基于深度学习的图像-文本匹配研究综述

刘 萌[1] 齐孟津[1] 詹圳宇[2] 曲磊钢[2] 聂秀山[1] 聂礼强[3]

[1](山东建筑大学计算机科学与技术学院 济南 250101)

[2](山东大学（青岛）计算机科学与技术学院 山东 青岛 266000)

[3](哈尔滨工业大学（深圳）计算机科学与技术学院 广东 深圳 518055)

摘 要 图像-文本匹配任务旨在衡量图像和文本描述之间的相似性，其在桥接视觉和语言中起着至关重要的作用. 近年来，图像与句子的全局对齐以及区域与单词的局部对齐研究方面取得了很大的进展. 本文对当前先进的研究方法进行分类和描述. 具体地，本文将现有方法划分为基于全局特征的图像-文本匹配方法、基于局部特征的图像-文本匹配方法、基于外部知识的图像-文本匹配方法、基于度量学习的图像-文本匹配方法以及多模态预训练模型，对于基于全局特征的图像-文本匹配方法，本文依据流程类型划分为两类：基于嵌入的方法和基于交互的方法；而对于基于局部特征的图像-文本匹配方法，依据其交互模式的不同，则被细分为三类：基于模态内关系建模的方法、基于模态间关系建模的方法以及基于混合交互建模的方法. 随后，本文对当前图像-文本匹配任务的相关数据集进行了整理，并对现有方法的实验结果进行分析与总结. 最后，对未来研究可能面临的挑战进行了展望.

关键词 图像-文本匹配；跨模态图像检索；多模态预训练模型；综述；深度学习；人工智能

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2023.02370

A Survey on Deep Learning Based Image-Text Matching

LIU Meng[1] QI Meng-Jin[1] ZHAN Zhen-Yu[2] QU Lei-Gang[2] NIE Xiu-Shan[1] NIE Li-Qiang[3]

[1](*Department of Computer Science and Technology, Shandong Jianzhu University, Jinan* 250101)

[2](*Department of Computer Science and Technology, Shandong University, Qingdao, Shandong* 266000)

[3](*Department of Computer Science and Technology (Shenzhen), Harbin Institute of Technology, Shenzhen, Guangdong* 518055)

Abstract Recent years have witnessed the rapid growth of multimedia data, such as texts and images, inducing many researchers to work on multimodal representation, understanding, and reasoning. As a fundamental task of multimodal interaction, image-text matching, focusing on measuring the semantic similarity between an image and a text, has attracted extensive research attention. It indeed facilitates various applications, such as cross-modal retrieval, visual question answering, and multimedia understanding, and plays a critical role in bridging vision and language. Recently, deep learning techniques have emerged as powerful methods for various tasks. This motivates many researchers to resort to deep learning approaches to tackle the image-text matching task. Particularly, great progress has been made by exploiting the global alignment between images and sentences, or local alignments between image regions and textual words. They can be roughly divided into the following categories: global representation-based image-text matching methods, local representation-based image-text matching methods, external knowledge-based image-text matching methods, metric learning-based image-text matching methods, and multimodal pre-training mod-

收稿日期：2022-03-03；在线发布日期：2022-12-29. 本课题得到国家自然科学基金项目（No. 62006142、No. U1936203）、山东省杰出青年基金项目（No. ZR2021JQ26）、山东省基金重大基础研究项目（No. ZR2021ZD15）、山东省高等学校青年创新科技创新计划（No. 2021KJ036）、山东建筑大学特聘教授专项基金资助. **刘 萌**，博士，教授，中国计算机学会（CCF）会员，主要研究领域为多媒体计算和信息检索. E-mail: mengliu.sdu@gmail.com. **齐孟津**，硕士研究生，主要研究领域为多媒体内容分析和信息检索. **詹圳宇**，硕士研究生，主要研究领域为跨媒体信息检索. **曲磊钢**，硕士研究生，主要研究领域为跨媒体信息检索. **聂秀山**，博士，教授，主要研究领域为人工智能、机器学习与数据挖掘. **聂礼强**（通信作者），博士，教授，主要研究领域为多媒体计算与信息检索. E-mail: nieliqiang@gmail.com.

els. To be specific, global representation-based image-text matching methods usually realize cross-modal matching by measuring the semantic similarity between the global image and text representations; local representation-based image-text matching methods focus on modeling fine-grained correlations between visual and textual entities; external knowledge-based image-text matching methods are devoted to acquire certain prior knowledge from external sources, such as scene graph, to improve the accuracy of image-text matching; metric learning-based image-text matching methods try to explore a better constraint or similarity measurement to improve the discriminability between unpaired samples and the relevance between the paired samples; as well as the multimodal pre-training models including single stream and two stream frameworks have strong generalization ability. To give a comprehensive overview of this field, including models, datasets, and future directions, we summarize the work on image-text matching and present this survey. Specifically, to perform a deeper analysis of existing approaches, we establish the fine-grained taxonomy of each category. For instance, for global representation-based image-text matching methods, we further divide them into two categories according to their architectures: embedding-based methods and interaction-based methods, respectively. Thereinto, embedding-based methods directly constrain the representation learning of images and text in the common space, while interaction-based methods exploit the cross-modal interactive information for better semantic matching. As to local feature-based image-text matching methods, we further divide them into three categories according to interaction patterns: intra-modal modeling, inter-modal modeling, and hybrid interaction modeling-based approaches. More concretely, intra-modal modeling-based image-text matching methods independently explore relationships between entities within a particular modality, and inter-modal modeling-based image-text matching methods explore cross-modal relationships to better align visual and textual semantic information. Differently, hybrid interaction modeling-based approaches consider both cross-modal interaction information modeling and intra-modal correlation modeling, to simultaneously enhance the modeling of intra-modal and inter-modal relationships. Subsequently, we summarize several benchmark image-text matching datasets and analyze the experimental results of existing models. In addition, we also introduce some related research tasks, including weakly-supervised cross-modal matching, zero-shot cross-modal matching, cross-linguistic image retrieval, and scene-text aware cross-modal retrieval. Finally, we discuss promising future directions for this task, in particular standard dataset partitioning, interpretable image-text matching models, and efficient image-text matching models.

Keywords　image-text matching; cross-modal image retrieval; multimodal pre-training model; survey; deep learning; artificial intelligence

1　引　言

移动便携设备的日益普及和 Web 应用的日新月异，致使图像和文本数据呈爆炸式增长. 举例来讲，在 LOCALIQ 公司[①]2022 年 5 月发布的“What Happens in an Internet Minute in 2022: 90 Fascinating Online Stats”数据表明：在 Instagram 网站上，平均一分钟约有 65 972 张图像或视频被分享；与此同时，Facebook 用户也非常活跃，一分钟约有 51 万条评论和 29 万 3 千条状态更新，以及 24 万张照片被上传. 面对这些复杂的大规模异构多模态数据，精准且高效地搜索用户感兴趣的信息，变得越来越困难. 如，给定图 1(a)图像作为查询信息，来检索具有相似语义内容的特定句子（句子 1 和句子 2），同时剔除部分语义相似（playing volleyball）的句子 3；反之，给定图 1(b)的复杂文本语句作为查询，来检索具有相似语义信息的特定图像 1，同时排除部分语义匹配的图像 2（kid, red and black coat）和图像 3（kid）. 显然地，与传统的基于关键词或主题的图像检索任务相比，跨模态图像-文本检索任务中文本查询语句更

① https://localiq.com/blog/what-happens-in-an-internet-minute

为复杂，其往往包含多个实体以及多种实体关系语. 由此可见，准确地为用户反馈检索结果并非易事，不仅需要充分挖掘文本查询和图像中蕴含的语义信息，还需对捕获的语义信息进行综合全面的匹配.

鉴于此，图像-文本匹配方法的研究，即如何更好地且全面地衡量图像与句子之间的语义相似性，在学术界和工业界引起了广泛的关注. 图像-文本匹配任务研究具有很多的现实应用意义，如可以为给定图像查找合适语句，用于图像标注，以减少人工标记的成本；也可以用于基于句子描述的图像搜索，提升用户的搜索体验；还可以通过自然语言来查询视觉内容，拓展智能机器人的问答功能. 所以，图像-文本匹配任务的不断发展，会为相关研究领域带来一定的帮助.

近几年，深度学习方法被广泛地应用于计算机视觉等领域且均取得了优异的性能表现. 受此启发，许多基于深度学习的图像-文本匹配方法被提出，它们的模型框架图可简化为图 2. 其中，实线框标识的视觉编码器、文本编码器、跨模态匹配以及损失函数模块为现有方法的通用模块，而虚线框标识的外部知识和度量学习模块为部分相关方法所特有. 下面将对通用模块进行简单介绍.

文本编码器模块 该模块旨在捕获文本语句描述中蕴含的语义信息，以得到文本模态的特征表示. 早期的图像-文本匹配方法多采用整体文本编码方式，如一些方法[1,2]直接使用 Skip-Gram[3]或者费舍尔向量[4]（Fisher Vector，缩写为 FV）对文本进行表示. 由于上述编码方式忽略了上下文信息的重要作用，后续多数方法[5]先采用 word2vector 提取文本中的词嵌入表示，然后利用循环神经网络模型（如，LSTM 和 GRU）建模序列上下文信息，再输出文本表示. 随着 Transformer 模型的提出和它在自然语言处理领域的成功应用，一些方法将多头注意力机制引入文本编码器中，对文本语句进行建模，以得到相应表示[6]. 除此之外，也有一些方法尝试通过引入外部知识，来增强文本语义理解，如共识知识[7].

视觉编码器模块 此模块主要负责编码图像中蕴含的视觉语义信息，以得到视觉特征表示. 相对于文本编码器模块，视觉编码器模块的设计较为多样. 一些方法采用卷积神经网络，如 AlexNet，对图像进行整体建模，来得到图像的全局视觉表示. 由于全局编码会损失细粒度语义信息，后续一些方法则提出基于局部建模的编码策略，如基于物体检测的区域建模方法[8,9]和基于特征图的像素建模方

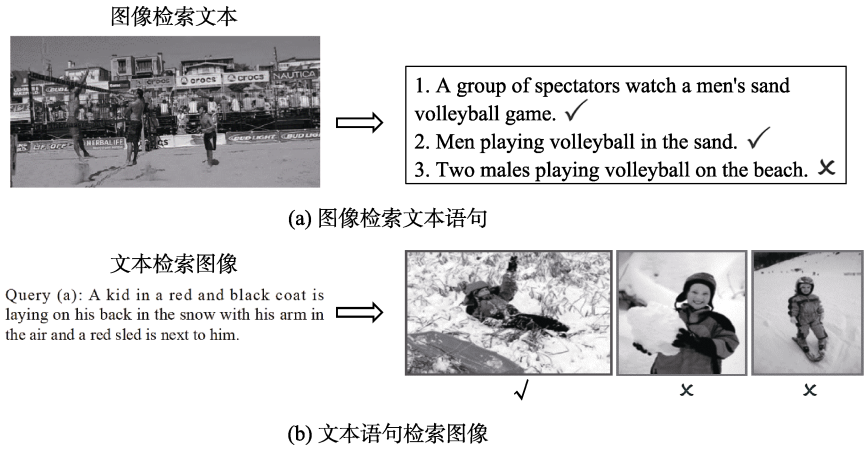

图 1 双向跨模态图像-文本检索示意图

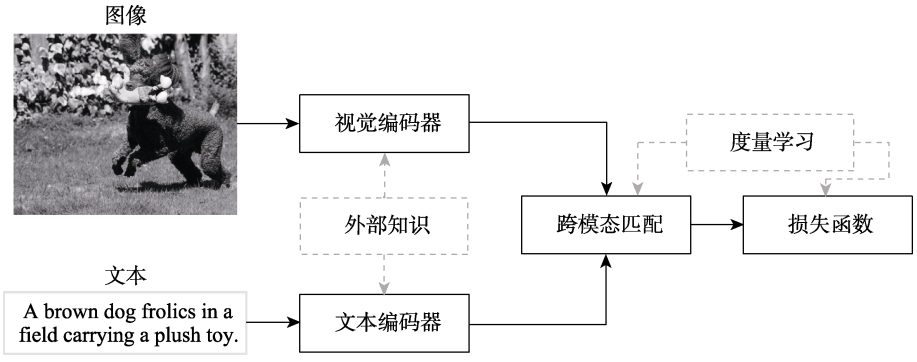

图 2 图像-文本匹配方法简化框架图

法[10,11]. 除此以外，也有一些方法采用自注意机制或图卷积网络，来强化视觉模态内实体间关系，以进一步增强视觉语义理解[12,13]. 为了增强对于视觉语义信息的理解，近期一些方法尝试引入外部知识，如场景图信息[14,15].

跨模态匹配模块　这个模块旨在消除模态间的语义鸿沟，以便更好地对齐两个模态的语义信息，从而更加精准地进行相似性估计. 其中，基于全局特征的图像-文本匹配方法[1,16]，主要将视觉特征和文本特征联合嵌入到共空间中进行跨模态语义相似性估计. 而在基于局部特征的图像-文本匹配方法中[17,18]，部分侧重于直接探究视觉区域和词语对之间的潜在语义对齐，而后利用平均或 LogSumExp 聚合局部相似性，以得到图像和文本的匹配得分；另一部分方法则倾向于设计不同跨模态交互机制，以期更好地实现跨模态语义匹配. 除上述方法外，也有一些方法通过设计不同度量方式，如 NAAF 利用不匹配的实体信息[19]，CMCAN 则为匹配区域-词语对赋值不同置信度[20]，来衡量跨模态语义相似性.

损失函数模块　该部分旨在通过不同的目标函数，来约束图像-匹配模型的学习. 大多数方法采用相同的排序函数来学习，如铰链排序损失（也称三元组排序损失）和双向铰链排序损失. 由于很难直接找到最合适的三元组，可能会致使视觉-语义嵌入学习效果不理想. 鉴于此，许多基于度量学习的图像-文本匹配方法被提出，它们通过引入新的损失函数，来提升学习性能，如枢纽感知损失[21]、多项式损失[22]、五元组损失函数[23]等. 此外，考虑到目前相关损失函数需要人为预先定义边距信息，一些自适应边距的三元组损失函数[24]以及边距无关的损失函数[25]被提出.

图 3 对基于深度学习的图像-文本匹配方法进行分类展示，这些方法从不同方面推动图像-文本匹配任务的发展. 为便于相关学者更好地对该领域进行了解，本文旨在对图像-文本匹配工作进行全面的分析与总结. 虽然国内外已有不少跨模态检索相关文献及综述[26-28]，但这些文献大多把图像-文本匹配看作一种用于解决跨模态检索的技术，并未对涉及的相关方法进行细致地分类与分析. 如文献[27]仅对 BRNN、SCAN、CAAN、IMRAM 等方法进行介绍，而文献[26]只对 GXN 和 ACMR 等方法进行简单介绍. 不同于这些文献综述，本文对图像-文本匹配方法进行了系统地总结与介绍，同时也对该任务的数据集、评价标准、模型性能等方面进行介绍. 具体地，本文的第 2 章对图像-文本匹配任务及其挑战进行介绍；本文的第 3 章对图像-文本匹配任务方法进行分类总结和分析；本文的第 4 章介绍图像-文本匹配任务相关的数据集和评价标准；第 5 章则对模型的实验结果进行简单的分析和总结；第 6 章对图像-文本匹配任务进行展望.

2　问题与挑战

给定一张图像以及一个文本描述语句，图像-文本匹配任务旨在判断两者之间是否具有语义一致性. 如图 1(a)所示，给定图像以及文本查询 1，图像-文本匹配模型需要推断出两者具有语义一致性；而当给定图像和文本查询 3 时，其需要精准地判定两者不具备语义一致性. 虽然图像-文本匹配具有广泛的应用场景，但它是一个具有挑战性的研究任务，主要体现在如下方面：

模态内语义理解和建模：图像-文本匹配任务需要同时对图像和文本的内容进行全面地建模，以充分理解两个模态中的语义信息. 如图 1(a)，只有充分理解图像中包含多个男人，才能判别出第 3 个文本语句与图像的语义是不匹配的. 然而，图像中往往蕴含复杂的场景信息，即包含多种不同类别物体且这些物体间关系繁多；而文本语句描述通常也比较繁琐冗长（如图 1(b)的查询文本），故如何充分挖掘并理解两种模态的语义信息，是一个极具挑战的问题.

跨模态语义对齐和匹配：图像和文本为两种异构模态信息，两者间存在巨大的语义鸿沟. 所以，该任务还需要精准匹配两个模态间的对应语义关系. 尽管两个模态间许多局部语义信息是匹配的，如图 1(b)中“red and black coat”和“snow”，但若仅对这些局部信息进行匹配，很容易得到错误的匹配结果. 故模型需要对具有区分性的语义信息，如“a red sled is next to”和“arm in the air”进行准确地对齐，才能精准地返回匹配结果. 由此可见，如何精准地建模与匹配两个模态间的语义对应关系，也是一个亟待探究的问题.

