软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.006774]

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

视觉语言预训练综述 *

殷烱[1], 张哲东[3], 高宇涵[2,3], 杨智文[1], 李亮[4], 肖芒[5], 孙 棋[3], 颜成钢[3]

[1](杭州电子科技大学 计算机学院,浙江 杭州 310018)

[2](杭电(丽水)研究院,浙江 丽水 323000)

[3](杭州电子科技大学 自动化学院,浙江 杭州 310018)

[4](中国科学院计算技术研究所,北京 100190)

[5](浙江大学 邵逸夫医院,浙江 杭州 310020)

通讯作者: 肖芒, E-mail: joelxm@zju.edu.cn

摘 要: 近年来深度学习在计算机视觉(CV)和自然语言处理(NLP)等单模态领域都取得了十分优异的性能.随着技术的发展，多模态学习的重要性和必要性已经慢慢展现.视觉语言学习作为多模态学习的重要部分，得到国内外研究人员的广泛关注.得益于 Transformer 框架的发展，越来越多的预训练模型被运用到视觉语言多模态学习上，相关任务在性能上得到了质的飞跃.本文系统地梳理了当前视觉语言预训练模型相关的工作，首先介绍了预训练模型的相关知识，其次从两种不同的角度分析比较预训练模型结构，讨论了常用的视觉语言预训练技术，详细介绍了五类下游预训练任务，最后介绍了常用的图像和视频预训练任务的数据集，并比较和分析了常用预训练模型在不同任务下不同数据集上的性能.

关键词: 多模态学习;预训练模型;Transformer;视觉-语言学习

中图法分类号: TP311

中文引用格式: 殷烱,张哲东,高宇涵,杨智文,李亮,肖芒,孙垚棋,颜成钢.视觉语言预训练综述.软件学报. http://www.jos.org.cn/1000-9825/6774.htm

英文引用格式: Yin J, Zhang ZD, Gao YH, Yang ZW, Li L, Xiao M, Sun YQ, Yan CG. A survey on visual language pre-training. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/6774.htm

A survey on visual language pre-training

YIN Jiong[1], ZHANG Zhe-Dong[3], GAO Yu-Han[2,3], YANG Zhi-Wen[1], Li Liang[4], XIAO Mang[5], SUN Yao-Qi[3], YAN Cheng-Gang[3]

[1](School of Computer Science, Hangzhou Dianzi University, Hangzhou 310018, China)

[2](Lishui Institute of Hangzhou Dianzi University, Lishui 323000, China)

[3](School of automation, Hangzhou Dianzi University, Hangzhou 210016, China)

[4](Institute of computing technology, CAS, Beijing 100190, China)

[5](Sir Run Run Shaw Hospital, Zhejiang University, Hangzhou 310020, China)

Abstract: In recent years, deep learning has achieved excellent performance in unimodal areas such as computer vision (CV) and natural language processing (NLP). With the development of technology, the importance and necessity of multimodal learning has been shown. As an important part of multimodal learning, visual language learning has received a lot of attention from researchers in China and abroad. Thanks to the development of Transformer framework, more and more pre-trained models have been applied to visual language multimodal learning, and the performance of related tasks has been improved qualitatively. In this paper, we systematically review the current work on visual language pretraining models, firstly, we introduce the knowledge about pretraining

* 基金项目: 国家重点研发计划项目（2020YFB1406604），国家自然科学基金（61931008,62071415，U21B2024）
收稿时间: 2022-04-18; 修改时间: 2022-05-29, 2022-08-03; 采用时间: 2022-08-24; jos 在线出版时间: 2022-09-20

models, secondly, we analyze and compare the structure of pretraining models from two different perspectives, discuss the commonly used visual language pretraining techniques, detail the five types of downstream pretraining tasks, and finally, we introduce the datasets of commonly used image and video pretraining tasks, and compare and analyze the commonly used pretraining models on different datasets under different tasks.

Key words: Multimodal learning; Pre-training; Transformer; Visual-Language Learning

机器学习的目标是让机器像人一样感受世界和理解世界。正如人的感官能去感知一样，多模态机器学习旨在处理和理解不同模态(诸如视觉、语言、听觉等)交织融合的信息。从过去到现在，研究者们已经做出了很多单模态学习的工作，诸如人脸识别、目标检测等，并从科学研究扩展到产业落地，最后服务于生活。但是随着深度学习技术的发展，多模态学习慢慢展现出其重要性和必要性[1]。作为人类生活中最重要的文化载体，视觉和语言在多模态学习领域承载着十分重要的一部分，在近几年里，视觉语言多模态学习也得到了广泛的关注和飞速的发展。通常，参数较大的模型往往需要大量的标注数据来进行训练，但由于多模态标注技术、标注成本等一系列因素的制约，高质量的标签数据始终比较缺乏，这也给模型的性能提升带来了瓶颈。

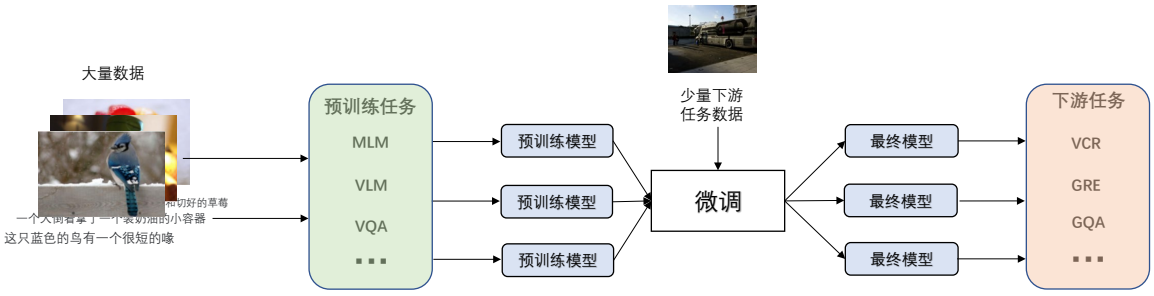

图 1 模型预训练流程图

2017 年美国谷歌公司研究人员提出 Transformer[2]的基础框架，用于解决这个问题。Transformer 模型首先通过自监督学习进行预训练，通过一系列的任务来从大规模的无标注数据中挖掘监督信息以训练模型，从而来学习数据的一般化表征。然后对于不同的下游任务只需要采用少量的人工标注的数据进行微调就能达到优异的效果，预训练流程见图 1 所示。在自然语言处理(NLP)领域中，BERT[3]的出现后，各种预训练任务便如雨后春笋般涌现出来，诸如 GPT[4]系列，MASS[5]等。不仅仅局限在 NLP 领域，计算机视觉(CV)领域中也出现了许多杰出的预训练方法，比如 ViT[6]等。与此同时，模型预训练技术也在多模态领域得到了研究人员越来越多的关注，特别是在视觉-语言联合表征学习方面，预训练模型在各种下游任务上都取得了优异的性能。

如图 2 所示，本文将围绕视觉语言预训练模型展开介绍，并通过以下六个重要方面详细介绍和讨论视觉语言预训练模型的最新进展：首先我们介绍视觉语言预训练模型的相关知识，包括 Transformer 框架、模型预训练范式和视觉语言预训练模型常见网络结构；其次我们介绍三类模型预训练任务，通过这些任务，网络模型可以在无标注的情况下进行跨模态的语义对齐；然后我们将从图像-文本预训练和视频-文本预训练两个方面分别来介绍最新的工作进展，并介绍预训练模型的下游任务；接着我们将介绍广泛使用的图像文本和视频文本的多模态数据集，并比较和分析了常用预训练模型在不同任务下不同数据集上的性能；最后我们对视觉语言预训练进行总结和展望。

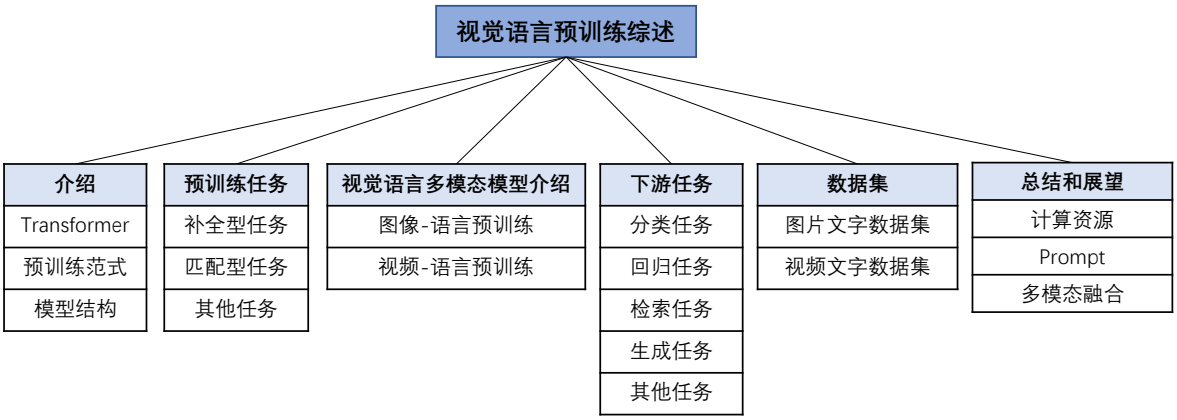

图 2 视觉语言预训练综述结构框图

1 介绍

在本节中，我们将介绍与视觉、语言预训练相关的背景基础知识。第 1.1 节我们将介绍 Transformer 的关键机制和结构；第 1.2 节我们将介绍当前比较流行的预训练范式，包括预训练-微调学习和预训练-提示语学习；第 1.3 节我们从两个不同的角度介绍了当前视觉语言预训练的模型结构。

1.1 Transformer

Transformer[2]最早在自然语言处理(NLP)领域提出，并在各种任务上表现出很好的性能。在此之后，它也被成功应用于其他领域，从语言再到视觉领域。如图 3 所示，一个标准的 Transformer 由几个编码器块和解码器块组成。每个编码器块包含一个自注意(Self-Attention[2])层和一个前馈(Feed Forward)层。不同于编码器块，每个解码器块除了自注意力层和前馈层外，还包含一个编解码注意力层。

