习之上，提出了前向无监督卷积神经网络的人脸表示学习方法；李志宇等[31]提出了基于动态阻尼正负采样的社会网络结构特征嵌入模型（DNPS），构建了针对新增节点的动态特征学习方法。在特征抽取方面，李志义等[32]在条件随机场模型（Conditional Random Fields，CRFs）的基础上，提出了基于依存语法的<评价特征，评价词>对抽取方法。不难看出，关于表示学习的研究已从最初的简单图像、语音识别转向到了基于深度学习的情感分类、人体行为识别、跨模态检索、信息推荐等更复杂的领域发展。

3.2　深度学习的模型、方法

按照学习过程是否存在人工干预，深度学习方法可分为监督学习与无监督学习两类。其中，目前在表示学习研究中常见的自动编码机（AutoEncoder）、稀疏编码（Sparse Coding）、限制玻尔兹曼机（Restricted Boltzmann Machine）、深度信念网络（Deep Belief Network）等模型均为典型的无监督模型；而广泛应用于人工智能领域的多层感知机（Multilayer Perceptron）、卷积神经网络（Convolutional Neural Networks，CNNs）等模型则属于监督学习模型。

1）自动编码器

自动编码器是 Rumelhart 等[33]于 1986 年提出的一种用于高维复杂数据处理任务的算法，其主要可分为输入、隐藏、输出共三层，从输入层到隐藏层，再从隐藏层到输出层可分别理解为一个“压缩”的学习过程与 “解压”的输出过程，进而获取样本的关键特征，以数据降维的形式缩短训练时间，实现性能优化。2006 年，Hinton 等[24]提出了三隐含层深度自编码器模型。如图 1 所示，从输入 x 至输出，中间包含了三层隐含表示 h。这就大幅提高了自动编码器的非线性拟合与分布式表示数据特征能力，并能够完成数据潜在分层特征的抽取，获得原始数据的“层次型分组”结构。此后，Vincent 等[34]提出了具备去噪能力的自动编码器（Denoising Autoencoder）；Rifai 等[35]进一步对学习约束项进行修改，提出了隐层的维度远小于输入层的压缩自动编码器（Contractive Autoencoder）。卷积自动编码器、RBM 自动编码器[36-42]在表示学习降噪的成功应用，为自动编码器在跨模态检索及特征抽取建模方面的应用夯实了基础。

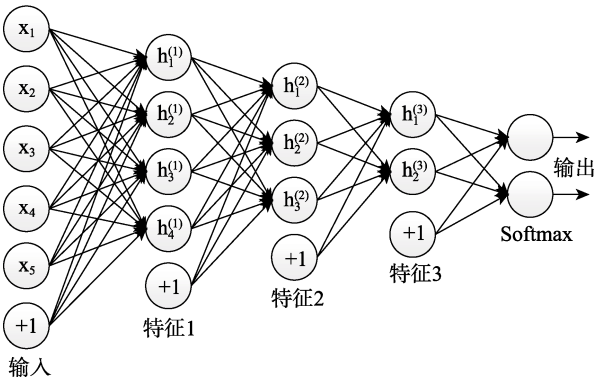

图 1　三隐含层深度自编码器结构

2）稀疏编码

稀疏编码算法是 Bengio 等[27]于 2007 年提出的一种无监督学习方法，目前被广泛应用于图像的去噪、分类、人脸识别、语音识别、多模态表示学习

等方面。例如，刘菲等[43]提出了一种基于稀疏编码的多模态信息交叉检索算法，使交叉检索的平均准确率（MAP）提高了 18.7%，同时增强了稀疏表示的鲁棒性；赵仲秋等[44]提出了针对图像语义构成的基于稀疏编码的多尺度空间潜在语义分析方法，发现其与支持向量机算法相结合能够共同提升图像表征和分类性能；万源等[45]针对稀疏编码的不稳定性以及图像表示与分类相互独立的问题，提出了 NLLSC-CI 的图像分类算法，进一步提高了图像分类效果等。此外，稀疏编码还可将 DAE 与无监督特征学习联系起来，然后采用不同代价函数训练优化策略，使 DAE 具有大规模并行、分布式处理、自组织和自学习等优点。

3）限制玻尔兹曼机

限制玻尔兹曼机（Restricted Boltzmann Machine，RBM）是由 Smolensky[46]针对玻尔兹曼机（Boltzmann Machine, BM）因训练时间过长所导致的难以得到服从自身表示分布的随机样本问题所提出的两层结构、对称连接且无自反馈的随机神经网络模型[47]。与其他算法相比，RBM 算法模型主要具有两大优势[48]：在给定可见层单元状态时，各隐藏节点的激活条件独立；在给定隐藏单元状态时，可见层单元的激活条件相互独立。因此，当隐藏节点的数目足够多时，RBM 能够拟合任意离散分布[49]。2002 年，RBM-CD 快速学习算法的提出进一步推动了随机近似推理、基于能量的模型与未归一化的统计模型的发展[50]。目前，RBM 被广泛地应用于分类[51-53]、回归、降维时间序列建模[54]、图像特征提取[55]、协同过滤[56]等机器学习领域，成为人工智能研究中最常用的算法之一。

4）深度信念网络

深度信念网络是一种具有高效学习算法的神经网络模型，其网络结构如图 2 所示：由若干层 RBM 和一层 BP 堆叠而成，以自底向上的顺序逐层完成 RBM 的训练。

DBN 的训练过程主要分为预训练、微调、反向传播共三步。从功能上而言，RBM 网络模型的训练过程可视为对一个深层 BP 网络权值参数的初始化，使 DBN 克服了 BP 网络容易因随机初始化权值参数而造成的仅局部最优化与训练时间长的缺点。此外，Hinton[50]提出的 RBM 的对比散度快速学习方法（Contrastive Divergence，CD）解决了从整体上对 DBN 进行训练的高复杂度问题，大幅提升了模型的训练速度，并通过产生更合适的参数初始值提升了

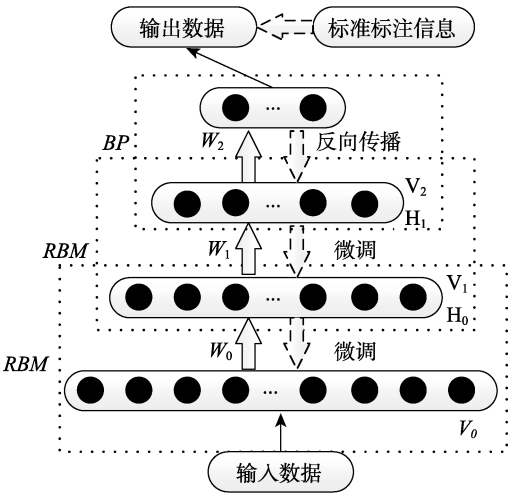

图 2 深度信念网络结构

模型的实际建模能力。DBN 在网络模型训练的过程中，隐含层单元的数量必须要手工进行调整，才能达到最佳的训练状态[57]，基于此问题，潘广源等[58]根据隐含层和误差的关系，提出一种基于重构误差的网络深度判断方法，有效提高了运算效率；何俊等[59]对 DBN 参数采用动态搜索的方法，在一个大范围内搜索确定 RBM Mini-batch、BP Mini-batch 与 RBM 隐层数量的最优值，从而确定 DBN 深度与迭代次数最佳值，为 DBN 深度的确定提供了另一种解决途径。值得提出的是，由于 DBN 等相关算法的应用，由机器学习领域衍生出了深度学习研究方向，更进一步提出了面向人工智能的机器学习算法的设计目标[60]。

