软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(12):5822−5847 [doi: 10.13328/j.cnki.jos.006764]

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

跨模态数据实体分辨研究综述*

曹建军[1], 聂子博[1], 郑奇斌[2], 吕国俊[2], 曾志贤[1]

[1](国防科技大学 第六十三研究所, 江苏 南京 210007)

[2](中国人民解放军陆军工程大学 指挥控制工程学院, 江苏 南京 210007)

通信作者: 聂子博, E-mail: niezibo233@nudt.edu.cn

摘　要: 实体分辨广泛地存在于数据质量控制、信息检索、数据集成等数据任务中. 传统的实体分辨主要面向关系型数据, 而随着大数据技术的发展, 文本、图像等模态不同的数据大量涌现催生了跨模态数据应用需求, 将跨模态数据实体分辨提升为大数据处理和分析的基础问题之一. 对跨模态实体分辨问题的研究进展进行回顾, 首先介绍问题的定义、评价指标; 然后, 以模态内关系的保持和模态间关系的建立为主线, 对现有研究进行总结和梳理; 并且, 通过在多个公开数据集上对常用方法进行测试, 对出现差异的原因和进行分析; 最后, 总结当前研究仍然存在的问题, 并依据这些问题给出未来可能的研究方向.

关键词: 实体分辨; 跨模态数据处理; 深度学习; 相似性度量

中图法分类号: TP393

中文引用格式: 曹建军, 聂子博, 郑奇斌, 吕国俊, 曾志贤. 跨模态数据实体分辨研究综述. 软件学报, 2023, 34(12): 5822–5847. http://www.jos.org.cn/1000-9825/6764.htm

英文引用格式: Cao JJ, Nie ZB, Zheng QB, Lü GJ, Zeng ZX. Survey on Cross-modal Data Entity Resolution. Ruan Jian Xue Bao/Journal of Software, 2023, 34(12): 5822–5847 (in Chinese). http://www.jos.org.cn/1000-9825/6764.htm

Survey on Cross-modal Data Entity Resolution

CAO Jian-Jun[1], NIE Zi-Bo[1], ZHENG Qi-Bin[2], LÜ Guo-Jun[2], ZENG Zhi-Xian[1]

[1](The Sixty-third Research Institute, National University of Defense Technology, Nanjing 210007, China)

[2](College of Command and Control Engineering, Army Engineering University of PLA, Nanjing 210007, China)

Abstract: Entity resolution widely exists in data tasks such as data quality control, information retrieval, and data integration. Traditional entity resolution methods mainly focus on relational data, while with the development of big data technology, the application requirements of cross-modal data are generated due to the proliferation of different modal data including texts and images. Hence, cross-modal data entity resolution has become a fundamental problem in big data processing and analysis. In this study, the research development of cross-modal entity resolution is reviewed, and its definition and evaluation indexes are introduced. Then, with the construction of inter-modal relationships and the maintenance of intra-modal relationships as the main line, existing research results are surveyed. In addition, widely used methods are tested on different open datasets, and their differences and reasons behind them are analyzed. Finally, the problems in the present research are concluded, on the basis of which the future research trends are given.

Key words: entity resolution; cross-modal data processing; deep learning; similarity measure

在大数据时代, 随着信息技术、移动网络的迅速发展, 日益丰富的文本、图像、音频、视频等多媒体数据, 使广大用户面临跨模态 (cross-modal), 又称为跨媒体 (cross-media) 数据处理及分析应用的现实需求, 其中跨模态对象之间的实体分辨是一个重要的基础性问题. 实体分辨 (entity resolution) 的任务是从单个或多个数据源中识别出描述同一客观实体的不同数据对象, 是数据质量 (data quality) 领域的 19 个研究主题之一[1].

* 基金项目: 国家自然科学基金 (61371196); 中国博士后科学基金 (2015M582832); 国家科技重大专项 (2015ZX01040201-003)
收稿时间: 2022-05-16; 修改时间: 2022-06-23, 2022-07-28; 采用时间: 2022-08-15; jos 在线出版时间: 2023-03-02
CNKI 网络首发时间: 2023-03-03

现实世界中, 每个事物都有着自己在世界上独一无二的地位, 在信息系统中, 这一事实被提升为实体身份完整性 (entity identity integrity): 每一个实体在系统中被描述且仅被描述一次; 系统对不同的实体有不同的描述[2]. 实体分辨是提高信息系统实体身份完整性的重要手段. 在数据集成中, 该问题称为记录链接 (record linkage)[3], 旨在快速准确地匹配和汇聚表示同一客观实体的数据对象[4,5].

在传统信息检索中, 该问题称为消歧 (disambiguation)[6], 最典型的是名称消歧 (name disambiguation), 名称的混淆导致信息检索和融合领域多方面应用的性能下降, 其中包括文献检索、网页搜索、信息推荐等. 跨模态信息检索是当前的研究热点, 用户输入任意一种模态类型的查询数据, 要求检索出所有模态类型中的语义相关数据, 其核心是在描述相同或相似客观实体的不同模态数据对象之间建立关联, 其本质仍然是实体分辨或记录链接[7].

传统关系型数据实体分辨可以利用数据集的多种信息, 如属性信息、上下文信息、关联关系信息, 基于以上3类信息的实体分辨分别称为基于特征 (属性) 相似度 (feature-based similarity, FBS) 的方法、基于上下文的方法 (context-based methods) 和基于关联关系的方法 (relationship-based methods)[8]. 与传统关系型数据实体分辨不同, 在跨模态数据实体分辨中, 不同模态数据具有低层数据表征异构、高层语义相关、描述粒度不同等特点, 使不同模态数据之间难以直接进行相似性度量; 因此, “多模态数据的统一表示和跨模态数据关联”是跨模态数据实体分辨的核心问题[8].

本文主要对跨模态数据实体分辨研究做详细总结, 为相关研究者了解领域进展提供参考. 本文第1节给出了关系型数据实体分辨和跨模态数据实体分辨的定义, 分析了跨模态数据实体分辨的特点, 归纳了实体的评价方法; 第2节从特征提取、语义嵌入两个层面回顾了模态内关联关系保持的相关工作; 第3节归纳梳理了模态间关系建立的代表性传统方法和深度神经网络方法, 给出了跨模态数据实体分辨的深度神经网络构架; 第4节对不同特征提取方法和数据集对跨模态实体分辨方法的影响进行了实验对比, 并对实验结果进行分析; 第5节总结当前跨模态实体分辨方法仍存在的问题, 并基于此展望未来研究方向.

1 问题描述

1.1 关系型数据实体分辨

传统关系型数据实体分辨根据其过程中所利用的信息层次可分为基于特征相似度的方法、基于上下文的方法和基于关联关系的方法[9]. 基于特征相似度的方法依据记录各属性值的相似程度进行实体分辨, 此类方法是最基本的实体分辨手段. 该类方法通常使用两级相似度函数: 首先用属性相似度函数度量两条记录各对应属性值的相似程度; 然后, 再用以属性相似度为变量的记录相似度函数计算两条记录的整体相似程度[10−13]. 继基于特征相似度的方法之后, 出现了基于上下文的方法, 这类方法不但考虑记录本身的属性, 还考虑上下文的特征或上下文记录的属性. 如, 实体分辨之前, 先依据某准则对记录排序, 使可能匹配的记录聚集, 然后设定滑动检测窗口仅对窗口内的记录检测, 可以大大提高检测效率[13−15]. 为了获取更多的可用信息, 又出现了超出上下文范围的基于关系的方法, 其中最具代表性的是 Kalashnikov 研究组提出的“基于关联关系的数据清洗 (relationship-based data cleaning, RelDC)”[16]. RelDC 的基本思想是用无向图对关系数据库建模, 得到数据库的实体关系图, 通过分析实体关系图, 挖掘实体之间的关联关系进行实体分辨[13,17−19].

关系型数据实体分辨的形式化描述见定义1.

定义 1. 实体分辨. 给定描述 k 个客观实体的 N ($N\geqslant k$) 条数据记录集合 $R=\{r_1, r_2, \ldots, r_N\}$, 实体分辨的目的是将 R 划分为一组子集 $R_1, R_2, \ldots, R_k$, 满足$\forall m, n$ ($m, n=1, 2, \ldots, k$ 且 $m\neq n$), $R_m\cap R_n=\varnothing$ 且 $R_1\cup R_2\cup\ldots\cup R_k=R$, 使得$\forall m$, 若 $r_i, r_j\in R_m$ ($i, j=1, 2, \ldots, N$ 且 $i\neq j$), 则 r_i 和 r_j 描述的是同一客观实体. 例如微博等社交媒体上存在很多用户留下的影评, 我们希望对这些评论进行聚类, 让每聚类中的评论都是针对同一部电影而不同聚类之间不存在交叉.

关系型数据实体分辨的一般流程如图1所示.

图1中的流程主要包括数据预处理、数据分块、记录对匹配3个步骤. 其中数据预处理是对数据进行必要的名称规范化、缺失数据补全、特征选择等处理; 数据分块通过某种分块技术, 如双分块键、迭代分块、滑动窗口、

Canopy、各种自适应分块技术等, 将可能对应同一客观实体的数据记录分到同一块内, 然后进行块内记录间的相似度计算、匹配, 提高分辨的效率[4,20,21]; 记录对的匹配通过计算比较各属性值的相似程度, 判断实体对匹配或者不匹配. 图 1 中的实体分辨流程适用于其他半结构化、非结构化数据的单模态数据实体分辨.

图 1 实体分辨的一般流程

1.2 跨模态数据实体分辨

跨模态数据实体分辨, 其目的就是分辨出同一个客观实体 (类) 在不同模态数据中的不同表示, 其形式化描述见定义 2.

定义 2. 跨模态数据实体分辨 (cross-modal data entity resolution). 给定分别描述相同 k 个客观实体 (类) 的两个不同模态数据集 $X=\{x_1, x_2, \ldots, x_N\}$ 和 $Y=\{y_1, y_2, \ldots, y_M\}$, 其中 $N, M \geqslant k$, 以及实体类集 $C=\{c_1, c_2, \ldots, c_k\}$, 目标是找到 $X \times Y$ 上的映射关系 $f=\{<x_i, y_j> | x_i \in X, y_j \in Y, i=1, 2, \ldots, N, j=1, 2, \ldots, M\}$, 其中 x_i 与 y_j 具有相同的类标 $c_l \in C$, $1 \leqslant l \leqslant k$. 例如有若干打乱的图像和文本, 我们希望建立一种方法能让关于同一事物的图像和文本划归一类, 就像 Wiki 百科中的条目一样. 现实中跨模态方法常用领域有: 跨模态检索, 即通过一个模态的数据去检索另一模态的数据; 行人再识别, 即从不同机位摄影设备所得影响中识别出同一个人等.

跨模态数据实体分辨面向的数据对象属于不同模态, 且不同模态数据呈现出低层特征异构、高层语义相关、描述粒度不同的特点, 传统的针对结构化数据的实体分辨方法并不能适用于跨模态数据, 跨模态数据的实体分辨面临新挑战.

如定义 2, 跨模态数据实体分辨的目的是通过计算分别来自两个模态的数据对象间的相似度, 利用相似关系, 找到 X 和 Y 中描述同一客观实体 (类) 的数据对象. 其中, 跨模态相似性的度量是其核心. 然而, 由于不同模态的数据分属于完全不同的特征空间, 特征表示的不一致性造成难以直接度量跨模态数据间的相似度, 给跨模态实体分辨带来挑战[7,22].

如图 2 所示, 为解决异构性等问题, 最常见的方法是构建共同表征空间并将不同模态的对象映射到其中, 使得该空间中的对象之间的距离能够尽量体现其相似程度 (图 2(a)); 或者将其中一个模态的表征作为目标空间, 直接将另一个模态的对象映射到其中 (图 2(b)).

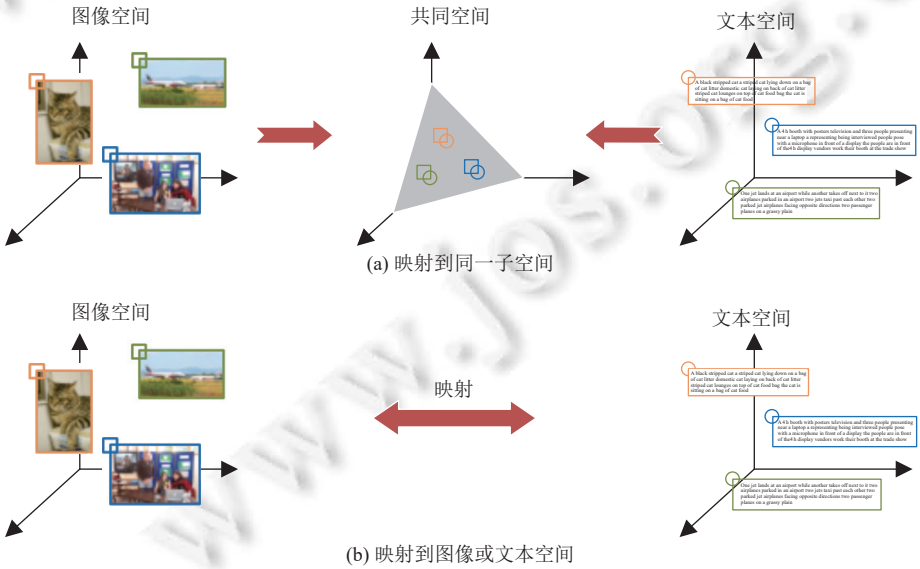

图 2 构建共同表征空间

现有研究中提出了不同的学习模型来构建共同表征空间, 这些模型都需要在共同空间的建立过程中 (隐式或显式地) 保持模态内关联关系 (intra-modal relationship) 并建立模态间关联关系 (inter-modal relationship), 如图 3 所示.

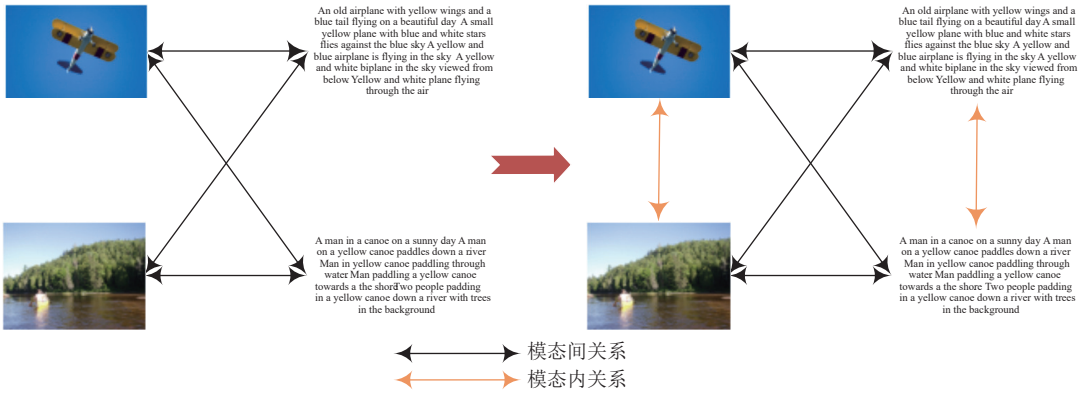

图 3 构建共同表征空间需要保持或建立的两种关联关系

图 3 中的实线代表模态间关联关系, 即来自不同模态的对象之间的关联关系, 如图片和其关联的文本描述之间的对应关系; 橙色线代表模态内关联关系, 即单一模态内对象之间的关联关系, 如模态内对象间的相似关系、近邻关系、类别关系等[23]. 早期的跨模态相似性度量主要考虑如何更好地关联两个模态 (即模态间关系建立). 然而, 自然语言处理和图像处理等领域的发展使得我们可以更加有效地在共同空间中保持模态内关联关系, 并极大地提高了跨模态相似性度量的精确度, 这也同时说明了保持模态内关系的重要性.