3　研究现状分析

迄今为止，已有大量基于深度学习的图像-文本匹配方法被提出，它们分别发表在计算机视觉和信息检索等不同领域. 考虑到信息编码和理解方式的多样性，如对信息进行整体理解的全局编码方式、对信息进行细粒度理解的局部编码方式以及借助额

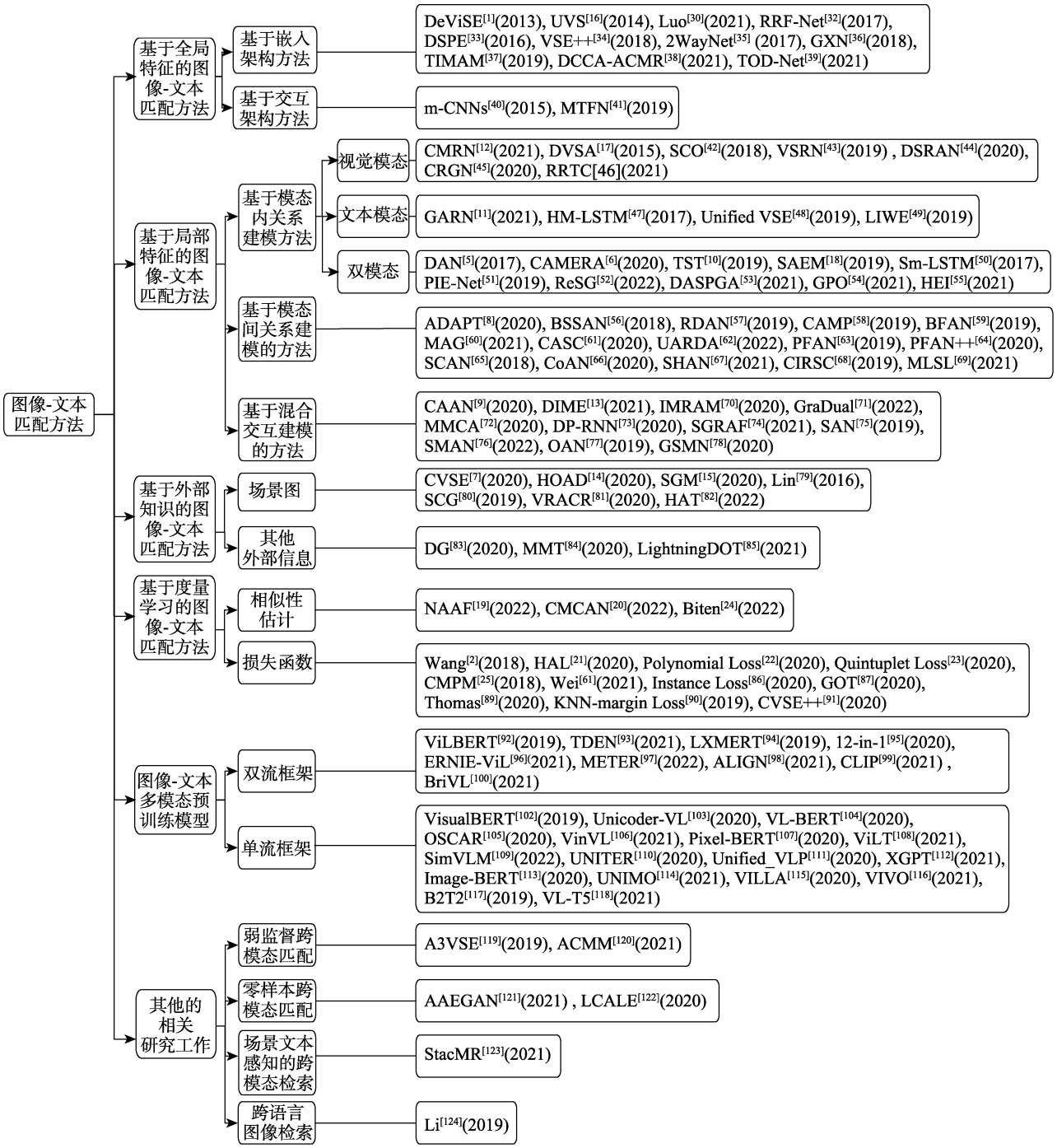

图 3 现有基于深度学习的图像-文本匹配方法

外知识增强对信息理解的编码方式，现有图像-文本匹配方法可分为：（1）基于全局特征的图像-文本匹配方法；（2）基于局部特征的图像-文本匹配方法；以及（3）基于外部知识的图像-文本匹配方法. 不同于从视觉或文本编码器方面着手解决图像-文本匹配任务，一些方法从损失函数设计（度量学习）角度出发，通过约束高质量的视觉和文本表示学习，来解决图像-文本匹配任务；也有一些方法借助于预训练多模态模型，来提升图像-文本匹配的性能. 最近，一些方法对图像-文本任务进行拓展并提出不同的解决策略，如弱监督跨模态匹配、零样本跨模态匹配、场景文本感知的跨模态检索. 故本文在上述三个维度之外，增加了三个新的维度进行研究进展介绍，即基于度量学习的图像-文本匹配方法、图像-文本多模态预训练模型以及其他相关研究工作. 总体来说，本文将现有图像-文本匹配方法划分为六大类，分别为基于全局特征的图像-文本配方法、基于局部特征的图像-文本匹配方法、基于外

部知识的图像-文本匹配方法、基于度量学习的图像-文本匹配方法、图像-文本多模态预训练模型以及其他相关任务. 下文将依次展开对这些类别方法的介绍.

3.1　基于全局特征的图像-文本匹配方法

基于全局特征的图像-文本匹配方法通常采用卷积神经网络，从整张图片来捕获全局的视觉语义信息；或者利用循环神经网络等模型，从文本语句提取整体语义信息. 然后，利用不同的模型架构实现图像-文本的跨模态匹配. 如图 4 所示，现有基于全局特征的图像-文本匹配方法主要分为两大类：基于嵌入架构的方法和基于交互架构的方法. 其中，基于嵌入架构的方法直接在共空间中约束图像和文本表示的学习；而基于交互式架构方法则通过挖掘两种模态间的交互信息，以更好地实现跨模态语义匹配.

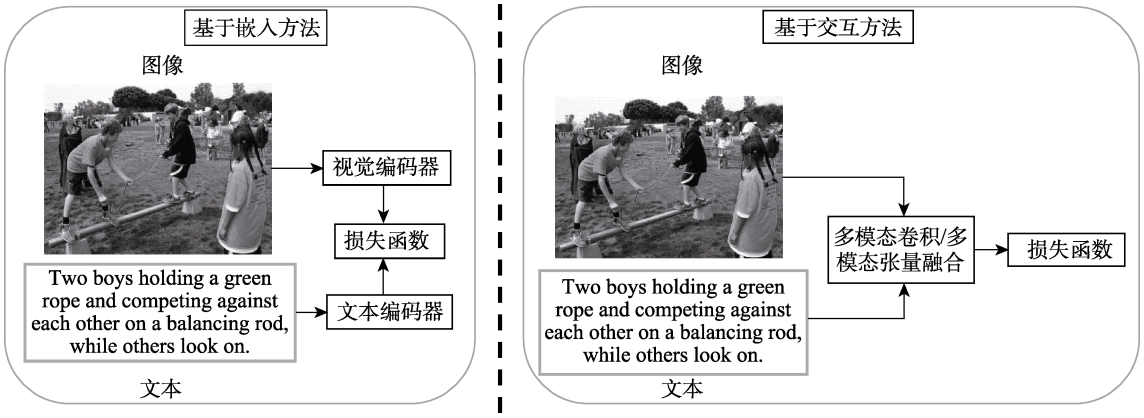

图 4　基于全局特征图像-文本匹配方法的分类研究框架图

3.1.1　基于嵌入架构的全局图像-文本匹配方法

基于嵌入架构的全局图像-文本匹配方法通常先将整个图像和文本描述的特征转换到一个公共嵌入空间中，然后通过度量嵌入特征间的相似度来衡量图像与文本是否匹配. 换而言之，该类方法重在探究如何学习鲁棒的多模态嵌入表示. 其中，部分方法直接采用传统共空间表示学习方法；一些方法则在传统共空间表示学习基础上，通过引入其他约束，如保结构约束、跨模态相关性约束以及对抗约束，来增强两种模态信息的表示.

举例来说，Frome 等人[1]提出的深度视觉语义嵌入模型（Deep Visual-Semantic Embedding，简称 DeViSE）是最早尝试将图像和文本映射到共空间中的方法. 该方法首先用在 ImageNet 上预训练的卷积神经网络 AlexNet[29]提取图像特征表示；同时采用预训练的 Skip-Gram 模型，提取文本表示. 然后，将上述两个模态特征表示转换到维度相同的共空间中，计算余弦相似度. 最后，采用铰链损失函数（Hinge Loss）进行模型训练. 不同于 DeViSE 的文本和视觉编码方式，骆等人[30]分别利用 Doc2vec 模型和 VGG16 模型提取图像和文本特征. 而 Kiros 等人[16]则采用 LSTM 网络和 OxfordNet[31]分别提取全局文本特征和全局视觉特征，同时提出统一视觉-语义嵌入（Unifying Visual-Semantic Embeddings，简称 UVS）模型用于跨模态语义匹配. 为增强共空间中特征区分性，Liu 等人[32]提出基于递归残差融合的深度匹配网络（Recurrent Residual Fusion Network，简称 RRF-Net），其在共空间投影过程中插入递归残差连接，以生成表达能力更强的视觉和文本表示①.

虽然上述方法能较好地对齐视觉和文本语义信息，但其忽略了保持各个模态内邻域结构的重要性. 为此，Wang 等人[33]提出深度结构保持的嵌入学习方法（Deep Structure-Preserving Embedding，简称 DSPE），其在目标函数中考虑了模态内邻域结构保持约束②. 而 VSE++[34]则将难分样本挖掘（Hard Negative Mining）的思想融入到损失函数中，即铰链损失中只有一个最难区分的负样本，以此优化计算效率且提高模型学习能力. Eisenschtat 等人[35]提出一种双通道神经网络结构（命名为 2WayNet），通过强制各个通道将一个模态信息转换为另一个模态，并约束中间表示相关性最大化，将两个模态特征投射到一个共同的且最相关的空间中.

考虑到生成对抗网络（Generative Adversarial Networks，简称 GANs）学习判别表示的强大能力，Gu 等人[36]提出名为 GXN 网络，其通过鼓励视觉嵌入生成与输入文本相似的句子以及文本嵌入生成与

① https://github.com/yuLiu24/RRF-Net

② https://github.com/BryanPlummer/two_branch_networks

输入图像相似的图像，以更好地对齐视觉与文本语义信息. 类似地，Sarafianos 等人[37]提出文本-图像模态对抗式匹配方法（Text-Image Modality Adversarial Matching，简称 TIMAM），希望同时利用鉴别器的对抗损失、识别损失和跨模态嵌入匹配损失，来学习模态不变的嵌入表示. 不同于上面两种方法，刘等人[38]提出一种融合深度典型相关分析的对抗式跨模态检索方法，名为 DCCA-ACMR. 其在图像和文本表示层增加深度典型相关约束，来挖掘图像和文本数据的关联关系. 而 Matsubara 等人[39]提出的面向目标的变形网络（Target-Oriented Deformation Network，简称 TODNet），则通过在给定条件下，将嵌入空间变换为新的嵌入空间，以更好地实现跨模态匹配. 特别地，TOD-Net 可串联到现有嵌入空间学习模型后面，以进一步提高其嵌入学习能力.

3.1.2 基于交互架构的全局图像-文本匹配方法

不同于基于嵌入的全局图像-文本匹配方法，基于交互的全局图像-文本匹配方法则通过捕获模态间交互信息，来预测匹配分数. 相较于基于嵌入的全局匹配方法，基于交互的全局图像-文本匹配工作较少，它们的不同之处主要集中在交互机制的设计上面. 例如，Ma 等人[40]提出名为 m-CNNs 的多模态卷积神经网络，将图像表示与不同粒度文本表示进行拼接和非线性映射操作，得到不同粒度的交互信息，用于预测匹配分数. 而 Wang 等人[41]提出的多模态张量融合网络（Multi-modal Tensor Fusion Network，简称 MTFN）则通过带有秩约束的张量融合模块，将多模态间交互信息编码为一个向量表示，再通过一个全卷积层预测相似度分数①.

3.1.3 小 结

基于全局特征的图像-文本匹配方法仅仅对全局语义信息进行编码与处理，并不涉及细粒度语义信息的建模，故其匹配效率通常较高. 但也由于它们未能对视觉和文本模态内细粒度语义信息进行挖掘，致使其不能充分地理解各个模态内蕴含的丰富语义信息，甚至引入噪声信息，这无疑会对跨模态语义匹配带来负影响.

3.2 基于局部特征的图像-文本匹配方法

基于全局特征的图像-文本匹配方法无法很好地建模细粒度的实体间关联关系，如图像对象和句子单词间的关系，这极大地限制了图像-文本匹配的准确性. 因此，基于局部视觉特征的图像-文本匹配方法被提出. Karpathy 等人[17]提出的深度视觉-语义对齐（Deep Visual-Semantic Alignment，简称 DVSA）模型为该分支的先驱工作，其将图像区域特征和单词特征间最大的匹配分数当作图像-文本匹配分数，并通过多示例学习约束图像-文本间匹配. 虽然该工作的性能优于基于全局特征的图像-文本匹配方法，但其存在以下弊端：（1）未能充分建模各个模态内的关联关系，如图像区域间相关性，致使不能充分挖掘各个模态的语义信息，实现精准的图像-文本匹配；以及（2）未能全面探索不同模态间的相关性关系，如图像和词语之间的相关性，导致图像和文本间语义不能很好地对齐.

为解决上述问题，其他基于局部特征的图像-文本匹配方法陆续地被提出. 如图 5 所示，它们可大致划分为 3 类：（1）基于模态内关系建模的方法，

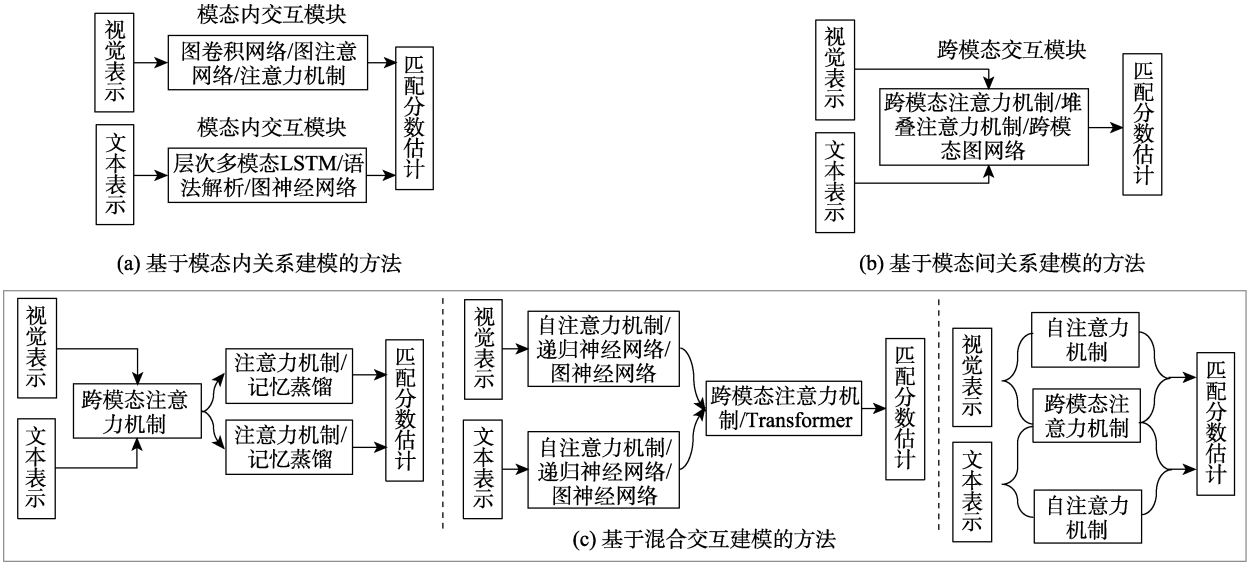

图 5 基于局部特征的图像-文本匹配方法分类研究框架

① https://github.com/Wangt-CN/MTFN-RR-PyTorch-Code

这类方法侧重于探索各个模态内局部信息间的关联关系，以增强视觉或者文本模态的表示；（2）基于模态间关系建模的方法，这类方法侧重于探究跨模态交互信息的挖掘，以更好地对齐两个模态信息；和（3）基于混合交互建模的方法，这类方法兼顾上述两类方法的优势. 下面，本文将依次对各类方法进行展开介绍.

3.2.1　基于模态内关系建模的方法

基于模态内关系建模的图像-文本匹配方法仅独立地对视觉或文本模态内部信息进行建模，以探索特定模态内实体之间的关系. 自然地，依据建模模态信息的不同，这类方法可进一步细分为三小类：（1）侧重于视觉模态实体关系建模的方法；（2）侧重于文本模态实体关系建模的方法；以及（3）同时考虑视觉和文本模态实体关系建模的方法.

（1）视觉模态交互关系建模方法

为充分挖掘图像中局部语义信息以及这些信息间的关系，现有方法设计了不同种模态内关系挖掘方式，如语义重组[42]、图卷积网络[12,43,44]和注意力机制[45,46].

举例来说，Huang 等人[42]提出语义增强图像-文本匹配模型（被称作 SCO），通过迫使模型按照正确顺序组织学习得到的局部语义概念的方式，来隐式地增强视觉表示. Li 等人[43]则提出视觉语义推理模型（Visual Semantic Reasoning Network，简称 VSRN），通过利用图卷积网络对图像区域间的关系进行推理，同时利用门控和记忆机制对关系增强后的区域特征进行全局语义推理，以增强视觉表示①. 类似地，Wen 等人[44]提出双语义关系注意网络（Dual Semantic Relations Attention Network，简称 DSRAN），也通过图注意网络来增强区域-区域和区域-全局关系，为最终的视觉表示提供更多信息②. 不同于上述方法采用图网络来建模局部区域之间的关系，Zhang 等人[45]提出跨模态关系导向网络（Cross-modal Relation Guided Network，简称 CRGN），其利用注意力机制来平衡全局视觉表示和区域特征的相对贡献，输出最终的图像表示③. 虽然上述方法取得不错的跨模态匹配效果，但均侧重于语义关系建模，忽略了图像区域之间的空间位置关系. 鉴于此，Zhang 等人[12]设计了跨模态多关系感知推理网络（Cross-modal Multi-relationship aware Reasoning Network，简称 CMRN），其同时考虑图像区域间几何位置关系和语义交互关系，来学习视觉表示. 考虑到不同区域-词的关联关系对图像-文本匹配贡献度不一致，Wu 等人[46]提出基于主题约束的区域增强网络（Region Reinforcement Network with Topic Constraint，简称 RRTC），该算法通过考虑视觉区域间关系，为不同区域-词对分配权值;并通过对视觉主题进行约束，避免图像原始语义漂移，以此提升图像-文本的语义匹配.

（2）文本模态交互关系建模方法

由于链式 RNN 结构很难构建如短语级的细粒度文本表示，故基于文本模态交互建模的方法多致力于细粒度语义关联关系建模，以增强文本嵌入表示. 例如，Niu 等人[47]提出层次多模态长短期记忆模型（Hierarchical Multimodal Long Short-Term Memory，简称 HM-LSTM），通过树解析器将句子分解为若干短语，然后利用句子和短语之间的层次关系，学习文本语句和短语的嵌入表示. Wu 等人[48]则提出统一视觉-语义嵌入学习模型（Unified Visual-Semantic Embedding，简称 Unified VSE），其将文本解析为对象名词、名词前修饰语和关系依赖，并用特定类型的编码器对不同类型的语义组件进行编码，通过把得到的语义组件嵌入组合，得到最终的文本嵌入. 为进一步对文本模态内关系进行探索，Jing 等人[11]提出图注意关系网络（Graph Attentive Relational Network，简称 GARN），其将文本中的名词组织成图结构，并利用跳跃图神经网络，来学习更有效的文本表示. 不同于上述方法得到词嵌入向量方式，Wehrmann 等人[49]提出一种语言不变的词嵌入学习方法（Language-Invariant Word Embeddings，简称 LIWE），该方法用一个可学习的函数代替传统的词嵌入矩阵学习词嵌入④.

（3）双模态交互关系建模方法

为同时提升视觉或文本模态的特征表示，多数方法采用注意力机制建模不同实体间关系[5,6,18,50,51]，部分方法则采用图卷积神经网络增强实体间关系建模[52,53]，也有方法通过多模态一致性关系约束来捕获视觉和文本中的一致性语义信息[10]. 除此以外，也有一些方法致力于多模态信息编码效率方面的研究[54,55].