5）卷积神经网络

卷积神经网络（CNN）是 LeCun 等[61]于 1989 年在 BP 神经网络的基础上提出了一种包含卷积层的深度神经网络模型，其结构如图 3 所示，通常包括非线性卷积层、子采样层、全连接层和隐藏层四部分。而 CNN 模型一般包含 2 个可以通过训练产生的非线性卷积层、2 个固定的子采样层、1 个全连接层与至少 5 个隐藏层，其算法流程包括滤波及反向传输两个流程[62]。

其中，滤波流程包括输入层、卷积层与采样层的处理，CNN 模型的反向传输流程则通过最小化残差来调整权重和偏置，其中输出层的残差是输出值与类标值的误差值：

$$\delta_i^{(n_l)} = \frac{\partial}{\partial z_i^{(n_l)}} \frac{1}{2} \left\| y - h_{W,\ b}(x) \right\|^2 = -(y_i - a_i^{(n_l)}) \cdot f'(z_i^{(n_l)}) \tag{2}$$

当一个卷积层 L 的下一层（L+1）为采样层，采样层

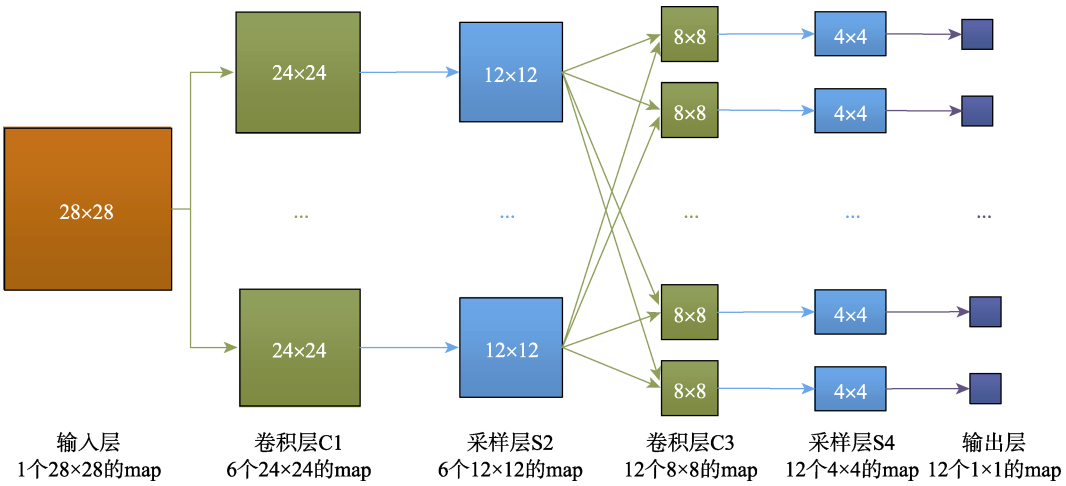

图 3　卷积神经网络结构

（L+1）的 map 大小是卷积层 L 的 1/(scale× scale)；当某个采样层 L 的下一层是卷积层（L+1），采样层到卷积层直接的连接是有权重和偏置参数的，因此假设 L 层第 j 个 map M_j 与 L+1 层的 M_{2j} 关联，按照 BP 的原理，L 层的残差 D_j 是 L+1 层残差 D_{2j} 的加权和。目前，CNN 已被广泛应用于图像分类、目标检测、图像语义分割等领域，其相关应用领域的精度也得到了迅速的提高。以图像分类领域的研究为例，在 AlexNet 将 ImagNet 的图像分类准确度大幅提升到 84.7%之后，针对多个特定领域进行改进的 CNN 模型相继被提出，并不断突破 AlexNet 的纪录，具有代表性的网络包括：VGG[63]、GoogLeNet 和 PReLU-net[64]等。李彦东等[65]指出，CNN 将逐渐向更完备的理论指导、更具体领域中的参数量化问题，以及与迁移学习相结合的一系列研究方向发展。

以上深度学习模型和方法都具有共同特点，即都是由多层感知器构成的多层神经网络，且层数一般都在三层以上，均能通过组合低层特征，形成更加抽象的高层表示属性类别或特征，以发现数据的高层语义表示。

3.3　面向信息检索的跨模态系统建模研究现状与趋势

跨模态检索的核心在于构建不同模态信息之间的关联，检索的质量直接取决于这种关联建模的质量。目前跨模态检索的研究主要分为构思探索、算法改进及模型构建三个阶段。这三个阶段的主要标志分别是基于共享层建立各模态数据之间的关联、以深度学习为基础的跨模态检索算法，以及在表示空间中建立不同模态间的关联。

3.3.1　基于共享层建立各模态数据之间的关联

基于共享层而建立各模态数据之间的关联，需要学习一个多模态数据的共享层。该类模型的典型代表是 Amir 等[5]于 2004 年提出的基于内容的自动交互系统模型，其针对图像提出了基于特征抽取的学习算法，建立了基于多模态的视频内容语义检索系统，并在算法中使用 0-1 损失函数对模型预测值与真实值的相异程度进行估量：

$$\lambda(\alpha_i \varpi_j)=\begin{cases}0, & i=j\\ 1, & i\neq j\end{cases} \tag{3}$$

随后，Zheng 等[66]将文本及图像间的跨模态检索技术应用于生物图像自动注释中，建立了细胞迁移自动注释及检索模型；Jia 等[67]提出了一种结合马尔科夫随机场和 LDA 模型，提出将话题分布看作是图像和文本的共享层，针对图文之间关联不够紧密的语料进行了优化，推动了文本-图像跨模态检索技术的进一步发展。然而，由于不同模态数据之间具有异构性，利用这种策略所构建的跨模态检索模型仍难以充分有效地学习不同模态数据之间的关联，而 LDA 模型在面对较大的语料库进行训练时，若语料库质量较低，其训练的文本特征有效性也会随之大幅降低。该阶段的跨模态数据的关联仍处于构思与探索的阶段，跨模态检索相关研究尚未涉及深度学习算法，也未能形成标准化的跨模态检索模型。

3.3.2　以深度学习为基础的特征抽取算法

国内外学者在深度特征学习阶段中提出了大量公开算法，其中较典型的算法包括：文本对象建模的 Word2Vec 算法、深度残差网络算法（DRN 算法）、关联学习阶段的 CCA 算法等。