1.3 实体分辨结果评价

对于实例层粒度的跨模态相似性度量, 将每个图像-文本对看成一个实例, 通常为一一对应的关系, 其目的是判断图像-文本对之间是否匹配, 粒度相较类别层更加细致, 主要适用于跨模态实体匹配等场景. 实验通过使用有监督或者无监督的方法对每个图像-文本对是否匹配进行判断, 并以其结果的精确率 (P)、召回率 (R) 和 $F1$ 指标进行实例层粒度的度量. 由公式 (1)–公式 (3) 分别计算.

$$P = \frac{TP}{TP+FP} \tag{1}$$

$$R = \frac{TP}{TP+FN} \tag{2}$$

$$F1 = \frac{2\times P\times R}{P+R} = \frac{2TP}{2TP+FP+FN} \tag{3}$$

其中, TP 表示对于匹配的图像-文本对分类正确的个数, FP 表示对于不匹配的图像-文本对分成匹配的个数, FN 表示对于匹配的图像-文本对分成不匹配的个数. 一言蔽之, 精确率是模型认为是对的数据中有多少确实是对的, 召回率是那些确实是对的数据中模型找出了多少. 因此精确率和召回率又叫查准率和查全率, 即模型查得有多精准和模型查得有多全面.

当数据集中每个实例至多只有一个实例与其匹配 (如 MSCOCO 数据集中每个图片只有一段文字与之对应), 通常只考虑相似度最高的前 K 个实例的召回率, 即 $Recall@K$:

$$Recall@K = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{L_i}\sum_{j=1}^{K}\phi(i,j)\right) \tag{4}$$

其中, N 表示样本总数, L_i 为所有与样本 i 匹配的样本数, 表示如果第 i 个样本与相似度排名第 j 的样本匹配, 则 $\phi(i,j)$=1, 否则 $\phi(i,j)$=0. 在一些情况下 (具体地, $Precision@K$=$Recall@1$) 精确率和召回率是高度相关的, 召回率本身即可同时反映精确率和召回率[24].

跨模态方法中最常用的评价指标为类别层粒度的跨模态数据实体分辨度量, 其目的是完成同类别图像-文本

对之间的相似性匹配. 实验通过 TOP-k 下的平均精度均值 (mean average precision, *MAP*) 和混淆矩阵进行类别层粒度的度量. *MAP* 指标常用来度量跨模态检索效果, 以输入输出是否属于相同类别判断是否相似, 由公式 (5) 进行计算.

$$MAP=\frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{k=1}^{M}P(k)r(k)}{L_i} \tag{5}$$

其中, N 为所有待匹配样本总数, M 为对每个样本返回 M 个匹配结果, L_i 为数据集中所有与第 i 个待匹配样本匹配的结果数量, $P(k)$ 表示相似度排名前 k 个样本的精确率. 类别相同则 $r(k)$=1, 反之 $r(k)$=0. 而混淆矩阵则是对 k=1 时, *MAP* 值的分类结果进行更加直观的展示.

2 模态内关联关系保持

对于模态内关联关系 (如相似关联关系、近邻关联关系、类别关联关系等) 的保持, 主要体现在模态内数据的特征提取 (feature extraction) 和语义嵌入 (semantic embedding) 的过程中. 跨模态学习涉及的对象往往是非结构化数据, 其原始数据或者低层特征都只停留在描述层, 为了能够充分挖掘模态内语义层关系, 首先需要通过特征提取得到文本和图像的描述层信息; 通过特征提取获得了文本和图像的特征描述, 但是特征层的描述中语义过于稀疏, 仍然不足以准确描述模态内语义关联关系, 因此需要通过语义嵌入将特征层表示进一步抽象, 得到文本和图像的高层表示 (如主题, 字典等).

2.1 特征提取

所谓特征即是事物异于其他事物的特点, 特征提取的目的是从描述事物的数据中找出这些特点. 跨模态实体分辨面向的对象主要是图片、文本、声音等非结构化数据, 其原始数据中的语义层次低且稀疏. 文献 [24] 的工作证明, 合适的特征提取对于模态内关系的建模具有十分重要的意义. 本节回顾“文本-图像”跨模态学习研究中使用的特征提取方法.

2.1.1 文本特征提取

文本是语言的符号表示, 人类的绝大部分知识是以语言文字的形式记载和流传下来的. 文本较图像更加接近于抽象概念, 但是要让计算机能够理解文本却仍然是一项非常困难的任务. 下面介绍常用的文本数据特征提取方法, 主要分为文本的词级别表示和句级别表示.

词是语言的基本单元, 按照词的表示方式可以将文本特征提取分为独热表示 (one-hot representation) 方法和分布式表示 (distributed representation) 方法. 独热表示将词进行符号化, 用 0-1 字符进行文本编码 (在词典中, 文本中包含的词为 1, 不包含的词为 0). 由于独热表示仅能表示词出现与否, 其语义表征能力很有限, 基于分布式假说[25]的分布式表示取代它成为了主流文本表示方法. 词的分布式表示称为词嵌入 (word embedding), 所谓词嵌入是将一个词的表示嵌入到一个连续的向量空间中, 也就是用更紧凑的向量表示词. 经嵌入后, 词之间可以通过诸如余弦相似度的方法计算相似度, 优秀的嵌入方法可以使得相似的词之间距离较近而不相似的词之间距离较远. 根据不同的建模方式, 研究者们提出了多种词嵌入方案, 如著名的 Word2Vec[26]、GloVe[27].

由于句子是跨模态学习中更为常见的文本形式, 大部分研究 (如文本分类等) 中, 文本的表示还需要在词表示的基础上计算句子表示. 词袋模型 (bag-of-words, BoW) 基于传统统计方法, 忽略文本的顺序和语法等属性, 以一组无序的词对文本进行表示, 词袋模型除记录词的出现与否外还记录词的出现频率, 对文本有较好的表示能力; 传统的基于统计的方法还有主题模型 (如潜在狄利克雷分布 (latent Dirichlet allocation, LDA))、TF-IDF (term frequency-inverse document frequency)、平滑逆频率 (smooth inverse frequency, SIF)[28]等.

值得注意的是, 文本模态具有序列特性, 即含有上下文信息, 传统方法对文本上下文信息处理能力不足, 而神经网络的发展为文本处理带来了新的可能, 其中之一是循环神经网络. 循环神经网络 (recurrent neural network, RNN)[29]以文本经分词 (tokenize) 操作所得分词序列为输入, 得到文本的特征表示, 其结构如图 4 所示.

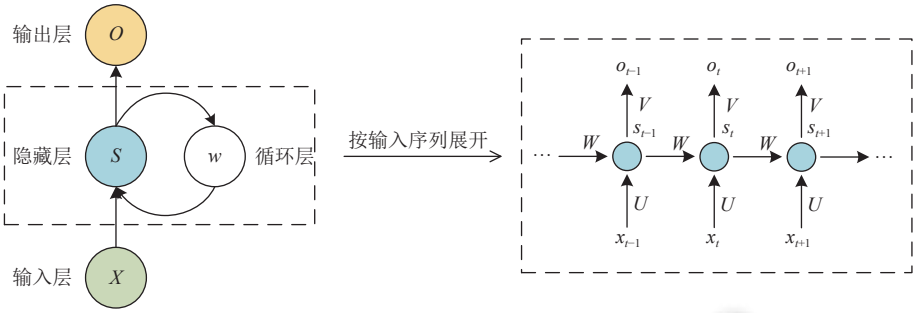

图 4 循环神经网络结构图

图 4 中左侧隐藏层 S 为全连接网络, w 为循环层, 右侧为将循环层与隐藏层按时间线展开所得示意图, W、U 和 V 均为权重矩阵. 可以看出, 对时刻 t 的输入 x_t, 其全连接结果 s_t 不仅取决于 x_t 本身, 还取决于上一时刻全连接结果 s_{t-1}, 同时 s_t 还将影响下一时刻的全连接结果 s_{t+1}. 如此, RNN 能够"记住"输入序列的前文信息. 然而, 由于梯度消失和爆炸的问题, 若文本较长, 位于前方的内容会随序列输入逐渐被"遗忘". 针对此问题, 长短时记忆网络 (long short-term memory, LSTM)[30]通过将遗忘门和记忆门引入循环神经网络, 对输入序列有选择地进行记忆和遗忘, 从而记住重要信息并遗忘无用信息, 最终实现对更大范围的上下文信息的覆盖. LSTM 继承 RNN 仅将序列前方的信息向后方传递的特点, 导致后方信息无法向前传递, 对此, 双向长短时记忆网络 (bi-directional long short-term memory, BiLSTM)[31]将前向 LSTM 与后向 LSTM 结合, 使得序列信息可向前后两方向传递. 文献 [32] 提出 EMLo 模型, 利用双层 BiLSTM, 第 1 层处理语法信息, 第 2 层处理语义信息, 对词的语法和语义特征以及不同语境下的使用进行建模, 解决以往词嵌入方法无法处理多义词的问题.

RNN 的另一局限是要求输入输出序列长度相同, 而在自然语言处理中会出现输入输出不定长的情况, 简单 RNN 对此很难处理. 受编码-解码器机制 (encoder-decoder) 启发, 序列到序列方法 (sequence to sequence, seq2seq)[33]利用 RNN 作为编码器将输入文本编码为定长向量, 而后解码为任意长度输出结果, 实现对不定长输入输出的处理. 由于 seq2seq 编码所生成向量长度固定, 当文本长度过长时会出现信息丢失, 对此, 文献 [34] 提出 Transformer 模型, 完全摒弃 RNN 与 CNN 而仅依靠注意力机制实现编码器和解码器, 使得模型能够保留重要信息, 同时, 由于摒弃了以序列为输入的 RNN, Transformer 模型并行程度更高, 利用 GPU 并行计算优势, 速度更快的同时效果更好. 但是 Transformer 需要进行大量训练, 针对这一问题, 文献 [35] 基于 Transformer 提出 BERT (bidirectional encoder representations from Transformers), 以正序和倒序的方式利用无标记文本数据在每一层对深度双向表示模型进行预训练, 最终得到的训练结果只需通过增加一个输出层进行微调即可达到很好的效果. 双向 Transformer 的 BERT 与单向 Transformer 的 GPT 模型结构如后文图 5 所示.

可以看出图 5(a) 中 BERT 模型中每个输入 E 以及第 1 层 Transformer 可与下一层中位于其左侧和右侧的 Transformer 都进行连接, 因此数据可以向左和右两个方向传递; 而图 5(b) 所示 GPT 中数据则只能向右传递.

需要指出的是, 虽然自然语言处理技术的发展为跨模态学习提供了更多的选择, 词袋模型仍然经常被用于跨模态学习中的文本特征提取. 一方面这是因为大部分跨模态学习的基准数据集默认提供词袋特征; 另一方面也是词嵌入和句子嵌入技术相对传统方法的性能优势还不够显著, 相对其更高的复杂性, 词袋模型性能尚还够用, 因此研究者在跨模态学习中使用这些技术的动力不足. 而近年来 BERT 模型及其变体不断推高文本特征提取的最好成绩, 可以预见深度方法将越来越多地应用于跨模态方法中.

2.1.2 图像特征提取

跨模态学习中常见的视觉特征提取方式主要包括人工设计特征和卷积神经网络 (convolution neural network, CNN) 特征.

人工设计特征尝试通过颜色、形状、纹理等来表示图片, 包括局部特征如尺度不变特征转换 (scale-invariant feature transform, SIFT)[36]和方向梯度直方图 (histogram of oriented gradient, HOG)[37], 全局特征如空间包络特征

(GIST)[38]. 人工设计的特征具有可解释性强等特点, 在涉及图像的机器学习任务 (如图像识别、图像理解等) 中发挥了重要作用, 在初期的跨模态学习任务中, SIFT、GIST 和 HOG 等特征被广泛应用于图像数据的表示.

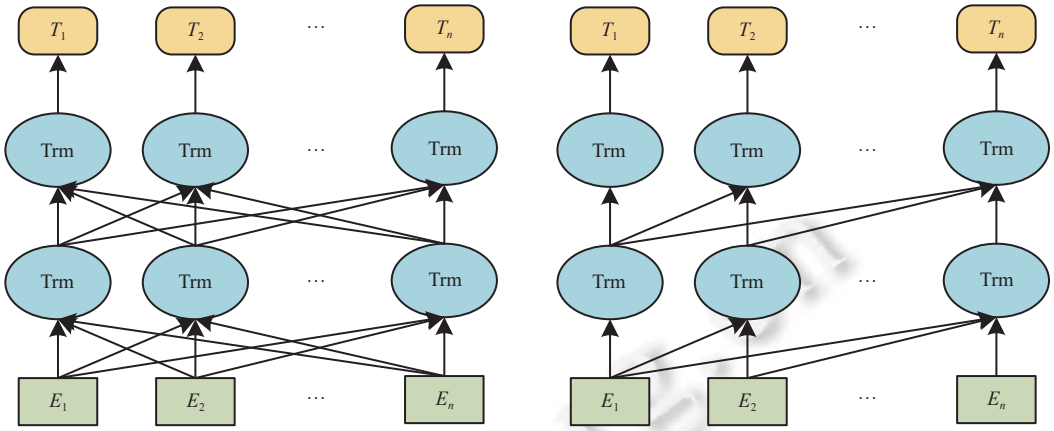

图 5 BERT 与 GPT 结构[35]

随着卷积神经网络 (CNN) 的发展, 其准确性在各种机器视觉任务中都超越了传统方法和其他机器学习方法, 卷积神经网络已经渐渐成为了图像特征提取的主要方法. 卷积神经网络一般由 3 种结构组成: 卷积层、池化层和全连接层. 卷积层 (convolutional layer) 利用卷积核在保留图像像素间关系的同时减小图像大小; 池化层 (pooling layer) 常位于卷积层后, 对卷积层输出结果的每个区域计算出一个值对这个区域进行表示, 用于缩小卷积层输出结果; 全连接层 (fully connected layer) 常被用于分类.

随着迁移学习的发展, 预训练模型开始崭露头角. 在通过预训练模型进行图像特征提取时, 一般固定预训练模型中大多数的权重, 只微调靠近输出的网络高层的权重. 预训练能成功的原因尚无定论, 主流观点之一是不同应用场景和数据集虽然高层语义不同, 但图像低层特征却是相近的, 因而在大规模的视觉数据集上 (如 ImageNet[39]、PASCAL VOC[40]) 训练得到的模型在特征提取上有较好的迁移性, 不需要重新训练或仅需要较少的训练即可用于其他机器视觉任务的特征提取中. VGG[41]模型证明了加深神经网络的层数可以对模型最终结果产生影响, VGG 分为 VGG19 和 VGG16, 二者主要是层数上有所不同, VGG19 的模型结构如图 6 所示.

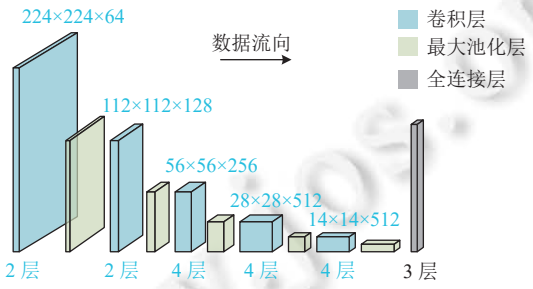

图 6 VGG19 网络结构

可以看出 VGG 网络是标准的含有卷积、池化和全连接层的卷积神经网络. VGG19 的 19 仅包含卷积层和全连接层即 16 层卷积层和 3 层全连接层, 而未对池化层进行计数, 从图 6 可以看出, VGG19 使用的卷积核大小为 3×3, 通过堆叠较小卷积核来提高感受野同时增加模型深度. 针对随着模型加深出现的梯度消失、梯度爆炸和模型退化的问题, ResNet[42]基于 VGG, 提出残差 (residual) 模块, 人为地使一些层跳过相邻的下一层而与更下一层相连以减弱相邻层之间的相关性, 缓解模型退化的问题, 同时将批标准化 (batch-normalization) 引入数据预处理中, 使

数据分布更接近激活函数较为“敏感”的中间区域, 缓解了梯度消失和梯度爆炸的问题, 使网络层数得以超过 1 000 层. Transformer 预训练模型对注意力机制的应用及其在 NLP 领域的成功启发了计算机视觉, 此前注意力机制在计算机视觉中的应用主要是连接卷积层或是替换卷积的部分组成, 文献 [43] 提出 Vision Transformer (ViT), 将 Transformer 应用至计算机视觉, ViT 图像特征提取部分结构如图 7 所示.