在基于注意力机制的方法中，Huang 等人[50]提出选择性多模态长短时记忆网络（Selective multimodal Long Short-Term Memory，简称 Sm-LSTM），通过利用多模态上下文调制注意力机制，来强化局

① https://github.com/KunpengLi1994/VSRN

② https://github.com/kywen1119/DSRAN

③ https://github.com/zyfsa/CRGN

④ https://github.com/jwehrmann/lavse

部视觉和文本语义表示. 考虑到注意力机制的良好性能, DAN[5]和 SAEM[18]①也利用自注意力机制, 来探索各个模态不同实体间关系, 以此增强视觉和文本表示. 随着 Transformer 模型的提出和它在自然语言处理领域的成功应用, Qu 等人[6]提出上下文感知的多视角总结网络（Context-Aware Multi-viEw summaRizAtion network, 简称 CAMERA）, 将多头注意力机制引入文本编码器以及视觉编码器中, 以期同步增强两个模态语义信息的理解②. 为了有效地处理图像-文本匹配任务中的多义实例问题, Song 等人[51]提出多义实例嵌入网络(Polysemous Instance Embedding Network, 简称 PIE-Net）, 通过使用多头自注意模块, 提取单个实例的多个嵌入, 用于跨模态匹配③.

不同于基于注意力机制的方法, Liu 等人[52]提出关系增强语义图模型（Relationship-enhanced Semantic Graph, 简称 ReSG）, 其分别利用视觉关系增强图和文本关系增强图对对应实例的高级语义概念及其上下文语义关系进行编码, 并将图像-文本匹配问题转化为图匹配问题. 为明确地将不同的模态语义信息转化到一个共空间中, Yan 等人[53]提出基于离散连续行动空间策略梯度的注意力（Discrete-continuous Action Space Policy Gradient-based Attention, 简称 DASPGA）, 其分别利用图卷积神经网络建模视觉与文本模态的语义关系, 而后将区域特征输入到所提出的离散连续策略梯度算法中生成注意图, 并利用注意图对区域特征进行调整和融合, 确保将嵌入的图像和文本投射到一个公共空间.

上述方法虽取得不错的效果, 但未能排除嵌入中模态特有特征为匹配带来的负面影响. 为此, Guo 等人[10]引入两阶段训练策略（Two-Stage Training, 简称 TST）, 来学习更有效和可解释的图像和文本表示. 具体地, 其将互信息估计引入到图像-文本匹配任务中, 为视觉语义嵌入保留更多有用的信息; 同时提出跨模态解缠表示学习, 从已学习的模态共享特征中, 排除模态专属信息带来的影响.

为提升视觉和文本模态编码效率, Tu 等人[55]提出基于哈希的高效推理模块（Hashing based Efficient Inference, 简称 HEI）. Chen 等人[54]提出一种广义池化操作（Generalized Pooling Operator, 简称 GPO）, 其可自动学习适应不同特性的最佳池化策略, 来聚合细粒度特征到整体嵌入, 并且该策略不需要手动调优④.

3.2.2 基于模态间关系建模的方法

不同于基于模态内关系建模方法, 基于模态间关系建模的局部图像-文本匹配方法致力于探索跨模态实体间关系建模, 以更好地对齐视觉和文本语义信息,增强图像-文本匹配精准度.但类似于模态内关系建模方法, 该分支多数方法仍采用注意力机制或改进注意力机制建模跨模态语义关联关系, 少数方法采用一致性语义约束或图网络等方式捕获跨模态关联关系.

受注意力机制在计算视觉等领域取得优异性能的启发, 许多基于模态间关系建模的方法采用注意力机制来对齐跨模态语义信息. 例如, Huang 等人[56]提出双向空间语义注意网络（Bi-directional Spatial-Semantic Attention Network, 简称 BSSAN）, 其分别利用词语-区域注意力和图像对象-词语注意力, 来突出视觉和文本中的显著特征. Hu 等人[57]则凭借引入关系感知的双重注意网络（Relation-wise dual attention network, 简称 RDAN）来推断多层次的视觉-语义对齐, 以期提升图像-文本匹配性能. Wang 等人[58]提出一种跨模态自适应消息传播模型（Cross-modal Adaptive Message Passing, 简称 CAMP）, 该模型既可以通过词语和图像区域之间的跨模态注意力将单词对应的显著视觉信息以及各个区域对应的显著文本信息进行聚合; 还可以利用跨模态门控融合模块, 来自适应地控制跨模态特征融合⑤. 由于传统的注意力机制不仅关注与目标内容相关的信息, 还关注与目标内容无关的信息. 为了解决该问题, Liu 等人[59]提出一种双向焦点注意网络（Bidirectional Focal Attention Network, 简称 BFAN）, 其将所有的注意力转移到相关片段（图像区域或词语）上, 从而消除不相关的片段信息的影响⑥. 此外, 由于双重注意力往往会导致对齐不一致的问题, 宫等人[60]提出基于一致性协议匹配的跨模态图像文本检索方法（Matching with AGreement, 简称 MAG）, 其将对齐层的对齐分数和协议层的协议分数结合起来, 计算给定图像-文本对的相似度, 以提升跨模态检索性能. Xu 等人[61]将图像区域和句子单词之间的全局语义一致性作为局部对齐的补充信息, 提出一种语义一致性跨模态注意力方法

① https://github.com/yiling2018/saem
② https://github.com/LgQu/CAMERA
③ https://github.com/yalesong/pvse
④ https://github.com/woodfrog/vse_infty
⑤ https://github.com/ZihaoWang-CV/CAMP_iccv19
⑥ https://github.com/CrossmodalGroup/BFAN

（Cross-modal Attention with Semantic Consistency，简称 CASC），用于图像文本匹配. 具体地，该方法利用跨模态注意力实现图像和文本的局部语义对齐，同时引入多标签预测用以保持全局语义一致性. 虽然取得不错的效果，但上述方法不能在训练中自适应准确区分相关和不相关词语-图像区域相似度的变化分布，这不可避免地限制了语义对齐学习. 为此，Zhang 等人[62]提出统一的自适应相关性可区分注意机制（Unified Adaptive Relevance Distinguishable Attention Network，简称 UARDA），该机制首次将相关性阈值和特征学习纳入一个统一的框架，通过最大限度地区分相关和不相关的跨模态实体对相似性分布来提高语义对齐. 鉴于上面方法均忽略了位置信息的重要性，Wang 等人[63]提出位置聚焦注意网络（Position Focused Attention Network，简称 PFAN），其利用区域的位置特征来增强原始的图像区域特征，而后通过一个视觉-文本注意力机制模块，来对齐图像区域信息和文本词语信息①. 之后，他们提出改进版的 PFAN++ 模型[64]，其通过集成全局特性，来进一步增强共享子空间学习，使匹配性能得到进一步提高.

由于执行固定步的注意力推理模型仅能对齐有限的跨模态语义信息，Lee 等人[65]提出堆叠交叉注意力网络（Stacked Cross Attention Network，简称 SCAN），其可以同时对所有可能的跨模态语义信息进行对齐②. 邓等人[66]提出文本-图像协同注意网络（Co-Attention Network，简称 CoAN），其通过数次相互指导的注意力机制，得到关键的视觉与文本表示，有效缩减异构模态数据间的语义鸿沟. Ji 等人[67]提出一种逐级分层对齐网络（Step-wise Hierarchical Alignment Network，SHAN），将图像-文本匹配分解为多步跨模态推理过程. 这种渐进式对齐策略为模型提供了更多互补的语义线索，以理解图像和文本之间的层次关系.

为更好地对齐图像和文本，一些方法采用非注意力机制的交互策略. 如，Wehrmann 等人[8]提出基于一个模态全局信息改进另一个模态实例的嵌入表示的方法，称为 ADAPT③. Chen 等人[68]提出基于语义一致性的图像-文本检索模型（Cross-Modal Image-Text Retrieval with Semantic Consistency，简称 CIRSC），其在排序目标函数中加入了语义一致性约束，使得图像和文本嵌入空间可以同时学习、互相受益，从而提高匹配性能④. 而 Li 等人[69]提出的多层次表示学习方法（Multi-Level Similarity Learning，简称 MLSL），则通过多标签卷积神经网络编码语义级信息，同时利用图网络捕获图像-文本对中对象和词的关系，以实现结构级信息编码.

3.2.3　基于混合交互建模的方法

为同时增强模态内和模态间关系建模，一些基于混合交互建模的方法被提出，它们既考虑文本-视觉交互信息建模，也考虑视觉和文本内部关联性关系建模. 如图 5 所示，一些方法采用串行混合模式，即先进行模态间交互，再进行模态内交互[9,70,71]；或者先进行模态内交互，再进行模态间交互[72-76]. 也有一些方法，同步进行模态内和模态间交互[13,77]. 类似于基于模态间和基于模态内交互方法，混合交互策略大多也是采用注意力机制进行相关信息的建模，少数工作采用图网络建模模态内或模态间关联关系.

以一些代表性的研究工作为例，Huang 等人[77]提出一种面向对象的跨模态注意网络（Object-oriented Attention Network，简称 OAN），其同时采用跨模态和模态内注意网络，建模模态间以及模态内依赖关系. 此外，它还设计了一个新的多模态保结构目标函数，用来强调模态内的难区分负样本. 而 Zhang 等人[9]提出的上下文感知注意网络（Context-Aware Attention Network，简称 CAAN），则先利用模态间注意力来发现词语-区域对之间的所有可能对齐，再结合模态内注意力来学习单个模态片段的语义关联. 类似地，Chen 等人[70]提出基于反复注意记忆的迭代匹配方法（Iterative Matching with Recurrent Attention Memory，简称 IMRAM），其也是先利用跨模态注意力单元进行跨模态信息对齐，然后采用记忆蒸馏单元调整各个模态内部信息，用于下一轮对齐匹配⑤. 与上述两种策略皆不同，Diao 等人[74]提出的名为 SGRAF 的图文匹配网络，先利用自注意力机制建模视觉和文本模态内的细粒度交互信息，然后采用跨模态注意力机制实现跨模态语义对齐⑥. Wei 等人[72]提出多模态交叉注意网络（Multi-Modality Cross Attention，简称 MMCA），其虽利用自注意机制模块建模模态内关系，但采用 Transformer 对由图像区域和句子单词堆叠的特征序列进行

① https://github.com/HaoYang0123/Position-Focused-\\Attention-Network

② https://github.com/kuanghuei/SCAN

③ https://github.com/jwehrmann/retrieval.pytorch

④ https://github.com/HuiChen24/MM_SemanticConsistency

⑤ https://github.com/HuiChen24/IMRAM

⑥ https://github.com/Paranioar/SGRAF

处理，以同时对图像区域和句子单词的模态间和模态内关系进行建模. 为了加强语义相关的对象之间的连接关系，Chen 等人[73]提出双径递归神经网络（Dual Path Recurrent Neural Network，简称 DP-RNN），通过递归神经网络，对称地处理图像和文本，来增强各个模态实例的表示. 而后利用交叉模态引导注意力机制和自注意力机制，聚集物体与词语之间的相似性，得到图像-文本的相似度. 不同于 DP-RNN 采用对称结构来表示两种模态信息，Ji 等人[75]提出显著性引导的注意力网络（Saliency-Guided Attention Network，简称 SAN），它先通过轻量级显著性检测器选择性地关注局部视觉特征，再通过多模态引导（融合视觉显著性、全局视觉信息和文本信息）注意力机制得到文本特征，以实现图文匹配. 与上述单步推理方法不同，Ji 等人[76]提出一种堆叠式多模态注意网络（Stacked Multimodal Attention Network，简称 SMAN），其依次以模内信息和多模态信息为指导，进行多步注意推理，实现图像和文本之间的细粒度关联建模.

虽然上述方法均取得不错的匹配效果，但它们的图像表示缺乏对象和关系语义信息来识别其对应的文本，同时文本表示涵盖了相当有限的视觉细节信息. 鉴于此，Long 等人[71]提出基于图的双模态表示模型（Graph-based Dual-modal representation，简称为 GraDual），其首先利用双模态图表示机制，将丰富的视觉语义信息集成到文本表示以及丰富的上下文语义信息集成到图像表示，而后利用一个模态实例的信息来生成另一个模态的过滤表示，以进行跨模态匹配. 为了加强细粒度匹配，Liu 等人[78]提出图结构化匹配网络（Graph Structured Matching Network，简称 GSMN），其将图像区域特征以及文本词语特征分别组建为图结构，通过衡量异构图相似性，得到图像-文本匹配得分①. 此外，现有混合交互建模方法均为静态模型架构，即对于任何输入数据，均经过同样复杂的处理流程. 但是，对于一些简单的数据对而言，并不需要复杂的交互机制处理. 鉴于此，Qu 等人[13]提出动态模态交互建模网络（DynamIc Modality intEraction，简称 DIME），其可以依据输入数据信息，动态自适应地探索不同的交互方式②，用于图像-文本匹配.

3.2.4 小结

为了有效解决图像-文本匹配的模态内语义理解与建模挑战，基于局部特征的匹配方法设计不同的注意力机制以及图卷积网络，来挖掘各个模态内有用的细粒度实体信息以及实体间关联关系. 同时，为了提升模态间语义的对齐和匹配，许多基于注意力机制的模态间关联性关系建模策略被提出，通过对关键性跨模态语义信息对进行对齐和聚集，以得到精准的图像-文本匹配分数估计. 虽然这些方法从不同层面尝试解决图像-文本匹配任务的两个挑战且取得了一定效果，但这些复杂的关联性关系建模设计致使匹配效率过低. 另外，这些方法的底层特征往往由其他数据集上预训练模型所提取（如，Faster R-CNN 和 word2vec），导致模型对实体信息理解较为局限，易受上游预训练模型影响.

3.3 基于外部知识的图像-文本匹配方法

虽然基于局部特征的图像-文本匹配方法捕获了细粒度语义信息用于跨模态语义匹配，但它们对一些语义信息的理解仍不全面，且对一些语义关联关系的挖掘不够充分.为缓解该问题，基于外部知识的图像-文本匹配方法被提出. 这类方法大多从外部获取一定的先验知识，如预训练任务或场景图，用于提升对视觉或者文本的语义理解，继而提升图像-文本匹配的精准度.

以一些代表性工作为例，Lin 等人[79]将视觉问答（Visual Question Answering，简称 VQA）视为一个特征提取模块，来提取图像和文本的特征表示，并将这些表示用于图像-文本匹配任务. Shi 等人[80]旨在利用从大量的图像场景图中提取共现常识知识——场景概念图（Scene Concept Graph，简称 SCG），来增强图像表示. HOAD[14]则采用场景图来表示高度结构化的视觉或文本语义，并将视觉-语义匹配任务转化为异构图匹配问题. 虽然 SGM[15]也提出用场景图来表示图像和文本，但其利用模态场景图来共同表征相应模态的对象信息和关系信息，并基于对象层、关系层以及全局层的跨模态特征，来评估图像和文本的相似度. 为了避免关系信息的丢失，Guo 等人[81]提出了一种结合场景图和图卷积网络的视觉关系建模框架（被命名为 VRACR），该方法同时考虑图像视觉特征和关系特征，这两种特征分别与文本特征在两个嵌入空间上对齐，最终的相似度评分为两个嵌入空间中相似度求和. 受 Transformer 在不同领域被成功应用的影响，Dong 等人[82]提出基于 Transformer 的分层特征聚合算法（Hierarchical feature Aggregation algorithm based on Transformer，简称 HAT），其利用现成的场景图

① https://github.com/CrossmodalGroup/GSMN

② https://github.com/LgQu/DIME

提取器分别生成图像和文本场景图，同时利用层次图卷积网络逐层生成融合属性信息和关系信息的对象表示；然后，将对象表示和来自其他模态的全局特征同时输入 Transformer 进行跨模态融合；最后，将融合后的特征映射到公共空间中，用于度量图像-文本匹配程度.

不同于基于场景图方法，Wang 等人[7]提出一种共识感知的视觉语义嵌入（Consensus-aware Visual-Semantic Embedding，简称 CVSE）模型，它将共识信息，即两种模态间共享的常识性知识，整合到图像-文本匹配中①. Zhang 等人[83]则将外延图（Denotation Graphs，简称 DG）融入到图像和文本表示学习中，这里外延图可以看作是概念及其对应视觉目标之间语义知识的层次组织，其中节点是比目标标签语义丰富的复合短语，且节点之间的关系也更丰富. 此外，鉴于神经机器翻译中的编码器-解码器结构能够丰富单语言和多语言文本多样性这一事实，Huang 等人[84]利用多模态神经机器翻译（Multimodal neural Machine Translation，简称 MMT），进行基于显著性视觉对象的正向和反向翻译，以生成额外的文本-图像对，从而提升单语言跨模态检索和多语言跨模态检索模型性能. Sun 等人[85]则关注匹配效率问题，提出一种简单且高效的匹配方法，即 LightningDOT. 它通过在三个新学习目标上进行预训练，可以离线提取特征并采用点积匹配，这显著加快了匹配检索过程. 在不牺牲准确性的前提下，它将图像-文本匹配时间加快了数千倍②.

总的来讲，通过引入实体语义信息以及实体间关联关系的先验知识，一定程度上可以增强模态内语义信息的理解，并且有助于跨模态语义的匹配. 但是，由于先验知识数据来源与图像文本匹配任务数据集之间的域差异，可能会导致部分语义理解偏差且引入噪声，继而影响匹配结果.

3.4　基于度量学习的图像-文本匹配方法

不同于上述三种类别方法，基于度量学习的图像-文本匹配方法旨在探究更好的模型约束[21,22,86]或相似性度量机制[19,87]，以提升不同样本对间的可区分性和同一样本对间的关联性. 特别地，这些方法往往可嵌入到上述三种类别方法的模型中，进一步提升它们的性能.