1）基于连续词袋（CBOW）的文本特征抽取算法

Word2Vec 是 Google 公司[68]于 2013 年提出的一种用于计算词向量(Distributed Representation)模型，其能够捕捉语境信息的同时压缩数据规模。同年，CBOW 及 Skip-gram 作为 Word2Vec 的配套算法为 Google 官方公开发布[15]。CBOW 及 Skip-gram 均包含输入层、投影层和输出层三部分，主要应用于自然语言处理领域中的文本特征抽取方面。CBOW 的核心思想包括：①模型中的第一层输入时以将词所对应的词向量相加的方式替代排序组合，降低计算复杂度。②取消三层神经网络中的第二层，进一步降低计算复杂度。③由于 CBOW 的目的是训练出词的向量表述，利用 Word2Vec 能够更加周全地考虑上下文信息，使文本特征抽取达到更好的效果[68]。

在 CBOW 模型中，对于词汇中的每个词 $w \in W$ 均利用其周边窗口范围内的词作为上下文信息 c_w 来预测其本身。例如，(w,c_w) 是从训练数据集 T 提取出来的词 w，在上下文信息 c_w 确定的情况下，CBOW 的训练目标就是使出现在训练数据 T 当中的信息对 $(w,c_w) \in T$ 的概率尽可能大，而没有出现在训练数据 T 中的 $(w,c_w) \notin T$ 信息对概率尽可能小。其中以上下文信息的向量为特征预测 c_w 的概率为

$$p(w|c_w) = \frac{1}{1+\mathrm{e}^{-v_{c_w}W_w}}$$

其中，W 为神经网络隐藏层与 Softmax 层输出参数。上下文中每个词对应的向量之和表示为 $v_{c_w} = \sum_{w \in c_w} v_w$。

CBOW 的训练目标即其本身的极大似然估计函数：

$$\mathrm{OBJ}_{\mathrm{cbow}} = \arg\max_{v_{w^2}W_1}\left(\prod_{(w_2c_w)\in T}\log p(w|c_w)\cdot \prod_{w_2c_w\notin T}(1-\log(p(w|c_w)))\right) \tag{4}$$

其中，$(w,c_w) \notin T$ 通过负采样的方法得到，利用随机梯度下降法对语料进行训练，训练后得到的词向量使用基于神经元的非线性作用函数进行二进制处理，得到与对应图片相同维数的二值检索向量，最后采取有监督的学习方法对得到的图文二值向量进行分类，并存储于跨模态数据库中。此外，结合 Hierarchy Softmax 和 Negative Sampling[69]优化技术，Word2Vec 可以快速高效地将词语表达成向量。同时，因词向量捕获了自然语言中词语之间的语义特征，通过保存到文件中，词向量便可以供其他相关应用研究使用。由于基于 CBOW 的文本特征抽取算法在自然语言处理方面表现出色，目前已被广泛应用于中文分词、POS Tagging、情感分类、句法依存分析中。

2）基于深度残差网络（DRN）的图像特征抽取算法

深度残差网络（Deep Residual Network，DRN）模型近年来在计算机视觉、多媒体数据处理等领域引起了广泛的关注。在深度学习模型中，随着网络深度的增加，针对图像特征抽取准确率迅速达到饱和并逐渐趋于下降的“退化问题”，DRN 在所增加的层次上采用了恒等映射的方法，以达到控制训练结果误差的效果[70]。构建深度残差网络以提高无监督学习的效率及图像特征抽取的准确率，构建时可利用多层网络拟合一个残差映射，用 $H(x)$表示所期望得到的实际映射，使得堆叠的非线性多层网络去拟合另一个映射关系 $F(x) = H(x)–x$，实际的映射关系可表示为 $F(x)+x$。其中，可利用卷积神经网络（CNN）模型和最近邻分类器（Nearest Neighbor Classifier，NNC）完成图像语料及检索输入图像的处理，并根据输入图像的像素值，计算和训练集中的图像距离，采用两个图像像素向量之间的曼哈顿距离进行距离评定，根据距离最近的类来将其归类。

需要指出的是，对于映射寻优问题，残差映射寻优比直接对原始的参考映射寻优更加方便。如果一个识别映射已被优化，则使识别映射的残差值趋于 0 值会更加容易，此时则不需要利用一个堆叠的非线性组合拟合一个恒等映射（Identify Mapping）。映射关系 $F(x)+x$ 可以通过已添加捷径连接的前向神经网络实现，其中所使用的捷径连接均为恒等映射，其输出加入到堆叠层次的输出中，并没有引入新的参数，因此也没有增加计算的复杂度。

3）典型相关分析（CCA）的关联算法

典型相关分析（Canonical Correlation Analysis，CCA）是 Vía 等[71]于 2005 年提出的一种用于寻求同一对象的两组变量之间最大相关性的多元统计方法。给定两个数据集 X 和 Y，CCA 的目标是求得一对基向量，使得两数据集之间的相关性最大。如，若需要考察一个人的快速学习的能力 Y(阅读能力为 y_1，独立思考能力为 y_2) 与其教育水平 X (知识广度 x_1，知识深度 x_2)之间的关系，则形式化的表示为 $u = a_1y_1 + a_2y_2$ 和 $v = b_1x_1 + b_2x_2$，然后使用 Pearson 相关系数 $\rho X,Y = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\delta x \delta y} = \frac{E[(X-\mu x)(Y-\mu y)]}{\sigma x \sigma y}$ 来

度量 u 和 v 的关系，而 CCA 的目标就是寻求一组最优解 a 和 b 使得 corr(X,Y)最大,其中 a 和 b 就是使 U 和 V 有最大的关联权重[72]。对于给定的两组变量维度为 w_1 的 x_1 与维度为 w_2 的 x_2，若 $w_1 \leqslant w_2$，x 可表示为：

$$x=\begin{bmatrix}x_1\\x_2\end{bmatrix}E[x]=\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix}\sum\mathrm{Var}(x)=\begin{bmatrix}\sum 11\sum 12\\\sum 21\sum 22\end{bmatrix} \quad (5)$$

其中，$\sum$是 x 的协方差矩阵，$\sum 11$ 是 x_1 本身的协方差矩阵；$\sum 12$ 是 $\mathrm{cov}(x_1,x_2)$；$\sum 21$ 是 $\mathrm{cov}(x_2,x_1)$，也是$\sum 12$ 的转置；$\sum 22$ 是 x_2 的协方差矩阵。从 x_1 和 x_2 的整体入手，定义 $\mu=a^{\mathrm{T}}x_1v=b^{\mathrm{T}}x_2$ 可以计算出 u 和 v 的方差和协方差。目前，CCA 算法被应用于人工智能与自动化的多个领域，应用的实际问题也包括多个方面，如高维小样本数据在核化图嵌入过程中出现的复杂度问题[73]、人脸识别问题[74]及笔迹识别问题[75]等。

3.3.3　基于表示空间的跨模态检索模型

在表示空间中建立不同模态间的关联的跨模态检索模型主要有两类，一类是经典的基于典型关联分析（CCA）的非神经网络结构跨模态检索模型，另一类是 Feng 等[73]于 2014 年提出的基于神经网络的跨模态检索模型（Correspondence Autoencoder，Corr-AE）。