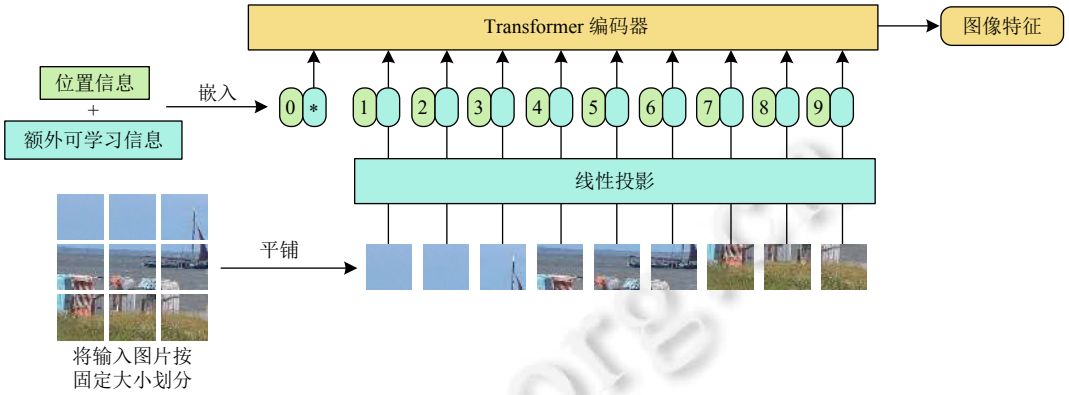

图 7 Vision Transformer 图像特征提取部分结构图[43]

ViT 首先将图片按固定大小进行划分, 将划分结果平铺后进行线性映射进行图片嵌入, 并添加位置信息和下游任务所决定的额外可学习的信息 (如分类信息), 而后将附带位置信息的嵌入结果送至标准 Transformer 编码器, 得到图片的特征表示. 文献 [44] 通过实验验证和理论分析的方式对 Transformer 在计算机视觉任务中的鲁棒性进行了证明, 认为注意力机制发挥了不可或缺的作用.

预训练模型在跨模态领域受到了广泛关注, 例如: 文献 [45] 利用预训练的 VGGNet 来进行图像的特征提取; 文献 [24] 在跨模态检索采用了 ImageNet 数据集上预训练的 DeCAF 特征, 提高了跨模态检索的准确性; 文献 [46] 在跨模态排序学习中使用 ImageNet 数据集上预训练的 GoogLeNet 作为图像特征提取工具; 文献 [22] 使用了 AlexNet 预训练模型作为跨模态检索中的图像特征提取工具; 文献 [47] 使用了 4 096 维的 VGGNet 特征作为模型的输入.

值得指出的是, 一些研究者认为使用预训练模型限制了模型的学习能力, 而完全端到端的训练可以在具体的任务上得到更高的精确率. 例如: 文献 [48] 提出使用一种区域卷积神经网络 (region convolutional neural network, RCNN) 来提取图像特征, 首先将图像分割为多个切片, 然后利用文献 [49] 中的 CNN 网络提取每个切片的特征; 文献 [50] 提出使用深度典型相关分析 (deep canonical correlation analysis, DCCA) 进行“文本-图像”匹配, 其中图片特征提取也是通过类似文献 [49] 中的网络结构进行的. 虽然理论上端到端的训练方式可以学习到更贴近于目标数据集的特征表示, 但是在跨模态相似度性度量中, 其结果的准确性结果往往不如预训练模型. 例如, 文献 [51] 在实验中发现使用 VGG、GoogLeNet 等预训练模型的效果优于其提出的端到端的训练模型. 其主要原因是完全端到端的学习模型的训练过程需要大量的训练样本, 而对跨模态实体分辨任务来说, 由于成本等原因, 获取大量的配对的多模态训练样本是十分困难的.

2.2 语义嵌入

本节中回顾了现有跨模态学习中使用到的语义嵌入方法, 其中一些相对具有较为明确的语义信息, 如主题、字典等, 另外也有一些没有明确的物理意义, 如神经网络中的隐变量等.

2.2.1 稀疏字典学习

一种表征学习方法, 旨在找到一组基本元素使得输入可以映射到这组元素的稀疏表达式, 包含两个阶段: 构建字典以及通过字典进行表示. 字典学习可以学习到数据背后最本质的特征, 被广泛应用于信号处理等领域. 由于其得到的稀疏表示具有可解释性强等特点, 稀疏字典学习也用在不同模态数据的语义表征中[1−4]. 文献 [52] 假设不

同模态的稀疏系数间存在线性映射, 并基于此将单模态成对字典学习拓展到跨模态检索中. 文献 [53] 提出跨模态子模块字典学习 (cross-modality submodular dictionary learning, CmSDL), 并得到模态自适应的字典对以及共享跨模态同构表示空间. 文献 [54] 提出一种共同标签增强的区分字典学习 (discriminative dictionary learning), 加强不同标签的模态内数据的区分度以及具有相同标签的模态间数据关联度.

2.2.2 主题模型

一种原用来发现文本中抽象主题的统计模型, 将文本表示为多个主题上的分布, 常用主题模型为隐狄利克雷分布 (latent Dirichlet allocation, LDA), 将文档可能的主题以概率分布的形式给出[55]. 而实际上“主题”这一概念不止用于文本, 也适用于图像等其他类型的数据, 一些研究者将其进行拓展并应用在多模态数据检索中. 文献 [56] 对文献 [57] 中的二维多模态 LDA 模型进行了扩展, 通过图像的无监督聚类将低层视觉特征抽象为 LDA 模型, 其效果要比只使用文本 LDA 的模型更好.

2.2.3 切 片

不同于将图像和文本整体作为单元进行分析, 部分研究工作以图像和文本中的实体作为基本分析单元进行细粒度分析. 例如: 文献 [48] 将图像和文本表示为实体的集合, 分别使用区域卷积神经网络 (region-based convolutional neural network, RCNN) 和依赖树关系 (dependency tree relation) 对图像和文本进行划分; 但是对文本进行切片会丢失词的上下文信息, 文献 [51] 使用双向递归神经网络来替换文献 [5] 中的词切片; 文献 [58] 为了强调局部的重要性以及实现更为准确的跨模态检索, 对各模态 (包括文本、图像、声音、3D 模型) 对象进行切分.

2.2.4 哈希编码

为了提高在大规模数据上进行跨模态学习的速度, 经常需要使用基于哈希的方法. 通过将数据的原始特征描述映射到海明空间中的哈希编码, 进而提高跨模态学习的速度. 原始的哈希方法并不能保证海明空间的相似性与原始数据的相似性的一致, 为此局部敏感哈希[59]以及更为精确的谱哈希[60]、图哈希等被提出. 通过将这些方法与类标、模态内相似性等约束结合, 可以将单模态数据映射到更为紧凑的海明空间中, 此时具有相近哈希码的对象具有语义相似性.

2.2.5 深度神经网络

深度神经网络具有强大的学习能力, 当有足够的训练数据时基于深度神经网络的方法往往表现出较为明显的性能优势. 这些模型虽然不会显式地执行语义嵌入, 但是在模型训练的过程中模型的中间层隐变量即是对输入数据的抽象. 目前, 在基于深度神经网络构建共同空间的方法中, 一般通过损失函数显式地进行模态内关联关系保持. 文献 [47] 利用模态内样本类别信息的保持, 对投影到类别空间中的样本进行 Softmax 来判断样本类别, 构建交叉熵损失实现模态内样本在映射前后类别语义信息的一致性保持. 文献 [61] 为了保持映射前后模态内样本的 k-近邻关系, 设计了结构保持条件, 即利用一个强制距离参数, 在共同空间中, 控制样本到 k 个近邻样本和到其余样本之间的距离, 实现非线性映射过程中模态内样本间结构关系的保持.

3 模态间关联关系建立

模态间关联关系是跨模态实体分辨的重要依据, 然而由于不同模态数据间异构鸿沟 (heterogeneity gap) 的存在, 无法直接将输入的不同模态间高层语义相关的数据进行关联, 因此需要对不同模态间数据进行处理, 使之映射至共同表征空间 $\mathfrak{R}$ 中, 进而对模态间关联关系 (如成对关联关系、类别关联关系等) 进行建立. 根据整个过程是否有深度网络参与, 将映射方法分为浅层方法与深度方法.

3.1 浅层方法

为了在共同空间 $\mathfrak{R}$ 中更好地建立数据的模态间关联关系, 通常需要利用已知对应关系的训练数据, 即匹配的跨模态样本对 $\mathcal{S}=\{(x_i, y_j)|x_i \approx y_i\}$ 以及不匹配的跨模态样本对 $\mathcal{D}=\{(x_i, y_j)|x_i \neq y_i\}$, 将其转换为公式 (6) 中的优化问题:

$$\min L(\mathfrak{R}|\mathcal{S}, \mathcal{D}) = f(\mathcal{S}|\mathfrak{R}) + \omega g(\mathcal{D}|\mathfrak{R}) + \lambda r(\mathfrak{R}) \tag{6}$$

其中, f 和 g 为 $\mathcal{S}$ 和 $\mathcal{D}$ 在 $\mathfrak{R}$ 中的损失, r 为正则化项, ω 和 λ 为平衡因子. 或者一起计算 $\mathcal{S}$ 和 $\mathcal{D}$ 在 $\mathfrak{R}$ 中的损失:

$$\min L(\mathfrak{R}|\mathcal{S},\mathcal{D}) = h(\mathcal{S},\mathcal{D}|\mathfrak{R}) + \lambda r(\mathfrak{R}) \tag{7}$$

现有的浅层跨模态学习研究中大部分的跨模态关联模型与公式 (6) 或公式 (7) 具有相似的形式, 根据具体实现形式将其归纳为如下几种类型.

3.1.1 基于统计关联分析的方法

其中最经典的方法是典型相关分析 (canonical correlation analysis, CCA), 该方法利用两组变量之间的相关关系来反映两组指标之间整体相关性, 并通过最大化两组变量之间线性组合的相关性, 将它们投影到同一共同表征空间[62]. 文献 [63] 将 CCA 方法成功引入跨模态检索, 将文本特征和图像特征分别看作不同的特征空间, 通过最大化投影向量的相关性, 得到一个共同表征空间; 文献 [64] 在进行共同空间学习时利用到了多标签标注的高层语义信息, 提出多标签典型相关分析 (multi-label CCA). 受到 CCA 本身的限制, 如公式 (8), 这些方法只考虑公式 (6) 中匹配对象对 $\mathcal{S}$ 的损失 f.

$$f = \frac{\sqrt{\mathcal{W}_{\mathcal{X}}^{\mathrm{T}}\Sigma_{\mathcal{X}}\mathcal{W}_{\mathcal{X}}}\sqrt{\mathcal{W}_{\mathcal{Y}}^{\mathrm{T}}\Sigma_{\mathcal{Y}}\mathcal{W}_{\mathcal{Y}}}}{\mathcal{W}_{\mathcal{X}}^{\mathrm{T}}\Sigma_{\mathcal{X}\mathcal{Y}}\mathcal{W}_{\mathcal{Y}}} \tag{8}$$

其中, $\mathcal{W}_{\mathcal{X}}$ 和 $\mathcal{W}_{\mathcal{Y}}$ 指 $\mathfrak{R}^{\mathcal{X}}$ 和 $\mathfrak{R}^{\mathcal{Y}}$ 的线性组合, $\Sigma_{\mathcal{X}}$、$\Sigma_{\mathcal{Y}}$ 分别是 $\mathfrak{R}^{\mathcal{X}}$ 和 $\mathfrak{R}^{\mathcal{Y}}$ 的协方差矩阵, $\Sigma_{\mathcal{X}\mathcal{Y}}$ 为 $\mathfrak{R}^{\mathcal{X}}$ 和 $\mathfrak{R}^{\mathcal{Y}}$ 的交叉协方差矩阵. 最小化公式 (8) 本质上最大化了 $\mathfrak{R}^{\mathcal{X}}$ 和 $\mathfrak{R}^{\mathcal{Y}}$ 的线性组合的线性相关性, 由多个 $\mathcal{W}_{\mathcal{X}}$ 和 $\mathcal{W}_{\mathcal{Y}}$ 组成的线性映射矩阵定义了如何将 $\mathfrak{R}^{\mathcal{X}}$ 和 $\mathfrak{R}^{\mathcal{Y}}$ 映射到公共空间 $\mathfrak{R}$ 中.

传统的 CCA 方法只能用于发现线性关系, 无法用来寻找更复杂的非线性关系. 为了使 CCA 能够用于发现更为复杂的相关关系, 研究者提出了对 CCA 的非线性拓展, 文献 [65] 提出的核典型相关分析 (kernel canonical correlation analysis, KCCA), 文献 [66] 提出的深度典型相关分析 (deep canonical correlation analysis, DCCA). KCCA 和 DCCA 相当于先将 $\mathcal{S}$ 映射到一个高维空间, 然后在高维空间中执行 CCA. 非线性版本的 CCA 可以发现更为复杂的相关关系[66], 因此可以被用在发现跨模态对象的关系中, 如文献 [67] 和文献 [68] 都提出了基于 KCCA 的跨模态检索模型; 文献 [50] 通过 DCCA 方法来发现文本和图像之间的复杂关联关系. 由于 CCA 的无监督特性, 该方法忽略了跨模态数据中的类别信息, 因此, 文献 [69] 在 CCA 的基础上提出了 3V-CCA 模型, 通过引入类别信息并进行聚类分析来获得更好的投影空间, 文献 [70] 提出了基于聚类的核 CCA 方法, 将基于聚类的 CCA 方法利用核函数实现非线性映射来表达非线性关系. 这些方法相当于将公式 (8) 中的线性映射替换为非线性映射.

偏最小二乘法 (partial least squares, PLS)[71]可以对两组变量进行线性回归建模, 因此也可以用来分析跨模态数据之间的关联关系. 相对于 CCA, PLS 直接进行观测变量到预测变量的映射, 并且更适合预测矩阵比观测矩阵有更多变量的情况. 但是同样由于其基于线性假设, 当相关关系为非线性时难以得到有效的结果[72]. 因此, 一些研究者对 PLS 方法进行拓展使其能够适用于非线性场景, 例如 INLR[73], Kernel PLS[74,75].

3.1.2 基于哈希的方法

由于数据量的快速增长, 在空间中搜索近邻的效率不容忽视. 基于哈希的方法将不同模态的异构数据统一映射为海明空间中的二值向量, 进而可以快速检索相似对象. 传统哈希使用人工设计的特征将数据从原始空间映射至二进制码.