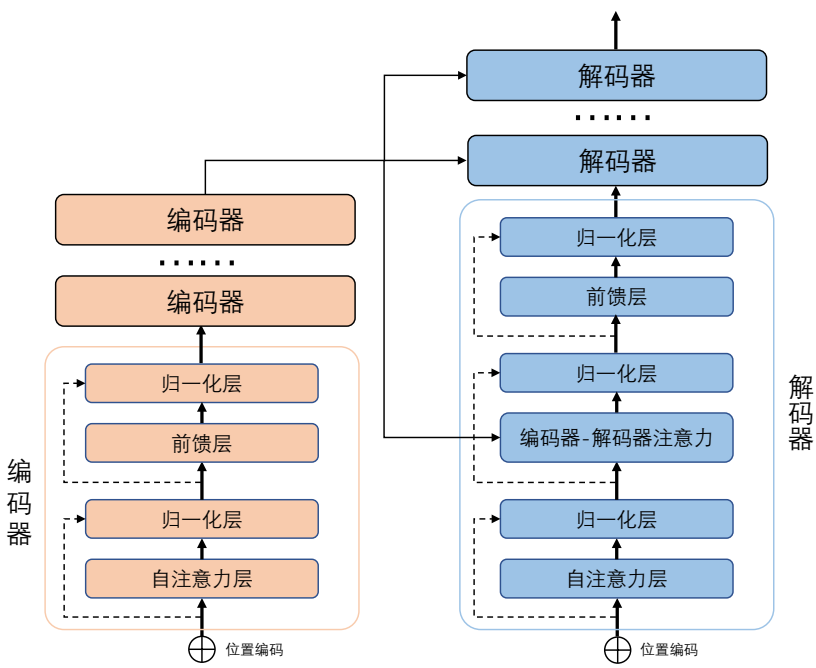

图 3 Transformer 结构图

1.1.1 自注意力机制(Self-Attention)

自注意力机制是 Transformer 的核心机制之一。在自注意力层中，词元序列$X=\{x_0,x_1,...,x_n\}$作为输入，该序列可以是 NLP 领域中的单词序列，也可以是视频和多模态领域的图像特征或视频片段。自注意力层首先将输入的词元序列转换为三个不同的向量，分别命名为：Key($K\in\mathbb{R}^{n\times d^K}$)，Query($Q\in\mathbb{R}^{n\times d^Q}$)，Value($V\in\mathbb{R}^{n\times d^V}$)。注意力的公式如公式(1)所示：

$$Att(X)=soft\max\left(\frac{Q\cdot K^T}{\sqrt{d^Q}}\right)\times V. \tag{1}$$

其中，$Q\cdot K^T$用来获取不同词元之间的相关性得分，$\sqrt{d^Q}$用来使训练过程中相关性得分具有更加稳定的梯度。$soft\max$让获得的概率分布正则化，最后和V相乘，获得相关性加权之后的注意力矩阵。

在解码器中，编解码注意力与自注意力类似，Key 向量和 Query 向量来自于编码器模块，Value 向量来自于前一个解码器模块的输出。但是，并不是所有的词元都能参与自注意力训练。比如，在 BERT[3]的训练阶段，15%的词元被随机掩码，被掩码的词元就不应该参与自注意力进行训练。当在下游任务中进行语句生成的过程中，使用 BERT 生成下一个单词词元时，解码器模块中的自注意力模块只会关注到之前生成的词元，这也是使用掩码来实现的，相应的掩码位置则设置为 0。于是掩码的自注意力公式可以由原来的自注意力公式调整为与如下公式(2)所示：

$$MaskedAtt(X)=soft\max\left(\frac{Q\cdot \mathrm{K}^T}{\sqrt{d^Q}}\circ M\right)\times V. \tag{2}$$

其中，$Q\cdot K^T$计算出的词元相关性得分与随机掩码M进行哈达玛积，未被掩码的元素保留相关性得分，而被掩码的元素则归零。最后经过$soft\max$归一化之后与V相乘，得到掩码注意力矩阵。

1.1.2 多头注意力机制(multi-head attention)

多头注意力机制在 2017 年被 Vaswani[2]等人提出，其旨在从不同方面来对复杂的序列进行建模以助于模型捕捉到更加丰富的特征和信息。具体来讲，输入序列X被线性转换成h个$\{K_i,Q_i,V_i\}_{i=0}^{h-1}$的组，每组重复自注意力过程。最终输入是由$h$个组的输出串联而成，整个过程可以表示为公式(3)(4)：

$$MultiHeadAtt(X)=[Att_0(X),Att_1(X),...,Att_{h-1}(X)]W, \tag{3}$$

$$Att_i(X)=soft\max\left(\frac{Q_i\cdot K_i^T}{\sqrt{d_i^Q}}\right)\times V_i. \tag{4}$$

其中，Att_i表示对第i个组的元素进行注意力操作，$MultiHeadAtt(X)$为最终的多头注意力的输出，其将每个Att矩阵进行拼接，W为权重矩阵。

1.1.3 位置编码

与 CNN[7]或 RNN[8]相比，自注意力机制缺乏捕捉序列位置信息的能力。为了解决这个问题，Vaswani[2]在编码器和解码器的输入中加入了位置编码。位置编码如公式(5)(6)所示：

$$PE(pos,2i)=\sin\left(\frac{pos}{10000^{2i/d_{model}}}\right), \tag{5}$$

$$PE(pos,2i+1)=\cos\left(\frac{pos}{10000^{2i/d_{model}}}\right). \tag{6}$$

其中，pos指词元的位置信息，i指词元的维度。另一种常用的引入位置信息的方法是可学习的位置编码[9]。实验表明[2]，这两种位置编码方法取得了相近的性能。

1.1.4 Transformer 网络结构

最初，Transformer 遵循编码器-解码器结构，由 6 个编码器模块和 6 个解码器模块堆叠而成。编码器模块包含一个多头自注意力层和一个位置前馈层，其中位置前馈层包含两个线性层和一个 ReLU 激活层。相比

于编码器模块，解码器模块多了一层编解码注意力层。为了进一步提升性能，模型中每个模块中都加入了残差结构和 Layer Normalization(LN)层。因此，相比于 CNN 或者 RNN，Transformer 能够更好地捕捉全局信息和进行并行计算。此外，Transformer 简洁明了和可堆叠的结构使其能够在更大的数据集上进行训练，这也促进了预训练的发展。Transformer 的核心是自注意力机制，在传统自注意力模块下的时间复杂度和空间复杂度都是 $O(n^2)$。现有的视觉语言预训练模型采用 Transformer 框架对视觉和语言特征进行编码和对齐，并在尽可能降低时间复杂度的同时提升对不同下游任务的表现。

1.2 预训练范式

1.2.1 预训练-微调(Pretrain Fine-tuning)

预训练-微调已经成了经典的预训练范式。其做法是：首先以监督或无监督的方式在大型数据集上预训练模型，然后通过微调将预训练的模型在较小的数据集上适应特定的下游任务。这种模式可以避免为不同的任务或数据集从头开始训练新模型。越来越多的实验证明，在较大的数据集上进行预训练有助于学习通用表征，从而提高下游任务的性能。GPT[4]在对有 7000 本未出版书籍的 BooksCorpus 数据集[10]进行预训练后，在 9 个下游基准数据集(如 CoLA[11]、MRPC[12]上获得平均 10%的性能大提升。视觉模型 ViT-L/32[6]在对拥有 3 亿张图像的 JFT-300M[13]进行预训练后，在 ImageNet[14]的测试集上获得了 13%的准确率提升。

目前，预训练微调范式在 NLP 和 CV 领域都在如火如荼展开工作，多模态领域也不例外，大量优秀的工作在此诞生，包括图像-文本和视频-文本领域。

1.2.2 预训练-提示(Pretrain Prompt)

提示学习起源于 NLP 领域，随着预训练语言模型体量的不断增大，对其进行微调的硬件要求、数据需求和实际代价也在不断上涨。除此之外，丰富多样的下游任务也使得预训练-微调阶段的设计变得繁琐复杂，提示学习就此诞生。在预训练-提示范式中通常使用一个模板来给预训练模型提供一些线索和提示，从而能够更好地利用预训练语言模型中已有的知识，以此完成下游任务。

在 GPT-3[15]中，所有任务都可以被统一建模，任务描述与任务输入视为语言模型的历史上下文，而输出则为语言模型需要预测的未来信息，通过给予模型一些提示语，让模型根据提示语来生成所需要的输出，这种方式也被称为是情景学习(in-context learning)。Prefix-Tuning[16]摒弃了人工设计模板或自动化搜索模板的方式，提出了任务特定的可训练前缀。P-tuning V1[17]首次提出了用连续空间搜索的嵌入来做提示语。P-tuning V2[18]引入深度提示编码(Deep Prompt Encoding)和多任务学习(Multi-task Learning)等策略进行优化，解决 v1 版本在一些复杂的自然语言理解任务上任务不通用和规模不通用的问题。

提示学习相对于微调的优势在于：1)计算代价非常低。由于整个模型的参数都是固定的，并不需要对模型中所有的参数进行微调。2)非常节省空间。在使用预训练模型进行微调时，每个不同的下游任务的参数都会相应改变，因此每个任务都需要进行存储，而提示学习则不需要。基于这些优势，提示学习已经称为了 NLP 领域的又一大研究热点，预训练-提示也作为继预训练-微调的又一大范式，处处崭露头角。在多模态领域也慢慢燃起了提示学习之火，诸如 CLIP[19]，CPT[20]等出色的工作应运而生。

1.3 模型结构

在本小节中，我们从两个不同的角度介绍视觉语言预训练模型的体系结构:(1)从多模态融合的角度对比单流结构与双流结构，(2)从整体架构设计的角度对比仅编码结构和编码-解码结构。

1.3.1 单流与双流的对比

单流结构：单流结构指一种将文本和视觉特征连接到一起，然后输入进单个 Transformer 模块中，如图 4(左)所示。单流结构利用注意力来融合多模态输入，因为对不同的模态都使用了相同形式的参数，其在参数方面更具效率。

双流结构：在双流结构中文本和视觉特征没有连接在一起，而是单独输入到两个不同的 Transformer 模块中，如图 4(右)所示。这两个 Transformer 没有共享参数。为了达到更高的性能，双流结构使用交叉注意

力的方式(如图 4(右)中的虚线所示)来实现不同模态之间的交互。为了达到更高的效率，处理不同模态信息的 Transformer 模块之间也可以不存在交叉注意。

1.3.2 仅编码结构与编码-解码结构

许多视觉语言预训练模型采用仅编码的体系结构，其中跨模态表示被直接输入到输出层以生成最终输出。而其他视觉语言预训练模型使用转换器编码-解码体系结构，在这种体系结构中，交叉模态表示首先被输入解码器，然后再输入输出层。

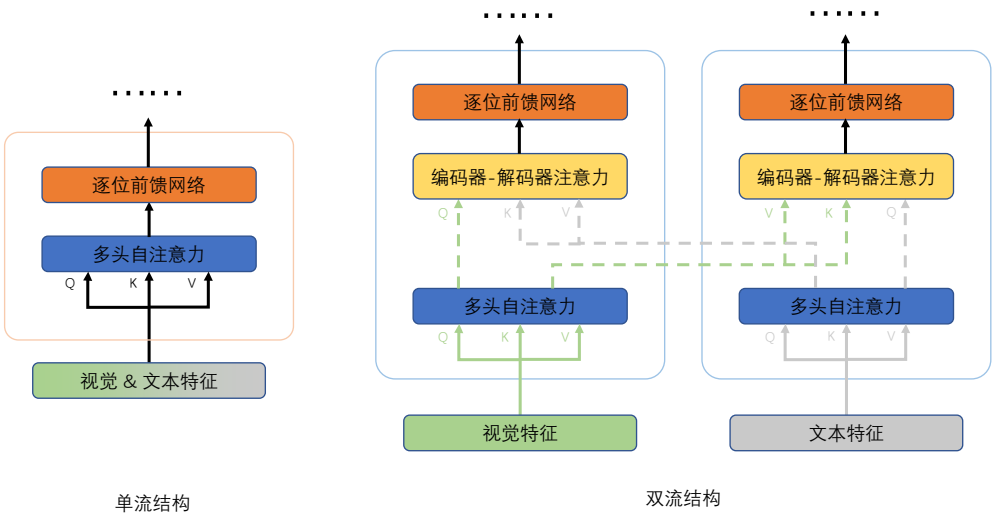

图 4 单流结构(左)和双流结构(右)