1）基于典型关联分析的跨模态检索模型

基于典型关联分析的非神经网络结构跨模态检索模型能够将图像的底层特征及通过深度学习方法得到的文本主题分布特征映射到一个共同的空间中，建立基于语义层面的深度学习跨模态检索模型。自 2010 年起，典型关联分析在图像检索应用的问题上陆续取得了突破性进展，其中包括针对图像检索优化的 LSI 模型[76]、基于 PLSA 的新型多模态集成与扩展模型（MMIP）[75]与基于关联图像正则化的多模态子空间学习算法（JGRMSL）[76]。受到 CCA 算法关联思想启发，国内学者也相继提出了语义模型来构建跨模态检索的语义关联[77]、跨模态模型 CCSS[78]、基于半监督耦合字典学习与耦合特征映射的跨模态检索模型[79]等。

基于典型关联分析的各种非神经网络结构跨模态检索模型成功应用到了图像处理、人脸识别、数据分析与基于内容的图像检索、文本挖掘等各领域中，但其仍具有一定的局限性。首先，CCA 是一种无监督方法，其仅关注成对样本之间的相关性，以相关性作为不同空间中样本之间的相似度，并未对样本的类信息加以利用，类信息的作用未能得到体现。为此，彭岩等[80]提出的半监督的典型相关分析方法（Semi-CCA），在一定程度上解决了 CCA 算法中先验知识的充分应用的问题。其次，CCA 算法本质上是一种线性子空间的学习算法，其学习到的是一种全局线性情况下的线性特征。对于非线性的场景，CCA 的学习效果往往差强人意。针对该问题，Akaho [81]提出了核 CCA（Kernel CCA），在一定程度上克服了 CCA 在非线性情况下的不足；Yin 等[82]提出了局部线性嵌入（Local linear embedding, LLE）的非线性降维方法，将其与流形学习（Manifold learning, ML）联系结合起来。然而，由于不同模态之间的异构性，即使是同一模态的数据，其特征向量的联合分布复杂度也可能相差巨大。基于 CCA 的各种模型均属于由一到两层结构组成的“浅层”模型，在面对复杂数据集时，浅层结构对多模态数据进行高层语义表示及关联建模仍存在较大的不足。

2）基于神经网络的跨模态检索模型

2011 年，Ngiam 等[7]提出了基于联合表示的模型——Bimodal Deep Autoencoder（双模态深度自编码器）。如图 4 所示，它能通过学习得到不同模态的中间表示，通过编码层和联合表示向量，重构文本或图像等单模态与联合模态之间的关联，实现不同模态表示的相互间关联建模。2014 年，Feng 等[83]提出了跨模态对应自编码器（Correspondence Autoencoder），并提出在分别学习不同模态的表示的同时，通过特征关联约束算法使得各个模态的表示之间产生某种关联，由此使得前一阶段学习得到的各个模态的中间表示蕴含了模态间的关联关系。这两种基于神经网络结构的模型在多个数据集上的实验结果均超过了目前传统的典型关联分析方法所得的效果，将跨模态检索研究推向一个新的高度。

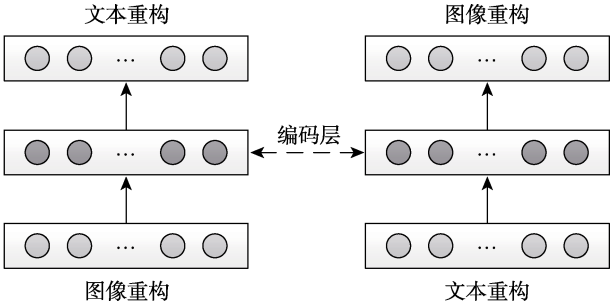

图 4　基于双模态的深度自动编码器

2012 年，Kim 等[84]首次将 DAE 和 CCA 两种方法组合起来，实现了跨模态信息之间的匹配，为深

度学习算法与关联算法的结合提供了良好的理论基础和研究方向。随后，Verma 等[85]在此基础上提出了“或结构”支持向量机，为图像与文本之间的跨模态检索提供了新思路；Wang 等[86]提出以无监督堆叠自编码器（Stacked Auto-Encoder，SAE）、有监督深度卷积神经网络（DCNN）、神经语言模型（NLM）三个算法结合来完成映射函数的学习过程，成功构建了有效图像的模态和文本模态的映射函数，并取得了显著的效果；Cao 等[87]提出了一种新的哈希方法架构，利用稀疏编码器（SAE）描述图像的高级显著结构；邓正恒[72]分别利用 LDA 算法与 SIFT 算法对文本与图像的表示进行建模，然后搭建异构网络，并提出 RandomWalk 算法实现跨模态信息匹配；2016 年，董永亮等[88]构建了一种双层的多模态语义网络模型，对每个单模态的数据分别建立子语义网络，再把子语义网络中的节点聚类成不同的分组以完成建模；丁恒等[89]分别采用 LDA 算法模型与词袋算法作为文本和图像资源的特征表达方式，首次使用最小二乘法替代典型相关性分析法学习特征子空间投影函数，并取得了稳定的跨模态检索结果。

3.3.4 跨模态检索模型构建的趋势和设想

不难看出，以上无论是基于共享层而建立各模态数据之间的关联，还是将不同模态的数据经过抽象后都映射到一个公共的表示空间，在该表示空间中建立不同模态间的关联；抑或将表示学习与关联学习合并、分离，将不同模态信息统一整合为相对低维、易于机器识别与检索的形式，都难以精确满足建模的需求与最优化，以及平衡其识别率、匹配准确度等多方面因素。基于表示学习的跨模态系统建模需要综合多类策略的优势，将继续向不同模态信息的特征抽取、跨模态信息匹配算法的改进等方面拓展、深入，以深度学习为代表的表示学习算法为此提供了较佳的理论基础和研究方向。

未来将探索在大规模数据集上的跨模态检索问题，以便解决在多个层次建立不同模态数据之间的关联，从而实现多模态检索。因此，本文在研究现有检索模型的基础上，预测未来的跨模态检索模型将是如图 5 所示的多模态检索模型 Cross-RBM，其模型分为表示层、隐藏层、输入层、输出层四部分，并设定顶层对不同模态间的数据关联进行学习。为了提升跨模态检索的速度，实验设置可限定表示层的神经元数目，使学习到的公共表示空间为一个低维空间。

设 $f_I(\cdot)$、$f_T(\cdot)$ 和 $f_S(\cdot)$ 分别是从图像 RBM、文本 RBM 和语音 RBM 的输入层到隐藏层的映射函数。Θ 代表为其中的所有参数，包括三个 RBM 的输入层与表示层之间的权重 W、输入层的偏置 c 和表示层的偏置 b，即 $\Theta=\{w^I,c^I,b^I,w^T,c^T,b^T,w^S,c^s,b^s\}$，参数中的上标 I、T 和 S 分别代表图像（Image）、文本（Text）和语音（Speech）。给定一个包含 m 个对齐的图像文本的训练集 $\left\{\left(v_i^I,v_i^T\right)\right\}_{i=1}^m$，模型的训练目标是获得满足目标函数公式的参数 Θ。设 Minimize $\theta L_p+\alpha L_I+\beta L_T$，其中，$L_D=\sum_{i=1}^{m}\left\|f_I\left(v_i^I\right)-f_T\left(v_i^T\right)\right\|_2^2$，