文献 [76] 提出一种多视角哈希方法, 通过最小化相似对象间的距离并最大化非相似对象的距离来产生哈希码. 文献 [77] 提出通过协同正则化来得到一致的多模态数据哈希码. 文献 [78] 首先为每个模态构建各自的海明空间, 然后通过逻辑回归将不同的模态关联起来. 文献 [79] 为跨模态检索提出一种基于联合多模态字典学习的稀疏哈希方法, 首先通过将模态内关系和模态间关系构建为一个超图, 然后通过超图拉普拉斯稀疏编码 (hypergraph Laplacian sparse coding) 学习多模态字典. 文献 [80] 提出一种基于关系的异构哈希, 综合利用了数据特征, 域内同构关系以及域间的异构关系, 首先学习得到同构域内的哈希函数, 然后假设异构域之间的哈希码可以通过回归方法来映射并基于训练数据得到映射关系, 通过哈希码的匹配可以快速检索到可能匹配的实体. 文献 [81] 提出语义关联的多模态哈希算法, 结合图的半监督学习来增强训练样本的语义信息, 构造所有样本的语义关联并保存在哈希函数中, 然后将所有模态映射到统一的哈希空间中, 该方法有效解决了多模态数据表现形式不同和数据规模大

的问题, 缺点是其语义学习的过程拥有较高的计算复杂度. 文献 [82] 通过对有监督、无监督和半监督 3 种基于哈希方法的跨模态表示方法进行研究, 分析了各自的优缺点, 并进行实验对比, 其中有监督的方法因为利用了类别信息能够获得更好的跨模态表示, 但获取大规模的标注数据是困难的, 而对于无监督方法, 充分利用无标注数据的信息是很有意义的.

与其他的跨模态实体分辨方法相比, 基于哈希的方法是对两个海明空间中的表示进行关联. 因此在公式 (6) 的具体实现中通常利用到一些适用于二值变量的方法, 如逻辑回归, 海明距离等. 此外, 其正则项也有其特殊性. 如在文献 [78] 中, 正则项为:

$$r=\lambda_1\sum_{p=1}^{P}\left\|\mathfrak{R}\odot\mathfrak{R}-E\right\|_F^2+\lambda_2\sum_{p=1}^{P}\left\|\mathfrak{R}O\right\|^2+\lambda_3\sum_{p=1}^{P}\left\|\mathfrak{R}\mathfrak{R}^{\mathrm{T}}-m_pI\right\|_F^2 \tag{9}$$

其中, E 为一个全 1 矩阵, O 为一个全 1 向量, I 为单位矩阵. 该正则化项保证了哈希位在全体数据上的平衡, 以及哈希位之间的信息互补并包含最大的信息量.

3.1.3 基于图正则化的方法

图正则化 (graph regularization), 通过探索图中顶点和边丰富的关系, 能够将半监督学习建模为在部分标记图上的标记问题, 边上的权重表示数据间的亲和度, 其目的是对未标记的顶点进行标记; 同时, 图可以用来有效地表示跨模态学习中复杂的关联关系, 例如语义关联、模态内相似度和模态间相似度等[83].

文献 [84] 提出联合图正则化异构度量学习 (joint graph regularized heterogeneous metric learning, JGRHML), 为了更好地利用结构信息, 通过联合图正则化整合不同模态的结构信息; 文献 [85] 将图正则化引入到跨模态检索问题中, 并提出联合表示学习 (joint representation learning, JRL) 方法学习相关性和语义信息的联合表示. 文献 [86] 提出的联合特征选择子空间 (即共同空间) 学习 (joint feature selection and subspace learning, JFSSL) 中, 图正则化被用来保持模态内和模态间的相似度. 同样, 文献 [79] 提出的稀疏多模态哈希中也同样通过图正则化来建模模态内和模态间相似度. 为了解决特征投影时的过拟合问题, 文献 [87] 在多模态子空间学习的目标函数中加入 Dropout 正则化项. 文献 [8] 提出了群组不变跨模态共同空间学习方法, 该方法在学习投影共同空间的同时, 学习不同模态间的群组共生关系, 这种细粒度的语义对应关系提高了潜在共同空间的鲁棒性.

文献 [88] 提出一个基于图的半监督深度模型, 将类标空间作为公共子空间, 通过使用标签图捕获不同模态的内在流形, 使用标签链接损失和图正则化的组合来联合预测未标记数据并有效地学习公共子空间. 不同于文献 [88] 和文献 [89] 首先对未标记数据进行标签初始化, 然后通过文献 [90] 提出的标签噪声更正的方法, 利用已标签数据输入网络进行训练, 利用得到的网络模型预测未标记数据, 通过对数据进行标记, 结合有监督的深度网络模型实现跨模态数据的统一表示.

基于图正则化的跨模态关联方法在公式 (6) 的正则化项 $r(\mathfrak{R})$ 中引入线性映射 W 的每一列 W_i 的平滑函数:

$$\mathcal{S}(W_i)=\sum_j A_{ij}(W_i-W_j)^2=W_i^{\mathrm{T}}LW_i \tag{10}$$

其中, L=D–A 为其图拉普拉斯矩阵 (graph Laplacian matrix), D 为其次数矩阵 (degree matrix), A 为其通过类标定义的邻接矩阵 (adjacency matrix). 平滑函数的引入有利于在原始空间中相似的对象在映射空间中保持较近的距离.

图正则化方法计算复杂度较高并只适用于直推式学习, 在实际应用中受到一定限制[22].

3.1.4 主题模型

为了充分挖掘高层语义信息, 基于主题模型的方法通过生成式模型来挖掘跨模态数据中的隐含主题空间, 将跨模态数据的低层特征映射到一个“隐性语义空间”. 许多跨模态主题模型基于对跨模态主题分布的假设, 例如跨模态主题一一对应, 但是这些假设在实际中并不一定总是成立[22].

文献 [91] 首先将 LDA 主题生成模型推广到跨模态检索中. 文献 [92] 提出了主题回归的多模态 Dirichlet 模型, 首先对每个模态分别学习一个潜在的主题模型, 然后在不同模态之间运用回归的方法建立不同的关联关系; 文献 [93] 在训练目标中引入类标信息, 对其进行了推广, 进一步增强了潜在主题特征的分类能力. 文献 [94] 通过引入下游监督主题模型 (downstream supervised topic model, DSTM), 提出一种监督式多模态互主题强化建模 (multi-

modal mutual topic reinforce modeling, M^{3}R) 方法, 构建联合跨模态概率图模型来发现相互一致的语义主题. 文献 [95] 通过对不同模态数据的类别信息联合建模, 提出了语义主题关联模型, 较其他主题模型更加简单高效. 文献 [96] 定义文档级马尔可夫随机场, 进而学习模态间的共享主题, 解决了当模态间不一一对应时的跨模态文档相似度计算问题.

3.2 深度神经网络方法

深度学习源于人工神经网络的研究, 其目标是建立一个多层的网络结构, 通过逐层表示, 深度挖掘数据的本质信息[94]. 基于深度学习方法的基本思想是利用深度学习的抽取能力, 在低层抽取不同模态的有效表示, 并建立非线性的映射将不同模态数据映射到共同空间中进行数据关联. 文本和图像数据间具有复杂的隐式关联, 在训练数据足够的条件下, 基于深度神经网络的方法可以发现不同模态表征之间更为复杂的关联.

3.2.1 多层前馈神经网络

多层前馈神经网络 (multilayer feedforward neural network), 也称多层感知器 (multilayer perceptron, MLP), 是较为经典的深度学习模型, 广泛地应用于分类等机器学习任务中. 由于非线性映射的引入, 前馈网络具有学习复杂非线性相关关系的能力. 文献 [66] 使用前馈网络替换 CCA 中的线性变换, 使得将 CCA 拓展为具有发现非线性关联能力的深度典型相关分析 DCCA. 文献 [50] 提出一种端到端的方式来构建一个基于 DCCA 的跨模态关联模型, 在深度网络表征的基础上, 通过优化 CCA 目标, 即迹范数 (trace norm) 来学习模型的参数, 并在时空复杂度等方面进行了优化. 文献 [23] 认为尽管深度神经网络具有强大的能力来建模跨模态关系, 但是存在局部最优等问题, 并为此提出了基于预训练和微调的两路神经网络, 其中非线性映射也是通过前馈网络实现的.

文献 [97] 将深度神经网络引入跨模态哈希, 其提出的 CMSSH 结构如图 8 所示.

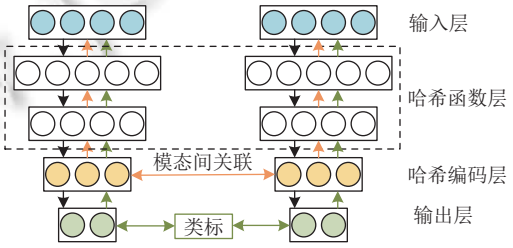

图 8 CMSSH 结构图[97]

CMSSH 使用深度神经网络限制两大训练目标: 针对跨模态相似数据的哈希码与用于预测标签的有区分性的哈希码, 输出层和哈希编码层所得结果与真实结果对比后逐层向上传递对前方的层进行优化. 文献 [98] 提出基于多尺度融合的语义保持哈希算法, 并提出图像特征训练网络和文本特征训练网络的多尺度融合池化模型, 可以提取图像数据集的多尺度特征, 解决文本向量的稀疏性问题. 文献 [99] 提出了基于层次语义交互的深度哈希网络, 首先对网络的每一层应用多尺度和融合操作, 然后, 利用双向哈希交互设计不同层之间的线性交互, 实现不同层之间的语义交互. 针对跨模态方法往往无法很好地对训练过程中不同模态之间的相互影响进行处理, 文献 [100] 基于双向学习提出跨模态检索框架, 方法利用多层监督网络学习所生成异构表征的跨模态关联, 将判别一致性和双向交叉损失函数整合为目标函数进行学习. 知识蒸馏 (knowledge distillation)[101]通过将大模型的监督信息应用至轻量化小模型上, 提高小模型性能和精度, 进而实现模型压缩, 受此启发, 文献 [102] 提出基于语义对齐的无监督知识提取跨模态哈希方法, 利用预训练的无监督教师模型中隐藏的相关信息重构相似度矩阵, 进而指导有监督的学生模型. 文献 [103] 利用老师-学生模型优化跨模态子空间构建, 利用现有较成熟的图像-文本互相生成算法作为老师模型, 指导跨模态学习的学生模型, 以提升模型训练效率并改善最终效果. 文献 [104] 提出基于神经网络的图嵌入学习框架将深度神经网络引入基于图的跨模态方法, 所学得的嵌入结果直接对跨模态一致表达进行近似, 以实现跨模态检索与结合了文本信息的图像分类. 该框架从图模型中抽取所学得表示, 同时在半监督情况下训练分类器. 图卷积神经网络 (graph convolutional network, GCN) 从图数据中提取特征, 进而可以对图数据进行节点分

类、图分类与边预测等操作. 文献 [105] 基于 GCN 提出图特征生成器以及全连接网络, 基于局部图重构节点特征并将两模态的特征映射至公共空间. 由于图卷积网络会受到很多目标的影响, 文献 [106] 提出 DREA 对相似度高的冗余特征进行过滤, 同时对显著目标进行特征增强.

3.2.2 受限玻尔兹曼机

受限玻尔兹曼机 (restricted Boltzmann machine, RBM) 是一种生成式随机神经网络, 普遍应用于降维、分类、特征学习等机器学习任务. 文献 [107] 设计了一种联合表示视频和音频的 RBM 模型, 用于一系列多模态学习任务. 文献 [108] 提出了使用 RBM 学习文本和图像输入的生成模型, 该模型可以为不同模态的数据提供统一的表示并用来进行多模态数据的分类和检索, 具体地, 模型首先学习多模态输入空间的概率密度, 然后同样使用隐变量的状态来表示输入, RBM 学习概率密度的过程也就是寻找模态间关联的过程. 文献 [109] 提出混合表示学习, 利用堆叠受限玻尔兹曼机 (stacked restricted Boltzmann machine, SRBM) 为每个模态提取模态友好的表达, 该表达之间的相似度就统计特性而言优于来自两模态的原始输入数据.

3.2.3 自编码器

自编码器 (auto-encoder) 作为编码-解码器的特殊形式, 其结构如图 9 所示.

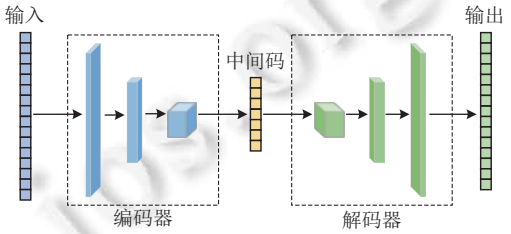

图 9 自编码器结构

自编码器由编码器和解码器两个部分组成, 通过最小化重建损失来完成降维等任务. 文献 [110] 使用堆叠自编码器学习文本和视觉输入的高层映射. 文献 [111] 同样利用堆叠自编码器来获得跨模态检索中的模态内和模态间关系. 文献 [112] 和文献 [113] 提出基于跨模态对应自编码器 (correspondence auto-encoder, Corr-AE) 的深度学习模型, 该模型通过最小化单模态自编码器的重构误差和不同模态表示层的相关性误差之和, 从而将单模态表示学习和模态间的相关性学习集成到一个框架下. 文献 [114] 使用去噪自编码器保持原始表示中的信息并去除噪声, 通过最小化重建损失和关联损失来构建多模态共同空间, 该空间可以同时保持模态内距离和模态间距离. 文献 [115] 和文献 [116] 在估计“文本-图像”的跨模态关系时也使用自编码器来获得文本和图像的联合嵌入. 文献 [117] 提出多模态卷积自编码器方法 MUCAE, 对每个模态, 将卷积操作整合入自编码器框架, 以学习原始图像和文本内容的联合表示. 通过开发卷积自编码器所得隐藏表示间的关系, 对不同模态的卷积自编码器进行优化. 文献 [118] 将自编码器与生成对抗网络相结合, 以联合合并公共潜在子空间学习、知识传递以及特征合成, 自编码器还被用于将所有多模态数据映射至所学得的潜在空间.

3.2.4 注意力机制

注意力机制 (attention mechanism) 源于人类大脑在面向大量信息时能聚焦并选择信息的能力[119], 深度学习中的注意力机制本质上讲和人类的选择性视觉注意力机制相类似, 其核心目标是从众多信息中选择出对于当前任务目标更关键的信息. 文本处理领域注意力机制示意图如图 10.

图 10 中, 给定固定的元素集合由一系列的<元素, 数值> 的数据对构成, 为计算目标元素的注意力数值, 分别计算目标元素与给定元素集合中每个元素的相似度来作为对应数值的权重系数, 最后通过元素集合中所有元素数值的加权求和来得到目标元素的注意力数值. 元素集合中元素的权重代表了信息的重要性, 某一元素的权重越大, 表示目标元素的信息越聚焦于该元素. 这种思想能够让注意力机制在帮助解决多模态任务 (如跨模态数据实体分辨, 图像理解) 时通过细粒度的跨模态关联取得更好的结果, 其目的是让模型能够关注特征中的关键部分. 根据不同的权重更新方式, 注意力机制可被分为自注意力机制和联合注意力机制.

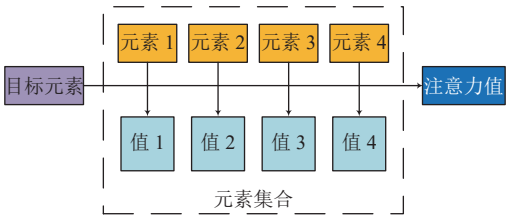

图 10 注意力机制示意图

自注意力机制更新模态表示权重的计算来源于模态本身, 广泛用于解决单模态数据的分类问题. 文献 [120] 通过引入自注意力机制, 提出了一种句向量表示方法, 实验表明该方法在情感分类、文本蕴含等文本任务上取得较好的结果. 为了在图像表示的过程中关注到图像中令人感兴趣的部分, 文献 [121] 提出了卷积块注意力模型 (convolutional block attention module, CBAM), 能够在利用 CNN 方法提取图像特征时, 通过计算通道 (channel) 注意力和空间 (spatial) 注意力权重改进图像的特征表示, 实验结果证明该方法能够改进很多基线方法的精确率.