举例来说，现有工作大多采用排序损失（Ranking Loss）学习视觉-语义嵌入表示，由于很难一开始就找到合适的三元组，故直接在异构特征上执行排序损失会使效果会变差，且损害网络对多模态关系的学习. 为此，一些方法提出加权策略，通过调整样本对的权值，提升模型性能. 例如，Zheng 等人[86]提出实例损失函数（Instance Loss）为排序损失提供更好的权重初始化，从而产生更具鉴别性和鲁棒性的图像和文本表示③. 类似地，Wei 等人[22]提出一种通用加权框架，能够有效地为信息对分配适当的权重，即更大的权重被分配给信息更丰富的样本对，同时引入多项式损失函数（Polynomial Loss）④. Liu 等人[21]提出枢纽感知损失函数（Hubness-Aware Loss，简称 HAL），其利用集散中心的信息来自动调整样本的权重⑤. Wei 等人[88]引入自相似多项式损失函数和相对相似多项式损失函数，以有效地抽样有效信息对，并分配适当的权重值给它们用于模型训练. 不同于权重调整策略，一些方法直接创新性地引入新的损失函数约束. 例如，Wang 等人[2]在双向三元组损失之外，引入了结构约束损失函数⑥. Zhang 等人[25]提出跨模态投影匹配（Cross-Modal Projection Matching，简称 CMPM）损失，来学习具有判别性的图像-文本嵌入表示⑦. Thomas 等人[89]提出一种新颖的模态损失，鼓励文本和图像子空间的语义一致性，确保语义相似的内容（如，图像-图像对）在子空间中也非常接近. Liu 等人[90]提出 kNN 边界损失（KNN-margin Loss），CVSE++[91]设计了阶梯损失⑧. 为提高模型区分正负的泛化能力，Chen 等人[23]提出一种离线负样本采样方案和五元组损失函数（Quintuplet Loss），其从整个训练集中离线取样负样本，使得得到的负样本更具区分性⑨. 不同于基于损失函数设计的策略，一些方法从相似性度量设计角度出发，解决图像-文本匹配任务. 如，Chen 等人[87]提出图最优传输（Graph Optimal Transport，简称 GOT）框架，将图像-文本匹配任务转化为跨模态图匹配问题，并提出两种最优传输度量，即用于进行节点匹配的 Wasserstein 距离和用于边匹配的 Gromov-Wasserstein 距离. 考虑到不匹配的片段中蕴含的重要线索，Zhang 等人[19]提出消极感知注意框架（Negative-Aware Attention Framework，简称

① https://github.com/BruceW91/CVSE

② https://github.com/intersun/LightningDOT

③ https://github.com/layumi/Image-Text-Embedding

④ https://github.com/wayne980/PolyLoss

⑤ https://github.com/hardyqr/HAL

⑥ https://github.com/lwwang/Two_branch_network

⑦ https://github.com/labyrinth7x/Deep-Cross-Modal-Projection-Learning-for-Image-Text-Matching

⑧ https://github.com/cdluminate/ladderloss

⑨ https://github.com/sunnychencool/AOQ

NAAF），它明确地利用匹配片段的积极作用和不匹配片段的消极作用来联合推断图像-文本的相似性①. Zhang 等人[20]提出跨模态信心感知网络（Cross-Modal Confidence-Aware Network，简称 CMCAN），该网络考虑了匹配区域-词对的置信度，并将其与局部语义相似度相结合，以准确衡量跨模态相关性. 此外，由于目前广泛采用的 Recall@K 不能完全评估给定图像与检索到的句子的准确性和覆盖率的程度，Biten 等人[24]将图像描述度量引入到图像文本任务评估中②.

总体来讲，基于度量学习的图像-文本匹配方法通常在不影响匹配效率的情况下，可进一步提升匹配精度. 因为它们仅仅对模型优化的目标函数进行改进，未引进过多的额外参数以及交互模块设计. 但是，相较于前面三类方法，它们并未从本质上解决模态内语义理解以及模态间语义对齐两个挑战.

表 1 视觉-文本多模态预训练模型总结

方法	结构	预训练数据集	预训练任务	下游任务	发表时间
ViLBERT[92]③	双流框架	Conceptual Captions、BooksCorpus、Visual Genome	掩码语言模型、掩码区域模型、图像文本匹配	视觉问答、视觉常识推理、指称表达理解、图文检索	NIPS 2019
TDEN[93]④	双流框架	Visual Genome、Conceptual Captions	掩码语言模型、掩码区域模型、图像文本匹配、掩码句子生成	视觉问答、视觉常识推理、图像描述、图文检索	AAAI 2021
LXMERT[94]⑤	双流框架	COCO、Visual Genome、VQA、GQA、Visual7W	图像文本匹配、掩码语言模型、掩码区域模型、视觉问答	视觉问答、自然语言视觉推理	EMNLP 2019
12-in-1[95]⑥	双流框架	Conceptual Captions	多任务学习	视觉问答、自然语言视觉推理、指称表达理解、图文检索、视觉蕴含	CVPR 2020
ERNIE-ViL[96]⑦	双流框架	Conceptual Captions、SBU Captions	掩码语言模型、掩码区域模型、图像文本匹配、场景图预测	视觉问答、视觉常识推理、指称表达理解、图文检索	AAAI 2021
METER[97]⑧	双流框架	Visual Genome、Conceptual Captions\\COCO、SBU Captions	掩码语言模型、图像文本匹配、掩码图像预测	视觉问答、视觉常识推理、视觉蕴含、图文检索	CVPR 2022
ALIGN[98]	双流框架	A Large-Scale Noisy Image-Text Dataset	图像文本匹配	图文检索、视觉分类、语义文本相似、语义图像相似性、语义图文相似性	ICML 2021
CLIP[99]⑨	双流框架	WebImageText	图像文本匹配	OCR、动作识别、地理定位、身份识别等多种视觉分类任务	ICML 2021
BriVL[100]⑩	双流框架	RUC-CAS-WenLan	图像文本匹配	图文检索、图像描述	Arxiv 2021
GilBERT[101]	单流框架	Conceptual Captions、COCO、SBU Captions、Flicker30k、GQA	掩码语言模型、掩码视觉表示分类	图文检索、图像描述、视觉问答、新对象图像描述、自然语言视觉推理	SIGIR 2021
Visual-BERT[[102]⑪	单流框架	COCO Captions	文本视觉对齐、掩码语言模型	视觉问答、视觉常识推理、自然语言视觉推理、短语定位	Arxiv 2019
Uni-coder-VL[103]	单流框架	Conceptual Captions、SBU Captions	文本视觉对齐、掩码语言模型、掩码视觉表示分类	图像文本检索、零样本图像文本检索	AAAI 2020

① https://github.com/CrossmodalGroup/NAAF
② https://github.com/furkanbiten/ncs_metric
③ https://github.com/jiasenlu/vilbert_beta
④ https://github.com/YehLi/TDEN
⑤ https://github.com/airsplay/lxmert
⑥ https://github.com/facebookresearch/vilbert-multi-task
⑦ https://github.com/PaddlePaddle/ERNIE/tree/repro/ernie-vil
⑧ https://github.com/zdou0830/METER
⑨ https://github.com/openai/CLIP
⑩ https://github.com/BAAI-WuDao/BriVL
⑪ https://github.com/uclanlp/visualbert

续表 1　视觉-文本多模态预训练模型总结

方法	结构	预训练数据集	预训练任务	下游任务	发表时间
VL-BERT[104]①	单流框架	Conceptual Captions、BooksCorpus、English Wikipedia	掩码语言模型、掩码视觉表示分类	视觉问答、视觉常识推理、指称表达理解	ICLR 2020
OSCAR[105]②	单流框架	SBU Captions、Conceptual Captions、COCO、Flickr30K、GQA	掩码语言模型、掩码区域建模	视觉问答、自然语言视觉推理、图文检索、图像描述、新对象图像描述、图像问答	ECCV 2020
VinVL[106]③	单流框架	SBU Captions、COCO、Flickr30K、OpenImages、VQA、Conceptual Captions、GQA、Visual Genome	掩码语言模型、掩码区域建模	视觉问答、自然语言视觉推理、图文检索、图像描述、新对象图像描述、图像问答	CVPR 2021
Pixel-BERT[107]	单流框架	COCO、Visual Genome	掩码语言模型、图像文本匹配	视觉问答、自然语言视觉推理、图文检索	Arxiv 2020
ViLT[108]④	单流框架	SBU Captions、COCO、Conceptual Captions、Visual Genome	掩码语言模型、图像文本匹配	视觉问答、自然语言视觉推理、图文检索	ICML 2021
SimVLM[109]	单流框架	A large-scale Noisy Image-Text Dataset	前缀语言预测	视觉问答、自然语言视觉推理、图像描述、视觉蕴含、多模态翻译	ICLR 2022
UNITER[110]⑤	单流框架	COCO、Visual Genome、Conceptual Captions、SBU Captions	掩码语言模型、掩码区域建模、图像文本匹配、词语区域对齐	视觉问答、视觉常识推理、自然语言视觉推理、视觉蕴含、图像文本检索、指称表达理解	ECCV 2020
Unified-VLP[111]⑥	单流框架	Conceptual Captions	双向掩码语言模型、序列到序列掩码语言模型	视觉问答、图像描述	AAAI 2020
XGPT[112]	单流框架	Conceptual Captions	图像条件掩码语言模型、图像条件去噪编码、文本条件图像特征生成、图像描述	图像描述	NLPCC 2020
Image-BERT[113]	单流框架	SBU Captions、Conceptual Captions、LAIT	掩码语言模型、掩码区域建模、图像文本匹配	图文检索	Arxiv 2020
UNIMO[114]	单流框架	SBU Captions、COCO、BooksCorpus、OpenImages、Visual Genome、English Wikipedia	掩码区域建模、图像文本匹配、双向文本预测、序列到序列文本生成	视觉问答、图像描述、视觉蕴含、文本分类、文本摘要、问题生成	ACL 2021
VILLA[115]⑦	单流框架	SBU Captions、COCO、Visual Genome、Conceptual Captions	掩码语言模型、掩码区域建模、图像文本匹配	视觉问答、自然语言视觉推理、视觉常识推理、指称表达理解、图文检索、视觉蕴含	NIPS 2020
VIVO[116]	单流框架	OpenImages	区域标签匹配	新对象图像描述、图像分类	AAAI 2021
B2T2[117]⑧	单流框架	Conceptual Captions	掩码语言模型、图像文本匹配	视觉问答、自然语言视觉推理、视觉常识推理、图文检索	EMNLP 2019
VL-T5[118]⑨	单流框架	COCO Captions、COCO、Visual Genome、VQA v2.0、GQA、Visual7W	掩码语言模型、图像文本匹配、视觉定位、文本定位、视觉问答	视觉问答、自然语言视觉推理、视觉常识推理、指称表达理解、图像描述、图形问答、多模态翻译	ICML 2021
M6[119]	单流框架	M6-Corpus	文本-文本传输、图像-文本传输、多模态-文本传输	视觉问答、图文检索、图像描述、文本生成图像、长文本答案生成、诗歌生成	SIGKDD 2021

① https://github.com/jackroos/VL-BERT
② https://github.com/microsoft/Oscar
③ https://github.com/pzzhang/VinVL
④ https://github.com/dandelin/vilt
⑤ https://github.com/ChenRocks/UNITER
⑥ https://github.com/LuoweiZhou/VLP
⑦ https://github.com/zhegan27/VILLA
⑧ https://github.com/google-research/language/tree/master/language/question_answering/b2t2
⑨ https://github.com/j-min/VL-T5

3.5 图像-文本多模态预训练模型

上述几种类别方法仅在有限数据上进行训练，故泛化性能较弱. 受计算机视觉和自然语言处理预训练模型成功的启发，近期许多视觉-文本多模态预训练模型被提出. 它们可以被分类为两类：（1）单流框架，视觉信息和文本信息由一个多模态编码器联合处理；（2）双流框架，单独地对视觉信息和文本信息进行建模，而后再进行多模态联合建模.

如表 1 所示，双流框架工作，如 ViLBERT[92]、TDEN[93]、LXMERT[94]、12-in-1[95]、ERNIE-ViL[96]、METER[97]、ALIGN[98]、CLIP[99]以及 BriVL[100]，通常用两个单模态网络分别处理输入的文本和图像，然后利用一个跨模态融合网络对不同模态进行关联. 作为该分支的代表性工作，ViLBERT 分别采用 Faster-RCNN 进行视觉信息处理和类 BERT 进行文本信息处理，而后采用协同注意 Transformer 进行跨模态特征融合. 考虑到视觉以及文本语义关系的重要性，LXMERT 在视觉编码分支引入对象关系编码器，而 ERNIE-ViL 在文本编码过程中引入场景图. 与此同时，12-in-1 提出多任务的预训练方法来削弱模型对数据集规模的依赖性，TDEN 则利用不同任务间的相关性来提升预训练的鲁棒性. 不同地，ALIGN、CLIP 和 BriVL 分别构建了大规模图文数据集，并基于对比学习进行图像文本匹配预训练. METER 以端到端方式设计和预训练一个完全基于 Transformer 的视觉-语言模型. GilBERT[101]在学习图像-文本数据的通用表示同时，还可以补全缺失模态.

不同于双流框架的视觉-文本预训练模型，单流框架的视觉-文本预训练模型结构较为简单，可以更早且更加自由地对两种模态信息进行融合. Visual-BERT[102]作为该分支的先驱性工作，其将文本单词信息以及图像区域信息共同输入到 Transformer 中进行预训练. 除了利用 Faster-RCNN 提取的区域视觉特征外，Unicoder-VL[103]还利用区域的标签信息以及位置信息进行视觉编码，VL-BERT[104]则新增一个视觉特征编码器. 而 OSCAR[105]不直接使用 Transformer 的自注意机制实现图文匹配，而是将相同语义下的物体作为图像和语言对齐的连接点，从而简化图像和文本之间的语义对齐学习任务. VinVL[106]通过对 OSCAR 的物体检测模型进行改进，进一步提升预训练效果. 不同于基于区域图像特征的预训练模型，Pixel-BERT[107]使用了传统的像素级特征进行预训练，实现了一种端到端的多模态训练方式，在一些任务上取得了更好的结果. ViLT[108]提出了一个极简的视觉-文本预训练模型，其把视觉编码模块变为无卷积的处理方式，提升了模型的效率. SimVLM[109]采用类似的视觉处理方式，利用大规模弱标记数据集进行预训练，故具有强大的泛化和迁移能力. 不同于上述工作侧重于对输入的编码器进行优化，UNITER[110]、Unified-VLP[111]和 XGPT[112]通过改进预训练任务来增强模型的学习能力；Image-BERT[113]、UNIMO[114]、VILLA[115]和 VIVO[116]通过优化训练策略来提升模型性能，而 B2T2[117]为针对视觉常识推理任务设定的预训练模型. 不同于现有的视觉和语言学习方法通常需要为不同任务设计特定于任务的架构和目标，VL-T5[118]提出了一个统一的框架，其使用相同的语言建模目标来学习不同任务. 考虑到中文多模态领域的蓬勃发展，达摩院智能计算实验室认知智能团队推出了大规模中文多模态预训练数据集和模型[119].

总体而言，图像-文本多模态预训练模型的泛化性能要优于前面几类方法. 由于训练任务以及数据的多样性，它们对于各个模态的语义理解也更加充分. 但是，这些模型往往复杂度和训练成本颇高.

3.6 相关研究任务

近几年，随着弱监督学习以及零样本学习等技术的兴起，一些研究人员将传统图像-文本匹配任务进行了扩展，提出了弱监督跨模态匹配任务、零样本跨模态匹配任务、场景文本感知的跨模态检索以及跨语言图像检索任务. 下文将逐一对这些任务进行介绍.

3.6.1 弱监督跨模态匹配

现有基于深度学习的图像-文本匹配方法极度依赖大规模人类标注的图像-文本对信息，即为全监督学习方法. 但收集大量高质量的人工注释图像-文本对通常非常昂贵和不切实际. 鉴于此，Huang 等人[120]针对稀疏标注场景（对于大规模图像数据，只有一小部分图像被标注了语义匹配的文本描述信息），提出对抗性注意对齐模型（Adversarial Attentive Alignment model for learning Visual-Semantic Embedding，简称 A3VSE），用于跨模态匹配. 该模型联合利用了来自标注的图像-文本对的强监督信号和从无注释图像自动提取的区域语义得到的弱监督信号，来学习视觉语义嵌入. 与此同时，该模型采用注意对抗目标函数，选择性地对齐视觉和文本输入中部分的实体，以缩小二者之间的域差距. 类似地，Huang 等人[121]关注少样本图像和句子匹配问

题（图像和文本中所包含区域和单词属于少样本类别），并提出一种对齐交叉模态记忆（Aligned Cross-Modal Memory，简称 ACMM）模型，该方法不仅能以弱监督方式对少样本内容进行对齐和记忆，还能自适应地平衡其与普通内容关系.

3.6.2　零样本跨模态匹配

Xu 等人[122]针对零样本跨模态检索问题，提出一种集成自动编码器和生成式对抗网络方法（Assembling AutoEncoder and Generative Adversarial Network，简称 AAEGAN），该方法可同时学习共同的潜空间和合成多模态特征. 为了加强共同潜空间的学习，AAEGAN 提出了一种分布对齐约束，以保持模态间的语义兼容性. Lin 等人[123]提出一种新的零样本跨模态检索框架，被称为 LCALE（Learning Cross-Aligned Latent Embeddings），该方法通过相关变分自编码器在低维潜在空间中生成潜在嵌入，实现了稳定的训练和卓越的检索性能. 与此同时，LCALE 利用跨模态重建和跨模态对齐的方式，来匹配不同模态数据的潜在嵌入，有效地增强了潜在嵌入空间学习.

3.6.3　场景文本感知的跨模态检索

Mafla 等人[124]关注于场景文本感知的跨模态检索（Scene-Text Aware Cross-Modal Retrieval,简称 StacMR）任务，该任务将图像场景中文本信息作为辅助信息，用以实现跨模态检索①.

3.6.4　跨语言图像检索

跨语言图像检索，即给定一个中文句子作为查询，目标是在一组图像中找到与查询最匹配的图像，该图有一个英文描述语句. 跨语言图像检索与单语言图像检索不同，前者不仅要考察视觉和文本模式之间的相似性，而且还要考察两种不同语言呈现的文本之间的相似性. Li 等人[125]用手写中文句子对 MSCOCO 数据集进行了扩充，提出新的数据集 COCO-CN②. 同时，基于 W2VV[126]，他们提出一个跨语言图像检索模型，其用对比损失替换 MSE 损失函数对模型进行训练.

4　数据集和评价标准

4.1　数据集

对于图像-文本匹配任务而言，目前多在图像-文本检索任务数据集上对其进行测试，主要采用的公开数据集如下:

（1）CUHK-PEDES 数据集[127]：该数据集是一个带有文本语句标注的行人数据集，共包含 40 206 张图像和 80 412 个文本语句.

（2）Flickr8K 数据集[128]：该数据集是一个由 8 092 张图片组成的众包数据集，单图片有 5 个对应的人类生成的文本语句描述.

（3）Flickr30K 数据集[129]：该数据集是 Flickr8K 数据集的扩展，共包含 31 783 张图像. 与 Flickr8K 一样，单张图片被 5 个文本句子所标注.

（4）MSCOCO 数据集[130]：该数据集起初是为图像目标识别、目标检测和标题生成任务而创建，目前也经常用于图像-文本匹配任务. 它由 123 287 张图像组成，并且单张图像标注 5 个文本句子.