2 预训练任务

本节将介绍如何使用不同的预训练任务对视觉语言预训练模型进行预训练，这对于模型学习视觉语言的一般化表征至关重要。我们将预训练任务归纳为三类:补全型、匹配型、其他型。

补全型任务通过利用未被掩码的剩余信息来理解模态，从而重建补全被掩码的元素。

匹配型任务是将视觉和语言统一到一个共同的潜在空间中来生成一个一般化的视觉-语言表达。

其他型任务的内容中包含了其他预训练任务。

2.1 补全型任务

掩码语言建模(Masked Language Modeling, MLM)在 1953 年首次由 Talyor 在文献[3]中提出，因 BERT 模型将其作为一种新颖的预训练方式而广为人知。视觉语言预训练模型中的 MLM 与预训练语言模型中的 MLM 相似，但视觉语言预训练模型中的 MLM 在预测掩码文本词元时不仅可以使用剩余的文本词元，也可以同时使用视觉词元。通常来讲，视觉语言预训练模型遵从 BERT 模型的掩码方式，在输入的文本词元中随机掩码其中 15%，然后将其中 80%用一个特殊的词元***[mask]***代替，10%用随机词元代替，剩余 10%保持不变。

前缀语言建模(Prefix Language Modeling, PrefixLM)是掩码语言建模和语言建模(LM)的统一。前缀语言建模的提出是为了使该模型具有实质性的生成能力，从而在不进行微调的情况下实现文本导向的零样本学习。前缀语言建模不同于标准语言建模，它可以对前缀序列进行双向注意力和仅对剩余词元执行自回归因式分解。在序列间(Sequence-to-Sequence)框架下的前缀语言建模不仅具有与掩码语言建模相同的双向上下文表征，也具有类似于标准语言建模文本生成的能力。

与掩码语言建模一样，掩码视觉建模(Masked Vision Modeling, MVM)对视觉(图像或视频)区域或色块进行采样，通常掩码其 15%的视觉特征。掩码视觉建模需要在剩余的视觉特征和所有文本特征的基础上重建被掩码的视觉特征。被掩码的视觉特征设为零矩阵。由于视觉特征是高维和连续的，视觉语言预训练模型提出了两种掩码视觉建模变体。

(1)掩码特征回归通过学习将掩码特征的模型输出回归到其原始视觉特征。模型首先将掩码特征的模型输出转换为与原始视觉特征相同维度的向量，并对此向量与原始视觉特征进行 L2 回归来恢复掩码特征。

(2)掩码特征分类通过学习预测掩码特征的目标语义类别。模型首先将掩码特征的输出反馈到全连接层，以预测对象类的分数，然后通过 softmax 函数将其转换为正态分布。模型的训练有两种方法。一种是模型将目标检测模型中最可能的目标类作为硬标签(概率是 0 或者 1)，假设检测到的目标类是掩码特征的真值标签，使用交叉熵损失来最小化预测结果和伪类之间的差距。另一种是模型使用软标签作为监督信号，也就是检测器的原始输出(即对象类的分布)，并最小化两个分布之间的 K-L 散度。

2.2 匹配型任务

视觉-语言匹配(Vision-Language Matching, VLM)是最常用的视觉和语言一致性预训练目标。在单流模型中，使用特殊词元*[CLS]*的表示作为两种模态的融合表示。在双流模型中，将特殊视觉词元$[CLS_V]$的视觉表示和特殊文本词元$[CLS_T]$的文本表示串联起来，作为两种模态的融合表示。模型将这两种模态关系的融合表示输入给全连接层和 sigmoid 函数，预测出一个 0 到 1 之间的匹配度，其中 0 表示视觉和语言不匹配，1 表示视觉和语言匹配。在训练过程的每一步中，模型都会从数据集中提供匹配或不匹配的样本对，其中不匹配的样本对由随机替换匹配样本对中的视觉或者语言部分生成。

视觉语言对比学习(Vision-Language Contrastive Learning, VLC)在一个训练批次 N 个视觉-语言对的 N×N 个可能的视觉语言对中预测出匹配的视觉-语言对。注意，在一个训练批中有 N 到 N^2 个不匹配视觉-语言对。模型分别使用特殊视觉词元$[CLS_V]$的视觉表示和特殊文本词元$[CLS_T]$的文本表示来表达视觉和语言两种模态的融合表示。模型通过 softmax 函数归一化视觉(图像或视频)到文本和文本到视觉的相似性，并利用这些相似性的交叉熵损失函数进行训练和更新，相似度通常用点积来实现。

文字-区域对齐(Word-Region Alignment, WRA)是一种无监督的预训练方式，用于对齐视觉区域(vision patches)和文字。模型运用最优运输(Optimal Transport)来学习视觉和语言之间的对齐。因为精确最小化(the exact minimization)是在计算中是难以处理的，所以一般使用 IPOT 算法来近似 OT 距离。求出最小值后，以 OT 距离作为 WRA 损耗来训练模型。

2.3 其他任务

为了更好地对视频的时序进行建模，模型随机打乱一些输入帧的顺序，然后预测每一帧的实际位置。在具体的应用中，帧时序建模(FOM)会被设计成一个分类器。视觉语言预训练模型有时也使用一些下游任务的训练对象，如视觉问答(VQA)[21]和视觉描述(VC)来作为预训练对象。在视觉问答方面，模型采用上述融合的表示方法，使用一个全连接层，利用转换后的表示方法对预定义的答案进行分类预测。除此之外，还可以直接生成原始文本格式的答案。在视觉描述方面，为了重构输入句子赋予模型生成能力，模型使用自回归解码器生成对应图像或视频的文本描述。

3 视觉语言多模态模型介绍

视觉和语言是人类感知世界的两个重要方面，因此训练神经网络模型处理多模态信息对于人工智能的发展有着重要的意义。近年来，许多研究工作通过对其视觉和语言的语义信息实现了各种跨模态任务。其中图像文本预训练和视频文本预训练得到了最广泛的研究。本节我们将介绍图像-文本预训练和视频-文本预训练两个方面近年来的最新进展。

3.1 图像-文本预训练

2019 以来，有关图像-文本预训练的研究慢慢展开。Lu 等人提出了基于双流结构的 ViLBERT[22]，输入的文本和经过 Fast-RCNN[23]处理后的图像特征分别经过 Transformer 的编码器进行编码后，通过共注意力机制模块将语言信息和视觉信息相融合。该共注意力机制模块基于 Transformer 中自注意力模块的结构，在每个模态中都用自身的 Query 和另一个模态的 Value 和 Key 计算注意力，以此来融合多模态信息。Alberti 等人提出了 B2T2 模型[24]，进行了详细的对照实验，讨论了双编码器结构中的早期融合结构和晚期融合结构的优劣，得出早期融合结构效果更优的结论。Tan 等人提出了 LXMERT[25]，该模型与 ViLBERT 同样使用了双流结构，即图像和文本分别经过独立的编码器进行编码，然后通过跨模态编码器进行模态信息的融合。该跨模态编码器采用多层堆叠的方式，每一层中包含有两个自注意力层，两个前馈层和一个双向交叉注意力层，分别对视觉到语言和语言到视觉进行了交叉注意力。模型可以输出视觉，文本和跨模态三种信息。

Li 等人提出了基于单流结构的 VisualBERT 模型[26]，希望通过自注意力机制来挖掘图像和文本中的对应关系。与 BERT 类似，该模型直接将文本与图像信息通过 Transformer 进行对齐和融合。语言部分经过 BERT 得到文本特征，即词向量编码+位置编码+模态分割编码；而视觉部分采用了经过 Fast-RCNN 特征提取的区域特征，以及与之对应的位置编码作为输入。Li 等人提出了基于单流的 Unicoder-VL 模型[27]，该模型与 VisualBERT 最大的不同在于对视觉信息的输入处理上。输入的图像首先经过 Fast-RCNN 提取区域特征，将区域图像特征和其对应的边界框特征分别通过全连接层映射到和语言编码维度相同的向量空间上，加上对应区域的文本类别标签向量，与文本向量一起输入到单流模型中。

Su 等人提出了单流的 VL-BERT[28]，该模型在输入上分为四层，其中词嵌入层使用原始的 BERT 的设定；视觉特征层由视觉外部特征和视觉几何特征拼接而成，视觉外部特征是由 Faster-RCNN 提取，而视觉几何特征是根据位置信息做正余弦处理，经过全连接层得到的特征。分割层用于区分不同来源的信息输入。位置嵌入层与 BERT 类似，通过对文本添加一个可学习的位置特征来表示文本输入的顺序和相对位置。由于输入的图像没有相对的位置，所以图像的位置信息都是相同的。为了打造一个端到端的多模态生成和理解模型，Zhou 等人提出了 VLP[29]。在此之前，多模态预训练工作只包含编码器，需要根据不同的下游任务设计不同的解码器。该模型采用单 Transformer 结构，在预训练任务中引入掩码语言模型(MLM)，对于不同的下游任务，只需要对解码器进行微调训练。

但由于之前的预训练工作很少考虑到图像描述(Image Caption)等生成任务，Xia 等人[30]专门设计了针对生成任务模型结构。该模型借鉴了 NLP 领域的 MASS 模型[5]，对于文本，在 encoder 端连续掩码屏蔽掉一个连续序列的词，在 decoder 端只输入前 k-1 个词且屏蔽 encoder 中提供的词，以此来迫使 decoder 通过 encoder 来获取语义和视觉信息。Yu 等人提出了 ERNIE-VIL[31]，该模型使用了双流架构，提出了三个多模态场景预测任务：目标预测、属性预测和关系预测。在目标预测任务中，模型需要根据文本上下文和图像对文本掩码部分进行预测；在属性预测中，模型需要根据上下文和图像对物体的属性进行预测；在关系预测中，模型需要根据上下文和图像中的<物体，关系，物体>三元组进行物体与物体之间关系的预测。2021 年 Ramesh 等人提出了 DALL-E[32]，该模型主要用于文本生成图像任务，含有 120 亿的参数量，整体包含三个阶段，在第一个阶段，DALL-E 将一张256×256图像分为32×32个图像块，再使用 VQVAE[33]将经过编码的每个图像块映射到一个 8192 维的词表中，最终将一个图像转换为 1024 的词元序列；在第二个阶段，用 BPE 编码器对文本进行编码，得到最多 256 个文本词元，再将文本和图像词元进行拼接，输入到 120 亿参数量的 Transformer 中；最后，对生成的图像进行采样，并用 CLIP 模型对采样进行排序，得到与文本最匹配的图像。