联合注意力机制更新其中一个模态表示权重的计算一般来源于另一个模态. 文献 [122] 在生成对抗网络的基础上, 通过引入视觉-语义的注意力机制进行多模态的联合表示, 在多类别分类问题和标签推荐问题上均取得较优的结果. 文献 [123] 提出了基于栈的联合注意力网络来探索不同模态特征之间的语义相关性. 文献 [124] 将表示学习和相似性度量结合, 提出了一个端到端的跨模态相似性度量方法, 使用 CNN 网络完成特征提取并结合 LSTM 深度网络实现联合注意力机制, 完成不同模态独立语义空间的构建. 为了解决基于注意力的方法计算量较大的问题, 文献 [125] 利用联合注意力机制对图像和句子进行联合表示, 通过对句子粒度的划分 (词级别、短语级别和句子级别) 实现不同粒度的跨模态对齐, 提高图像-句子检索的准确性, 该模型能够在端到端的条件下进行训练.

文献 [126] 指出, 在引入注意力之前, 前人所提出的跨模态哈希方法在生成哈希码时会用到对象的所有信息. 这将导致重要性低甚至无用的信息被引入所生成哈希码, 降低哈希码的紧凑性和区分能力. 针对此问题, 文献 [127] 将注意力机制引入跨模态哈希, 通过生成注意力遮罩, 筛选出被“注意到”的特征表示, 从而有选择地关注多模态数据中更能提供有用内容的信息. 文献 [124] 首次将自注意力机制引入跨模态哈希, 在全局视角增强来自网络不同层的哈希表示中重要部分的显著程度, 同时将自注意力机制引入对抗学习. 文献 [128] 使用局部和全局双重注意力机制, 局部注意力机制用于提取每个模态局部的关键信息, 并提高模态特征的表达能力; 全局注意力机制用于增强不同模态间的联系, 进而生成更具一致性且准确的哈希码. 文献 [129] 提出堆叠多模态注意力网络, 利用堆叠注意力以充分利用图像-文本间的细粒度互相依赖关系.

受注意力机制启发, 文献 [130] 提出图注意力网络, 用以学习节点与其近邻之间的重要性, 对节点进行分类时则将近邻进行合并. 文献 [131] 综合了自注意力网络和图注意力网络, 前者用于捕获模态间特征级关联, 后者聚合不同模态间匹配项嵌入内容以辅助构建公共嵌入空间.

文献 [29] 认为由于不同模态中信息是不平衡的, 如果直接将不同模态的数据映射到一个公共空间中会丢失每个模态各自的特性, 因此首先利用循环注意力网络充分挖掘模态内特性, 得到图像 x_i 的划分 $x_i^{(k)}$ (k=1, …, n), 以及文本 y_j 的表示 $\sigma(y_j)$, 然后可以通过注意力权重 ω_k 计算它们之间的相似性, 如公式 (11):

$$sim(x_i, y_i) = \sum_{k=1}^{n} \omega_k x_i^{(k)} \cdot \sigma(y_i) \tag{11}$$

其中, ω_k 的权重通过最小化公式 (5) 的成对损失计算, 具体的形式为:

$$h(\mathcal{S}, \mathcal{D}|\Re) = \sum_{(x_i,y_i)\in\mathcal{S},(x_i,y_c)\in\mathcal{D}} \max(0, \theta - sim(x_i, y_i) + sim(x_i, y_c)) + \sum_{(x_i,y_i)\in\mathcal{S},(x_c,y_i)\in\mathcal{D}} \max(0, \theta - sim(x_i, y_i) + sim(x_c, y_i)) \tag{12}$$

其中, θ 表示边界.

3.2.5 对抗式网络

以生成对抗网络 (generative adversarial network, GAN) 为代表的对抗网络 (adversarial networks) 模型, 被广泛地运用到了计算机视觉领域 (图像生成等), 它通过生成器和判别器的对抗机制来生成“真假图像”的一致表示[132].

生成对抗网络核心部分包括一个生成器和一个判别器, 生成器根据输入的随机噪声产生样本交由判别器进行判断, 根据判别器所得结果计算出损失函数, 而后根据损失函数调整生成器和判别器, GAN 结构如图 11 所示.

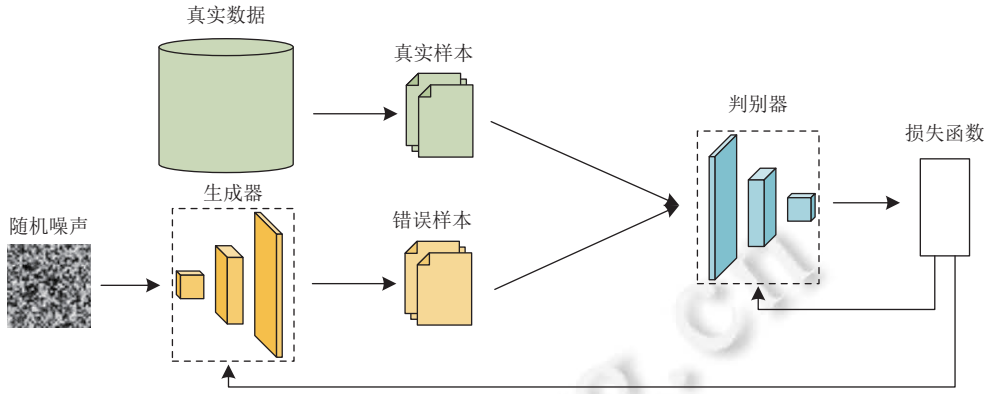

图 11 生成对抗网络结构

近年来, 使用对抗学习来获得跨模态数据之间更一致的表示得到广泛关注. 文献 [35] 考虑到在构建跨模态检索学习模型时需要的模态内和模态间关联关系, 设计了一种由特征映射器和模态分类器组成的对抗式学习框架, 利用类别信息实现模态间的对齐关联, 取得了较好的跨模态相似性度量. 文献 [133] 提出一个新的 GAN 框架, 设计了双层对抗辨别器来实现域变换并获得域不变表示, 完成图像之间的转换. 文献 [134] 提出跨模态生成式对抗网络 (cross-modal GAN, CM-GAN) 解决跨模态检索中的异构性问题, 通过对模态间关联和模态内重建信息建模, 该方法可以有效地估计异构数据的联合分布. 此外, 条件生成式对抗网络 (conditional generative adversarial network, CGAN) 也被用在跨模态数据的互相生成中. 如文献 [135] 提出 TAC-GAN, 根据文本描述产生相应的图像; 文献 [136] 提出 CMCGAN, 该模型支持视频和声音的互相生成. 如同经典的 GAN (文献 [123]) 一样, 上述这些方法都通过 MinMax 博弈来逼近跨模态数据的联合分布. 文献 [137] 提出无监督的跨模态语音和文本联合嵌入对齐方法, 利用生成对抗网络学习语音模态到文本模态的非线性映射, 并通过构建合成平行词典表达语音和文本之间的关联关系, 取得了较好的实体分辨结果. 除此之外深度信念网络 (deep belief network, DBN)[138]等结构也被应用于跨模态学习, 但并非主流. 文献 [139] 提出了跨模态多层深度网络模型 CMDN, 该模型通过联合保留模态内和模态间的信息, 为每种模态的数据生成互补的表示, 然后按层次组合, 通过堆叠学习的方式学习公共空间.

3.2.6 跨模态数据实体分辨的深度神经网络框架

以上讨论中, 研究者提出各式各样的模型来在构建共同空间时对模态内和模态间关联关系进行保持和建立, 从中可以发现无论是模态内关系还是模态间关系的建模过程, 都正在受益于深度学习技术的发展. 综合已有工作, 可以将基于深度神经网络的跨模态数据实体分辨模型归纳为如图 12 所示的架构.

图 12 中, 首先通过在大规模图像数据集上预训练的卷积神经网络提取视觉特征, 同文本特征分别输入多层前馈神经网络中从而实现对特征的非线性映射. 前馈神经网络的优化通过最小化模态内损失和模态间损失实现. 图中模型训练好之后, 可以直接通过文本和图像共同空间来计算其相似度, 进而实现跨模态实体分辨.

深度学习技术确实提高了跨模态数据实体分辨的精确度, 在引入深度神经网络后, 人类先验经验的作用越来越被弱化, 建立并训练一个端到端的神经网络似乎成了“万能”的做法. 这种方式避免了不准确先验知识带来的偏差, 但是“天下没有免费的午餐”, 至少在当前技术条件下, 只依赖端到端的深度神经网络模型并不能解决跨模态学习任务中的所有问题. 其主要原因是不同模态信息表征方式的差别过大, 自然语言处理中尝试捕捉的是离散单词的序列信息, 而计算机视觉中利用到的更多是全局信息. 为了同时训练文本、图像处理模型, 以及二者的关联模型需要大量的配对训练数据, 而这种训练数据的获取成本很高. 因此尽管有研究者, 如文献 [50] 中基于深度网络提出完全端到端的学习架构进行相似度计算, 但是性能却并不理想.

图 12 中, 深度神经网络被引入跨模态数据实体分辨之后, 同时在特征提取, 语义嵌入以及跨模态关联等方面发挥作用. 但是深度神经网络是如何在跨模态数据实体分辨的各个阶段中发挥作用的, 在哪个阶段引入的深度神

经网络对实体分辨效果提升最有意义, 现有研究并没有清楚地回答这些问题. 而这些问题对进一步优化现有架构, 提升训练样本利用率等方面具有重要意义.

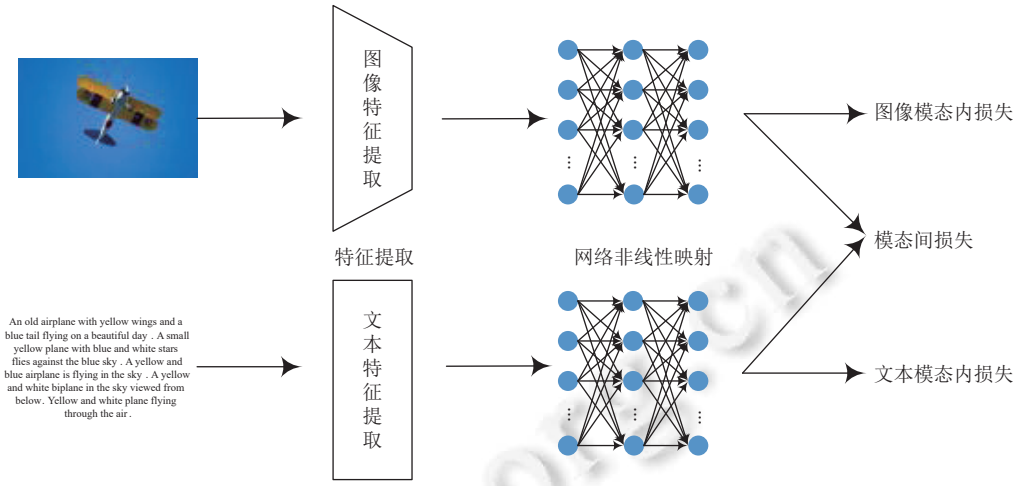

图 12　跨模态数据实体分辨的深度神经网络框架

4　实验分析

为了分析各种因素对跨模态方法性能的影响, 本章对经典的跨模态学习方法进行了分析. 实验主要关注如下方面: 采用对比实验的手段, 对深度神经网络在跨模态相似性连接问题中发挥的作用进行分析; 通过对比不同设定下各种方法的性能, 证明“抽象”和“关联”两种机制对提高跨模态相似性连接的重要作用; 通过对比不同数据集对跨模态方法准确性的影响, 讨论数据集中数据分布等特点对跨模态方法性能的影响.

4.1　数据集及特征提取

Wikipedia[63]: 收集自维基百科中的 29 个类别的 2 700 篇文章, 每篇文章都包含一张或者多张来自维基共享资源 (Wikimedia commons) 的图片. 由于有些类别样本较少, 因此只采用样本最多的 10 个类别. 每篇文章根据标题被划分为多个章节, 每个章节配以一副所在位置处的图片. 最终得到 10 个类别, 2 866 个文本-图像对.

Pascal-Sentences[140]: PASCAL VOC 数据集的子集, 包含来自 20 个类别的 1 000 张图片 (每个类别 50 张), 并且每张图片配有一段文本描述 (通常是 5 个独立的语句).

XMedia[85]: 由 5 种类型的数据 (文本、图像、视频、音频和 3D 模型) 组成的公开数据集. 本文仅使用其中的图像和文本数据, 即来自 20 个类别的 5 000 对图片和文本.

为了验证深度神经网络在单模态抽象中特征提取阶段发挥的作用, 实验中采取对每个数据集采取了 3 种特征提取的方案: 低层图像特征 (通过 SIFT 方法提取)+低层文本特征 (通过 BoW 方法提取)、深度图像特征 (通过 VGG19 预训练模型第 2 个全连接层提取)+低层文本特征 (BoW 特征)、深度图像特征 (VGG19)+高层文本特征 (通过 SIF 方法提取). 训练集和测试集按照 80% 和 20% 的比例划分.

4.2　对比方法和评价指标

(1) 对比方法

CCA[63]: 利用典型相关分析, 找到具有最大相关性的线性组合, 利用该线性组合将不同模态对象映射到共同空间中.

JFSSL[86]: 利用图正则化来保持模态内相似度和模态间相似度.

HSNN[141]: 跨模态相似度是由两个跨模态对象属于同一语义范畴的概率来度量的, 通过分析每个对象的模态内近邻来实现.

CMCP[142]: 一种考虑多模态对象间正、负向关联的多模态关联传播算法.

JRL[85]: 通过半监督正则化和稀疏正则化, 利用语义信息学习公共空间.

ACMR[47]: 通过对抗学习的方法综合考虑模态内关系和模态间关系.

CDPAE[114]: 综合距离保持自编码器 (comprehensive distance-preserving autoencoders).

这些方法中, CCA、CMCP、JFSSL、JRL 为非深度学习方法, HSNN、ACMR 和 CDPAE 为基于深度学习的方法.

(2) 评价指标

实验覆盖两种跨模态实体分辨任务, 即由图像检索文本 (image to text, I2T) 和由文本检索图像 (text to image, T2I). 实验以结果精确率作为指标, 验证各项因素对跨模态方法性能的影响, 因此选择使用最多的平均精确度 (mean average precision, MAP) 来度量跨模态实体分辨结果的精确度.

4.3 实验结果

表 1–表 3 分别为常见基准方法在 Wikipedia、Pascal-Sentences 以及 Xmedia 数据集上的性能对比.