（5）MUGE 数据集[119]：考虑到中文多模态领域的蓬勃发展，达摩院智能计算实验室认知智能团队推出了大规模中文多模态评测基准牧歌③（Multi-modal Understanding and Generation Evaluation，简称 MUGE），它是业界首个大规模中文多模态评测基准，包括图文描述、基于文本的图像生成、跨模态检索等. 同时，其构建了中文多模态预训练数据集，包含超过 1.9TB 图像和 292GB 文本.

（6）Wukong 数据集[131]：该数据集是一个大型中文跨模态数据集，包含了 1 亿个来自互联网的中文图像-文本对④.

由于 Flickr8K 和 Flickr30K 两个数据集的数据来源相同，仅在数据规模方面不同，现有方法普遍采用数据量更大的 Flickr30K 数据集进行实验. 而 CUHK-PEDES 数据集主要用行人-文本检索任务，在图像-文本匹配任务中极少被采用. 至于 Wukong 与 MUGE 为最近提出的中文数据集，目前在上面进行实验评估的方法非常少. 总结来说，目前在图像-文本匹配研究方向，大家常用数据集 Flickr30K 和 MSCOCO 进行性能评估. 但是，现有图像-文本匹配方法在利用上述两个数据集进行性能评估时，对于该数据集的训练集、验证集和测试集的划分策略不尽相同，如表 2 所示. 注意，对于 MSCOCO 而言，若测试集包含 5 000 张图像，一般设置两种测试方式：（1）MSCOCO1K，即将 5 000 张图像划分为 5 部分，分别包含 1 000 张图像,最终测试结果为在 5 个测试集结果的平均值；（2）MSCOCO5K，直接对 5 000 张图像进行测试.

① https://github.com/AndresPMD/StacMR

② https://github.com/li-xirong/coco-cn

③ https://tianchi.aliyun.com/muge

④ https://wukong-dataset.github.io/wukong-dataset/index.html

4.2 评价指标

对于图像-文本匹配任务，通常采用 Recall@K（R@K）作为评估指标，即正确结果排在前 K 个检索结果的查询的比例，这里 K 的取值一般为 1、5 以及 10. 除上述指标外，也有些方法采用 RSUM 作为评估指标，即将图像-文本检索和文本-图像检索两个方向的结果累加，即

$$RSUM = \underbrace{R@1+R@5+R@10}_{i-t}+\underbrace{R@1+R@5+R@10}_{t-i}\cdot$$

5 现有方法的性能及分析

我们总结了现有图像-文本匹配方法在两个大规模数据集：Flickr30K 和 MSCOCO 上的实验结果. 实验结果分别总结在表 3~表 12 中，其中表 3~表 5 总结了基于全局特征的图像-文本匹配方法在 Flickr30K 和 MSCOCO 数据集上实验结果；表 6~表 8 展示了基于局部特征的图像-文本匹配方法在 Flickr30K 和 MSCOCO 数据集上的结果；表 9 总结了多模态预训练模型在图像-文本匹配任务上的结

表 2 现有方法对 Flickr30K 和 MSCOCO 数据集的不同划分策略

名称	数据规模	模型数据	训练集	测试集	验证集	方法
Flickr30K	31,783	31,014	29,000	1,014	1,000	文献[82]
	31,783	31,783	30,000	1,000	1,000	文献[10]
	31,783	30,000	28,000	1,000	1,000	文献[1, 12, 16, 35, 34, 41, 42, 43, 49, 50, 52, 53, 61, 81]
	31,783	31,783	29,783	1,000	1,000	文献[5, 7, 11, 14, 15, 18, 25, 32, 37, 56, 58, 69, 73-76, 80, 91, 120]
	31,783	31,000	29,000	1,000	1,000	文献[6, 8, 9, 13, 17, 19, 20, 33, 40, 44-47, 54, 57, 59, 60, 62-65, 67, 68, 70-72, 77, 78, 119]
MSCOCO	123,287	91,783	82,783	4,000	5,000	文献[121]
	123,287	118,287	82,783	30,504	5,000	文献[25]
	123,287	123,287	82,783	40,504	1,000	文献[32]
	123,287	92,287	82,783	5,000	5,000	文献[34]
	123,287	88,783	82,783	1,000	5,000	文献[79]
	123,287	123,287	113,287	1,000	5,000	文献[71]
	123,287	123,287	82,783	35,504	5,000	文献[82]
	123,287	123,287	123,287	1,000	5,000	文献[7, 12]
	123,287	87,783	82,783	4,000	1,000	文献[40, 42, 50, 56]
	123,287	123,000	113,000	5,000	5,000	文献[17, 33, 35, 47, 63, 64]
	123,287	123,287	113,287	5,000	5,000	文献[6, 8-10, 13-15, 18-20, 36, 41, 43-46, 48, 49, 51-54, 57-62, 65, 67-70, 72-78, 80, 81, 91, 119]

表 3 基于全局特征的图像-文本匹配方法在 Flickr30K 数据集上的结果

方法	图像	文本	Image-to-Text(i-t)			Text-to-Image(t-i)			RSUM	发表时间	类别
			R@1	R@5	R@10	R@1	R@5	R@10			
UVS[16]	OxfordNet	LSTM	23.0	50.7	62.9	16.8	42.0	56.5	251.9	NIPS 2014	嵌入架构
DSPE[33]	VGG-19	Fisher Vector	40.3	68.9	79.9	29.7	60.1	72.1	351.0	CVPR 2016	嵌入架构
RRF-Net[32]	ResNet-152	HGLMM	47.6	77.4	87.1	35.4	68.3	79.9	395.7	ICCV 2017	嵌入架构
2WayNet[35]	VGG-19	Fisher Vector	49.8	67.5	–	36.0	55.6	–	–	CVPR 2017	嵌入架构
VSE++[34]	ResNet-152	GRU	43.7	71.9	82.1	32.3	60.9	72.1	363.0	BMVC 2018	嵌入架构
TIMAM[37]	ResNet-101	BERT	**53.1**	**78.8**	**87.6**	**42.6**	**71.6**	**81.9**	**415.6**	ICCV 2019	嵌入架构
m-CNNs[40]	VGG-19	Skim-Gram	33.6	64.1	74.9	26.2	56.3	69.6	324.7	ICCV 2015	交互架构
MTFN[41]	Faster-RCNN	GRU	**63.1**	**85.8**	**92.4**	**46.3**	**75.3**	**83.6**	**446.5**	ACM MM 2019	交互架构

表 4 基于全局特征的图像-文本匹配方法在 MSCOCO 数据集 1K 测试集上的结果

方法	图像	文本	Image-to-Text(i-t)			Text-to-Image(t-i)			RSUM	发表时间	类别
			R@1	R@5	R@10	R@1	R@5	R@10			
DSPE[33]	VGG-19	Fisher Vector	50.1	79.7	89.2	39.6	75.2	86.9	420.7	CVPR 2016	嵌入架构
2WayNet[35]	VGG-19	Fisher Vector	55.8	75.2	–	39.7	63.3	–	–	CVPR 2017	嵌入架构
RRF-Net[32]	ResNet-152	HGLMM	56.4	85.3	91.5	43.9	78.1	88.6	443.8	ICCV 2017	嵌入架构

续表 4　基于全局特征的图像-文本匹配方法在 MSCOCO 数据集 1K 测试集上的结果

方法	图像	文本	Image-to-Text(i-t)			Text-to-Image(t-i)			RSUM	发表时间	类别
			R@1	R@5	R@10	R@1	R@5	R@10			
VSE++[34]	ResNet-152	GRU	64.6	90.0	95.7	52.0	84.3	92.0	478.6	BMCV 2018	嵌入架构
GXN[36]	ResNet-152	Bi-GRU	**68.5**	–	**97.9**	**56.6**	–	**94.5**	–	CVPR 2018	嵌入架构
m-CNNs[40]	VGG19	Skim-Gram	42.8	73.1	84.1	32.6	68.6	82.8	384.0	ICCV 2015	交互架构
MTFN[41]	Faster-RCNN	GRU	**71.9**	**94.2**	**97.9**	**57.3**	**88.6**	**95.0**	**504.9**	ACM MM 2019	交互架构

表 5　基于全局特征的图像-文本匹配方法在 MSCOCO 数据集 5K 测试集上的结果

方法	图像	文本	Image-to-Text(i-t)			Text-to-Image(t-i)			RSUM	发表时间	类别
			R@1	R@5	R@10	R@1	R@5	R@10			
VSE++[34]	ResNet-152	GRU	41.3	71.1	81.2	30.3	59.4	72.4	355.7	BMCV 2018	嵌入架构
GXN[36]	ResNet-152	Bi-GRU	**42.0**	**–**	**84.7**	**31.7**	**–**	**74.6**	**–**	CVPR 2018	嵌入架构
MTFN[41]	Faster-RCNN	GRU	44.7	76.4	87.3	33.1	64.7	76.1	382.3	ACM MM 2019	交互架构

表 6　基于局部特征的图像-文本匹配方法在 Flickr30K 数据集上的结果

方法	图像	文本	Image-to-Text(i-t)			Text-to-Image(t-i)			RSUM	发表时间	类别
			R@1	R@5	R@10	R@1	R@5	R@10			
DVSA[17]	R-CNN	BRNN	22.2	48.2	61.4	15.2	37.7	50.5	235.2	CVPR 2015	先驱工作
HM-LSTM[47]	R-CNN	HM-LSTM	38.1	–	76.5	27.7	–	68.8	–	ICCV 2017	文本建模
Sm-LSTM[50]	VGG19	Bi-LSTM	42.5	71.9	81.5	30.2	60.4	72.3	358.8	CVPR 2017	双模态建模
DAN[5]	ResNet-152	Bi-LSTM	55.0	81.8	89.0	39.4	69.2	79.1	413.5	CVPR 2017	双模态建模
SCO[42]	VGG19	LSTM	55.5	82.0	89.3	41.1	70.5	80.1	418.5	CVPR 2018	视觉建模
VSRN[43]	Faster R-CNN	GRU	71.3	90.6	96.0	54.7	81.8	88.2	482.6	ICCV 2019	视觉建模
LIWE[49]	Faster R-CNN	Bi-GRU	66.4	88.9	94.1	47.5	76.2	84.9	458.1	ICCV 2019	文本建模
SAEM[18]	Faster R-CNN	BERT	69.1	91.0	95.1	52.4	81.1	88.1	476.8	ACM MM 2019	双模态建模
TST[10]	ResNet-152	Bi-GRU	57.7	–	89.2	42.9	–	79.3	–	ACM MM 2019	双模态建模
DSRAN[44]	Faster R-CNN and ResNet-152	BERT	77.8	95.1	97.6	59.2	86.0	91.9	507.6	IEEE TCSVT 2020	视觉建模
CRGN[45]	Faster R-CNN and ResNet-152	GRU	70.5	91.2	94.9	50.3	77.7	85.2	469.8	IEEE TIP 2020	视觉建模
CAMERA[6]	Faster R-CNN	BERT	78.0	95.1	97.9	60.3	85.9	91.7	508.9	ACM MM 2020	双模态建模
GARN[11]	ResNet-152	Bi-LSTM	60.1	84.6	90.4	44.2	71.2	80.3	430.8	IEEE TIP 2021	文本建模
CMRN[12]	Faster R-CNN	BERT-GRU	70.8	91.5	95.4	55.2	81.8	88.1	482.8	MTA 2021	视觉建模
DASPGA[53]	Faster R-CNN	FC	**82.8**	**95.9**	**97.9**	**62.2**	**89.3**	**93.8**	**521.9**	CVPR 2021	双模态建模
RRTC[46]	Faster R-CNN	Bi-GRU	72.7	93.8	96.8	54.2	79.4	86.1	483.0	IEEE TCSVT 2022	视觉建模
ReSG[52]	Faster R-CNN	Bi-GRU	77.2	94.2	98.2	58.0	83.1	88.7	499.4	IEEE T-Cybern 2022	双模态建模
BSSAN[56]	Faster R-CNN	Bi-LSTM	38.4	67.2	77.5	28.5	57.5	67.9	337.0	IEEE TIP 2018	跨模态交互建模
SCAN[65]	Faster R-CNN	Bi-GRU	67.4	90.3	95.8	48.6	77.7	85.2	465.0	ECCV 2018	跨模态交互建模
PFAN[63]	Faster R-CNN	Bi-GRU	70.7	91.8	95.0	50.4	78.7	86.1	472.7	IJCAI 2019	跨模态交互建模
RDAN[57]	Faster R-CNN	Bi-GRU	68.1	91.0	95.9	54.1	80.9	87.2	477.2	IJCAI 2019	跨模态交互建模
BFAN[59]	Faster R-CNN	Bi-GRU	68.1	91.4	–	50.8	78.4	–	–	ACM MM 2019	跨模态交互建模
CIRSC[68]	Faster R-CNN	Bi-GRU	69.7	91.7	96.4	54.0	79.7	87.2	478.7	ACM MM 2019	跨模态交互建模
CAMP[58]	Faster R-CNN	Bi-GRU	68.1	89.7	95.2	51.5	77.1	85.3	466.9	ICCV 2019	跨模态交互建模
IMRAM[70]	Faster R-CNN	Bi-GRU	74.1	93.0	96.6	53.9	79.4	87.2	484.2	CVPR 2020	跨模态交互建模
PFAN++[64]	Faster R-CNN	Bi-GRU	70.1	91.8	96.1	52.7	79.9	87.0	477.6	IEEE TMM 2020	跨模态交互建模
ADAPT[8]	Faster R-CNN	Bi-GRU	76.6	**95.4**	97.6	**60.7**	**86.6**	**92.0**	**508.9**	AAAI 2020	跨模态交互建模
GSMN[78]	Faster R-CNN	Bi-GRU	76.4	94.3	97.3	57.4	82.3	89.0	496.7	CVPR 2020	跨模态交互建模
CASC[61]	Faster R-CNN	Bi-GRU	68.5	90.6	95.9	50.2	78.3	86.3	469.8	IEEE TNNLS 2020	跨模态交互建模

续表 6 基于局部特征的图像-文本匹配方法在 **Flickr30K** 数据集上的结果

方法	图像	文本	Image-to-Text(i-t)			Text-to-Image(t-i)			RSUM	发表时间	类别
			R@1	R@5	R@10	R@1	R@5	R@10			
MLSL[69]	Faster R-CNN	Bi-LSTM	72.2	92.4	**98.2**	56.8	83.3	91.3	494.2	IPM 2021	跨模态交互建模
MAG[60]	Faster R-CNN	Bi-GRU	72.1	92.8	96.7	52.8	80.2	87.1	481.8	CAAI-TIS 2021	跨模态交互建模
SHAN[67]	Faster R-CNN	Bi-GRU	74.6	93.5	96.9	55.3	81.3	88.4	490.0	IJCAI 2021	跨模态交互建模
UARDA[62]	Faster R-CNN	Bi-GRU	**77.8**	95.0	97.6	57.8	82.9	89.2	500.3	IEEE TMM 2022	跨模态交互建模
OAN[77]	Faster R-CNN	CNN	53.3	80.1	87.1	**68.6**	**93.0**	**96.0**	478.1	ICMR 2019	模态内&模态间关系建模
SAN[75]	ResNet-152	Bi-GRU	75.5	92.6	96.2	60.1	84.7	90.6	499.7	ICCV 2019	模态内&模态间关系建模
CAAN[9]	Faster R-CNN	Bi-GRU	70.1	91.6	97.2	52.8	79.0	87.9	478.6	CVPR 2020	模态内&模态间关系建模
MMCA[72]	Faster R-CNN	BERT	74.2	92.8	96.4	54.8	81.4	87.8	487.4	CVPR 2020	模态内&模态间关系建模
DP-RNN[73]	Faster R-CNN	Bi-GRU	70.2	91.6	95.8	55.5	81.3	88.2	482.6	AAAI 2020	模态内&模态间关系建模
SGRAF[74]	Faster R-CNN	Bi-GRU	77.8	94.1	97.4	58.5	83.0	88.8	499.6	AAAI 2021	模态内&模态间关系建模
DIME[13]	Faster R-CNN	BERT	**81.0**	**95.9**	**98.4**	63.6	88.1	93.0	**520.0**	ACM SIGIR 2021	模态内&模态间关系建模
SMAN[76]	ResNet-152	Bi-LSTM	57.3	85.3	92.2	43.4	73.7	83.4	435.3	IEEE T-Cybern 2022	模态内&模态间关系建模
GraDual[71]	Faster R-CNN	Bi-GRU	78.3	96.0	98.0	60.4	86.7	92.0	511.4	WACV 2022	模态内&模态间关系建模

表 7 基于局部特征的图像-文本匹配方法在 **MSCOCO** 数据集 **1K** 测试集上的结果

方法	图像	文本	Image-to-Text(i-t)			Text-to-Image(t-i)			RSUM	发表时间	类别
			R@1	R@5	R@10	R@1	R@5	R@10			
DVSA[17]	R-CNN	BRNN	16.5	39.2	52.0	27.4	60.2	74.8	270.1	CVPR 2015	先驱工作
Sm-LSTM[50]	VGG19	Bi-LSTM	53.2	83.1	91.5	40.7	75.8	87.4	431.7	CVPR 2017	双模态建模
SCO[42]	VGG19	LSTM	69.9	92.9	97.5	56.7	87.5	94.8	499.3	CVPR 2018	视觉建模
Unified VSE[48]	ResNet-152	Caption Encoder	64.3	89.2	94.8	48.3	81.7	91.2	469.5	CVPR 2019	文本建模
PIE-Net[51]	ResNet-152	Bi-GRU	69.2	91.6	96.6	55.2	86.5	93.7	492.8	CVPR 2019	双模态建模
VSRN[43]	Faster R-CNN	GRU	76.2	94.8	98.2	62.8	89.7	95.1	516.8	ICCV 2019	视觉建模
LIWE[49]	Faster R-CNN	Bi-GRU	69.6	93.9	98.0	55.5	87.3	94.2	498.6	ICCV 2019	文本建模
SAEM[18]	Faster R-CNN	BERT	71.2	94.1	97.7	57.8	88.6	94.9	504.3	ACM MM 2019	双模态建模
TST[10]	ResNet-152	Bi-GRU	59.4	–	95.1	46.5	–	90.8	–	ACM MM 2019	双模态建模
DSRAN[44]	Faster R-CNN and Res-Net-152	BERT	78.3	95.7	98.4	**64.5**	90.8	**95.8**	523.4	IEEE TCSVT 2020	视觉建模
CRGN[45]	Faster R-CNN and Res-Net-152	GRU	73.8	95.6	98.5	60.1	88.9	94.5	511.4	IEEE TIP 2020	视觉建模
CAMERA[6]	Faster R-CNN	BERT	77.5	96.3	98.8	63.4	**90.9**	**95.8**	522.7	ACM MM 2020	双模态建模
CMRN[12]	Faster R-CNN	BERT-GRU	68.5	92.5	97.0	57.3	88.4	94.8	498.5	MTA 2021	视觉建模
DASPGA[53]	Faster R-CNN	FC	**84.0**	95.8	97.8	63.9	88.9	95.6	**526.0**	CVPR 2021	双模态建模
RRTC[46]	Faster R-CNN	Bi-GRU	76.2	96.3	**98.9**	61.6	89.3	94.6	516.9	IEEE TCSVT 2022	视觉建模
ReSG[52]	Faster R-CNN	Bi-GRU	79.3	**96.7**	98.3	**64.5**	90.0	**95.8**	524.6	IEEE T-Cybern 2022	双模态建模
BSSAN[56]	Faster R-CNN	Bi-LSTM	49.2	76.1	85.1	36.2	69.2	82.5	398.3	IEEE TIP 2018	跨模态交互建模