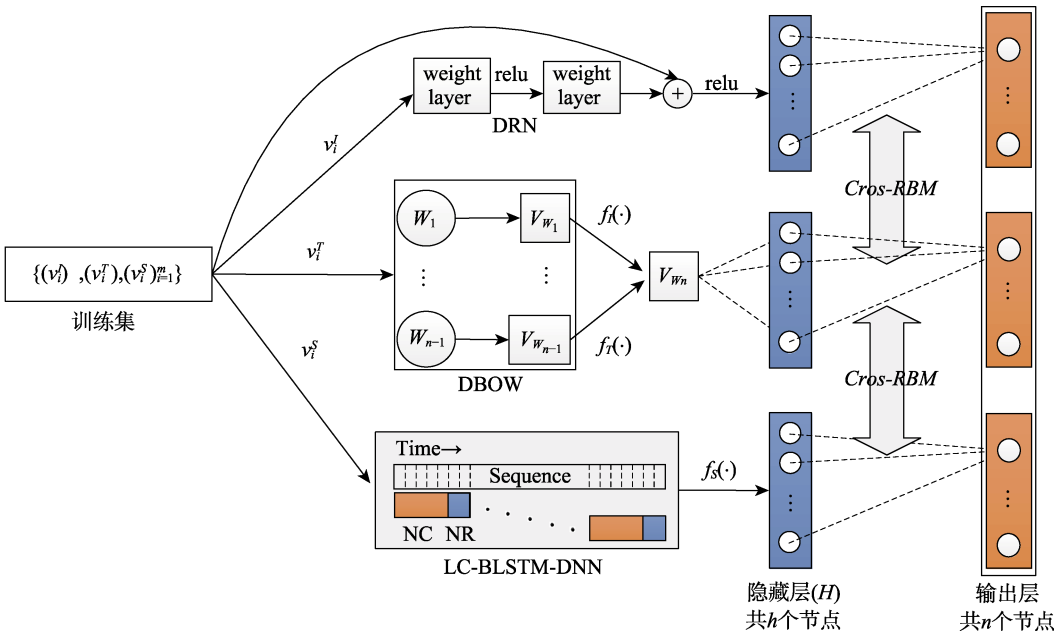

图 5 Cross-RBM 跨模态检索模型

$L_I=-\sum_{i=1}^{m}\log p\left(v_i^I\right)$，$L_T=-\sum_{i=1}^{m}\log p\left(v_i^T\right)$，$L_D$ 是训练集中图像和文本对在表示空间中的欧式距离之和，即多模态关联误差；L_I 与 L_T 分别是图像和文本 RBM 的优化目标函数，其中 $p\left(v_i^I\right)$ 和 $p\left(v_i^T\right)$ 分别是第 i 组图像和文本数据的似然；参数 α 和 β 用来决定图像和文本似然在目标函数中所占的比重。可见，模型的优化目标是最小化多模态关联误差和各自模态的似然之和。同理可构建出语音和文本的训练集 $\left\{\left(v_i^S,v_i^T\right)\right\}_{i=1}^{m}$。

4　评价指标

4.1　Brodatz 纹理库测试指标

纹理是人类视觉的一个重要组成部分，其能为景象深度和表面取向（Surface Orientation）提供线索[90]。在图像检索中，纹理是图像的识别与特征抽取的主要元素之一。目前大多数图像纹理的相关研究均集中在 Brodatz 纹理库中。在图像数据库中进行检索时，Brodatz 纹理库的测试指标能够在数据库中根据排序结果将一个基于相似度的排序列表返回给用户，并采用前 N 个结果中的准确率 P 对检索结果进行评价[91]。例如，对于一次以 $q_i\in R$ 为样例图片的检索过程，若 R 为某一具有特定语义含义的图像集合，用户提交 q_i 的目的就是要检索出 R 集合，系统返回的前 N 个结果为 p_j,j=1, 2, …N 准确率 $P_N(q_i)$ 定义为

$$P_N(q_i)=\sum_{k=1}^{N}\frac{\psi(p_k,R)}{N} \tag{6}$$

其中，$\psi(x,y)$ 进一步描述为

$$\psi(x,y)=\begin{cases}1,if\ x\in Y\\0,if\ x\notin Y\end{cases}$$

对于所有测试样例检索图像集得到的平均准确率为：

$$P_N=\sum_{i=1}^{\text{Total\_Query\_Count}}\frac{(P_N(q_i))}{\text{Total\_Query\_Count}} \tag{7}$$

在测试图库中，前 N 个结果的查全率 $P_N(q_i)$ 可表示为：

$$R_N(q_i)=\sum_{k=1}^{N}\frac{\psi(p_k,R)}{\|R\|} \tag{8}$$

其中，$\|R\|$ 为图像集 R 所含的图像数，对于所有测试样例检索图像集得到的平均查全率为

$$R_N=\sum_{i=1}^{\text{Total\_Query\_Count}}\frac{(R_N(q_i))}{\text{Total\_Query\_Count}} \tag{9}$$

其中，R_N 即在返回的前 N 个结果中正确的占整个系统拥有正确数量的比。此外，目前 Brodatz 纹理库的测试指标还包括全局颜色特征、边缘特征、纹理特征、GIST 和 CENTXIST 值等[92]。

4.2　马氏距离度量学习算法指标

基于深度学习的图像检索在提取图像的特征后需要形成特征向量，再根据特征向量来表征对应的图像。在图像检索中，判断图像之间的相似性主要通过比较两幅图像对应的特征向量之间的距离大小而确定[93]。恰当地完成特征向量的表示及合适的距离度量学习算法是图像检索的关键，马氏距离度量学习（Distance Metric Learning）是其中较经典的距离度量学习算法之一。传统马氏距离度量学习[73]的原理是从训练集 X 中寻找矩阵 $M\in R^{d\times d}$，计算两个样本 x_1、x_2 之间的马氏距离：

$$d_M(x_i,x_j)=\sqrt{(x_i-x_j)^TM(x_i-x_j)} \tag{10}$$

由于 M 为对称半正定矩阵，因此可以分解为：$M=W^TW$，其中 $M\in R^{p\times d}$，$p<d$，则：

$$d_M(x_i,x_j)=\sqrt{(x_i-x_j)^TW^TW(x_i-x_j)}=\left\|Wx_i-Wx_j\right\|_2 \tag{11}$$

综上所述，传统的马氏距离度量学习是通过寻找一个线性转换将每一个样本 x_i 投影到低维子空间中（因为 $p<d$），投影后样本间的欧式距离即为原空间中的马氏距离。除应用在图像检索外，马氏距离度量学习算法指标还可应用与语音评价，目前已被广泛应用于机器学习算法的相关研究中。

4.3　mAP 评价指标

常用于检索效果评价的指标有准确率、召回率及 F-measure 等评价指标，但仅使用简单的指数作为评价指标难免会带来单点值局限性。针对此问题，平均准确率法（Mean Average Precision，mAP）结合了以上三个常用的检索评价指标，可在检索召回结果中计算每一条检索结果的平均准确度，其在检索时先对相关文档进行排序，并根据排序结果反映检索系统中各文档的单个性能值[94]。对于一个查询词，首先确定相关文档在结果列表中的位置以及它是第几个相关文档，然后将所有相关文档的在所有