表 1 常见基准方法在 Wikipedia 数据集的 *MAP* 对比 (*表示基于机器学习的方法)

数据集	任务	CCA	HSNN*	JFSSL	CMCP	JRL	ACMR*	CDPAE*
Wikipedia-Shallow (SIFT+BoW)	I2T	0.1194	0.1194	0.1192	0.1197	0.1193	**0.1297**	0.1139
	T2I	0.1197	0.1187	0.1193	0.1188	0.1194	**0.1318**	0.1187
	平均	0.1196	0.1191	0.1193	0.1193	0.1194	**0.1306**	0.1163
Wikipedia-CNN (VGG19+BoW)	I2T	0.1460	0.3299	0.1372	0.3397	0.3342	**0.3597**	0.3430
	T2I	0.1273	0.3450	0.1186	0.3494	0.3350	0.3554	**0.3662**
	平均	0.1337	0.3375	0.1279	0.3447	0.3346	**0.3576**	0.3546
Wikipedia-CNN+SIF (VGG19+SIF)	I2T	0.1106	0.4604	0.1129	0.4624	0.4717	0.4897	**0.4953**
	T2I	0.1093	0.4103	0.1081	0.4233	0.4151	**0.4306**	0.4216
	平均	0.1010	0.4354	0.1105	0.4429	0.4434	**0.4602**	0.4585

表 2 常见基准方法在 Pascal-Sentences 数据集的 *MAP* 对比 (*表示基于机器学习的方法)

数据集	任务	CCA	HSNN*	JFSSL	CMCP	JRL	ACMR*	CDPAE*
Pascal-Sentences-Shallow (SIFT+ BoW)	I2T	0.0789	0.0800	0.0785	0.0815	0.0774	**0.1260**	0.1080
	T2I	0.0789	0.0776	0.0752	0.0839	0.0774	**0.1305**	0.1205
	平均	0.0789	0.0783	0.0769	0.0827	0.0774	**0.1283**	0.1197
Pascal-Sentences-CNN (VGG19+BoW)	I2T	0.0795	0.1790	0.0810	0.1641	0.1730	0.2455	**0.2555**
	T2I	0.0789	0.2081	0.0811	0.2073	0.2111	0.2130	**0.2162**
	平均	0.0792	0.1936	0.0811	0.1857	0.1920	0.2293	**0.2359**
Pascal-Sentences-CNN+SIF (VGG19+SIF)	I2T	0.0748	**0.4675**	0.0754	0.4579	0.4304	0.4561	0.4436
	T2I	0.0750	**0.4701**	0.0750	0.4583	0.3881	0.4505	0.4506
	平均	0.0749	**0.4688**	0.0752	0.4581	0.4093	0.4533	0.4471

表 3 常见基准方法在 XMedia 数据集的 *MAP* 对比 (*表示基于机器学习的方法)

数据集	任务	CCA	HSNN*	JFSSL	CMCP	JRL	ACMR*	CDPAE*
XMedia-Shallow (SIFT+ BoW)	I2T	0.1220	0.1841	0.0981	**0.2008**	0.1856	0.1956	0.1255
	T2I	0.1209	0.2069	0.0696	**0.2167**	0.1926	0.2026	0.0715
	平均	0.1215	0.1955	0.0839	**0.2088**	0.1891	0.1991	0.0985
XMedia-CNN (VGG19+BoW)	I2T	0.0678	0.3079	0.0646	0.3248	0.3059	**0.3595**	0.1579
	T2I	0.0604	0.4043	0.0579	0.3970	0.3959	**0.4078**	0.0721
	平均	0.0641	0.4011	0.0612	0.3609	0.3509	**0.3836**	0.1105
XMedia-CNN+SIF (VGG19+SIF)	I2T	0.0486	0.5130	0.0465	0.5269	0.6768	**0.7068**	0.5534
	T2I	0.0464	0.5288	0.0465	0.5309	0.6663	**0.6836**	0.5026
	平均	0.0475	0.5209	0.0465	0.5289	0.6716	**0.6952**	0.5280

4.4 结果分析

(1) 特征提取方法影响

对图像特征, 通过对比表 1–表 3 这 3 个数据集中每个方法在低层视觉特征和深度视觉特征上的精确度可以发现, 对除了 CCA 以外的所有方法来说, 引入通过预训练模型提取的视觉特征可以极大提高方法的精确度, 最高提高的 *MAP* 值达 0.6. 获得提升的方法中既有基于深度学习的方法, 也有非深度学习方法. 而 CCA 方法的 *MAP* 值却反而下降了, 原因可能是由于深度视觉特征维度太高 (4 096 维) 且十分稀疏, CCA 寻找其中的线性相关性太过困难.

对文本特征, 在同样使用 CNN 深度视觉特征的条件下, 同低层 BoW 文本特征相比, 大部分方法在使用了 SIF 文本特征后精确度都得到了显著提高. 但是在不同的数据集上, SIF 特征的表现并不相同. 例如在 Wikipedia 数据集上, SIF 特征带来的提升明显低于其他两个数据集, 这是由于该数据集的文本是篇章级的, 而 SIF 特征主要用于句子的表示. 此外, BoW 方法在 WikiPedia 数据集上效果优于 Pascal, 这是因为 BoW 方法本质是基于统计的方法, 对样本数量有一定要求, Pascal 中描述文本均为较短的句子, 词袋模型在这种情况下表征能力不如其在文本更长的 Wikipedia 上强.

通过对不同图像和文本特征提取方法的对比和分析我们可以得出结论: 特征提取的能力将影响跨模态方法的最终效果.

(2) 跨模态映射函数影响

在 ACMR 和 CDPAE 等方法中通过 tanh 非线性变换实现特征空间到公共表征空间的映射, 而 JRL、CMCP 等则使用线性变换作为跨模态映射函数, 通过实验可以看出, 在同样的特征表示前提下, 非线性变换相对于线性变换没有明显的优势. 其原因在于, 当语义嵌入部分能够较好地提升单模态表征的语义层次时, 不同模态之间的关联变得更为清晰, 理想的条件下甚至是简单的线性相关.

(3) 损失函数和优化方法影响

跨模态相似性连接的优化目标包含模态内损失和模态间损失, 二者并非一致, 因此如何进行优化也是重要的问题. 大部分方法通过权重来平衡二者, 而 ACMR 方法使用了对抗学习中的 MinMax 优化策略对其进行求解, 通过观察实验结果可以发现, 尽管提升较为有限, 但对抗式的优化方式对于跨模态相似性连接精确度的提高是有意义的.

(4) 数据集影响

对 3 种数据集原始数据进行分析: Wikipedia 数据集文本量丰富但部分数据文本-图像相关性隐藏较深, 例如对 1666 年伦敦大火的描述文本, 其对应图片为通过回忆描述了这场大火的作家约翰·伊夫林的肖像; Pascal 数据集中文本是对图像的直接描述, 每张图片用一句话从不同角度描述 5 次, 相关性较强但长度短. XMedia 数据集中文本源于 Wikipedia 数据集而图像源于 Flickr 数据集, 经筛选后其图像和文本关联性更强更明显, 同时保留了 Wikipedia 数据集文本量丰富的特点.

通过相同方法在不同数据集上的对比可以发现, 不同数据集对跨模态方法最终结果也存在影响. CCA 在 Wikipedia 数据集上表现优于 Pascal 数据集, 由于图像模态往往不会存在太大差异且跨模态数据集均经由人工构建, 数据分布平衡性有一定保障, 因此原因应在于文本模态. 如前文中所述, Wikipedia 中的文本模态是篇章级而 Pascal 中文本模态为句子级, CCA 本身所用文本特征提取方法 LDA 是典型 BoW 模型, 其目的是给出一篇文档主题的概率分布, 因而 CCA 对长文本的效果要优于短文本. 大多数方法在 XMedia 数据集表现均优于 Wikipedia 数据集和 Pascal 数据集, 其原因是 XMedia 文本和图像之间显示相关性更强, 同时文本量充足, 文本特征提取方法能够提取出更优秀的特征供后续跨模态方法使用.

通过整体对比分析发现, ACMR 在 Wikipedia 和 XMedia 数据集上表现较其他方法更好, 这是由于 ACMR 通过对抗学习优化共同子空间中表征的模态不变性, 使得其能够更充分地利用数据对自身进行训练. 而 HSNN 在 Pascal 上表现较其他方法更好, 这可能由于 HSNN 通过计算不同模态实体是否属于同一语义分类以进行跨模态检

索, Pascal 数据集中文本描述“短而精”的特点刚好适合 HSNN 进行处理.

通过以上 4 个方面的分析可以发现跨模态实体分辨近年来的进展主要受益于以下 3 个方面: (1) 机器视觉和自然语言处理领域技术的单点突破提供了图像和文本数据良好特征表示; (2) 基于合理的假设对单个模态的特征进行高层次抽象的思路, 即使在引入深度神经网络后, 仍然具有意义; (3) 对抗式的优化方法为同时优化模态内损失和模态间损失提供了新的思路. 此外数据集中数据分布会影响跨模态方法最终结果, 模态间显示关联越强, 数据越充足, 则跨模态方法效果越好, 同时不同方法最适合数据集也存在不同.

5 存在的问题与未来展望

跨模态实体分辨发展至今, 相较于最开始提出时已有长足进步, 取得了诸多瞩目的成果. 但目前研究成熟度尚有不足, 仍存在较大的进步空间. 同时, 随着跨领域等问题的出现, 也在给跨模态实体分辨技术提出新的挑战. 本章总结了一些尚待解决的问题, 可作为未来研究的方向指引.

(1) 弥合异构鸿沟仍具挑战性. 异构鸿沟是跨模态方法所面临的主要问题之一, 如何弥合异构鸿沟一直是研究的主要方向. 尽管研究人员已在其上取得了令人瞩目的成就, 异构鸿沟仍然横亘在人们面前, 阻碍着跨模态技术的发展. 如何更有效地解决异构鸿沟带来的问题, 将继续作为研究重点方向为研究人员所关注.

(2) 与上游领域最新进展结合不足. 近年来, 基于 Transformer 的 BERT 模型及其变体在文本特征提取上大放异彩, 同时 Transformer 在计算机视觉方向也逐渐崭露头角, 而跨模态研究对这些上游领域的最新的方法应用却不甚积极. 本文第 4 节通过实验证明, 特征提取方法能够对跨模态实体分辨性能产生重要影响, 因此将具有强大特征提取能力的 Transformer 应用至跨模态研究中将是未来的可能方向之一.

(3) 专业领域应用研究方兴未艾. 目前图像-文本跨模态方法多关注于自然图片与对应文本描述, 对专业领域的研究刚刚兴起. 如医疗领域, 目前多用跨模态技术提高医学图像分割方法性能, 且方法多集中于同为视觉模态的不同诊断器材所产生的结果上. 若能更好地将跨模态技术应用于医疗与工业生产等领域之中, 将有效推动社会向前发展.

(4) 实际场景与训练场景存在差异. 例如跨模态技术一大应用领域为社交媒体, 社交媒体的用户每时每刻都在上传海量多模态数据, 对这些多模态数据的应用一直是领域研究热点. 但是由于用户描述、拍摄等能力参差不齐, 添加的图片描述标签可能并不完整. 而跨模态技术往往假定所面对数据完备且充足, 这对跨模态技术在社交平台的应用造成了限制. 如何解决训练场景和实际场景之间差异性所带来的问题将成为未来的重点方向.

(5) 许多模态数据噪声较多. 由于摄像头与话筒等采集设备技术限制, 以及采集场景不理想等原因, 所采集的模态数据往往存在噪音, 这将对信息的提取造成影响. 如何提高面对噪声时的鲁棒性, 现有研究已取得一定成果, 未来亦会有更多研究关注此方向.

(6) 视觉模态无关内容多. 视觉模态中会包含丰富的实体与信息, 但所关注的重点内容往往只有其中几项, 其余非重点内容在某些情况甚至会成为干扰. 例如, 一张关于烹饪人员正在切菜的图片及其对应文本描述, 如果视觉模态方法错误地认为图片中出现的其他厨具内容为重点, 将导致跨模态方法性能降低甚至失效. 未来研究需关注如何在语义层面解决无关信息的干扰问题.

(7) 数据集匮乏. 数据集匮乏体现在两个方面, 专业领域数据集匮乏与一般领域大规模数据集匮乏. 专业领域跨模态研究如医疗, 所用数据集往往为研究人员于医疗机构采集所得, 而匮乏联合的数据集, 可能导致数据具有地方差异, 进而影响算法的效果. 一般领域所用数据集如 Wiki, 所含文档近 3 000 条, 不充分不平衡的数据难以充分发挥出深度神经网络的真正优势. 以迁移学习为代表的研究关注于如何利用好现有数据, 取得了不错的成果[143], 但大规模数据集的匮乏问题仍亟待解决.

(8) 在物联网领域应用不足. 5G 与物联网技术将大量各类型传感器连接起来, 实现实时上传, 带来规模更大, 范围更广泛的多模态数据. 如果能够将跨模态技术引入, 将极大提高生产生活便利性, 提高效率. 然而, 由于物联网传感器广泛分布所带来的在利用数据时可能对人隐私造成侵害等问题, 跨模态技术在物联网上的发展尚未完全兴

起. 未来, 如何解决隐私等问题, 将跨模态技术与物联网相结合应成为一个需要关注的方向.

6 结束语

日益丰富的文本、图像、音频、视频等多媒体数据, 使广大用户面临跨模态数据应用需求, 跨模态数据实体分辨是大数据处理和分析面临的基础性问题之一. 本文从关系型数据实体分辨入手, 对跨模态实体分辨问题的研究进展进行回顾, 首先介绍了问题的定义、评价指标; 然后, 以模态内关系的保持和模态间关系的建立为主线, 对现有研究进行总结和梳理; 并且, 在多个公开数据集上对主流方法关于不同图像、文本特征提取方法进行对比实验, 并对其间差异的原因进行了分析; 最后, 总结当前研究尚存在的问题, 并基于此给出未来可能的研究方向.

References:

[1] Madnick SE, Wang RY, Lee YW, Zhu HW. Overview and framework for data and information quality research. Journal of Data and Information Quality, 2009, 1(1): 2. [doi: 10.1145/1515693.1516680]

[2] Zhou YL, Talburt JR. Entity identity information management (EIIM). In: Proc. of the 16th Int'l Conf. on Information Quality (ICIQ2011). Adelaide: University of South Australia, 2011.

[3] Karapiperis D, Gkoulalas-Divanis A, Verykios VS. Efficient record linkage in data streams. In: Proc. of the 2020 IEEE Int'l Conf. on Big Data. Atlanta: IEEE, 2020. 523–532. [doi: 10.1109/BigData50022.2020.9378127]

[4] Christen P. A survey of indexing techniques for scalable record linkage and deduplication. IEEE Trans. on Knowledge and Data Engineering, 2012, 24(9): 1537–1555. [doi: 10.1109/TKDE.2011.127]

[5] Hung SK. High performance record linkage [Ph.D. Thesis]. University Park: The Pennsylvania State University, 2010.

[6] Loureiro D, Mário Jorge A, Camacho-Collados J. LMMS reloaded: Transformer-based sense embeddings for disambiguation and beyond. Artificial Intelligence, 2022, 305: 103661. [doi: 10.1016/j.artint.2022.103661]

[7] Peng YX, Qi JW, Huang X. Current research status and prospects on multimedia content understanding. Journal of Computer Research and Development, 2019, 56(1): 183–208 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2019.20180770]

[8] Liang J, He R, Sun ZN, Tan TN. Group-invariant cross-modal subspace learning. In: Proc. of the 25th Int'l Joint Conf. on Artificial Intelligence. New York: IJCAI/AAAI Press, 2016. 1739–1745. [doi: 10.5555/3060832.3060864]

[9] Cao JJ, Diao XC, Wang T, Wang FX. Research on domain-independent data cleaning: A survey. Computer Science, 2010, 37(5): 26–29 (in Chinese with English abstract).

[10] Wang QL, Guo YF, Yu LX, Chen XH, Li P. Deep *Q*-network-based feature selection for multisourced data cleaning. IEEE Internet of Things Journal, 2021, 8(21): 16153–16164. [doi: 10.1109/JIOT.2020.3016297]

[11] Cao JJ, Diao XC, Du Y, Wang FX, Zhang XY. Classification detection of approximately duplicate records based on feature selection using ant colony algorithm. Acta Armamentarii, 2010, 31(9): 1222–1227 (in Chinese with English abstract).