之前大多预训练工作都是先进行预训练，然后进行微调工作，各个下游任务之间相对独立，每一个下游任务都需要重新进行微调一个模型。由此 Lu 等人提出了 12-in-1 模型[34]。该模型是 ViLBERT 的拓展，将常用的 12 个数据集按对应的任务分类，相似的任务分为一组，共分为视觉问题回答、基于图像描述的图像检索，看图识物和多模态验证四组，进行多任务学习。Hu 等人提出了 UniT[35]，旨在多个领域的不同任务使

用同一个模型，在所有任务中共享相同的模型参数，而不是分别对特定任务的模型进行微调。对于每个任务的不同领域，UniT 采用不同的编码器，但都使用相同的解码器，并且在解码器之后加上一个特定任务的输出头。UniT 在尽可能减少参数量的同时，保证了效果，并且能在七个不同的下游任务中达到了不错的效果。Li 等人提出了 BLIP[36]，希望训练一个统一的多模态预训练模型来同时解决多模态理解和生成任务。BLIP 是个多模态的混合编码-解码器，可以实现：1)图像或文本的单模态编码；2)基于图像的文本编码；3)基于图像的文本解码三个功能。

多模态预训练的研究本质在于如何更好地对多种模态信息进行对齐和融合，以此来挖掘模态间对应信息，对此模型对多模态信息的细粒度融合是非常必要的。Li 等人指出，以往的视觉语言预训练方法没有将文本中的单词对应图中相应的区域，因此天然就是一个弱监督学习系统，因此提出了 Oscar[37]，将训练样本定义为一个三元组，每个三元组由单词序列，一组目标标记和一组图像区域组成。训练分两种角度，模态视角区分图像和文本表示，字典视角区别两个不同的语义空间。Xue 等人[38]认为把视觉内部的关系信息和跨模态对齐封装在一个 Transformer 网络中是不合理的，这种方式会忽略每个物体的特殊性，由此限制了 Transformer 中的多模态对齐学习。于是作者在视觉部分也采用了 Transformer，用自注意力来对视觉信息进行编码，以此来促进模态内的学习。Yao 等人指出，大多数现有的方法都是采用交叉/自注意力机制来进行跨模态的交互，以此感知模态间的相似性，但是交叉/自注意力在训练和推理方面的效率都较低，由此提出了 FILIP[39]，通过跨模态的晚期交互机制来实现更细粒度的对齐。FILIP 通过对比损失增强了图像块和文本单词之间的细粒度表达能力的同时，也保证了大规模预训练和推理的效率。Duan[40]等人认为，改善多模态信息的对齐部分将大大提高模型的性能，提出了较之前工作更为有效的对齐方式，使用聚类表示在更高更稳定的高层表征上进行模态对齐。其使用一个可学习的编码表将常见的文本-图像特征向量量化为编码词，与单模态特征相比，这些编码为对比推理提供了更加稳定的表现。实验结果表明，其在零样本跨模态检索和其他迁移学习任务上都取得了不错的效果。

在图像文本预训练中，一些工作也针对其中的目标检测进行改进。由于大多数预训练任务都采用目标检测模型来获取图像中感兴趣区域的视觉特征，然而区域特征提取器是根据特定视觉任务设计的，会造成其他重要视觉信息的缺失，对多模态任务很容易造成语义鸿沟。为此，Huang 等人提出了 Pixel-BERT[41]，对整张图像进行卷积池化后再进行随机采样，再与语义嵌入(semantic embedding)相加，得到像素级的特征编码后，与文本编码拼接输送给 Transformer 进行训练。Zhang 等人提出了 VinVL[42]，在其团队的前作 Oscar 模型上开发了一个新的目标检测模型，通过丰富视觉对象和属性类别，扩大模型尺寸并在一个更大的数据集上训练，建立一个新的目标检测模型，从而在更广泛的视觉语言任务上提高了性能。Huang 等人发现用 Fast-RCNN 提取的视觉区域特征存在上下文信息的丢失等问题，由此提出了 SOHO[43]。该模型以整张图像作为输入，以端到端的方式学习视觉表征，利用视觉字典把不同的视觉语义信息聚合成视觉词元，弥补了视觉特征和语言词元之间的鸿沟。

对比学习作为一种常用的自监督学习方法，在图像文本预训练中也表现出很出色的跨模态对齐和零样本学习的能力。Radford 等人提出了 CLIP[19]，CLIP 整体采用对比学习的方法，将图像和文本分别进行特征提取和编码后，计算图像文本对的余弦距离，相匹配的图像文本对距离趋向于 1，而不匹配的则趋向于 0，以此来对图像和文本建立联系。CLIP 在零样本学习上的效果足以媲美 ResNet50[44]，对之后的工作产生了很大的影响。Li 等人提出了 UNIMO[45]，该模型能够有效地同时进行的单模态和多模态的内容理解和生成任务，区别于其他模型只能采用有限的多模态图像文本对进行训练，该模型可以利用大量的开放域文本语料和图像进行训练。并且通过一系列的增强方式产生不同粗细粒度特征的正负样本，实现跨模态的对比学习。Li 等人提出了一个全新的视觉语言预训练框架 ALBEF[46]，首先通过图像编码器和文本编码器分别对图像和文本进行编码。然后使用多模态编码器通过跨模态注意力将图像特征与文本进行融合。ALBEF 在图像编码器和文本编码器之间加入了中间量的图像文本对比损失，使多模态编码器能够更好地进行跨模态对齐。2022 年 Yang 等人利用跨模态和模态内的自监督，提出了三重对比学习的视觉语言预训练 TCL[47]。之前的研究通

过跨模态对比损失简单地对齐图像和文本表示，TCL 进一步考虑模态内监督，以确保学习到的表示在每个模态中也有意义，进而有利于跨模态对齐和联合多模态嵌入学习。为了在表征学习中融入局部和结构信息，TCL 进一步引入了局部 MI，它最大化了全局表征和来自图像块或文本标记的局部信息之间的互信息。大量试验结果表明，TCL 性能有显著提高。

为了融合不同模态的任务，学习不同模态的信息，Wang 等人提出 VLMo[48]，将传统的 FFN 模块分为视觉、语言和跨模态三条不同的路径，分别构成双编码器结构和融合编码器结构以适用于不同的下游任务，在多模态检索等问题上用双编码器，在需要跨模态语义信息等问题上用融合编码器。该模型在多个下游任务中都取得了不错的效果。Shen 等人认为现有的视觉语言预训练方法太依赖于视觉编码器，但是高性能的视觉编码器往往被类别标签或边界框等标注信息制约，不具备良好的泛化性能，由此提出了 CLIP-ViL[49]。在 CLIP-ViL 的工作中，Shen 等人着重研究了 CLIP 带来的优势，并提出了在两种典型的场景中使用 CLIP 作为视觉的编码器：1)将 CLIP 插入到特定于任务的微调中；2)借助 CLIP 良好的零样本迁移学习的能力，将 CLIP 与视觉语言预训练相结合，并迁移到下游任务中。在下游任务中，模型获得了不错的效果。Dou 等人[50]用大量的实验尝试了端到端的视觉语言 Transformer 的效果以及各个部分的比较，得到了如下结论，1)VIT 在模型中起到的作用要高于语言 Transformer；2)cross-attention 相比于 self-attention 能更好的融合视觉语言信息；3)在视觉问答(VQA)和图像文本检索(image-text retrieval)中，只使用 encoder 效果要好于使用 encoder-decoder;4)Masked image modeling 这个预训练任务不重要。

数据集的质量和规模对于模型训练来说至关重要，Qi 等人提出了 ImageBERT[51]，设计了一种弱监督的方法，并从网络上搜集制作了一个千万级的图像文本数据集。由于数据集的来源不同，质量也就不同。于是作者将预训练过程分为了两个部分，首先用大量的域外数据集进行模型训练，然后再用小规模的域内数据进行训练，从而在目标任务上得到更好的效果。Jia 等人认为大多预训练工作还都是利用专业的多模态数据集(诸如 Conceptual Captions、MS COCO 等)，严重依赖于昂贵的专家知识，由此提出了 ALIGN[52]，ALIGN 利用了超过十亿个有噪声的图像文本对的数据集来训练，并且发现在这样的大规模噪声数据集上预训练的视觉语言表示在各种下游任务上取得了非常强的性能。Wang 等人提出了 SimVLM[53]，旨在大规模的 web 数据集上对图像文本和仅文本输入上进行预训练，用大规模弱监督学习来降低训练的复杂度。在预训练的方法上，不同于一般的多模态预训练模型使用 MLM，SimVLM 使用了 prefixLM 方法，即给定前缀(视觉信息)，生成后续内容，以此来保留视觉语言表征。

目前预训练-提示(Pretrain Prompt)在 NLP 领域已经成为了继预训练-微调(Pretrain Fine-tuning)之后的又一大预训练范式。Tsimpoukelli 等人提出 Frozen[54]，将 NLP 领域广泛应用的 Prompt 引入到了多模态领域，利用图像编码器把图像作为一种动态的提示词，和文本一起送入到语言模型中，以此能在语言模型中更好地获取先验知识。在训练时 Frozen 将选择冻结语言模型中的参数，仅训练图像编码器相关的参数。Yao 等人提出了 CPT[20]，CPT 主要在视觉描述定位(visual grounding)任务上进行。CPT 采用 Prompt 范式，其首先将图像用不同颜色来区分不同的实体模块，随后将问题文本和颜色块问题模板拼接，最后模型只需要预测描述在哪一块颜色块中即可，使视觉描述定位任务变为了一个为填空问题。

Transformer 因其优异的全局依赖关系建模能力，成为了多模态预训练的首选架构。但由于多模态预训练过于庞大的输入信息，当前来讲视觉语言预训练工作仍然需要极大的算力资源做支撑，致使部分研究人员无法展开相应研究。如何轻量化预训练模型以节省计算资源也是一个值得研究的内容。Transformer 因其优异的全局依赖关系建模能力，成为了多模态预训练的首选架构。然而在多模态领域中，捕捉局部信息对最终模型的推理也很重要，但是对于不同的目标需要配备不同大小的感受野，这大大增加了显存占用和计算量。Zhou 等人提出了一种轻量化的路由方案 TRAR[55]，在 Transformer 的每一层上都配备了一个路由控制器，根据上一层的输入来动态地选择每一步该采用的最优注意力。Kim 等人提出了 ViLT[56]，是一个参数量较小的多模态预训练模型。其通过块映射(patch projection)的多模态预训练方法，在保证效果的前提下大大减小了模型复杂度和运行时间。ViLT 采用了单流架构，相异于其他预训练模型需要在视觉模态上使用一个独立