相关文档中的序号与其在结果列表中的序号的比例相加，最后再对每个查询词的平均准确度求平均[95]。在检索效果评价中，如图 6 所示，准确率（P）指检索算法正确识别的个体总数与检索算法所识别出的个体总数之比，可表示为：$P=A/(A+B)$（A 为检索到与目标相关的内容，B 为检索到与目标不相关的内容）；召回率（R）指检索算法正确识别的个体总数与样本集中所存在的个体总数之比，可表示为 $R=A/(A+C)$（A 为检索到与目标相关的内容，C 为样本集存在的与目标相关的内容但未检索到）。

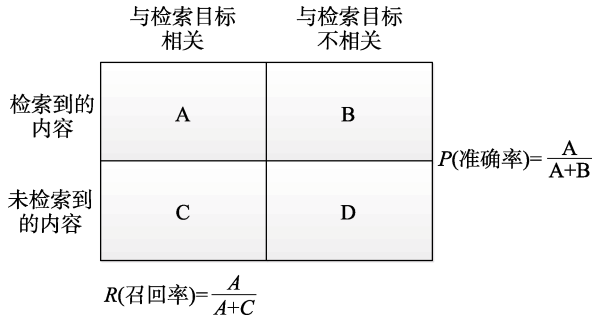

图 6 信息检索召回率与准确率评价指标

此外，还有 F 均值，即准确率 P 和召回率 R 加权调和平均值[96]，可表示为

$$F=\frac{(\alpha^2+1)P\times R}{\alpha^2(P+R)} \tag{12}$$

当参数 α=1 时，即：

$$F_1=\frac{2\times P\times R}{P+R} \tag{13}$$

显然，F_1 综合了准确率和召回率的结果。当 F_1 数值较大时，其检索方法所达到的效果较为理想。可见，F 值即为正确率和召回率的调和平均值，平均准确率 mAP 可定义为：

$$\text{mAP}=\int_0^1 P(R)\text{d}R \tag{14}$$

由于使用 mAP 进行评价相对更能突显检索算法的优劣，目前 mAP 已被广泛应用于图像检索效果评价工作中。

除以上外，11-point Precision-Recall[85]（11-PR）曲线同样可用于评测检索模型的准确率与召回率。其评价原理是将模型得到的准确率 P、召回率 R 和均值 F 各指标同基于深度学习的经典模型、基于 CCA 与基于 Hash 函数的已有模型所得结果进行比较，根据比较结果评价该模型的优越性与不足。

5 总结与展望

本文从信息抽取与表示、跨模态检索建模两个角度总结了目前基于表示学习的跨模态检索模型与特征抽取方面的研究成果，包括自动编码器、稀疏编码、限制玻尔兹曼机、深度信念网络、卷积神经网络等五个典型的表示学习算法；从基于共享层建立各模态间的关联、表示空间中各模态间的关联、以深度学习为基础的跨模态检索算法三个方面归纳了目前面向信息检索的跨模态系统建模研究现状，以及概括了针对跨模态检索的相关评价指标。作为目前数据挖掘、机器学习与人工智能等领域的研究热点，基于表示学习的跨模态检索相关问题已引起国内外学者的广泛关注。

虽然国内外学者在表示学习模型的构建及跨模态检索算法方面均已取得了一定的研究成果，但是当前仍有较多问题亟待解决：①将信息抽取表示的方法应用于跨模态检索系统模型时，目前在表示阶段中样本的特征维度仍然偏高，缺乏在保证检索精度的前提下高效合理、且能适应大规模图像集的检索机制；②各种跨模态检索模型均有其研究对象的针对性，面对各种不同情况时需要利用特定的模型以体现其优势，在解决问题的过程中如何根据各种算法与模型的优势与局限进行结合应用，目前仍处于探索阶段，缺乏较全面的选择方法与指引方案；③由于不同模态信息的表示具有异构性，通常文本表示是离散的，而图像和语音表示是连续的，目前对音频、图像与文本三个模态语料对齐的跨模态检索相关研究仍然较少，且大部分研究处于探索阶段，两个以上不同模态的共同表示问题仍是目前继续解决的难题之一。随着机器学习、人工智能研究的深入，信息检索的应用也逐渐向多模态检索、智能检索方向发展。目前以深度学习为代表的表示学习算法领域的研究与应用将有助于解决多模态信息对齐等问题，大幅促进信息检索应用工作的发展。

参 考 文 献

[1] 王剑. 基于深度学习的跨模态图像检索方法研究[D]. 北京: 中国科学院大学研究生院, 2016.

[2] 何泳澔. 跨模态关联学习及其在图像检索中的应用研究[D]. 北京:中国科学院大学自动化研究所, 2016.

[3] 张昭旭. CNN 深度学习模型用于表情特征提取方法探究[J]. 现代计算机, 2016(3): 41-44.

[4] 孙志军, 薛磊, 许阳明. 基于深度学习的边际 Fisher 分析特征提取算法[J]. 电子与信息学报, 2013, 35(4): 805-811.

[5] Amir A, Basu S, Iyengar G, et al. A multi-modal system for the retrieval of semantic video events[J]. Computer Vision & Image Understanding, 2004, 96(2): 216-236.

[6] Rasiwasia N, Costa Pereira J, Coviello E, et al. A new approach to cross-modal multimedia retrieval[C]// Proceedings of the International Conference on Multimedia. New York: ACM Press, 2010: 251-260.

[7] Ngiam J, Khosla A, Kim M, et al. Multimodal deep learning[C]// Proceedings of the International Conference on Machine Learning. Washington, USA, 2011: 689-696.

[8] 刘春丽, 李晓戈, 刘睿, 等. 基于表示学习的中文分词[J]. 计算机应用, 2016, 36(10): 2794-2798.

[9] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.

[10] Zhao Y, Liu Z Y, Sun M S. Phrase type sensitive tensor indexing model for semantic composition[OL]. [2017-07-25]. http://www.thunlp.org/~lzy/publications/aaai2015_tim.pdf.

[11] Hu B T, Lu Z D, Li H, et al. Convolutional neural network architectures for matching natural language sentences[OL]. [2017-07-25]. http://www.hangli-hl.com/uploads/3/1/6/8/3168008/hu-etal-nips2014.pdf.

[12] Le Q V, Mikolov T. Distributed representations of sentences and documents[OL]. [2017-07-25]. http://proceedings.mlr.press/v32/le14.pdf.

[13] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences[OL]. [2017-07-25]. http://www.cs.wayne.edu/~mdong/Kalchbrenner_DCNN_ACL14.pdf.

[14] Perozzi B, Al-Rfou R, Skiena S. DeepWalk: online learning of social representations[OL]. [2017-07-25]. http://www.perozzi.net/publications/14_kdd_deepwalk-slides.pdf.