[12] Liu Y, Diao XC, Cao JJ, Zhou X, Shang YL. A method for entity resolution in high dimensional data using ensemble classifiers. Mathematical Problems in Engineering, 2017, 2017: 4953280. [doi: 10.1155/2017/4953280]

[13] Tan MC, Diao XC, Cao JJ. Survey on entity resolution. Computer Science, 2014, 41(4): 9–12, 20 (in Chinese with English abstract). [doi: 10.3969/j.issn.1002-137X.2014.04.002]

[14] Yu RH, Tian ZP, Zhou AY. A synthetical approach for detecting approximately duplicate database records of multi-language data. Computer Science, 2002, 29(1): 118–121 (in Chinese with English abstract). [doi: 10.3969/j.issn.1002-137X.2002.01.036]

[15] Tan MC, Diao XC, Cao JJ, Zhou X, Liu Y, Zheng Q. Inverted index and space mapping based redundancies eliminating for data blocking in entity resolution. Journal of Computational Information Systems, 2015, 11(17): 6187–6198. [doi: 10.12733/jcis15067]

[16] Kalashnikov DV, Nuray-Turan R, Mehrotra S. Adaptive connection strength models for relationship-based entity resolution. ACM Journal of Data and Information Quality, 2012, 4(2): 1–22.

[17] Kalashnikov DV, Mehrotra S. Exploiting relationships for data cleaning. In: Proc. of the 2005 SIAM Int'l Conf. on Data Mining (SDM). 2005. 262–273.

[18] Tan MC, Diao XC, Cao JJ. Relationship type based connection strength model for relation-ship-based entity resolution. Journal of Computational Information Systems, 2015, 11(16): 5947–5957.

[19] Shang YL, Cao JJ, Li HM, Zheng QB. Co-author and affiliate based name disambiguation approach. Computer Science, 2018, 45(11): 220–225, 260 (in Chinese with English abstract). [doi: 10.11896/j.issn.1002-137X.2018.11.034]

[20] Zhang DX, Li DS, Guo L, Tan KL. Unsupervised entity resolution with blocking and graph algorithms. IEEE Trans. on Knowledge and

Data Engineering, 2022, 34(3): 1501–1515. [doi: 10.1109/TKDE.2020.2991063]

[21] Yu SQ. Entity resolution with recursive blocking. Big Data Research, 2020, 19–20: 100134. [doi: 10.1016/j.bdr.2020.100134]

[22] Peng YX, Huang X, Zhao YZ. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. IEEE Trans. on Circuits and Systems for Video Technology, 2018, 28(9): 2372–2385. [doi: 10.1109/TCSVT.2017.2705068]

[23] Mollell G, Moens MF. Do neural network cross-modal mappings really bridge modalities? In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018. 462–468. [doi: 10.18653/v1/P18-2074]

[24] Wei YC, Zhao Y, Lu CY, Wei SK, Liu LQ, Zhu ZF, Yan SC. Cross-modal retrieval with CNN visual features: A new baseline. IEEE Trans. on Cybernetics, 2017, 47(2): 449–460. [doi: 10.1109/TCYB.2016.2519449]

[25] Harris ZS. Distributional structure. Word, 1954, 10(2–3): 146–162.

[26] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv:1301.3781, 2013.

[27] Pennington J, Socher R, Manning C. GloVe: Global vectors for word representation. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing. Doha: Association for Computational Linguistics, 2014. 1532–1543. [doi: 10.3115/v1/D14-1162]

[28] Arora S, Liang YY, Ma TY. A simple but tough-to-beat baseline for sentence embeddings. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017. 1–16.

[29] Socher R, Karpathy A, Le QV, Manning CD, Ng AY. Grounded compositional semantics for finding and describing images with sentences. Trans. of the Association for Computational Linguistics, 2014, 2: 207–218. [doi: 10.1162/tacl_a_00177]

[30] Mao JH, Wei X, Yang Y, Wang J, Huang ZH, Yuille AL. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In: Proc. of the 2015 Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 2533–2541. [doi: 10.1109/ICCV.2015.291]

[31] Qiao YL, Lai YK, Fu HB, Gao L. Synthesizing mesh deformation sequences with bidirectional LSTM. IEEE Trans. on Visualization and Computer Graphics, 2022, 28(4): 1906–1916. [doi: 10.1109/TVCG.2020.3028961]

[32] Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. In: Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans: Association for Computational Linguistics, 2018. 2227–2237. [doi: 10.18653/v1/N18-1202]

[33] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: ACM, 2014. 3104–3112.

[34] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: ACM, 2017. 6000–6010.

[35] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2018. 4171–4186. [doi: 10.18653/v1/N19-1423]

[36] Lowe DG. Distinctive image features from scale-invariant keypoints. Int'l Journal of Computer Vision, 2004, 60(2): 91–110. [doi: 10.1023/B:VISI.0000029664.99615.94]

[37] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proc. of the 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. San Diego: IEEE, 2005. 886–893. [doi: 10.1109/CVPR.2005.177]

[38] Oliva A, Torralba A. Modeling the shape of the scene: A holistic representation of the spatial envelope. Int'l Journal of Computer Vision, 2001, 42(3): 145–175. [doi: 10.1023/A:1011139631724]

[39] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 248–255. [doi: 10.1109/CVPR.2009.5206848]

[40] Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A. The PASCAL visual object classes (VOC) challenge. Int'l Journal of Computer Vision, 2010, 88(2): 303–338. [doi: 10.1007/s11263-009-0275-4]

[41] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.

[42] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[43] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929, 2020.

[44] Paul S, Chen PY. Vision transformers are robust learners. Proc. of the 2022 AAAI Conf. on Artificial Intelligence, 2022, 36(2): 2071–2081. [doi: 10.1609/aaai.v36i2.20103]

[45] Mroueh Y, Marcheret E, Goel V. Asymmetrically weighted CCA and hierarchical kernel sentence embedding for image & text retrieval.

arXiv:1511.06267, 2017.

[46] Luo MN, Chang XJ, Li ZH, Nie LQ, Hauptmann AG, Zheng QH. Simple to complex cross-modal learning to rank. Computer Vision and Image Understanding, 2017, 163: 67–77. [doi: 10.1016/j.cviu.2017.07.001]

[47] Wang BK, Yang Y, Xu X, Hanjalic A, Shen HT. Adversarial cross-modal retrieval. In: Proc. of the 25th ACM Int'l Conf. on Multimedia. Mountain View: ACM, 2017. 154–162.

[48] Karpathy A, Joulin A, Fei-Fei L. Deep fragment embeddings for bidirectional image sentence mapping. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: ACM, 2014. 1889–1897.

[49] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proc. of the 25th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: ACM, 2012. 1097–1105.

[50] Yan F, Mikolajczyk K. Deep correlation for matching images and text. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 3441–3450. [doi: 10.1109/CVPR.2015.7298966]

[51] Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(4): 664–676. [doi: 10.1109/TPAMI.2016.2598339]

[52] Zhuang YT, Wang YF, Wu F, Zhang Y, Lu WM. Supervised coupled dictionary learning with group structures for multi-modal retrieval. In: Proc. of the 27th AAAI Conf. on Artificial Intelligence. Bellevue: AAAI Press, 2013. 1070–1076.

[53] Zhu F, Shao L, Yu MY. Cross-modality submodular dictionary learning for information retrieval. In: Proc. of the 23rd ACM Int'l Conf. on Information and Knowledge Management. Shanghai: ACM, 2014. 1479–1488. [doi: 10.1145/2661829.2661926]

[54] Deng C, Tang X, Yan JC, Liu W, Gao XB. Discriminative dictionary learning with common label alignment for cross-modal retrieval. IEEE Trans. on Multimedia, 2016, 18(2): 208–218. [doi: 10.1109/TMM.2015.2508146]

[55] Wei YC, Zhao Y, Zhu ZF. Modality-dependent cross-media retrieval. ACM Trans. on Intelligent Systems and Technology, 2016, 7(4): 1–13.

[56] Roller S, Walde SSI. A multimodal LDA model integrating textual, cognitive and visual modalities. In: Proc. of the 2013 Conf. on Empirical Methods in Natural Language Processing. Seattle: Association for Computational Linguistics, 2013. 1146–1157.

[57] Andrews M, Vigliocco G, Vinson D. Integrating experiential and distributional data to learn semantic representations. Psychological Review, 2009, 116(3): 463–498. [doi: 10.1037/a0016261]

[58] Peng YX, Zhai XH, Zhao YZ, Huang X. Semi-supervised cross-media feature learning with unified patch graph regularization. IEEE Trans. on Circuits and Systems for Video Technology, 2016, 26(3): 583–596. [doi: 10.1109/TCSVT.2015.2400779]

[59] Indyk P, Motwani R. Approximate nearest neighbors: Towards removing the curse of dimensionality. In: Proc. of the 13th ACM Symp. on Theory of Computing. Dallas: ACM, 1998. 604–613. [doi: 10.1145/276698.276876]

[60] Weiss Y, Torralba A, Fergus R. Spectral hashing. In: Proc. of the 21st Int'l Conf. on Neural Information Processing Systems. Vancouver: ACM, 2008. 1753–1760.

[61] Wang LW, Li Y, Lazebnik S. Learning deep structure-preserving image-text embeddings. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 5005–5013. [doi: 10.1109/CVPR.2016.541]

[62] Li XR, Xiu XC, Liu WQ, Miao ZH. An efficient newton-based method for sparse generalized canonical correlation analysis. IEEE Signal Processing Letters, 2022, 29: 125–129. [doi: 10.1109/LSP.2021.3129170]

[63] Rasiwasia N, Pereira JC, Coviello E, Doyle G, Lanckriet GRG, Levy R, Vasconcelos N. A new approach to cross-modal multimedia retrieval. In: Proc. of the 18th ACM Int'l Conf. on Multimedia. Firenze: ACM, 2010. 251–260. [doi: 10.1145/1873951.1873987]

[64] Ranjan V, Rasiwasia N, Jawahar CV. Multi-label cross-modal retrieval. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 4094–4102. [doi: 10.1109/ICCV.2015.466]

[65] Akaho S. A kernel method for canonical correlation analysis. arXiv:cs/0609071, 2006.

[66] Andrew G, Arora R, Bilmes J, Livescu K. Deep canonical correlation analysis. In: Proc. of the 30th Int'l Conf. on Machine Learning. Atlanta: ACM, 2013. 1247–1255.

[67] Pereira JC, Coviello E, Doyle G, Rasiwasia N, Lanckriet GGR, Levy R, Vasconcelos N. On the role of correlation and abstraction in cross-modal multimedia retrieval. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2014, 36(3): 521–535. [doi: 10.1109/TPAMI.2013.142]

[68] Wang LQ, Sun WC, Zhao ZC, Su F. Modeling intra- and inter-pair correlation via heterogeneous high-order preserving for cross-modal retrieval. Signal Processing, 2017, 131: 249–260. [doi: 10.1016/j.sigpro.2016.08.012]

[69] Gong YC, Ke QF, Isard M, Lazebnik S. A multi-view embedding space for modeling Internet images, tags, and their semantics. Int'l Journal of Computer Vision, 2014, 106(2): 210–233. [doi: 10.1007/s11263-013-0658-4]

[70] Rasiwasia N, Mahajan D, Mahadevan V, Aggarwal G. Cluster canonical correlation analysis. In: Proc. of the 17th Int'l Conf. on

Artificial Intelligence and Statistics. Reykjavik: PMLR, 2014. 823–831.

[71] Petter S, Hadavi Y. With great power comes great responsibility: The use of partial least squares in information systems research. ACM SIGMIS Database: The DATABASE for Advances in Information Systems, 2021, 52(SI): 10–23. [doi: 10.1145/3505639.3505643]

[72] Roman R. Nonlinear partial least squares: An overview. In: Lodhi H, Yamanishi Y, eds. Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques. Hershey: Medical Information Science Reference, 2011. 169–189. [doi: 10.4018/978-1-61520-911-8.ch009]

[73] Berglund A, Wold S. INLR, implicit non-linear latent variable regression. Journal of Chemometrics, 1997, 11(2): 141–156. [doi: 10.1002/(SICI)1099-128X(199703)11:2<141::AID-CEM461>3.0.CO;2-2]

[74] Rosipal R, Trejo LJ. Kernel partial least squares regression in reproducing kernel hilbert space. The Journal of Machine Learning Research, 2002, 2: 97–123.

[75] Rosipal R, Trejo LJ, Matthews B. Kernel PLS-SVC for linear and nonlinear classification. In: Proc. of the 20th Int'l Conf. on Machine Learning. Washington: AAAI Press, 2003. 640–647.

[76] Kumar S, Udupa R. Learning hash functions for cross-view similarity search. In: Proc. of the 22nd Int'l Joint Conf. on Artificial Intelligence. Barcelona: IJCAI/AAAI, 2011. 1360–1365. [doi: 10.5591/978-1-57735-516-8/IJCAI11-230]

[77] Zhen Y, Yeung DY. Co-regularized hashing for multimodal data. In: Proc. of the 25th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: ACM, 2012. 1376–1384.

[78] Ou M, Cui P, Wang F, Wang J, Zhu WW, Yang SQ. Comparing apples to oranges: A scalable solution with heterogeneous hashing. In: Proc. of the 19th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Chicago: ACM, 2013. 230–238. [doi: 10.1145/2487575.2487668]

[79] Wu F, Yu Z, Yang Y, Tang SL, Zhang Y, Zhuang YT. Sparse multi-modal hashing. IEEE Trans. on Multimedia, 2014, 16(2): 427–439. [doi: 10.1109/TMM.2013.2291214]

[80] Ou MD. Hashing representation learning for massive heterogeneous data [Ph.D. Thesis]. Beijing: Tsinghua University, 2016 (in Chinese with English abstract).

[81] Xiong HZ, Xie L. Semantic correlation multi-modal hashing for web image retrieval. Journal of Wuhan University of Technology, 2016, 38(8): 71–75 (in Chinese with English abstract). [doi: 10.3963/j.issn.1671-4431.2016.08.013]

[82] Fan H, Chen HH. Research on cross-modal retrieval based on hash method. Data Communication, 2018 (3): 39–45(in Chinese with English abstract). [doi: 10.3969/j.issn.1002-5057.2018.03.011]

[83] Zhang PF, Li Y, Huang Z, Xu XS. Aggregation-based graph convolutional hashing for unsupervised cross-modal retrieval. IEEE Trans. on Multimedia, 2022, 24: 466–479. [doi: 10.1109/TMM.2021.3053766]

[84] Zhai XH, Peng YX, Xiao JG. Heterogeneous metric learning with joint graph regularization for cross-media retrieval. Web Information Systems Engineering, 2013: 43–56.