续表 7　基于局部特征的图像-文本匹配方法在 MSCOCO 数据集 1K 测试集上的结果

方法	图像	文本	Image-to-Text(i-t)			Text-to-Image(t-i)			RSUM	发表时间	类别
			R@1	R@5	R@10	R@1	R@5	R@10			
SCAN[65]	Faster R-CNN	Bi-GRU	72.7	94.8	98.4	58.8	88.4	94.8	507.9	ECCV 2018	跨模态交互建模
RDAN[57]	Faster R-CNN	Bi-GRU	74.6	96.2	98.7	61.6	89.2	94.7	515.0	IJCAI 2019	跨模态交互建模
BFAN[59]	Faster R-CNN	Bi-GRU	74.9	95.2	–	59.4	88.4	–	–	ACM MM 2019	跨模态交互建模
CIRSC[68]	Faster R-CNN	Bi-GRU	73.8	95.3	98.3	59.9	88.9	94.9	511.1	ACM MM 2019	跨模态交互建模
CAMP[58]	Faster R-CNN	Bi-GRU	72.3	94.8	98.3	58.5	87.9	95.0	506.8	ICCV 2019	跨模态交互建模
PFAN[63]	Faster R-CNN	Bi-GRU	76.5	96.3	**99.0**	61.6	89.6	95.2	518.2	IJCAI 2019	跨模态交互建模
IMRAM[70]	Faster R-CNN	Bi-GRU	76.7	95.6	98.5	61.7	89.1	95.0	516.6	CVPR 2020	跨模态交互建模
PFAN++[64]	Faster R-CNN	Bi-GRU	77.1	**96.5**	98.3	62.5	89.9	95.4	513.3	IEEE TMM 2020	跨模态交互建模
ADAPT[8]	Faster R-CNN	Bi-GRU	76.5	95.6	98.9	62.2	90.5	96.0	519.7	AAAI 2020	跨模态交互建模
GSMN[78]	Faster R-CNN	Bi-GRU	78.4	96.4	98.6	63.3	90.1	95.7	522.5	CVPR 2020	跨模态交互建模
CASC[61]	Faster R-CNN	Bi-GRU	72.3	96.0	99.0	58.9	89.8	96.0	512.0	IEEE TNNLS 2020	跨模态交互建模
MLSL[69]	Faster R-CNN	Bi-LSTM	63.8	90.1	95.9	**77.1**	**96.3**	**98.6**	521.8	IPM 2021	跨模态交互建模
MAG[60]	Faster R-CNN	Bi-GRU	75.2	95.4	98.3	59.1	87.9	94.3	510.2	CAAI-TIS 2021	跨模态交互建模
SHAN[67]	Faster R-CNN	Bi-GRU	76.9	96.3	98.7	62.6	89.6	95.8	519.8	IJCAI 2021	跨模态交互建模
UARDA[62]	Faster R-CNN	Bi-GRU	**78.6**	**96.5**	98.9	63.9	90.7	96.2	**524.8**	IEEE TMM 2022	跨模态交互建模
OAN[77]	Faster R-CNN	CNN	60.2	88.6	94.5	**71.7**	**96.4**	**99.3**	510.7	ICMR 2019	模态内&模态间关系建模
SAN[75]	ResNet-152	Bi-GRU	**85.4**	**97.5**	**99.0**	69.1	93.4	97.2	**541.6**	ICCV 2019	模态内&模态间关系建模
CAAN[9]	Faster R-CNN	Bi-GRU	75.5	95.4	98.5	61.3	89.7	95.2	515.6	CVPR 2020	模态内&模态间关系建模
MMCA[72]	Faster R-CNN	BERT	74.8	95.6	97.7	61.6	89.8	95.2	514.7	CVPR 2020	模态内&模态间关系建模
DP-RNN[73]	Faster R-CNN	Bi-GRU	75.3	95.8	98.6	62.5	89.7	95.1	517.0	AAAI 2020	模态内&模态间关系建模
SGRAF[74]	Faster R-CNN	Bi-GRU	79.6	96.2	98.5	63.2	90.7	96.1	524.3	AAAI 2021	模态内&模态间关系建模
DIME[13]	Faster R-CNN	BERT	78.8	96.3	98.7	64.8	91.5	96.5	526.6	ACM SIGIR 2021	模态内&模态间关系建模
SMAN[76]	ResNet-152	Bi-LSTM	68.4	91.3	96.6	58.5	87.4	93.5	495.7	IEEE T-Cybern 2022	模态内&模态间关系建模
GraDual[71]	Faster R-CNN	Bi-GRU	77.0	96.4	98.6	65.3	91.9	96.4	525.6	WACV 2022	模态内&模态间关系建模

表 8　基于局部特征的图像-文本匹配方法在 MSCOCO 数据集 5K 测试集上的结果

方法	图像	文本	Image-to-Text(i-t)			Text-to-Image(t-i)			RSUM	发表时间	类别
			R@1	R@5	R@10	R@1	R@5	R@10			
DVSA[17]	R-CNN	BRNN	38.4	69.9	80.5	10.7	29.6	42.2	271.3	CVPR 2015	先驱工作
HM-LSTM[47]	R-CNN	HM-LSTM	43.9	–	87.8	36.1	–	86.7	–	ICCV 2017	文本建模
SCO[42]	VGG19	LSTM	42.8	72.3	83.0	33.1	62.9	75.5	369.6	CVPR 2018	视觉建模
Unified VSE[48]	ResNet-152	Caption Encoder	36.1	66.4	77.7	25.4	53.0	66.2	324.8	CVPR 2019	文本建模
PIE-Net[51]	ResNet-152	Bi-GRU	45.2	74.3	84.5	32.4	63.0	75.0	374.4	CVPR 2019	双模态建模
VSRN[43]	Faster R-CNN	GRU	53.0	81.1	89.4	40.5	70.6	81.1	415.7	ICCV 2019	视觉建模
TST[10]	ResNet-152	Bi-GRU	40.2	–	80.5	29.9	–	71.9	–	ACM MM 2019	双模态建模
DSRAN[44]	Faster R-CNN and ResNet-152	BERT	55.3	83.5	90.9	41.7	72.7	82.8	426.9	IEEE TCSVT 2020	视觉建模

续表 8 基于局部特征的图像-文本匹配方法在 MSCOCO 数据集 5K 测试集上的结果

方法	图像	文本	Image-to-Text(i-t)			Text-to-Image(t-i)			RSUM	发表时间	类别
			R@1	R@5	R@10	R@1	R@5	R@10			
CRGN[45]	Faster R-CNN and ResNet-152	GRU	51.2	80.6	89.7	37.4	68.0	79.5	406.4	IEEE TIP 2020	视觉建模
CAMERA[6]	Faster R-CNN	BERT	55.1	82.9	91.2	40.5	71.7	82.5	423.9	ACM MM 2020	双模态建模
DASPGA[53]	Faster R-CNN	FC	**68.7**	**88.7**	**93.0**	**46.2**	**77.8**	**85.5**	**459.9**	CVPR 2021	双模态建模
ReSG[52]	Faster R-CNN	Bi-GRU	55.8	83.0	91.0	42.0	72.4	82.1	426.3	IEEE T-Cybern 2022	双模态建模
SCAN[65]	Faster R-CNN	Bi-GRU	50.4	82.2	90.0	38.6	69.3	80.4	410.9	ECCV 2018	跨模态交互建模
CAMP[58]	Faster R-CNN	Bi-GRU	50.1	82.1	89.7	39.0	68.9	80.2	410.0	ICCV 2019	跨模态交互建模
IMRAM[70]	Faster R-CNN	Bi-GRU	53.7	83.2	91.0	39.7	69.1	79.8	416.5	CVPR 2020	跨模态交互建模
PFAN[63]	Faster R-CNN	Bi-GRU	50.8	83.9	89.1	39.5	69.5	80.8	413.6	IJCAI 2019	跨模态交互建模
PFAN++[64]	Faster R-CNN	Bi-GRU	51.2	84.3	89.2	**41.4**	70.9	79.0	416.0	IEEE TMM 2020	跨模态交互建模
CASC[61]	Faster R-CNN	Bi-GRU	47.2	78.3	87.4	34.7	64.8	76.8	389.2	IEEE TNNLS 2020	跨模态交互建模
MAG[60]	Faster R-CNN	Bi-GRU	52.0	81.3	90.0	37.2	65.4	77.9	404.8	CAAI-TIS 2021	跨模态交互建模
SHAN[67]	Faster R-CNN	Bi-GRU	**75.9**	**96.1**	–	**60.7**	**88.2**	–	–	IJCAI 2021	跨模态交互建模
UARDA[62]	Faster R-CNN	Bi-GRU	56.2	83.8	**91.3**	40.6	69.5	**80.9**	**422.3**	IEEE TMM 2022	跨模态交互建模
OAN[77]	Faster R-CNN	CNN	37.0	66.6	78.0	**47.8**	**81.2**	**90.4**	401.0	ICMR 2019	模态内&模态间关系建模
CAAN[9]	Faster R-CNN	Bi-GRU	52.5	83.3	90.9	41.2	70.3	82.9	421.1	CVPR 2020	模态内&模态间关系建模
SGRAF[74]	Faster R-CNN	Bi-GRU	57.8	–	91.6	41.9	–	81.3	–	AAAI 2021	模态内&模态间关系建模
DIME[13]	Faster R-CNN	BERT	**59.3**	**85.4**	**91.9**	43.1	73.0	83.1	**435.8**	ACM SIGIR 2021	模态内&模态间关系建模

表 9 多模态预训练模型在 MSCOCO 数据集和 Flickr30K 数据集上的结果

方法	MSCOCO							Flickr30K							发表时间	类别
	Image-to-Text(i-t)			Text-to-Image(t-i)			RSUM	Image-to-Text(i-t)			Text-to-Image(t-i)			RSUM		
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10			
1K Test																
ViLBERT[92]	–	–	–	–	–	–	–	–	–	–	58.2	84.9	91.5	–	NIPS 2019	双流框架
UNITER[110]	–	–	–	–	–	–	–	85.9	97.1	98.8	72.5	92.4	96.1	542.7	ECCV 2020	单流框架
Unicoder-VL[103]	84.3	97.3	99.3	69.7	93.5	97.2	–	86.2	96.3	99.0	71.5	90.9	94.9	538.3	AAAI 2020	单流框架
Image-BERT[113]	87.0	97.6	99.2	73.1	92.6	96.0	–	85.4	98.7	99.8	73.6	94.3	97.2	549.0	Arxiv 2020	单流框架
Pixel-BERT[107]	–	–	–	–	–	–	–	87.0	98.9	99.5	71.5	92.1	95.8	544.8	Arxiv 2020	单流框架
VILLA[115]	–	–	–	–	–	–	–	86.6	97.9	99.2	74.7	92.9	95.8	547.1	NIPS 2020	单流框架
OSCAR[105]	**88.4**	**99.1**	**99.8**	**75.7**	**95.2**	**98.3**	**565.5**	–	–	–	–	–	–	–	ECCV 2020	单流框架
12-in-1[95]	–	–	–	65.2	91.0	96.2	–	–	–	–	65.1	88.7	93.5	–	CVPR 2020	双流框架
TDEN[93]	–	–	–	–	–	–	–	–	–	–	63.6	88.2	92.9	–	AAAI 2020	双流框架

续表 9　多模态预训练模型在 MSCOCO 数据集和 Flickr30K 数据集上的结果

方法	MSCOCO							Flickr30K							发表时间	类别
	Image-to-Text(i-t)			Text-to-Image(t-i)			RSUM	Image-to-Text(i-t)			Text-to-Image(t-i)			RSUM		
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10			
ERNIE-ViL[96]	–	–	–	–	–	–	–	86.7	97.8	99.0	74.4	92.7	95.9	546.5	AAAI 2020	双流框架
ViLT[108]	–	–	–	–	–	–	–	83.5	96.7	98.6	64.4	88.7	93.8	525.7	ICML 2020	单流框架
ALIGN[98]	–	–	–	–	–	–	–	**95.3**	**99.8**	**100.0**	**84.9**	**97.4**	**98.6**	**576.0**	ICML 2020	双流框架
METER[97]	–	–	–	–	–	–	–	94.3	99.6	99.9	82.2	96.3	98.4	570.7	CVPR 2020	双流框架
5K Test																
UNITER[110]	64.4	87.4	93.1	50.3	78.5	87.2	460.9	–	–	–	–	–	–	–	ECCV 2020	单流框架
Unicoder-VL[103]	62.3	87.1	92.8	46.7	76.0	85.3	450.2	–	–	–	–	–	–	–	AAAI 2020	单流框架
Image-BERT[113]	66.4	89.8	94.4	50.5	78.7	87.1	466.9	–	–	–	–	–	–	–	Arxiv 2020	单流框架
Pixel-BERT[107]	63.6	87.5	93.6	50.1	77.6	86.2	458.6	–	–	–	–	–	–	–	Arxiv 2020	单流框架
OSCAR[105]	**70.0**	**91.1**	**95.5**	**84.0**	**80.8**	**88.5**	**509.9**	–	–	–	–	–	–	–	ECCV 2020	单流框架
VinVL[106]	74.6	92.6	96.3	58.1	83.2	90.1	494.9	–	–	–	–	–	–	–	CVPR 2021	单流框架
ViLT[108]	61.5	86.3	92.7	42.7	72.9	83.1	439.2	–	–	–	–	–	–	–	ICML 2021	单流框架
ALIGN[98]	77.0	93.5	96.9	59.9	83.3	89.8	500.4	–	–	–	–	–	–	–	ICML 2021	双流框架
METER[97]	76.2	93.2	96.8	57.1	82.7	90.1	496.1	–	–	–	–	–	–	–	CVPR 2022	双流框架

表 10　其他三类图像-文本匹配方法在 Flickr30K 数据集上的结果

方法	图像	文本	Image-to-Text(i-t)			Text-to-Image(t-i)			RSUM	发表时间	类别
			R@1	R@5	R@10	R@1	R@5	R@10			
SCG[80]	VGG19	LSTM	71.8	90.8	94.8	49.3	76.4	85.6	468.7	IJCAI 2019	外部知识
SGM[15]	Faster R-CNN	Bi-GRU	71.8	91.7	95.5	53.5	79.6	86.5	478.6	WACV 2020	外部知识
VRACR[81]	ResNet-152	Bi-GRU	41.7	71.5	80.4	55.1	81.8	88.5	419.0	ICMR 2020	外部知识
HOAD[14]	Faster R-CNN	Bi-LSTM	70.8	92.7	96.0	60.9	86.1	91.0	497.5	CVPR 2020	外部知识
CVSE[7]	Faster R-CNN	Bi-GRU	73.6	90.4	94.4	56.1	83.2	90.0	487.7	ECCV 2020	外部知识
HAT[82]	Faster R-CNN	GRU	**77.0**	**95.1**	**97.7**	**65.4**	**90.1**	**92.7**	**518.0**	IEEE TCSVT 2022	外部知识
CMPM[25]	MobileNet	Bi-LSTM	49.6	76.8	86.1	37.3	65.7	75.5	391.0	ECCV 2018	度量学习
CVSE++[91]	ResNet-152	GRU	50.2	78.8	87.3	37.1	66.9	76.4	396.7	AAAI 2020	度量学习
NAAF[19]	Faster R-CNN	Bi-GRU	81.9	96.1	98.3	61.0	85.3	90.6	513.2	CVPR 2022	度量学习
CMCAN[20]	Faster R-CNN	Bi-GRU	**79.5**	**95.6**	**97.6**	**60.9**	**84.3**	**89.9**	**507.8**	AAAI 2022	度量学习
A3VSE[120]	Faster R-CNN	LSTM	49.5	79.5	86.6	**65.0**	**89.2**	**94.5**	464.3	ACM MM 2019	弱监督跨模态匹配
ACMM[121]	Faster R-CNN	Skip-Gram	**85.4**	**96.9**	**98.4**	55.1	81.7	88.4	**505.9**	IEEE TPAMI 2021	弱监督跨模态匹配

表 11　其他三类图像-文本匹配方法在 MSCOCO 数据集 1K 测试集上的结果

方法	图像	文本	Image-to-Text(i-t)			Text-to-Image(t-i)			RSUM	发表时间	类别
			R@1	R@5	R@10	R@1	R@5	R@10			
VQA[79]	VGG19	LSTM	50.5	80.1	89.7	37.0	70.9	82.9	411.1	ECCV 2016	外部知识
SCG[80]	VGG19	LSTM	76.6	96.3	99.2	61.4	88.9	95.1	517.5	IJCAI 2019	外部知识
SGM[15]	Faster R-CNN	Bi-GRU	73.4	93.8	97.8	57.5	87.3	94.3	504.1	WACV 2020	外部知识
VRACR[81]	ResNet-152	Bi-GRU	56.6	87.9	94.7	67.7	91.7	96.7	495.3	ICMR 2020	外部知识
HOAD[14]	Faster R-CNN	Bi-LSTM	77.8	96.1	98.7	66.2	93.0	97.9	529.7	CVPR 2020	外部知识
CVSE[7]	Faster R-CNN	Bi-GRU	78.6	95.0	97.5	66.3	93.0	97.9	529.7	ECCV 2020	外部知识
HAT[82]	Faster R-CNN	GRU	**82.5**	**97.2**	**99.2**	**84.5**	**97.4**	**98.9**	**559.7**	IEEE TCSVT 2022	外部知识
CMPM[25]	MobileNet	Bi-LSTM	56.1	86.3	92.9	44.6	78.8	89.0	447.7	ECCV 2018	度量学习

续表 11 其他三类图像-文本匹配方法在 MSCOCO 数据集 1K 测试集上的结果

方法	图像	文本	Image-to-Text(i-t)			Text-to-Image(t-i)			RSUM	发表时间	类别
			R@1	R@5	R@10	R@1	R@5	R@10			
CVSE++[91]	ResNet-152	GRU	66.7	90.2	94.0	48.4	81.0	90.0	470.3	AAAI 2020	度量学习
NAAF[19]	Faster R-CNN	Bi-GRU	80.5	96.5	98.8	64.1	90.7	96.5	527.2	CVPR 2022	度量学习
CMCAN[20]	Faster R-CNN	Bi-GRU	**81.2**	**96.8**	**98.7**	**65.4**	**91.0**	**96.2**	**529.3**	AAAI 2022	度量学习
ACMM[121]	Faster R-CNN	Skip-Gram	84.4	97.9	99.4	63.4	90.3	95.7	531.1	IEEE TPAMI 2021	弱监督跨模态匹配

表 12 其他三类图像-文本匹配方法在 MSCOCO 数据集 5K 测试集上的结果

方法	图像	文本	Image-to-Text(i-t)			Text-to-Image(t-i)			RSUM	发表时间	类别
			R@1	R@5	R@10	R@1	R@5	R@10			
SCG[80]	VGG19	LSTM	56.6	84.5	92.0	39.2	68.0	81.3	421.6	IJCAI 2019	外部知识
SGM[15]	Faster R-CNN	Bi-GRU	50.0	79.3	87.9	35.3	64.9	76.5	393.9	WACV 2020	外部知识
HOAD[14]	Faster R-CNN	Bi-LSTM	51.4	81.8	89.1	40.5	73.5	84.1	420.4	CVPR 2020	外部知识
HAT[82]	Faster R-CNN	GRU	**58.5**	**86.1**	**92.9**	**58.6**	**86.6**	**92.9**	**475.6**	IEEE TCSVT 2022	外部知识
CMPM[25]	MobileNet	Bi-LSTM	31.1	60.7	73.9	22.9	50.2	63.8	302.6	ECCV 2018	度量学习
CVSE++[91]	ResNet-152	GRU	39.3	69.1	80.3	25.2	25.8	54.0	293.7	AAAI 2020	度量学习
NAAF[19]	Faster R-CNN	Bi-GRU	58.9	**85.2**	92.0	42.5	70.9	81.4	**430.9**	CVPR 2022	度量学习
CMCAN[20]	Faster R-CNN	Bi-GRU	**61.5**	–	92.9	44.0	–	82.6	–	AAAI 2022	度量学习
A3VSE[120]	Faster R-CNN	LSTM	39.0	68.0	80.1	**49.3**	**81.1**	**90.2**	407.7	ACM MM 2019	弱监督跨模态匹配
ACMM[121]	Faster R-CNN	Skip-Gram	**67.5**	**89.6**	**95.0**	42.1	70.9	81.3	**446.4**	IEEE TPAMI 2021	弱监督跨模态匹配

果;表 10~表 12 总结了其他类别图像-文本匹配方法在 Flickr30K 和 MSCOCO 数据集上实验结果. 注意,表格中实验结果均来自于相应工作发表论文中,符号-表示无对应的实验结果.