[15] Tang J, Qu M, Wang M Z, et al. LINE: Large-scale information network embedding[OL].[2018-04-10]. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/frp0228-Tang.pdf.

[16] Grubinger M, Clough P, Müller H, et al. The IAPR TC12 Benchmark: A new evaluation resource for visual information systems[C/OL]// Proceedings of the International Workshop OntoImage 2006 Language Resources for Content-Based Image Retrieval. [2017-07-25]. http://www-i6.informatik.rwth-aachen.de/publications/download/34/Grubinger-LREC-2006.pdf.

[17] Plummer B A, Wang L, Cervantes C M, et al. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models[C]// Proceedings of the International Conference on Computer Vision. Las Vegas: IEEE, 2016: 2

[18] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// Proceedings of the International Conference on Neural Information Processing Systems. Curran Associates, 2012: 1097-1105.

[19] David R. Signature analysis for multiple-output circuits[J]. IEEE Transactions on Computers, 1986, 35(9): 830-837.

[20] Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[21] Greene W H. Marginal effects in the bivariate probit model[J]. Social Science Electronic Publishing[OL]. [2017-07-25]. http://archive.nyu.edu/bitstream/2451/26254/2/EC-96-11.pdf.

[22] Bengio Y. Learning deep architectures for AI[J]. Foundations & Trends® in Machine Learning, 2009, 2(1): 1-127.

[23] 韩力群. 人工神经网络理论、设计及应用[M]. 北京: 化学工业出版社, 2002: 191-193.

[24] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527.

[25] Deng L, Li J Y, Huang J T, et al. Recent advances in deep learning for speech research at Microsoft[C]// Proceedings of the 38th IEEE International Conference on Acoustics, Speech, and Signal. Vancouver: IEEE, 2013: 8604-8608.

[26] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[C]// Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. Society for Artificial Intelligence and Statistics, 2010: 249-256.

[27] Bengio Y, Lamblin P, Popovici D, et al. Greedy layer-wise training of deep networks[C]// Proceedings of the 19th International Conference on Neural Information Processing Systems. Vancouver: MIT Press, 2006: 153-160.

[28] 吴海燕. 基于自动编码器的半监督表示学习与分类学习研究[D]. 重庆: 重庆大学, 2015.

[29] Andreas J, Rohrbach M, Darrell T, et al. Learning to compose neural networks for question answering[OL]. [2017-07-31]. http://www.stanfordlibraries.info/class/cs224n/lectures/cs224n-2017-lecture17-highlight.pdf.

[30] 朱陶, 任海军, 洪卫军. 一种基于前向无监督卷积神经网络的人脸表示学习方法[J]. 计算机科学, 2016, 43(6): 303-307.

[31] 李志宇, 梁循, 徐志明, 等. DNPS: 基于阻尼采样的大规模动态社会网络结构特征表示学习[J]. 计算机学报, 2017, 40(4): 805-823.

[32] 李志义, 王冕, 赵鹏武. 基于条件随机场模型的“评价特征-评价词”对抽取研究[J]. 情报学报, 2017, 36(4): 411-421.

[33] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323: 533-536.

[34] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders[C]// Proceedings of the International Conference on Machine Learning. New York: ACM Press, 2008: 1096-1103.

[35] Rifai S, Vincent P, Muller X, et al. Contractive auto-encoders: Explicit invariance during feature extraction[OL]. [2017-07-31]. http://www.iro.umontreal.ca/~lisa/bib/pub_subject/language/pointeurs/ICML2011_explicit_invariance.pdf.

[36] Masci J, Meier U. Stacked convolutional auto-encoders for hierarchical feature extraction[C]// Proceedings of the International Conference on Artificial Neural Networks. Springer-Verlag, 2011: 52-59.

[37] Vincent P, Larochelle H, Lajoie I, et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion[J]. Journal of Machine Learning Research, 2010, 11(12): 3371-3408.

[38] Mitchell B, Sheppard J. Deep structure learning: Beyond connectionist approaches[C]// Proceedings of the International Conference on Machine Learning and Applications. IEEE, 2013: 162-167.

[39] Erhan D, Bengio Y, Courville A, et al. Why does unsupervised pre-training help deep learning?[J]. Journal of Machine Learning Research, 2010, 11(3): 625-660.

[40] Deng L, Seltzer M L, Yu D, et al. Binary coding of speech spectrograms using a deep auto-encoder[C]// Proceedings of the Conference of the International Speech Communication Association, Makuhari, Chiba, Japan. DBLP, 2010: 1692-1695.

[41] Lee H, Ekanadham C, Ng A Y. Sparse deep belief net model for visual area V2[C]// Proceedings of the International Conference on Neural Information Processing Systems. Curran Associates, 2007: 873-880.

[42] 李海峰, 李纯果. 深度学习结构和算法比较分析[J]. 河北大学学报(自然科学版), 2012, 32(5): 538-544.

[43] 刘菲, 刘学亮. 基于稀疏编码的多模态信息交叉检索[J]. 中国图象图形学报, 2015, 20(9): 1170-1176.

[44] 赵仲秋, 季海峰, 高隽, 等. 基于稀疏编码多尺度空间潜在语义分析的图像分类[J]. 计算机学报, 2014, 37(6): 1251-1260.

[45] 万源, 史莹, 陈晓丽. 非负局部Laplacian稀疏编码和上下文信息的图像分类[J]. 中国图象图形学报, 2017, 22(6): 731-740.

[46] Smolensky P. Information processing in dynamical systems: Foundations of harmony theory[C]// MIT Press, 1986: 194-281.

[47] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]// Proceedings of the International Conference on Neural Information Processing Systems. Curran Associates, 2013: 3111-3119.

[48] Freund Y, Haussler D. Unsupervised learning of distributions on binary vectors using two layer networks[J]. Advances in Neural Information Processing Systems, 1999(4): 912-919.

[49] Le Roux N, Bengio Y. Representational power of restricted boltzmann machines and deep belief networks[J]. Neural Computation, 2008, 20(6): 1631-1649.

[50] Hinton G E. Training products of experts by minimizing contrastive divergence[J]. Neural Computation, 2002, 14(8): 1771-1800.

[51] Ashwin T S, Saran S, Reddy G R M. Video affective content analysis based on multimodal features using a novel hybrid SVM-RBM classifier[C]// IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics Engineering. IEEE, 2017: 416-421.

[52] 王曙. 深度学习算法研究及其在图像分类上的应用[D]. 南京: 南京邮电大学, 2016.

[53] 张阳, 刘伟铭, 吴义虎. 基于深信度网络分类算法的行人检测方法[J]. 计算机应用研究, 2016, 33(2): 594-597.

[54] Morère O, Lin J, Veillard A, et al. Nested invariance pooling and RBM hashing for image instance retrieval[C]// Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval. New York: ACM Press, 2017: 260-268.

[55] 刘兴旺, 王江晴, 徐科. 一种融合 AutoEncoder 与 CNN 的混合算法用于图像特征提取[J]. 计算机应用研究, 2017, 34(12): 3839-3843.

[56] 黎亚雄, 张坚强, 潘登, 等. 基于 RNN-RBM 语言模型的语音识别研究[J]. 计算机研究与发展, 2014, 51(9): 1936-1944.