[85] Zhai XH, Peng YX, Xiao JG. Learning cross-media joint representation with sparse and semisupervised regularization. IEEE Trans. on Circuits and Systems for Video Technology, 2014, 24(6): 965–978. [doi: 10.1109/TCSVT.2013.2276704]

[86] Wang KY, He R, Wang L, Wang W, Tan TN. Joint feature selection and subspace learning for cross-modal retrieval. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2016, 38(10): 2010–2023. [doi: 10.1109/TPAMI.2015.2505311]

[87] Cao GQ, Waris MA, Iosifidis A, Gabbouj M. Multi-modal subspace learning with dropout regularization for cross-modal recognition and retrieval. In: Proc. of the 6th Int'l Conf. on Image Processing Theory, Tools and Applications. Oulu: IEEE, 2017. 1–6. [doi: 10.1109/IPTA.2016.7821032]

[88] Zhang L, Ma BP, Li GR, Huang QM, Tian Q. Generalized semi-supervised and structured subspace learning for cross-modal retrieval. IEEE Trans. on Multimedia, 2018, 20(1): 128–141. [doi: 10.1109/TMM.2017.2723841]

[89] Mandal D, Rao P, Biswas S. Semi-supervised cross-modal retrieval with label prediction. IEEE Trans. on Multimedia, 2020, 22(9): 2345–2353. [doi: 10.1109/TMM.2019.2954741]

[90] Andreas V, Alldrin N, Chechik G, Krasin I, Gupta A, Belongie S. Learning from noisy large-scale datasets with minimal supervision. In: Proc. of the 2017 Int'l Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6575–6583. [doi: 10.1109/CVPR.2017.696]

[91] Blei DM, Jordan MI. Modeling annotated data. In: Proc. of the 26th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Toronto: ACM, 2003. 127–134. [doi: 10.1145/860435.860460]

[92] Putthividhy D, Attias HT, Nagarajan SS. Topic regression multi-modal latent Dirichlet allocation for image annotation. In: Proc. of the 2010 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. San Francisco: IEEE, 2010. 3408–3415. [doi: 10.1109/CVPR.2010.5540000]

[93] Zheng Y, Zhang YJ, Larochelle H. Topic modeling of multimodal data: An autoregressive approach. In: Proc. of the 2014 IEEE Conf. on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 1370–1377. [doi: 10.1109/CVPR.2014.178]

[94] Wang YF, Wu F, Song J, Li X, Zhuang YT. Multi-modal mutual topic reinforce modeling for cross-media retrieval. In: Proc. of the 22nd ACM Int'l Conf. on Multimedia. Orlando: ACM, 2014. 307–316. [doi: 10.1145/2647868.2654901]

[95] Qin ZC, Yu J, Cong YH, Wan T. Topic correlation model for cross-modal multimedia information retrieval. Pattern Analysis and Applications, 2016, 19(4): 1007–1022. [doi: 10.1007/s10044-015-0478-y]

[96] Jia YQ , Mathieu S, Trevor D. Learning cross-modality similarity for multinomial data. In: Proc. of the 2011 Int'l Conf. on Computer Vision. 2011. 2407–2414.

[97] Zhuang YT, Yu Z, Wang W, Wu F, Tang SL, Shao J. Cross-media hashing with neural networks. In: Proc. of the 2014 ACM Int'l Conf. on Multimedia. Orlando: ACM, 2014. 901–904. [doi: 10.1145/2647868.2655059]

[98] Zhang H, Pan M. Semantics-preserving hashing based on multi-scale fusion for cross-modal retrieval. Multimedia Tools and Applications, 2021, 80(11): 17299–17314. [doi: 10.1007/s11042-020-09869-4]

[99] Chen SB, Wu S, Wang L. Hierarchical semantic interaction-based deep hashing network for cross-modal retrieval. PeerJ Computer Science, 2021, 7: e552. [doi: 10.7717/peerj-cs.552]

[100] Li ZY, Lu HB, Fu H, Gu GH. Image-text bidirectional learning network based cross-modal retrieval. Neurocomputing, 2022, 483: 148–159. [doi: 10.1016/j.neucom.2022.02.007]

[101] Liu H, Qu Y, Zhang LQ. Multispectral scene classification via cross-modal knowledge distillation. IEEE Trans. on Geoscience and Remote Sensing, 2022, 60: 5409912. [doi: 10.1109/TGRS.2022.3174352]

[102] Li MY, Li QQ, Tang LR, Peng S, Ma Y, Yang DG. Deep unsupervised hashing for large-scale cross-modal retrieval using knowledge distillation model. Computational Intelligence and Neuroscience, 2021, 2021: 5107034. [doi: 10.1155/2021/5107034]

[103] Liu JH, Yang M, Li CM, Xu RF. Improving cross-modal image-text retrieval with teacher-student learning. IEEE Trans. on Circuits and Systems for Video Technology, 2021, 31(8): 3242–3253. [doi: 10.1109/TCSVT.2020.3037661]

[104] Zhang YC, Cao JY, Gu XD. Learning cross-modal aligned representation with graph embedding. IEEE Access, 2018, 6: 77321–77333. [doi: 10.1109/access.2018.2881997]

[105] Dong XF, Liu L, Zhu L, Nie LQ, Zhang HX. Adversarial graph convolutional network for cross-modal retrieval. IEEE Trans. on Circuits and Systems for Video Technology, 2022, 32(3): 1634–1645. [doi: 10.1109/TCSVT.2021.3075242]

[106] Yuan ZQ, Zhang WK, Tian CY, Rong XE, Zhang ZY, Wang HQ, Fu K, Sun X. Remote sensing cross-modal text-image retrieval based on global and local information. IEEE Trans. on Geoscience and Remote Sensing, 2022, 60: 5620616. [doi: 10.1109/TGRS.2022.3163706]

[107] Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY. Multimodal deep learning. In: Proc. of the 2011 Int'l Conf. on Machine Learning. Bellevue: ACM, 2011. 689–696.

[108] Srivastava N, Salakhutdinov R. Multimodal learning with deep Boltzmann machines. In: Proc. of the 25th Int 'l Conf. on Neural Information Processing Systems. Lake Tahoe: ACM, 2012. 2222–2230.

[109] Cao WM, Lin QB, He ZH, He ZQ. Hybrid representation learning for cross-modal retrieval. Neurocomputing, 2019, 345: 45–57. [doi: 10.1016/j.neucom.2018.10.082]

[110] Silberer C, Lapata M. Learning grounded meaning representations with autoencoders. In: Proc. of the 52nd Meeting of the Association for Computational Linguistics. Baltimore: Association for Computational Linguistics, 2014. 721–732. [doi: 10.3115/v1/P14-1068]

[111] Wang W, Ooi BC, Yang XY, Zhang DX, Zhuang YT. Effective multi-modal retrieval based on stacked auto-encoders. Proc. of the VLDB Endowment, 2014, 7(8): 649–660. [doi: 10.14778/2732296.2732301]

[112] Feng FX, Wang XJ, Li RF, Ahmad I. Correspondence autoencoders for cross-modal retrieval. ACM Trans. on Multimedia Computing, Communications, and Applications, 2015, 12(1s): 26. [doi: 10.1145/2808205]

[113] Feng FX, Wang XJ, Li RF. Cross-modal retrieval with correspondence autoencoder. In: Proc. of the 22nd Int'l Conf. on Multimedia. Orlando: ACM, 2014. 7–16. [doi: 10.1145/2647868.2654902]

[114] Zhan YB, Yu J, Yu Z, Zhang R, Tao DC, Tian Q. Comprehensive distance-preserving autoencoders for cross-modal retrieval. In: Proc. of the 26th ACM Int'l Conf. on Multimedia. Seoul: ACM, 2018. 1137–1145. [doi: 10.1145/3240508.3240607]

[115] Henning C, Ewerth R. Estimating the information gap between textual and visual representations. Int 'l Journal of Multimedia Information Retrieval, 2018, 7(1): 43–56. [doi: 10.1007/s13735-017-0142-y]

[116] Otto C, Holzki S, Ewerth R. "Is this an example image?"—Predicting the relative abstractness level of image and text. In: Proc. of the 41st European Conf. on IR Research on Advances in Information Retrieval. Cologne: Springer, 2019. 711–725. [doi: 10.1007/978-3-030-15712-8_46]

[117] Liu XL, Wang M, Zha ZJ, Hong RC. Cross-modality feature learning via convolutional autoencoder. ACM Trans. on Multimedia Computing, Communications, and Applications, 2019, 15(1s): 7. [doi: 10.1145/3231740]

[118] Xu X, Tian JL, Lin KY, Lu HM, Shao J, Shen HT. Zero-shot cross-modal retrieval by assembling autoencoder and generative adversarial network. ACM Trans. on Multimedia Computing, Communications, and Applications, 2021, 17(1s): 3. [doi: 10.1145/3424341]

[119] Baars BJ, Gage NM, Wrote; Wang ZX, Ku YX, Li CX, Trans. Cognition, Brain, and Consciousness: Introduction to Cognitive Neuroscience. Shanghai: Shanghai People's Publishing House, 2015. 262–263.

[120] Lin ZH, Feng MW, Dos Santos CN, Yu M, Xiang B, Zhou BW, Bengio Y. A structured self-attentive sentence embedding. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017. 1–15.

[121] Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional block attention module. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 3–19. [doi: 10.1007/978-3-030-01234-2_1]

[122] Huang FR, Zhang XM, Li ZJ. Learning joint multimodal representation with adversarial attention networks. In: Proc. of the 26th ACM Int'l Conf. on Multimedia. Seoul: ACM, 2018. 1874–1882. [doi: 10.1145/3240508.3240614]

[123] Lu YH, Yu J, Liu YB, Tan JL, Guo L, Zhang WF. Fine-grained correlation learning with stacked co-attention networks for cross-modal information retrieval. In: Proc. of the 11th Int'l Conf. on Knowledge Science, Engineering and Management. Changchun: Springer, 2018. 213–225. [doi: 10.1007/978-3-319-99365-2_19]

[124] Peng YX, Qi JW, Yuan YX. Modality-specific cross-modal similarity measurement with recurrent attention network. IEEE Trans. on Image Processing, 2018, 27(11): 5585–5599. [doi: 10.1109/TIP.2018.2852503]

[125] Wang SH, Chen YY, Zhuo JB, Huang QM, Tian Q. Joint global and co-attentive representation learning for image-sentence retrieval. In: Proc. of the 26th ACM Int'l Conf. on Multimedia. Seoul: ACM, 2018. 1398–1406. [doi: 10.1145/3240508.324053510.1145/3240508.3240535]

[126] Zhang X, Zhou SY, Feng JS, Lai HJ, Li B, Pan Y, Yin J, Yan SC. HashGAN: Attention-aware deep adversarial hashing for cross modal retrieval. arXiv:1711.09347, 2017.

[127] Chen SB, Wu S, Wang L, Yu ZY. Self-attention and adversary learning deep hashing network for cross-modal retrieval. Computers & Electrical Engineering, 2021, 93: 107262. [doi: 10.1016/J.COMPELECENG.2021.107262]

[128] Wu JG, Weng WW, Fu JX, Liu LF, Hu B. Deep semantic hashing with dual attention for cross-modal retrieval. Neural Computing and Applications, 2022, 34(7): 5397–5416. [doi: 10.1007/s00521-021-06696-y]

[129] Ji Z, Wang HR, Han JG, Pang YW. SMAN: Stacked multimodal attention network for cross-modal image-text retrieval. IEEE Trans. on Cybernetics, 2022, 52(2): 1086–1097. [doi: 10.1109/TCYB.2020.2985716]

[130] Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. arXiv:1710.10903, 2017.

[131] Wu HC, Guan ZY, Zhi T, Zhao W, Xu C, Han H, Yang YM. Adversarial graph attention network for multi-modal cross-modal retrieval. In: Proc. of the 2019 IEEE Int'l Conf. on Big Knowledge (ICBK). Beijing: IEEE, 2019. 265–272. [doi: 10.1109/ICBK.2019.00043]

[132] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: ACM, 2014. 2672–2680.

[133] Hu LQ, Kan MN, Shan SG, Chen XL. Duplex generative adversarial network for unsupervised domain adaptation. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Lake City: IEEE, 2018. 1498–1507. [doi: 10.1109/CVPR.2018.00162]

[134] Peng YX, Qi JW. CM-GANs: Cross-modal generative adversarial networks for common representation learning. ACM Trans. on Multimedia Computing, Communications, and Applications, 2019, 15(1): 22. [doi: 10.1145/3284750]

[135] Dash A, Gamboa JCB, Ahmed S, Liwicki M, Afzal MZ. TAC-GAN-Text conditioned auxiliary classifier generative adversarial network. arXiv:1703.06412, 2017.

[136] Hao WL, Zhang ZX, Guan He. CMCGAN: A uniform framework for cross-modal visual-audio mutual generation. arXiv:1711.08102, 2017.

[137] Chung YA, Weng WH, Tong S, Glass J. Unsupervised cross-modal alignment of speech and text embedding space. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: ACM, 2018. 7354–7364.

[138] Jiang B, Yang JC, Lv ZH, Tian K, Meng QG, Yan Y. Internet cross-media retrieval based on deep learning. Journal of Visual Communication and Image Representation, 2017, 48: 356–366. [doi: 10.1016/j.jvcir.2017.02.011]

[139] Peng YX, Huang X, Qi JW. Cross-media shared representation by hierarchical learning with multiple deep networks. In: Proc. of the 25th Int'l Joint Conf. on Artificial Intelligence. New York: ACM, 2016. 3846–3853.

[140] Rashtchian C, Young P, Hodosh M, Hockenmaier J. Collecting image annotations using Amazon's mechanical Turk. In: Proc. of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Los Angeles: Association for

Computational Linguistics, 2010. 139–147.
[141] Zhai XH, Peng YX, Xiao JG. Effective heterogeneous similarity measure with nearest neighbors for cross-media retrieval. In: Proc. of the 18th Int'l Conf. on Advances in Multimedia Modeling. Klagenfurt: Springer, 2012. 312–322. [doi: 10.1007/978-3-642-27355-1_30]
[142] Zhai XH, Peng YX, Xiao JG. Cross-modality correlation propagation for cross-media retrieval. In: Proc. of the 2012 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). Kyoto: IEEE, 2012. 2337–2340. [doi: 10.1109/ICASSP.2012.6288383]
[143] Zhuang FZ, Qi ZY, Duan KY, Xi DB, Zhu YC, Zhu HS, Xiong H, He Q. A comprehensive survey on transfer learning. Proc. of the IEEE, 2021, 109(1): 43–76. [doi: 10.1109/JPROC.2020.3004555]

附中文参考文献:

[7] 彭宇新, 綦金玮, 黄鑫. 多媒体内容理解的研究现状与展望. 计算机研究与发展, 2019, 56(1): 183–208. [doi: 10.7544/issn1000-1239.2019.20180770]
[9] 曹建军, 刁兴春, 汪挺, 王芳潇. 领域无关数据清洗研究综述. 计算机科学, 2010, 37(5): 26–29.
[11] 曹建军, 刁兴春, 杜鹢, 王芳潇, 张潇毅. 基于蚁群特征选择的相似重复记录分类检测. 兵工学报, 2010, 31(9): 1222–1227.
[13] 谭明超, 刁兴春, 曹建军. 实体分辨研究综述. 计算机科学, 2014, 41(4): 9–12, 20. [doi: 10.3969/j.issn.1002-137X.2014.04.002]
[14] 俞荣华, 田增平, 周傲英. 一种检测多语言文本相似重复记录的综合方法. 计算机科学, 2002, 29(1): 118–121. [doi: 10.3969/j.issn.1002-137X.2002.01.036]
[19] 尚玉玲, 曹建军, 李红梅, 郑奇斌. 基于合作作者与隶属机构信息的同名排歧方法. 计算机科学, 2018, 45(11): 220–225, 260. [doi: 10.11896/j.issn.1002-137X.2018.11.034]
[80] 欧明栋. 面向大规模异构数据的哈希表征学习研究 [博士学位论文]. 北京: 清华大学, 2016.
[81] 熊昊哲, 谢良. 面向Web图像检索的语义关联多模态哈希方法. 武汉理工大学学报, 2016, 38(8): 71–75. [doi: 10.3963/j.issn.1671-4431.2016.08.013]
[82] 樊花, 陈华辉. 基于哈希方法的跨模态检索研究进展. 数据通信, 2018 (3): 39–45. [doi: 10.3969/j.issn.1002-5057.2018.03.011]
[119] Baars BJ, Gage NM, 著; 王兆新, 库逸轩, 李春霞, 译. 认知、大脑和意识——认知神经科学导论. 上海: 上海人民出版社, 2015. 262–263.

曹建军(1975－), 男, 博士, 副研究员, CCF 高级会员, 主要研究领域为数据质量控制与数据治理, 智能数据分析与应用.

吕国俊(1995－), 男, 硕士, 主要研究领域为数据质量控制与数据治理.

聂子博(1998－), 男, 硕士, CCF 学生会员, 主要研究领域为数据质量控制与数据治理.

曾志贤(1996－), 男, 硕士, CCF 学生会员, 主要研究领域为数据质量控制与数据治理.

郑奇斌(1990－), 男, 博士, 主要研究领域为数据质量控制与数据治理.