的视觉编码器，ViLT 使用预训练的 ViT 来处理视觉特征后仅仅用了简单的线性映射，大大降低了视觉编码器的参数量。

2022 年也诞生出很多不错的工作。Zhou 等人提出了无监督的视觉语言预训练模型 UVLP[57]，其根据检索的方式构建了一个弱监督视觉语言语料库，然后通过基于检索的多粒度对齐来学习非对齐文本和图像源的强视觉和语言联合表示。实验表明 UVLP 在 VQA、NLVR2 等任务上都有不错的表现。Wang 等人提出了一个任务无偏和模态无偏的框架，OFA[58]，以达到任务全面性的效果。OFA 通过人为指定的预训练和微调任务来达到模型的任务无偏，仅使用 transformer 编码器作为模态无偏的模型框架而不针对任何下游任务添加可学习的模态组件。OFA 通过在 2 千万个公开的图像-文本对上进行了预训练，在图像描述、文本图像生成、VQA，视觉蕴含等多个下游任务上达到了非常不错的效果。

图像-文本预训练模型汇总见表 1。

表 1 图像-文本预训练模型汇总表

模型	预训练任务	预训练数据	下游任务
单流模型			
VisualBERT[26]	MLM+VLM	COCO	GRE+NLVR+VCR+VQA
B2T2[24]	MLM+VLM	CC3M	VCR
Unicoder-VL[27]	MLM+VLM+MVM	CC3M+SBU	VLR+VCR
VL-BERT[28]	MLM+MVM	CC3M	GRE+VCR+VQA
UNITER[59]	MLM+VLM+MVM+WRA	COCO+VG+SBU+CC3M	GRE+VLR+NLVR+VCR+VE+VQA
12-IN-1[34]	MLM+MVM	MTL	GQA+GRE+VC+NLVR+VE+VQA
ImageBERT[51]	MLM+VLM+MVM	LAIT+CC3M+SBU	VLR
PixelBERT[41]	MLM+VLM	COCO+VG	VLR+NLVR+VQA
OSCAR[37]	MLM+VLM	COCO+SBU+CC3M+FLKR+VQA+GQA+VGQA	GQA+VC+VLR+NLVR+NoCaps+VQA
UNIMO[45]	VLC	COCO+VG+CC+SBU	VC+VQA+VLR+VE
ERNIE-ViL[31]	MLM+MVM	CC3M+SBU	GRE+VLR+VCR+VQA
VinVL[42]	MLM+VLM	COCO+CC3M+SBU+FLKR+VQA+GQA+VGQA	GQA+VC+VLR+NLVR+NoCaps+VQA
VL-T5[60]	MLM+VLM+VQA+GRE+VC	COCO+VG+VQA+GQA+VGQA	GQA+GRE+VC+MMT+NLVR+VCR+VQA
UniT[35]	MLM	COCO+VG+VQAv2+SNLI-VE	VQA+VE
ViLT[56]	MLM+VLM	COCO+VG+SBU+CC3M	VLR+NLVR+VQA
SOHO[43]	MLM+VLM+MVM	COCO+VG	VLR+NLVR+VE+VQA
CLIP-ViL[49]	MLM+VLM+VQA	COCO+VG+VQA+GQA+VGQA	VE+VLN+VQA
SimVLM[53]	PrefixLM	AltText	VC+NLVR+VE+VQA
CPT[20]	MLM+VLC	COCO	VG
VLMO[48]	MLM+VLC+VLM	COCO+VG+CC3M+SBU	VQA+NLVR+VLR
双流模型			
ViLBERT[22]	MLM+VLM+MVM	COCO+VG	VLR+NLVR+VE+VQA
LXMERT[25]	MLM+VLM+MVM+VQA	COCO+VG+VQA+GQA+VGQA	GQA+NLVR+VQA
VLP[29]	MLM+LM	CC3M	VC+VQA
XGPT[30]	MLM+IDA+VC+TIFG	CC3M	VC+VLR
ALIGN[52]	VLC	AltText	VLR
ALBEF[46]	MLM+VLM+VLC	COCO+VG+CC3M+SBU	VLR+NLVR+VQA
CLIP[19]	VLC	SC	OCR
TRAR[55]	-	VQAv2+COCO+CLVER+Metric	VQA+GRE
METER[50]	MLM+VLM	COCO+VG+CC3M+SBU	VLR+NLVR+VE+VQA
BLIP[36]	VLC+MLM+VLM	COCO+VG+CC3M+CC12M+LAION	VLR+VC+VD+NLVR+VQA
FILIP[39]	-	CC+FLKR+COCO	VLR
TCL[47]	MLM+VLC+VLM	COCO+VG+CC+SBU	VQA+VE+NLVR
UVLP[57]	MLM+VLM	CC	VQA+NLVR+VE
OFA[58]	VLM	SBU+COCO+VG+CC+VQAv2+GQA	VC+VQA+VE

3.2 视频-语言预训练

表 2 视频-文本预训练模型汇总表

模型	预训练任务	预训练数据	下游任务
单流模型			
VideoBERT[61]	MLM+VLM+MVM	SC	AC+VC
CBT[62]	VLC	Kinetics	AC+AS+VC
ActBERT[63]	MLM+VLM+MVM	HT100M	AS+ASL+VC+VQA+VLR
HERO[64]	MLM+VLM+MVM+FOM	HT100M+TV	VC+VLI+VQA+VLR
VATT[65]	VLC	AudioSet+HT100M	AC+VLR
DeCEMBERT[66]	MLM+VLM+VLC	MSR-VTT+Youcook2	VC+VLR+VQA
CLIPBERT[67]	MLM+VLM	TGIF+MSR-VTT	VQA+VLR+VC
双流模型			
CLIP[62]	VLC	SC	OCR
VLM[68]	MMM+MFM	HT100M+MSR-VTT+Youcook2+COIN+CrossTask	VLR+AS+ASL+VQA+VC
Frozen[54]	VLC	WebVid2M+CC3M	VLR
UniVL[69]	MLM+VLM+VC	HT100M	AS+ASL+MSA+VC+VLR

Sun 等人提出了 VideoBERT[61]，该模型是第一个基于 Transformer 的视频语言预训练模型。在视频方面，模型将 n 个连续帧构成一个片段并对其进行特征提取，将特征向量做分层矢量量化(hierarchical vector quantization)处理，得到视频特征词元。语言方面首先用语音识工具提取视频文本，再沿用 BERT 的文本处理方式。最后将视频信息和语言信息拼接，通过 BERT 学习视频与语言之间的关联性。该模型以 YouTube 上大量无标签的视频作为数据集，在视频动作分类，视频描述等任务上都取的了很好的结果。Sun 等人认为 VideoBERT 中使用的矢量量化会丢失很多细粒度的细节，提出了 CBT[62]，该模型采用双流结构，摒弃了 VideoBERT 中的矢量量化操作，直接使用了视觉特征向量向量。CBT 将 BERT 结构扩展到多流结构，并验证了 NCE 损失[70]对于学习跨模式特征的有效性。Luo 等人提出了 UniVL[69]，该模型使用双流结构，用单模态编码器对文本和视频数据分别进行建模，再使用跨模态编码器对两个模态的表征进行联合编码。训练的过程中采用了四个预训练任务，分别是：条件掩码语言建模(CMLM，用于语言损坏)、条件掩码帧建模(CMFM，用于视频损坏)、视频-文本对齐和语言重建。在此基础上，作者还设计了两种预训练策略，包括分阶段预训练策略(StagedP)和增强视频表示策略(EnhanceDV)来促进 UniVL 的预训练。模型取得了很好的效果。Li 等人提出了 HERO[64]。在之前的工作中，视频语言预训练只是简单的改造了来自 NLP 领域的掩码语言建模(MLM)和视觉语言匹配(VLM)的预训练任务。考虑到视频在时间序列上的特殊性，在 HERO 中首先设计了局部视频语言匹配(LVLM)和帧时序建模(FOM)。实验表明，FOM 可以有效优化时间依赖性任务(诸如问答任务)，全局或局部的 VLM 可以优化检索任务。

由于视频相比于图像特征多了时间维度，提取视频特征非常耗时且计算量巨大。Lei 等人提出了一种新的端到端的学习框架，ClipBERT[67]，该框架采用稀疏采样，在每个训练步骤中仅采用少量采样的视频片段，并指出端到端训练策略中使用单个或几个(较少)稀疏采样的视频片段通常比使用密集提取视频特征的传统方法更精确。Akbari 等人提出了端到端的框架，VATT[65]，用于从视频、音频和文本中提取多模态表示。为了获得三种模态的内在共现关系，VATT 中采用了 ViT[6]，而不是分别为每种模态分别保留词元和线性层映射。VATT 通过匹配视频-音频对和视频-文本对的共同空间投影做噪声对比估计(NCE)来进行训练优化。

现有的预训练都是针对特定任务的，单流结构限制了模型对检索式任务的使用，双流结构限制了模型的早期跨模态融合，Xu 等人提出了一个任务无关的多模态预训练模型，VLM[68]。为了不牺牲可分离性，该模型在训练过程中引入了新的预训练任务，掩码模态建模(MMM)，来更好地进行跨模态融合。实验结果表明，VLM 以较少的参数达到了有竞争力地性能。为了解决大规模无标签视频数据自动生成的描述有噪声、不匹配等问题，Tang 等人提出了 DeCEMBERT[66]模型。该模型采用单流结构，首先使用由 ASR[71]生成的文本描述作为模型的文本输入。为了更好地匹配视频和与之对应的生成描述文本，DeCEMBERT 提出了一

个约束性的注意力损失机制，鼓励模型从描述候选池中选择最匹配的 ASR 描述。实验表现出 DeCEMBERT 在三个下游任务中都有不错的性能表现。

视频-文本预训练模型汇总见表 2。

4 下游任务

多样化的任务需要视觉和语言的融合知识。在本节中，我们将介绍此类任务的基本细节和目标，并将其分为四类：分类、检索、生成和其他任务。

4.1 分类任务

表 3 分类型下游任务模型性能表(节选)