5.1 基于全局特征图像-文本匹配方法分析

表 3、表 4 和表 5 中总结了基于全局特征的图像-文本匹配方法在 Flickr30K 以及 MSCOCO 数据集上面的实验结果. 其中,表 4 中汇总了这些方法在 MSCOCO1K 测试集上的结果,而表 5 汇总的则是不同方法在 MSCOCO5K 测试集上的实验结果.

通过对表 3 进行分析,我们可以发现:(1)基于嵌入架构的匹配方法中,UVS 取得优于 DeViSE 的性能,这表明采用双向铰链损失函数模型的性能一定程度上要优于基于单向铰链损失函数的模型. 与此同时,这也表明 LSTM 提取的全局文本特征比 Skip-Gram 提取得到的文本特征中蕴含更加丰富的语义信息. (2)虽然 UVS 和 DSPE 均采用双向铰链损失函数,但后者性能显著优于前者,其在 R@1 指标上分别提升 17%和 13%. 这表明保持各个模态内邻域嵌入的相似性十分重要. (3)VSE++取得了优于上述方法的结果,这反映出难分样本挖掘对于提升匹配性能而言,是十分必要的. (4)相较于 VSE++而言,TIMAM 在 R@1 上取得了进一步提升,这表明引入对抗损失对于促进跨模态匹配是有益的. (5)在基于交互架构的方法中,MTFN 取得最优结果,这表明精细化设计的融合模块比直接的非线性拼接,更加能够捕获两个模态间交互信息,促进跨模态语义对齐.

对于 MSCOCO 数据集而言,无论 1K 还是 5K 测试集上面,各个方法呈现的结果趋势与其在 Flickr30K 上类似. 具体地,VSE++取得了优于 DSPE、2WayNet、RRF-Net 的实验结果,这再一次验证损失函数中集成难分样本挖掘的有效性. 此外,基于嵌入的匹配方法中,GXN 性能要优于 VSE++,这同样反映出将生成过程融入到嵌入学习中,可以增进视觉与文本的语义对齐.

总体来说,在两个数据集上,基于交互框架的匹配方法 MTFN 取得了最好的实验结果. 这表明相较于直接进行嵌入表示学习,建模模态间的交互信息可能会对跨模态语义匹配起到一定的促进作用.

5.2 基于局部特征图像-文本匹配方法分析

表 6、表 7 以及表 8 中总结了基于局部特征的图像-文本匹配方法在两个数据集上的实验结果,其中表 7 中汇总了这些方法在 MSCOCO1K 测试集上

的结果,而表 8 汇总的则是不同方法在 MSCOCO5K 测试集上的实验结果. 通过对三个表格中数据进行分析，我们发现:（1）基于模态内交互的匹配方法中，GARN 和 Unified VSE 性能要优于 HM-LSTM. 这表明无论显式（关系依赖）还是隐式（图网络）地对实体间关系进行充分建模，均有利于文本语义的理解. 对于关注视觉内实体关系建模的方法中，DSRAN 取得最佳的实验结果,这进一步表明图网络在建模实体间关系方面的强大能力. 而兼顾两种模态内实体关系建模方法中，DASPGA 性能表现最理想，且整体结果也要优于 DSRAN 和 LIWE. 这与我们的认知是一致的，即同时提升视觉和文本表示比仅提升其中之一性能要好. 与此同时，这也再一次反映出图注意力网络在实体关系建模方面的有效性.（2）在基于模态间交互的局部匹配方法中，UARAD 在 MSCOCO1K 和 MSCOCO5K 上取得最优的实验结果，而 ADAPT 在 Flickr30K 上取得最好的结果. 这两个方法的共同之处在于均采取策略抑制不相关信息对于匹配的作用，其中 UARAD 利用自适应阈值的注意力机制，而 ADAPT 采用的是类注意力机制的调节机制. 这表明有效抑制不相关图像区域-词语对信息，对于提升跨模态语义匹配至关重要.（3）基于混合交互的局部特征匹配方法中，DIME 在 Flickr30K 和 MSCOCO5K 上面取得最优结果，这表明固定的模态交互建模策略并不适用于全部数据，而自适应的交互策略选择对于跨模态语义对齐更加有效. SAN 在 MSCOCO1K 上取得最优性能，这可能要归功于其采用显著性检测算法得到了更加精准的关注局部视觉区域.

总体来说，基于局部特征的图像-文本匹配方法性能要优于基于全局特征的图像-文本匹配方法，这是很容易理解的. 因为细粒度的跨模态语义对齐精准度要高于粗粒度的语义对齐. 此外，并不存在一种现有方法在三个数据集上一致取得最优结果. 这表明现有方法泛化性能较弱，亟待探究普适性更广的新交互模式.

5.3　图像-文本多模态预训练模型分析

预训练模型在所有数据集上性能均优于基于全局或者局部特征的图像-文本匹配方法. 这主要归因于这些方法在大规模数据上被训练，故其可以更好地表征图像以及文本内容，有效地削减跨模态语义鸿沟. 此外，在这些预训练方法中，ALIGN 在 Flickr30K 数据集上表现最优，而 OSCAR 在 MSCOCO 数据集上效果最好. 主要原因可能是（1）ALIGN 本身的预训练任务就是图像文本匹配任务，故其对于同任务的下游任务具有更好的泛化性能；（2）OSCAR 的预训练数据集中涵盖了 Flickr30K 和 COCO，故其与图像文本匹配任务之间不存在数据域差异. 在 MSCOCO5K 上，ALIGN 的性能略低于 OSCAR，这可能是由于 ALIGN 的预训练数据集与 MSCOCO 间存在数据域差异.

5.4　其他图像-文本匹配方法分析

表 10、表 11、表 12 中分别总结了其他三类方法在两个数据集上的实验结果. 通过对实验结果进行分析，我们可以发现:（1）在基于外部知识的图像-文本匹配方法中，HAT 在两个数据集上取得一致最优的结果. 这表明引入场景图知识有助于提升视觉与文本的表示，同时 Transformer 可以促进跨模态语义对齐.（2）在基于度量学习的图像-文本匹配方法中，CMCAN 整体表现更好. 这表明在评估相似性指标时，考虑不同区域-词语对的置信度十分必要.（3）对于弱监督信息下的图像-文本匹配任务，即小部分图像标注文本信息或者图像/文本包含少样本语义，A3VSE 和 ACMM 均取得不错效果，甚至优于部分基于外部知识和度量学习方法. 由此可见，弱监督的图像-文本匹配是一个值得探究的任务，其对于实际应用更有意义.

6　未来研究方向

在过去几年里，图像-文本匹配任务无论从方法设计还是数据集构建等方面都取得了一系列进展. 具体地，图像-文本匹配方法建立了以深度学习模型为基础的一系列架构模型，数据集方面则出现了以行人、动物等组建的不同规模大小的评测数据集. 基于现有图像-文本匹配方法和数据集的基础上，本文将对图像-文本匹配存在的问题以及未来发展方向进行探讨.

6.1　公开数据集的划分

对于数据集 Flickr30K 以及 MSCOCO 而言，现有方法均采用不同的数据集划分方式进行实验. 如表 2 所示，Flickr30K 数据集共存在 5 种训练集、验证集和测试集划分方式，而对于 MSCOCO 而言存在 11 种划分格式. 总的来讲，现有方法的实验设置不一致，这导致实验结果的可对比性减弱. 换而言之，我们无法确定当下方法的实验效果增益，来自于模型自身，还是数据集划分方式不一致所带来的. 由此可见，对于图像-文本匹配任务而言，为现有公开数据集提供一种统一的切分方式，以更加公平地

进行实验比较，是一个未来亟待解决的问题.

6.2 模态内语义信息建模

现有工作多利用注意力机制和图网络等技术，来挖掘各个模态内的语义信息. 从上面实验表格中可以看到，它们在文本检索图像部分的性能远低于其在图像检索文本部分的结果. 这表明现有方法在模态内语义信息理解方面仍存在缺陷. 虽然基于外部知识的图像-文本匹配方法在文本检索图像任务上性能有所提升，但它们较为依赖场景图提取技术. 所以，构建更为灵活的知识驱动的图像-文本匹配方法，如将知识图谱等结构化知识注入模型中，是一个未来值得探索的方向.

此外，预训练模型使用数据集与图像文本匹配任务数据集之间存在数据域差异，这可能会导致预训练模型迁移时性能下降. 尽管增大预训练模型的数据域分布，在一定程度上可缓解该问题，但会增加训练的成本. 因而，如何提升多模态预训练数据集多样性，以更好地适应图像文本匹配任务，也是一个可探究的方向.

6.3 跨模态关联语义建模

现有图像-文本匹配方法多采用静态模型架构进行跨模态语义对齐，使得它们不能最优地处理所有类型的数据，导致它们在不同数据集上面呈现性能不一致. 鉴于目前动态神经网络[132]在计算机视觉领域的成功应用，如何依据输入图像/文本语义信息的复杂程度，让模型自适应地选择合适的语义对齐方式，以进一步提升跨模态匹配性能，也是一个未来值得探究的问题.

现有多模态预训练模型虽通过设计不同的预训练任务，如文本视觉对齐，可捕捉不同模态间的关联信息. 但它们难以挖掘精细的模态间相关信息，如名词与物体对象之间的相关性，因此如何更巧妙的设计预训练任务，使得模型建立词语与物体对象之间的相关性，也很重要.

6.4 可解释性以及效率

对于图像-文本匹配模型而言，尤其是多模态预训练模型，提升其效率，以满足各种现实场景的应用需求（如，部署于资源受限的设备），是一件十分必要且重要的事情. 目前，只有少数方法尝试于提升图像-文本匹配的效率，如集成跨模态哈希表示学习. 据我们所知，模型压缩与加速是一个热点研究问题，具体方法包括参数共享、模型修剪和知识蒸馏等. 故基于当下模型压缩与加速技术，如何搭建高效的图像-文本匹配方法，仍是一个亟待探究的问题. 另外，现有图像-文本匹配方法大多进行隐式对齐，缺乏一定的可解释性. 所以，构建可解释的图像-文本匹配方法也值得关注.

7 结 论

本文主要总结了基于深度学习的图像-文本匹配算法，并且对这些方法进行了分类和描述，包括基于全局特征的匹配方法、基于局部特征的匹配方法、基于外部知识的匹配方法、以及基于度量学习的匹配方法. 对于基于全局特征的图像-文本匹配方法，本文根据其技术类型划分为两类：基于交互的方法和基于嵌入的方法. 而对于基于局部特征的图像-文本匹配方法，根据其交互类型划分为三类：基于模态内关系建模方法、基于模态间关系建模方法、基于混合交互建模的方法. 随后，本文对当前图像-文本匹配任务相关数据集进行介绍，并对现有方法的实验结果进行总结与分析. 最后，本文指出该任务未来可能的研究方向，供研究者参考.

参 考 文 献

[1] Frome A, Corrado G S, Shlens J, et al. Devise: A deep visual-semantic embedding model//Proceedings of the International Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2013: 2121-2129

[2] Wang L, Li Y, Huang J, et al. Learning two-branch neural networks for image-text matching tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(2): 394-407

[3] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space//Proceedings of the International Conference on Learning Representations. Scottsdale, USA, 2013: 1-12

[4] Klein B, Lev G, Sadeh G, et al. Associating neural word embeddings with deep image representations using fisher vectors//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 4437-4446

[5] Nam H, Ha J W, Kim J. Dual attention networks for multimodal reasoning and matching//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 299-307

[6] Qu L, Liu M, Cao D, et al. Context-aware multi-view summarization network for image-text matching//Proceedings of the ACM International Conference on Multimedia. Seattle, USA, 2020: 1047-1055

[7] Wang H, Zhang Y, Ji Z, et al. Consensus-aware visual-semantic embedding for image-text matching//Proceedings of the European Conference on Computer Vision. 2020: 18-34

[8] Wehrmann J, Kolling C, Barros R C. Adaptive cross-modal embeddings for image-text alignment//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 12313-12320

[9] Zhang Q, Lei Z, Zhang Z, et al. Context-aware attention network for image-text retrieval//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Online, 2020: 3536-3545

[10] Guo W, Huang H, Kong X, et al. Learning disentangled representation for cross-modal retrieval with deep mutual information estimation//Proceedings of the ACM International Conference on Multimedia. Nice, France, 2019: 1712-1720

[11] Jing Y, Wang W, Wang L, et al. Learning aligned image-text representations using graph attentive relational network. IEEE Transactions on Image Processing, 2021, 30: 1840-1852

[12] Zhang J, He X, Qing L, et al. Cross-modal multi-relationship aware reasoning for image-text matching. Multimedia Tools and Applications, 2021, (1): 1-23

[13] Qu L, Liu M, Wu J, et al. Dynamic modality interaction modeling for image-text retrieval//Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021: 1104-1113

[14] Li Y, Zhang D, Mu Y. Visual-semantic matching by exploring high-order attention and distraction//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2020: 12786-12795

[15] Wang S, Wang R, Yao Z, et al. Cross-modal scene graph matching for relationship-aware image-text retrieval//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Snowmass Village, USA, 2020: 1508-1517

[16] Kiros R, Salakhutdinov R, Zemel R S. Unifying visual-semantic embeddings with multimodal neural language models//Proceedings of the International Conference on Neural Information Processing Systems (Workshop). Montreal, Canada, 2014: 1-13

[17] Karpathy A, Fei-fei L. Deep visual-semantic alignments for generating image descriptions//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3128-3137

[18] Wu Y, Wang S, Song G, et al. Learning fragment self-attention embeddings for image-text matching//Proceedings of the ACM International Conference on Multimedia. Nice, France, 2019: 2088-2096

[19] Zhang K, Mao Z, Wang Q, et al. Negative-aware attention framework for image-text matching//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 15661-15670

[20] Zhang H, Mao Z, Zhang K, et al. Show your faith: Cross-modal confidence-aware network for image-text matching//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2022: 3262-3270

[21] Liu F, Ye R, Wang X, et al. Hal: Improved text-image matching by mitigating visual semantic hubs//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 11563-11571

[22] Wei J, Xu X, Yang Y, et al. Universal weighting metric learning for cross-modal matching//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2020: 13005-13014

[23] Chen T, Deng J, Luo J. Adaptive offline quintuplet loss for image-text matching//Proceedings of the European Conference on Computer Vision. 2020: 549-565

[24] Biten A, Mafla A, Gomex L, et al. Is an image worth five sentences? a new look into semantics for image-text matching// Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Hawaii, USA, 2022: 1391-1400

[25] Zhang Y, Lu H. Deep cross-modal projection learning for image-text matching//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 686-701

[26] Kaur P, Pannu H, Malhi A. Comparative analysis on cross-modal information retrieval: A review. Computer Science Review. 2021, 39: 100336

[27] Liu Y, Guo Y, Fang J, Fan J, Hao Y, Liu J. Survey of Research on deep learning image-text cross-modal retrieval. Journal of Frontiers of Computer Science and Technology, 2022, 16(3): 489-511(in Chinese)
(刘颖, 郭莹莹, 房杰, 范九伦, 郝羽, 刘继明. 深度学习跨模态图文检索研究综述. 计算机科学与探索, 2022, 16(3): 489-511)

[28] Yin Q, Huang Y, Zhang J, Wu S, Wang L. Survey on deep learning based cross-modal retrieval. Journal of Image and Graphics. 2021, 26(6): 1368-1388 (in Chinese)
(尹奇跃, 黄岩, 张俊格, 吴书, 王亮. 基于深度学习的跨模态检索综述. 中国图象图形学报, 2021, 26(6): 1368-1388)

[29] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. Proceedings of the Advances in Neural Information Processing Systems. Lake Tahoe, USA, 2012, 25: 1097-1105

[30] Luo Y, Zhu H, Liang S, Zhang T. Research on image and text retrieval method based on feature vector of heterogeneous data. Informatology, 2021, (4): 027-039 (in Chinese)
(骆有隆, 朱卉钰, 梁松宇, 张腾. 基于异构数据特征向量的图文检索方法研究.情报工程, 2021(4): 027-039)

[31] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556, 2014: 1-14

[32] Liu Y, Guo Y, Bakker E M, et al. Learning a recurrent residual fusion network for multimodal matching//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 4107-4116

[33] Wang L, Li Y, Lazebnik S. Learning deep structure preserving image-text embeddings//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 5005-5013