[57] 鲁铮. 基于T-RBM算法的DBN分类网络的研究[D]. 长春: 吉林大学, 2014.

[58] 潘广源, 柴伟, 乔俊飞. DBN 网络的深度确定方法[J]. 控制与决策, 2015, 30(2): 256-260.

[59] 何俊, 蔡建峰, 房灵芝, 等. 基于 LBP/VAR 与 DBN 模型的人脸表情识别[J]. 计算机应用研究, 2016, 33(8): 2509-2513.

[60] 吕启, 窦勇, 牛新, 等. 基于 DBN 模型的遥感图像分类[J]. 计算机研究与发展, 2014, 51(9): 1911-1918.

[61] LeCun Y, Bottou L, Bengio Y, et al. Gradient based learning applied to document recognition[C]// Proceedings of IEEE, 1998, 86(11): 2278-2324.

[62] Rasmusbergpalm/DeepLearnToolbox[OL]. [2017-07-12]. https://github.com/rasmusbergpalm/DeepLearnToolbox.

[63] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. [2017-11-16]. https://arxiv.org/pdf/1409. 1556.pdf.

[64] Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions[C]// Proceedings of the Conference on Computer Vision and Pattern Recognition. IEEE, 2015: 1-9.

[65] 李彦冬, 郝宗波, 雷航. 卷积神经网络研究综述[J]. 计算机应用, 2016, 36(9): 2508-2515.

[66] Zheng C X, Long A, Volkov Y, et al. A cross-modal system for cell migration image annotation and retrieval[C]// Proceedings of the International Joint Conference on Neural Networks. IEEE, 2007: 1738-1743.

[67] Jia Y Q, Salzmann M, Darrell T. Learning cross-modality similarity for multinomial data[C]// Proceedings of the International Conference on Computer Vision.Barcelona. IEEE Computer Society, 2011: 2407-2414.

[68] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[EB/OL]. [2017-07-12]. http://arxiv.org/pdf/1301.3781.pdf.

[69] Le T A. An exploration of the Word2vec algorithm: Creating a vector representation of a language vocabulary that encodes

meaning and usage patterns in the vector space structure[D]. University of North Texas, 2016.

[70] 张川. 面向图像分类的深度残差网络优化结构研究[D]. 北京: 中国科学院大学计算机技术研究所, 2016.

[71] Vía J, Santamaría I, Pérez J. A robust RLS algorithm for adaptive canonical correlation analysis[OL]. [2017-07-31]. http://pdfs.semanticscholar.org/59ef/40e0c8fd82c95b12f3aee38b57a65-3ab1ea1.pdf.

[72] 邓正恒. 跨模态信息检索方法的研究与实现[D]. 上海: 复旦大学, 2013.

[73] Feng F X, Wang X J, Li R F. Cross-modal retrieval with correspondence autoencoder[C]// Proceedings of the 22nd ACM International Conference on Multimedia. New York: ACM Press, 2014: 7-16.

[74] Chandrika P, Jawahar C V. Multi modal semantic indexing for image retrieval[C]// Proceedings of the ACM International Conference on Image and Video Retrieval. New York: ACM Press, 2010: 342-349.

[75] Lin W X, Lu T, Su F. A novel multi-modal integration and propagation model for cross-media information retrieval[C]// Proceedings of the International Conference on Advances in Multimedia Modeling. Springer-Verlag, 2012: 740-749.

[76] Wang K Y, Wang W, He R, et al. Multi-modal subspace learning with joint graph regularization for cross-modal retrieval[C]// Proceedings of the 2013 Second IAPR Asian Conference on Pattern Recognition. IEEE Computer Society, 2013: 236-240.

[77] Xie L, Pan P, Lu Y S. Analyzing semantic correlation for cross-modal retrieval[J]. Multimedia Systems, 2015, 21(6): 525-539.

[78] Wang S X, Pan P, Lu Y S, et al. Improving cross-modal and multi-modal retrieval combining content and semantics similarities with probabilistic model[J]. Multimedia Tools and Applications, 2015, 74(6): 2009-2032.

[79] Xu X, Yang Y, Shimada A, et al. Semi-supervised Coupled Dictionary Learning for Cross-modal Retrieval in Internet Images and Texts[C]// Proceedings of the ACM International Conference on Multimedia. New York: ACM Press, 2015: 847-850

[80] 彭岩, 张道强. 半监督典型相关分析算法[J]. 软件学报, 2008, 19(11): 2822-2832.

[81] Akaho S. A kernel method for canonical correlation analysis[C]// Proceedings of the International Meeting of the Psychometric Society. Springer, 2001: 263-269.

[82] Yin J S, Hu D W, Zhou Z T. Noisy manifold learning using neighborhood smoothing embedding[J]. Pattern Recognition Letters, 2008, 29(11): 1613-1620.

[83] Feng F X, Wang X J, Li R F. Cross-modal retrieval with correspondence autoencoder[C]// Proceedings of the 22nd ACM International Conference on Multimedia. New York: ACM Press, 2014: 7-16.

[84] Kim J S, Sim J Y, Kim C S. Multiscale saliency detection using random walk with restart[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2014, 24(2): 198-210.

[85] Verma Y, Jawahar C V. A support vector approach for cross-modal search of images and texts[J]. Computer Vision and Image Understanding, 2016, 154: 48-63.

[86] Wang W, Yang X Y, Ooi B C, et al. Effective deep learning-based multi-modal retrieval[J]. The VLDB Journal, 2016, 25(1): 79-101.

[87] Cao Y, Long M S, Wang J M, et al. Deep visual-semantic hashing for cross-modal retrieval[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2016: 1445-1454.

[88] 董永亮, 柴旭清. 基于潜在语义的双层图像-文本多模态检索语义网络[J]. 计算机工程, 2016, 42(7): 299-303.

[89] 丁恒, 陆伟. 基于相关性的跨模态信息检索研究[J]. 现代图书情报技术, 2016, 32(1): 17-23.

[90] 刘传才, 杨静宇. 一种新的图像纹理表示方法[J]. 计算机学报, 2001, 24(11): 1202-1209.

[91] 李瑞光, 姜锋霞. 基于内容图像检索的特征性能评价研究[J]. 电脑知识与技术, 2014(5): 922-923.

[92] Saracevic T. Evaluation of evaluation in information retrieval[C]// Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1995: 138-146.

[93] 江秋鑫. 基于 SIFT 特征的图像相似性度量及其应用研究[D]. 大连: 大连理工大学, 2012.

[94] 余锦秀. 基于用户行为分析的搜索引擎自动评价技术研究[D]. 北京: 北京邮电大学, 2013.

[95] Li K H, Huang Z, Cheng Y C, et al. A maximal figure-of-merit learning approach to maximizing mean average precision with deep neural network based classifiers[C]// Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2014: 4503-4507.

[96] 信息检索的评价指标(Precision, Recall, F-score, MAP)[EB/OL]. [2017-08-20]. http://blog.csdn.net/Lu597203933/article/details/41802155.

（责任编辑　马　兰）