指标 / 模型	VQA		NLVR^2		VCR			GQA	
	test-std	test-dev	test-P	dev	Q->A	QA->R	Q->AR	test-dev	test-std
ViLBERT[22]	70.92	70.55	-	-	73.30	74.60	54.80	-	-
B2T2[24]	-	-	-	-	74.00	77.10	57.10	-	-
VisualBERT[26]	71.00	70.80	67.00	67.40	71.60	73.20	52.40	-	-
Unicoder-VL[27]	-	-	-	-	73.40	74.40	54.90	-	-
VL_BERT-Base[28]	-	71.16	-	-	73.80	74.40	55.20	-	-
VL_BERT-Large[28]	72.22	71.79	-	-	75.50	77.90	58.90	-	-
UNITER-Base[59]	72.91	72.70	77.85	77.18	75.00	77.20	58.20	-	-
UNITER-Large[59]	73.82	74.02	79.98	79.12	77.30	80.80	62.80	-	-
12-in-1[34]	-	73.15	78.87		-	-	-	60.65	-
Pixel-BERT[41]	74.55	74.45	77.20	76.50	-	-	-	-	-
OSCAR-Base[37]	73.44	73.16	78.36	78.07	-	-	-	61.19	61.23
OSCAR-Large[37]	73.82	73.61	80.37	79.12	-	-	-	61.58	61.62
ERNIE-ViL[31]	-	-	-	-	78.98	83.70	66.44	-	-
UNIMO-Base[45]	74.02	73.79	-	-	-	-	-	-	-
UNIMO-Large[45]	75.06	75.27	-	-	-	-	-	-	-
VinVL-Base[42]	76.12	75.95	83.08	82.05	-	-	-	65.05	64.65
VinVL-Large[42]	76.60	76.52	83.98	82.67	-	-	-	-	-
ViLT-B/32[56]	-	71.26	76.13	75.70	-	-	-	-	-
VL-T5[60]	70.30	-	73.60	74.60	75.30	77.80	58.90	-	60.80
SOHO[43]	73.47	73.25	77.32	76.37	-	-	-	-	-
ALBEF[46]	76.04	75.84	83.14	83.14	-	-	-	-	-
Clip-ViL[49]	76.70	76.48	-	-	-	-	-	61.42	62.93
SimVLM[53]	80.34	80.03	85.15	84.53	-	-	-	-	-
TARA[55]	72.93	72.62	-	-	-	-	-	-	-
VLMo-Base[48]	76.64	76.89	82.77	83.34	-	-	-	-	-
VLMo-Large[48]	79.94	79.98	86.86	85.64	-	-	-	-	-
BLIP[36]	78.32	78.25	82.24	82.15	-	-	-	-	-
TCL[47]	74.92	74.90	83.08	82.05	-	-	-	-	-
UVLP[57]	-	72.50	75.90	-	-	-	-	-	-

视觉问答(Visual Question Answering, VQA)。给予视觉输入(图像或视频)，VQA 代表了正确提供一个问题的答案的任务。它通常被认为是一项分类任务，因为模型会从一个选择池中预测出最合适的答案。视觉推理和组合式问答(Visual Reasoning and Compositional Question Answering, GQA)。GQA 是 VQA 的升级版，旨在推进自然场景的视觉推理研究[72]。其数据集中的图像、问题和答案具有匹配的语义表示。这种结构化表示的好处是答案的分布可以更加均匀，我们可以从更多的维度分析模型的性能。自然语言视觉推理

(Natural Language for Visual Reasoning, NLVR)。NLVR 任务的输入是两张图像和一个文本描述，输出是图像和文本描述之间的对应关系是否一致(即真、伪两个标签)。视觉蕴涵(Visual Entailment, VE)。在视觉蕴含任务中，图像作为前提，文本作为假设，目的是判断前提是否能推理出假设，即预测视觉信息是否在语义上包含了文本信息。视觉常识推理(Visual Commonsense Reasoning, VCR)。VCR 类似于 VQA，但相比于 VQA，模型需要在选择出一个正确回答之后，还需要提供一个证明其答案的理由。看图识物(Grounding Referring Expressions, GRE)。GRE 的任务是给定一个文本参考，对一个图像区域进行定位。该模型可以为每个区域输出一个分数，其中具有最高分数的区域被定位用作预测区域。常见视觉语言预训练模型对应分类型下游任务如表 3 所示，包括视觉问答(VQA)，自然语言视觉推理(NLVR)，视觉常识推理(VCR)和视觉推理和组合式问答(GQA)，由于视觉语言预训练任务所包含的下游任务繁多，表 3 中仅节选出最为常见的下游任务进行性能的统计与比较。

表 4 检索型下游任务模型性能表

模型 \ 指标	Visual Retrieval						zero-shot Visual Retrieval					
	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10
Unicoder-VL(COCO)[27]	84.30	97.30	99.30	69.70	93.50	97.20	54.40	82.80	90.60	43.40	76.00	87.00
Unicoder-VL(Flickr30k)[27]	86.20	96.30	99.00	71.50	90.90	94.90	64.30	85.80	92.30	48.40	76.00	85.20
UNITER-Base(COCO)[59]	64.40	87.40	93.08	50.33	78.52	87.16	-	-	-	-	-	-
UNITET-Large(COCO)[59]	65.68	88.56	93.76	52.93	79.93	87.95	-	-	-	-	-	-
UNITER-Base(Flickr)[59]	85.90	97.10	98.80	72.52	92.36	96.08	80.70	95.70	98.00	66.16	88.40	92.94
UNITER-Large(Flickr)[59]	97.30	98.00	99.20	75.56	94.08	96.76	83.60	95.70	97.70	68.74	89.20	93.86
ImageBERT(Flickr30k)[51]	87.00	97.60	99.20	73.10	92.60	96.00	-	-	-	-	-	-
ImageBERT(COCO)[51]	85.40	98.70	99.80	73.60	94.30	97.20	-	-	-	-	-	-
XGPT(Flick30K)[30]	60.40	86.40	91.90	60.40	86.40	91.90	-	-	-	-	-	-
Pixel-BERT(Flickr30K)[41]	87.00	98.90	995.00	71.50	92.10	95.80	-	-	-	-	-	-
Pixel-BERT(COCO)[41]	84.90	97.70	99.30	71.60	93.70	97.40	-	-	-	-	-	-
OSCAR-Base[37]	70.00	91.10	95.50	54.00	80.80	88.50	-	-	-	-	-	-
OSCAR-Large[37]	73.50	92.20	96.00	57.50	82.80	89.80	-	-	-	-	-	-
UNIMO-Base[45]	89.70	98.40	99.10	74.66	93.40	96.08	-	-	-	-	-	-
UNIMO-Large[45]	89.40	98.90	99.80	78.04	94.24	97.12	-	-	-	-	-	-
VinVL-Base[42]	74.60	92.60	96.30	58.10	83.20	90.10	-	-	-	-	-	-
VinVL-Large[42]	75.40	92.90	96.20	58.80	83.50	90.30	-	-	-	-	-	-
ViLT-B/32(COCO)[56]	61.50	86.30	92.70	42.70	72.90	83.10	56.50	82.60	89.60	40.40	70.00	81.10
ViLT-B/32(Flickr30K)[56]	83.50	96.70	98.60	64.40	88.70	93.80	73.20	93.60	96.50	55.00	82.50	89.80
ALIGN(Flickr30k)[52]	95.30	99.80	100.00	84.90	97.40	98.60	88.60	98.70	99.70	75.70	93.80	96.80
ALIGN(COCO)[52]	77.00	93.50	96.90	59.90	83.30	89.80	58.60	83.00	89.70	45.60	69.80	78.60
SOHO(COCO)[43]	85.10	97.40	99.40	73.50	94.50	97.50	-	-	-	-	-	-
SOHO(Flickr30k)[43]	86.50	98.10	99.30	72.50	92.70	96.10	-	-	-	-	-	-
ALBEF(Flickr30k)[46]	95.90	99.80	100.00	95.60	97.50	98.90	94.10	99.50	99.70	82.80	96.30	98.10
CLIP(COCO)[19]	-	-	-	-	-	-	58.40	88.10	81.50	37.80	72.20	62.40
CLIP(Flickr30k)[48]	-	-	-	-	-	-	88.00	99.40	98.70	68.70	95.20	90.60
VLMo-Base(COCO)[48]	74.80	93.10	96.90	57.20	82.60	89.80	-	-	-	-	-	-
VLMo-Base(Flick30K)[48]	92.30	99.40	99.90	79.30	95.70	97.80	-	-	-	-	-	-
VLMo-Large(COCO)[48]	78.20	94.40	97.40	60.60	84.40	91.00	-	-	-	-	-	-
VLMo-Large(Flick30K)[48]	95.30	99.90	100.00	84.50	97.30	98.60	-	-	-	-	-	-
BLIP(Flickr30k)[36]	97.40	99.80	99.90	87.60	97.70	99.00	96.70	100.0	100.00	86.70	97.30	98.70
TCL(COCO)[47]	75.60	92.80	96.70	59.00	83.20	89.90	71.40	90.80	95.40	53.50	79.00	87.10
TCL(Flickr30k)[47]	94.90	99.50	99.80	84.00	93.70	98.50	93.00	99.10	99.60	79.60	95.10	97.40

表 3 中数据集 NLVR^2 保留了 NLVR 的语言多样性，同时也在 NLVR 的基础上采用了视觉上更为复杂的图像。在 VCR 任务中， Q->A 表示模型需要根据给出的视觉问题选择正确的答案，QA->R 表示模型需要根

据视觉问题和回答选择得出该答案的理由，Q->AR 则表示模型在给定的视觉问题之后，要先选择正确的答案，随后还需要对作答的理由进行选择。

4.2 检索任务

视觉-语言检索(Vision-Language Retrieval, VLR)。VLR 涉及对视觉(图像或视频)和语言的理解，以及适当的匹配策略。它包括两个子任务：从视觉到文本和从文本到视觉的检索，其中视觉到文本检索是根据视觉从更大的描述库中获取最重要的相关文本描述，反之亦然。常见视觉语言预训练模型对应检索型下游任务如表 4 所示，包括视觉-语言检索和零样本(zero-shot)的视觉-语言检索。其中，TR 表示从视觉到文本的检索，IR 表示从文本到视觉的检索。R@K(K=1,3,5)表示出现在排名前 K 个结果中与真值匹配的百分比。

4.3 生成任务

视觉描述(Visual Captioning, VC)。VC 旨在为给定的视觉(图像或视频)输入生成语义和句法上合适的文本描述。大规模新物体描述(Novel Object Captioning at Scale, NoCaps)。NoCaps[73]扩展了 VC 任务，以测试模型描述来自 Open Images 数据集的新物体的能力，这些物体都未曾在训练语料库中出现过。视觉对话(Visual Dialogue, VD)。VD 的任务形式是给定一个图像(或视频)、一个对话历史记录和一个用语言描述的问题，并让模型为问题生成一个答案。常见视觉语言预训练模型对应生成型下游任务如表 5 所示，包括视觉描述和大规模新物体描述。其中，CIDEr，BLEU-4，METEOR，SPICE 为四个评价生成语句的指标。**表 5** 生成型下游任务模型性能表