[34] Faghri F, Fleet D J, Kiros J R, et al. Vse++: Improving visual-semantic embeddings with hard negatives//Proceedings of the British Machine Vision Conference. Northumbria University, UK, 2018: 1-13

[35] Eisenschtat A, Wolf L. Linking image and text with 2-way nets//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 4601-4611

[36] Gu J, Cai J, Joty S R, et al. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7181-7189

[37] Sarafianos N, Xu X, Kakadiaris I A. Adversarial representation

learning for text-to-image matching//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 5814-5824

[38] Liu L, Gou T. Cross-modal retrieval combining deep canonical correlation analysis and adversarial learning. Computer Science. 2021, 48(9): 200-207 (in Chinese)
(刘立波, 苟婷婷. 融合深度典型相关分析和对抗学习的跨模态检索. 计算机科学, 2021, 48(9): 200-207)

[39] Matsubara T. Target-oriented deformation of visual semantic embedding space. IEICE Transactions on Information and Systems. 2021, 104(1): 24-33

[40] Ma L, Lu Z, Shang L, et al. Multimodal convolutional neural networks for matching image and sentence//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 2623-2631

[41] Wang T, Xu X, Yang Y, et al. Matching images and text with multi-modal tensor fusion and re-ranking//Proceedings of the ACM International Conference on Multimedia. Nice, France, 2019: 12-20

[42] Huang Y, Wu Q, Song C, et al. Learning semantic concepts and order for image and sentence matching//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 6163-6171

[43] Li K, Zhang Y, Li K, et al. Visual semantic reasoning for image-text matching//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 4654-4662

[44] Wen K, Gu X, Cheng Q. Learning dual semantic relations with graph attention for image-text matching. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(7): 2866-2879

[45] Zhang Y, Zhou W, Wang M, et al. Deep relation embedding for cross-modal retrieval. IEEE Transactions on Image Processing. 2020, 30: 617-627

[46] Wu J, Wu C, Lu J, et al. Region reinforcement network with topic constraint for image-text matching. IEEE Transactions on Circuits and Systems for Video Technology. 2021, 32(1): 388-397

[47] Niu Z, Zhou M, Wang L, et al. Hierarchical multimodal lstm for dense visual-semantic embedding//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 1881-1889

[48] Wu H, Mao J, Zhang Y, et al. Unified visual semantic embeddings: Bridging vision and language with structured meaning representations. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 6609-6618

[49] Wehrmann J, Souza D M, Lopes M A, et al. Language-agnostic visual-semantic embeddings//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 5804-5813

[50] Huang Y, Wang W, Wang L. Instance-aware image and sentence matching with selective multimodal lstm//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2310-2318

[51] Song Y, Soleymani M. Polysemous visual-semantic embedding for cross-modal retrieval//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 1979-1988

[52] Liu X, He Y, Cheung Y, et al. Learning relationship- enhanced semantic graph for fine-grained image-text matching. IEEE Transactions on Cybernetics. 2022 (Ealy Access)

[53] Yan S, Yu L, Xie Y. Discrete-continuous action space policy gradient-based attention for image-text matching. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Online, 2021: 8096-8105

[54] Chen J, Hu H, Wu H, et al. Learning the best pooling strategy for visual semantic embedding//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Online, 2021: 15789-15798

[55] Tu R, Ji L, Luo H, et al. Hashing based Efficient Inference for Image-Text Matching. Findings of the Association for Computational Linguistics: ACL-IJCNLP. Online, 2021: 743-752

[56] Huang F, Zhang X, Zhao Z, et al. Bidirectional spatial-semantic attention networks for image text matching. IEEE Transactions on Image Processing. 2018, 28(4): 2008-2020

[57] Hu Z, Luo Y, Lin J, et al. Multi-level visual-semantic alignments with relation-wise dual attention network for image and text matching//Proceedings of the International Joint Conference on Artificial Intelligence. Macao, China, 2019: 789-795

[58] Wang Z, Liu X, Li H, et al. Camp: Cross-modal adaptive message passing for text-image retrieval//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 5764-5773

[59] Liu C, Mao Z, Liu A A, et al. Focus your attention: A bidirectional focal attention network for image text matching//Proceedings of the ACM International Conference on Multimedia. Nice, France, 2019: 3-11

[60] Gong D, Chen H, Chen S, Bao Y, Ding G. Matching with agreement for cross-modal image-text retrieval. CAAI Transactions on Intelligent Systems. 2021, 16(6): 1143-1150 (in Chinese)
(宫大汉, 陈辉, 陈仕江, 包勇军, 丁贵广. 一致性协议匹配的跨模态图像文本检索方法. 智能系统学报, 2021, 16(6), 1143-1150)

[61] Xu X, Wang T, Yang Y, et al. Cross-modal attention with semantic consistence for image-text matching. IEEE Transactions on Neural Networks and Learning Systems. 2020, 31(12): 5412-5425

[62] Zhang K, Mao Z, Liu A, et al. Unified adaptive relevance distinguishable attention network for image- text matching. IEEE Transactions on Multimedia. 2022 (Early Access)

[63] Wang Y, Yang H, Qian X, et al. Position focused attention network for image-text matching//Proceedings of the International Joint Conference on Artificial Intelligence. Macao, China, 2019: 3792-3798

[64] Wang Y, Yang H, Bai X, et al. Pfan++: Bi-directional image-text retrieval with position focused attention network. IEEE Transactions on Multimedia. 2020, (1): 1-14

[65] Lee K H, Chen X, Hua G, et al. Stacked cross attention for image-text matching//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 201-216

[66] Deng Y, Zhang F, Chen X, Ai Q, Yu S. Collaborative attention network model for cross-modal retrieval. Computer Science.

2020, 47(4): 54-59 (in Chinese)
(邓一姣, 张凤荔, 陈学勤, 艾擎, 佘苏喆. 面向跨模态检索的协同注意力网络模型. 计算机科学, 2020, 47(4): 54-59)

[67] Ji Z, Chen K, Wang H. Step-wise hierarchical alignment network for image-text matching//Proceedings of the 30th International Joint Conference on Artificial Intelligence. Online, 2021: 765-771

[68] Chen H, Ding G, Lin Z, et al. Cross-modal image text retrieval with semantic consistency//Proceedings of the ACM International Conference on Multimedia. Nice, France, 2019: 1749-1757

[69] Li W H, Yang S, Wang Y, et al. Multi-level similarity learning for image-text retrieval. Information Processing & Management, 2021, 58(1): 102432

[70] Chen H, Ding G, Liu X, et al. Imram: Iterative matching with recurrent attention memory for cross-modal image text retrieval// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Online, 2020: 12655-12663

[71] Long S, Han S C, Wan X, et al. Gradual: Graph-based dual-modal representation for image-text matching//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Hawaii, USA, 2022: 3459-3468

[72] Wei X, Zhang T, Li Y, et al. Multi-modality cross attention network for image and sentence matching//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Online, 2020: 10941-10950

[73] Chen T, Luo J. Expressing objects just like words: Recurrent visual embedding for image-text matching//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 10583-10590

[74] Diao H, Zhang Y, Ma L, et al. Similarity reasoning and filtration for image-text matching//Proceedings of the AAAI Conference on Artificial Intelligence. Online, 2021: 1218-1226

[75] Ji Z, Wang H, Han J, et al. Saliency-guided attention network for image-sentence matching//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 5753-5762

[76] Ji Z, Wang H, Han J, et al. SMAN: Stacked multimodal attention network for cross-modal image-text retrieval. IEEE Transactions on Cybernetics. 2022, 52(2): 1086-1097

[77] Huang P Y, Chang X, Hauptmann A G. Improving what cross-modal retrieval models learn through object-oriented inter-and intra-modal attention networks//Proceedings of the International Conference on Multimedia Retrieval. Ottawa, Canada, 2019: 244-252

[78] Liu C, Mao Z, Zhang T, et al. Graph structured network for image-text matching//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Online, 2020: 10921-10930

[79] Lin X, Parikh D. Leveraging visual question answering for image-caption ranking//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 261-277

[80] Shi B, Ji L, Lu P, et al. Knowledge aware semantic concept expansion for image-text matching//Proceedings of the International Joint Conference on Artificial Intelligence. Macao, China, 2019: 5182-5189

[81] Guo Y, Chen J, Zhang H, et al. Visual relations augmented cross-modal retrieval//Proceedings of the International Conference on Multimedia Retrieval. Dublin Ireland, 2020: 9-15

[82] Dong X, Zhang H, Zhu L, et al. Hierarchical Feature Aggregation based on Transformer for Image-text Matching. IEEE Transactions on Circuits and Systems for Video Technology. 2022 (Early Access)

[83] Zhang B, Hu H, Jain V, et al. Learning to represent image and text with denotation graphs//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Online, 2020: 823-839

[84] Huang P Y, Chang X, Hauptmann A, et al. Forward and backward multimodal nmt for improved monolingual and multilingual cross-modal retrieval//Proceedings of the International Conference on Multimedia Retrieval. Dublin, Ireland, 2020: 53-62

[85] Sun S, Chen Y C, Li L, et al. Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online, 2021: 982-997

[86] Zheng Z, Zheng L, Garrett M, et al. Dualpath convolutional image-text embeddings with instance loss. ACM Transactions on Multimedia Computing, Communications, and Applications. 2020, 16(2): 1-23

[87] Chen L, Gan Z, Cheng Y, et al. Graph optimal transport for cross-domain alignment. International Conference on Machine Learning. Online, 2020: 1542-1553

[88] Wei J, Yang Y, Xu X, et al. Universal weighting metric learning for cross-modal retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021 (Early Access)

[89] Thomas C, Kovashka A. Preserving semantic neighborhoods for robust cross-modal retrieval//Proceedings of the European Conference on Computer Vision. Online, 2020: 317-335

[90] Liu F, Ye R. A strong and robust baseline for text image matching//Proceedings of the Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Florence, Italy, 2019: 169-176

[91] Zhou M, Niu Z, Wang L, et al. Ladder loss for coherent visual-semantic embedding. Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 13050-13057

[92] Lu J, Batra D, Parik D, et al. Vilbert: Pretraining task- agnostic visiolinguistic representations for vision-and-language tasks. Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2019: 13-23

[93] Li Y, Pan Y, Yao T, et al. Scheduled sampling in vision-language pretraining with decoupled encoder decoder network//Proceedings of the AAAI Conference on Artificial Intelligence. Online, 2021, 35(10): 8518-8526

[94] Tan H, Bansal M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing. Hong Kong, China, 2019: 5100-5111

[95] Lu J, Goswami V, Rohrbach M, et al. 12- in-1: Multi-task vision and language representation learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Online, 2020: 10437-10446

[96] Yu F, Tang J, Yin W, et al. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs//Proceedings of the AAAI Conference on Artificial Intelligence. Online, 2021, 35(4): 3208-3216

[97] Dou Z, Xu Y, Gan Z, et al. An empirical study of training end-to-end vision-and-language transformers//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 18166-18176

[98] Jia C, Yang Y, Xia Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision//Proceedings of the International Conference on Machine Learning. Online, 2021: 4904-4916

[99] Radford A, Kim J, Haallacy C, et al. Learning transferable visual models from natural language supervision//Proceedings of the International Conference on Machine Learning. Online, 2021: 8748-8763

[100] Huo Y, Zhang M, Liu G, et al. WenLan: Bridging vision and language by large-scale multi-modal pretraining. arXiv preprint arXiv: 2103.06561. 2021: 1-9

[101] Hong W, Ji K, Liu J, et al. Gilbert: Generative vision-language pre-training for image-text retrieval//Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval. Online, 2021: 1379-1388

[102] Li L, Yatskar M, Yin D, et al. Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv: 1908.03557. 2019: 1-14

[103] Li G, Duan N, Fang Y, et al. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34(07): 11336-11344

[104] Su W, Zhu X, Cao Y, et al. VL-BERT: Pre-training of Generic Visual-Linguistic Representations//Proceedings of the International Conference on Learning Representations. Online, 2020: 1-14

[105] Li X, Yin X, Li C, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks//Proceedings of the European Conference on Computer Vision. Online, 2020: 121-137

[106] Zhang P, Li X, Hu X, et al. Vinvl: Revisiting visual representations in vision-language models//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Online, 2021: 5579-5588

[107] Huang Z, Zeng Z, Liu B, et al. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. arXiv preprint arXiv: 2004.00849. 2020: 1-17

[108] Kim W, Son B, Kim I. Vilt: Vision and language transformer without convolution or region supervision. International Conference on Machine Learning. Online, 2021: 5583-5594

[109] Wang Z, Yu J, Yu A, et al. Simvlm: Simple visual language model pretraining with weak supervision//Proceedings of the International Conference on Learning Representations. Online, 2022, 1-17

[110] Chen Y, Li L, Yu L, et al. Uniter: Universal image-text representation learning. Proceedings of the European Conference on Computer Vision. Online, 2020: 104-120

[111] Zhou L, Palangi H, Zhang L, et al. Unified vision-language pre-training for image captioning and vqa//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34(07): 13041-13049

[112] Xia Q, Huang H, Duan N, et al. Xgpt: Cross-modal generative pre-training for image captioning. CCF International Conference on Natural Language Processing and Chinese Computing. Springer, Cham, 2021: 786-797

[113] Qi D, Su L, Song J, et al. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. arXiv preprint arXiv: 2001.07966. 2020: 1-12

[114] Li W, Gao C, Niu G, et al. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning// Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language. Bangkok, Thailand, 2021: 2592-2607

[115] Gan Z, Chen Y C, Li L, et al. Large-scale adversarial training for vision-and-language representation learning//Proceedings of the Advances in Neural Information Processing Systems. Online, 2020, 33: 6616-6628

[116] Hu X, Yin X, Lin K, et al. Vivo: Visual vocabulary pretraining for novel object captioning//Proceedings of the AAAI Conference on Artificial Intelligence. Online, 2021, 35(2): 1575-1583

[117] Alberti C, Ling J, Collins M, et al. Fusion of detected objects in text for visual question answering//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing. Hong Kong, China, 2019: 2131-2140

[118] Cho J, Lei J, Tan H, et al. Unifying vision-and-language tasks via text generation//Proceedings of the International Conference on Machine Learning. Online, 2021: 1931-1942

[119] Lin J, Men R, Yang A, et al. M6: Multi-modality to-multi-modality multitask mega-transformer for unified pretraining// Proceedings of the ACM SIGKDD Conference on Knowledge Discovery & Data Mining. Singapore, Singapore, 2021: 3251-3261

[120] Huang P Y, Kang G, Liu W, et al. Annotation efficient cross-modal retrieval with adversarial attentive alignment//Proceedings of the ACM International Conference on Multimedia. Nice, France, 2019: 1758-1767

[121] Huang Y, Wang J, Wang L. Few-shot image and sentence matching via aligned cross-modal memory. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021, (1): 1-16

[122] Xu X, Tian J, Lin K, et al. Zero-shot Cross-modal Retrieval by Assembling AutoEncoder and Generative Adversarial Network. ACM Transactions on Multimedia Computing, Communications, and Applications. 2021, 17(1s): 1-17

[123] Lin K, Xu X, Gao L, et al. Learning cross aligned latent embeddings for zero-shot cross-modal retrieval. Proceedings of the AAAI Conference on Artificial Intelligence. Online, 2020, 34(07): 11515-11522

[124] Mafla A, Rezende R S, GO´MEZ L, et al. Stacmr: Scene-text aware cross-modal retrieval//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Online, 2021: 2220-2230

[125] Li X, Xu C, Wang X, et al. COCO-CN for cross-lingual image tagging, captioning, and retrieval. IEEE Transactions on Multimedia. 2019, 21(9): 2347-2360

[126] Dong J, Li X, Snoek C G M. Predicting visual features from text for image and video caption retrieval. IEEE Transactions on Multimedia. 2018, 20(12): 3377-3388

[127] Li S, Xiao T, Li H, et al. Person search with natural language description//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 1970-1979

[128] Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research. 2013, 47: 853-899

[129] Young P, Lai A, Hodosh M, et al. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics. 2014, 2: 67-78

[130] Chen X, Fang H, Lin T Y, et al. Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv: 1504.00325. 2015: 1-7

[131] Gu J, Meng X, Lu G, et al. Wukong: 100 Million Large-scale Chinese Cross-modal Pre-training Dataset and A Foundation Framework. arXiv preprint arXiv: 2202.06767. 2022

[132] Han Y, Huang G, Song S, et al. Dynamic neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence.2022, 44(11): 7436-7456

LIU Meng, Ph.D., professor. Her main research interests include multimedia computing and information retrieval.

QI Meng-Jin, M.S. candidate. His main research interests include multimedia content analysis and information retrieval.

ZHAN Zhen-Yu, M.S. candidate. His research interest is cross-modal information retrieval.

QU Lei-Gang, M.S. candidate. His research interest is cross-modal information retrieval.

NIE Xiu-Shan, Ph.D., professor. His main research interests include artificial intelligence, machine learning and data mining.

NIE Li-Qiang, Ph.D., professor. His main research interests include multimedia computing and information retrieval.

Background

Recent years have witnessed the rapid growth of multimedia data, such as texts and images, inducing many researchers to work on the multimodal representation, understanding, and reasoning. Thereby, image-text matching, focusing on measuring the semantic similarity between an image and a text, has attracted extensive research attention. It benefits a variety of applications, such as cross-modal retrieval, visual question answering, and multimedia understanding. However, it is non-trivial to build an effective image-text matching model because of the difficult intra-modal reasoning and cross-modal alignment.

Recently, deep learning techniques have emerged as powerful methods for various tasks, such as image classification and visual object detection. This motivates many researchers to resort to deep learning approaches to tackle the image-text matching task, and achieve significant improvement. Although many deep learning-based methods have been proposed for image-text matching, we are unaware of comprehensive surveys of the subject. A thorough review and summarization of existing work is essential for further progress in image-text matching, particularly for researchers wishing to enter the field. Therefore, to provide advanced readers with a thorough overview of this field, including models, datasets, and future directions, we aim to summarize the work on image-text matching and present this survey. To the best of our knowledge, this article would be the first survey in the field of image-text matching.

This paper is supported by the National Natural Science Foundation of China, No. 62006142 and No. U1936203; the Shandong Provincial Natural Science Foundation for Distinguished Young Scholars, No. ZR2021JQ26; the Major Basic Research Project of Natural Science Foundation of Shandong Province, No. ZR2021ZD15; Science and Technology Innovation Program for Distinguished Young Scholars of Shandong Province Higher Education Institutions, No. 2021KJ036; as well as the special fund for distinguished Professors of Shandong Jianzhu University.

In this paper, we first introduce the definition and challenges of image-text matching task. And then we classify existing approaches into four categories and introduce the corresponding methods, including their motivations and method details. Afterwards, we would describe currently available datasets and evaluation metric used for image-text matching, as well conduct comparisons among different approaches. Finally, we would discuss possible future developments in the task of image-text matching.