模型	Visual Caption				Nocaps	
	CIDEr	BLEU-4	METEOR	SPICE	CIDEr	SPICE
VLP(COCO)[60]	116.9	36.5	28.4	21.2	-	-
VLP(Flick30K)[60]	67.4	30.1	23	17	-	-
XGPT(Flick30K)[30]	70.9	31.8	23.6	17.6	-	-
XGPT(COCO)[30]	120.1	37.2	28.6	21.8	-	-
OSCAR-Base[37]	137.6	40.5	29.7	22.8	78.8	11.7
OSCAR-Large[37]	140	41.7	30.6	24.5	80.9	11.3
UNIMO-Base[45]	124.4	38.8	-	-	-	-
UNIMO-Large[74]	127.7	39.6	-	-	-	-
VinVL-Base[42]	140.6	40.9	30.9	25.1	92.46	13.07
VinVL-Large[42]	140.9	41	31.1	25.2	-	-
VL-T5(180K image)[60]	116.5	34.5	28.7	21.9	-	-
SimVLM[53]	143.3	40.6	33.7	25.4	-	-
BLIP(Flickr30k)[36]	136.7	40.4	-	-	113.2	14.8
OFA[58]	154.9	44.9	32.5	26.6		

4.4 其他任务

多模态情感分析(Multi-modal Sentiment Analysis, MSA)旨在通过利用多模态信号(如视觉、语言等)来检测其中的情感。多模态机器翻译(Multi-modal Machine Translation, MMT)。多模态机器翻译是一项包含翻译和文本生成的双重任务，将文本从一种语言翻译成另一种语言，并加入来自其他模态的额外信息，即图像。视觉语言导航任务(Vision-Language Navigation, VLN)是让智能体跟着自然语言指令进行导航，这个任务需要同时理解自然语言指令与视角中可以看见的图像信息，然后在环境中对自身所处状态做出对应的动作，最终达到目标位置。光学字符识别(Optical Character Recognition, OCR)。OCR 一般是指检测和识别图像中的文本信息，它包括两个步骤：文字检测(类似于回归任务)和文字识别(类似于分类任务)。

此外，还有一些与视频相关的下游任务，用于评估视频-文本预训练模型，包括动作分类(AC)、动作分割(AS)和动作步骤定位(ASL)。

5 数据集

数据集是深度学习的基础，任何研究都离不开数据，任何优秀的工作都得益于优秀的数据集。本节将从图像-文本和视频-文本两个部分来分别介绍其领域常用的数据集。

5.1 图像-文本数据集

本小节将基于描述分为有描述数据集和无描述数据集。由于大多数视觉语言预训练工作大多是使用带有描述数据集上，但不乏部分采用无描述数据集，本节将以有描述数据集为主来介绍。

5.1.1 有描述数据集

SBU Captions (SBU)[75]包含 100 万个图像-标题对。SBU Captions 数据集的图像文本对的数量约为 0.8M。

Flickr30k 数据集[76]包含从 Flickr 收集的 31,000 张图像，以及由人类注释者提供的 5 个参考句子。

MS COCO(Microsoft Common Objects in Context)数据集[77]是一个大规模的物体检测、分割、关键点检测和描述数据集。该数据集由 164K 图像组成，分为训练集(83k)，验证集(41k)和测试集(41k)。

Flickr30K Entities 数据集[78]是 Flickr30K 数据集的一个扩展。它用 244k 核心参考链增加了原来的 158k 描述，将同一图像的不同描述中提到的相同实体联系起来，并将它们与 276k 人工标注的边界框联系起来。

Visual Genome[79]包含了多选题环境下的视觉答题数据。它包括来自 MSCOCO 的 101,174 张图像，有 170 万个 QA 对，平均每张图像有 17 个问题。

VQA[80]是一个包含关于图像的开放式问题的数据集。这些问题需要对视觉、语言和常识性知识的理解来回答。

Matterport3D 数据集[81]是一个大型室内场景数据集，它包含了 90 个真实建筑场景中的 10,800 个全景视图。

Fashion-Gen 数据集[82]由 293,008 张高清晰度的时尚图像和由专业造型师提供的物品描述组成。

CC3M[83]数据集有 300 多万张图像，与自然语言的标题相配。

GQA 数据集[72]是用于视觉问答的数据集。该数据集中包括了有关各种日常图像的近 2000 万条问题。每个图像都与一组场景图（scene graph）对应。每个问题都与其语义的结构化表示相关联在一起，并且约束应答者必须采用特定的推理步骤来回答它。

CC12M[84]是一个拥有 1200 万个图像-文本对的数据集，专门用于视觉和语言预训练。

相关数据集及其数据见表 6。

表 6 图像文本数据集

数据	图像	图像-文本对	年份
有描述数据集			
SBU Captions[75]	875K	875K	2011
Flickr30k[76]	29K	145K	2014
MS COCO[77]	113K	567K	2014
Flickr30k Entities[78]	31k	158k	2017
Visual Genome [79]	108K	5.4M	2017
VQA[80]	83K	444K	2017
Matterport3D[81]	104K	104K	2017
Fashion-Gen[82]	293K	293K	2018
CC3M[83]	3M	3M	2018
GQA[72]	110K	22M	2019
CC12M[84]	12M	12M	2021
无描述数据集			
CIFAR-100[85]	60k	-	2009
ImageNet[14]	14M	-	2009

5.1.2 无描述数据集

CIFAR-100 数据集[85]是 Tiny Images 数据集[86]的一个子集，由 60000 张 32x32 彩色图像组成。每个类别有 500 张训练图像和 100 张测试图像。

ImageNet[14]包含 14,197,122 张根据 WordNet[87]层次结构标注的图像。自 2010 年以来，该数据集被用于 ImageNet 大规模视觉识别挑战赛(ILSVRC)。

相关数据集及其数据见表 6。

5.2 视频-文本数据集

本小节将基于描述分为标签数据集、描述数据集和其他数据集来介绍。

5.2.1 标签数据集

HMDB51 数据集[88]是来自各种来源(包括电影和网络视频)的视频集合。该数据集由 6849 个视频片段组成，其中包含 51 个动作类别，每个类别至少包含 101 个片段。

UCF101 数据集[89]是 UCF50[74]的扩展，由 13320 个视频剪辑组成，分为 101 个类别。这 101 个类别可以分为 5 种类型(身体运动，人与人互动，人与物互动，演奏乐器和运动)。

MPII Cooking[90]，Kinetics400[91]，AVA[92]等其他相关数据集及其数据见表 7。

表 7 视频-文本数据集

数据集	视频	片段	注释	时长	来源	年份
基于标签的数据集						
HMDB51[88]	3.3k	6.8k	labels	24h	Web/Other Dataset	2011
UCF101[89]	2.5k	13.3k	labels	27h	YouTube	2012
MPII Cooking[90]	44	5.6k	labels	8h	Kitchen	2012
Kinetics400[91]	306k	306k	labels	817h	YouTube	2017
AVA[92]	430	230k	labels	717h	YouTube	2018
基于标题的数据集						
Howto100M[93]	1.22M	136M	136M captions	134,472h	YouTube	2019
Auto-captions on GIF[94]	163k	163k	164k	-	GIF Web	2020
ActivityNet[95]	20k	100k	100k captions	849h	YouTube	2015
Charades[96]	10k	18k	16k captions	82h	Home	2016
TGIF[97]	102k	102k	126k captions	103h	Tumblr GIFs	2016
YouCook2[98]	2k	14k	14k captions	176h	YouTube	2016
MSR-VTT[99]	7.2k	10k	200k captions	40h	YouTube	2016
DiDemo[100]	10k	27k	41k captions	87h	Flicker	2017
LSMDC[101]	200	128k	128k captions	150h	Movies	2017
How2[102]	13k	185k	185k captions	298h	YouTube	2018
TVR[103]	21.8k	21.8k	109k captions	460h	TV shows	2020
TVC[103]	21.8k	21.8k	262k captions	460h	TV shows	2020
VIOLIN[104]	6.7k	16k	95k captions	582h	Movie & TV shows	2020
其他数据集						
TVQA[105]	925	21.8K	152.5K QAs	460h	TV shows	2018
COIN[106]	12k	46k	segment labels	476h	YouTube	2019
CrossTask[107]	4.7k	20k	20k steps	376h	YouTube	2019

5.2.2 描述数据集

Activitynet[95]包含了 20 千个 YouTube 上未经修剪的视频，有 10 万个人工标注的描述语句，描述了相应视频片段的内容，由开始和结束的时间戳来注释。

Howto100M[93]是迄今为止最大的英语视频数据集，它包含了 1.36 亿个视频片段，并用 YouTube 上相配对的字幕进行标注(主要是教学视频)。

Auto-captions on GIF[94]一般用于基于 GIF 视频的视频理解类任务。该数据集中所有的视频-句子对都是通过自动提取和过滤数十亿网页上的视频字幕注释而创建的。

YouCook2[98]是目前最大的面向任务的教学视频数据集。它包含了来自 89 个烹饪食谱的 2000 个未经修剪的长视频。每个视频的程序步骤都有时间戳注释和描述语句。

Charades 数据集[96]由 9848 个平均长度为 30 秒的日常室内活动视频组成，涉及 15 种室内场景中的 46

个物体类别的互动，包含了 157 个动作类别的 66,500 个时间注释，46 个物体类别的 41,104 个标签，以及 27,847 个视频的文本描述。

DiDemo(Distinct Describable Moments)数据集[100]是用于对视频进行自然语言时间定位的最大和最多样化的数据集之一。数据集中的视频被分为 5 秒钟的片段，以减少注释的复杂性。该数据集分为训练集、验证集和测试集，分别包含 8,395、1,065 和 1,004 个视频。该数据集共包含 26,892 个时刻，一个时刻可能与多个注释者的描述相关。

LSMDC 数据集[101]中包含了从 202 部电影中提取的 118,081 个短视频片段，每一个片段都有一段描述。验证集中包含 7408 个电影片段，测试集包含 1000 个与训练集和评价集不相干的电影片段。

VIOLIN 数据集[104]中由 15,887 个视频片段的 95322 个视频假设对组成。用于给定一段匹配的描述和视频，来预测是否匹配的任务。

TGIF[97]、**MSR-VTT**[99]、**How2**[102]、**TVR**[103]、**TVC**[103]等相关数据集及其数据见表 7。

5.2.3 其他数据集

TVQA[105]是一个基于 6 个流行电视节目的视频问答数据集，共有 460 小时的视频和 152.5K 的问答对。每个问题提供 5 个候选答案，其中有一个正确答案，正确答案标有开始和结束时间戳，以便进一步推理。

COIN[106]是为综合教学视频分析而设计的，它以 3 个层次的结构来组织结构，从领域、任务到步骤。该数据集包含 11,827 个 12 个领域、180 个任务和 778 个步骤的教学视频。对于每一项任务，都提供了一个带有简短描述的有序步骤列表。

其他相关数据集及其数据见表 7。

6 总结和展望

在本文中，首先我们介绍了视觉语言预训练模型的相关知识，包括 Transformer 框架、预训练范式和视觉语言预训练模型常见网络结构；其次我们介绍了三类模型预训练任务，通过这些任务，网络模型可以在无标注的情况下进行跨模态的语义对齐；然后我们从图像-文本预训练和视频-文本预训练两个方面分别介绍了最新的工作进展，并介绍了预训练模型的下游任务；最后我们介绍了广泛使用的图像文本和视频文本的多模态数据集，并比较和分析了常用预训练模型在不同任务下不同数据集上的性能。视觉语言预训练在飞速发展的同时也取得了许多非常不错的成果，未来视觉语言预训练模型的发展方向可以借鉴如下：

1)计算资源。目前视觉语言预训练工作仍然需要极大的算力资源做支撑。2019 年以来，视觉语言预训练工作大部分都是产自于工业界，需要使用数十上百张显卡进行训练，导致部分研究人员没有足够的计算资源对其展开研究，而且难以对这些大规模工作进行验证。如何在资源受限的情况下进行视觉语言预训练研究，是一个很有研究价值的问题。

2)Prompt。预训练-提示范式在 NLP 领域引起了一波研究热潮，我们在 1.2.2 已经对其进行了介绍。提示相对于微调的优势在于：1)计算代价低。2)节省空间。目前已有少数工作对其进行展开了研究，诸如 CLIP，CPT 等，并且取得了不错的效果。预训练-提示范式目前还在探索阶段，未来将会有更多更有意义的工作出现。

3)多模态融合。之前大多数的多模态预训练工作都是强调视觉和语言这两个模态进行建模，但是忽略了其他模态(比如音频等)信息。其他模态信息往往也对跨模态学习有着重要的意义，因此研究更多模态信息建模的工作是具有研究价值和挑战性的。

References:

[1] 杜鹏飞, 李小勇, 高雅丽. 多模态视觉语言表征学习研究综述[J]. 软件学报, 2021, 32(2): 22.

[2] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł, Polosukhin I. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[3] Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[4] Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training[J], 2018.

[5] Song K, Tan X, Qin T, Lu J, Liu T-Y. Mass: Masked sequence to sequence pre-training for language generation[J]. arXiv preprint arXiv:1905.02450, 2019.

[6] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[7] Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[8] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv preprint arXiv:1412.3555, 2014.

[9] Gehring J, Auli M, Grangier D, Yarats D, Dauphin Y N. Convolutional sequence to sequence learning[C]. International Conference on Machine Learning, 2017: 1243-1252.

[10] Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, Fidler S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books[C]. Proceedings of the IEEE international conference on computer vision, 2015: 19-27.

[11] Warstadt A, Singh A, Bowman S R. Neural network acceptability judgments[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 625-641.

[12] Dolan B, Brockett C. Automatically constructing a corpus of sentential paraphrases[C]. Third International Workshop on Paraphrasing (IWP2005), 2005.

[13] Sun C, Shrivastava A, Singh S, Gupta A. Revisiting unreasonable effectiveness of data in deep learning era[C]. Proceedings of the IEEE international conference on computer vision, 2017: 843-852.

[14] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database[C]. 2009 IEEE conference on computer vision and pattern recognition, 2009: 248-255.

[15] Brown T, Mann B, Ryder N, Subbiah M, Kaplan J D, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[16] Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation[J]. arXiv preprint arXiv:2101.00190, 2021.

[17] Liu X, Zheng Y, Du Z, Ding M, Qian Y, Yang Z, Tang J. GPT understands, too[J]. arXiv preprint arXiv:2103.10385, 2021.

[18] Liu X, Ji K, Fu Y, Du Z, Yang Z, Tang J. P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks[J]. arXiv preprint arXiv:2110.07602, 2021.

[19] Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J. Learning transferable visual models from natural language supervision[C]. International Conference on Machine Learning, 2021: 8748-8763.

[20] Yao Y, Zhang A, Zhang Z, Liu Z, Chua T-S, Sun M. Cpt: Colorful prompt tuning for pre-trained vision-language models[J]. arXiv preprint arXiv:2109.11797, 2021.

[21] 包希港, 周春来, 肖克晶, 覃飙. 视觉问答研究综述[J]. 软件学报, 2021, 32(8): 23.

[22] Lu J, Batra D, Parikh D, Lee S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[J]. Advances in neural information processing systems, 2019, 32.

[23] Girshick R. Fast r-cnn[C]. Proceedings of the IEEE international conference on computer vision, 2015: 1440-1448.

[24] Alberti C, Ling J, Collins M, Reitter D. Fusion of detected objects in text for visual question answering[J]. arXiv preprint arXiv:1908.05054, 2019.

[25] Tan H, Bansal M. Lxmert: Learning cross-modality encoder representations from transformers[J]. arXiv preprint arXiv:1908.07490, 2019.

[26] Li L H, Yatskar M, Yin D, Hsieh C-J, Chang K-W. Visualbert: A simple and performant baseline for vision and language[J]. arXiv preprint arXiv:1908.03557, 2019.

[27] Li G, Duan N, Fang Y, Gong M, Jiang D. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 11336-11344.

[28] Su W, Zhu X, Cao Y, Li B, Lu L, Wei F, Dai J. Vl-bert: Pre-training of generic visual-linguistic representations[J]. arXiv preprint arXiv:1908.08530, 2019.

[29] Zhou L, Palangi H, Zhang L, Hu H, Corso J, Gao J. Unified vision-language pre-training for image captioning and vqa[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 13041-13049.

[30] Xia Q, Huang H, Duan N, Zhang D, Ji L, Sui Z, Cui E, Bharti T, Zhou M. Xgpt: Cross-modal generative pre-training for image captioning[C]. CCF International Conference on Natural Language Processing and Chinese Computing, 2021: 786-797.

[31] Yu F, Tang J, Yin W, Sun Y, Tian H, Wu H, Wang H. Ernie-vil: Knowledge enhanced vision-language representations through scene graph[J]. arXiv preprint arXiv:2006.16934, 2020.

[32] Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M, Sutskever I. Zero-shot text-to-image generation[C]. International Conference on Machine Learning, 2021: 8821-8831.

[33] Van Den Oord A, Vinyals O. Neural discrete representation learning[J]. Advances in neural information processing systems, 2017, 30.

[34] Lu J, Goswami V, Rohrbach M, Parikh D, Lee S. 12-in-1: Multi-task vision and language representation learning[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10437-10446.

[35] Hu R, Singh A. Unit: Multimodal multitask learning with a unified transformer[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 1439-1449.

[36] Li J, Li D, Xiong C, Hoi S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation[J]. arXiv preprint arXiv:2201.12086, 2022.

[37] Li X, Yin X, Li C, Zhang P, Hu X, Zhang L, Wang L, Hu H, Dong L, Wei F. Oscar: Object-semantics aligned pre-training for vision-language tasks[C]. European Conference on Computer Vision, 2020: 121-137.

[38] Xue H, Huang Y, Liu B, Peng H, Fu J, Li H, Luo J. Probing Inter-modality: Visual Parsing with Self-Attention for Vision-and-Language Pre-training[J]. Advances in Neural Information Processing Systems, 2021, 34.

[39] Yao L, Huang R, Hou L, Lu G, Niu M, Xu H, Liang X, Li Z, Jiang X, Xu C. FILIP: Fine-grained Interactive Language-Image Pre-Training[J]. arXiv preprint arXiv:2111.07783, 2021.

[40] Duan J, Chen L, Tran S, Yang J, Xu Y, Zeng B, Chilimbi T. Multi-modal alignment using representation codebook[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 15651-15660.

[41] Huang Z, Zeng Z, Liu B, Fu D, Fu J. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers[J]. arXiv preprint arXiv:2004.00849, 2020.

[42] Zhang P, Li X, Hu X, Yang J, Zhang L, Wang L, Choi Y, Gao J. Vinvl: Revisiting visual representations in vision-language models[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 5579-5588.

[43] Huang Z, Zeng Z, Huang Y, Liu B, Fu D, Fu J. Seeing out of the box: End-to-end pre-training for vision-language representation learning[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 12976-12985.

[44] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 770-778.

[45] Li W, Gao C, Niu G, Xiao X, Liu H, Liu J, Wu H, Wang H. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning[J]. arXiv preprint arXiv:2012.15409, 2020.

[46] Li J, Selvaraju R, Gotmare A, Joty S, Xiong C, Hoi S C H. Align before fuse: Vision and language representation learning with momentum distillation[J]. Advances in Neural Information Processing Systems, 2021, 34.

[47] Yang J, Duan J, Tran S, Xu Y, Chanda S, Chen L, Zeng B, Chilimbi T, Huang J. Vision-Language Pre-Training with Triple Contrastive Learning[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 15671-15680.

[48] Wang W, Bao H, Dong L, Wei F. VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts[J]. arXiv preprint arXiv:2111.02358, 2021.

[49] Shen S, Li L H, Tan H, Bansal M, Rohrbach A, Chang K-W, Yao Z, Keutzer K. How Much Can CLIP Benefit Vision-and-Language Tasks?[J]. arXiv preprint arXiv:2107.06383, 2021.

[50] Dou Z-Y, Xu Y, Gan Z, Wang J, Wang S, Wang L, Zhu C, Liu Z, Zeng M. An Empirical Study of Training End-to-End Vision-and-Language Transformers[J]. arXiv preprint arXiv:2111.02387, 2021.

[51] Qi D, Su L, Song J, Cui E, Bharti T, Sacheti A. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data[J]. arXiv preprint arXiv:2001.07966, 2020.

[52] Jia C, Yang Y, Xia Y, Chen Y-T, Parekh Z, Pham H, Le Q, Sung Y-H, Li Z, Duerig T. Scaling up visual and vision-language representation learning with noisy text supervision[C]. International Conference on Machine Learning, 2021: 4904-4916.

[53] Wang Z, Yu J, Yu A W, Dai Z, Tsvetkov Y, Cao Y. Simvlm: Simple visual language model pretraining with weak supervision[J]. arXiv preprint arXiv:2108.10904, 2021.

[54] Tsimpoukelli M, Menick J, Cabi S, Eslami S, Vinyals O, Hill F. Multimodal few-shot learning with frozen language models[J]. Advances in Neural Information Processing Systems, 2021, 34.

[55] Zhou Y, Ren T, Zhu C, Sun X, Liu J, Ding X, Xu M, Ji R. TRAR: Routing the Attention Spans in Transformer for Visual Question Answering[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 2074-2084.

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

[57] Zhou M, Yu L, Singh A, Wang M, Yu Z, Zhang N. Unsupervised Vision-and-Language Pre-training via Retrieval-based Multi-Granular Alignment[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 16485-16494.

[58] Wang P, Yang A, Men R, Lin J, Bai S, Li Z, Ma J, Zhou C, Zhou J, Yang H. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework[J]. arXiv preprint arXiv:2202.03052, 2022.

[59] Chen Y-C, Li L, Yu L, El Kholy A, Ahmed F, Gan Z, Cheng Y, Liu J. Uniter: Learning universal image-text representations[J], 2019.

[60] Cho J, Lei J, Tan H, Bansal M. Unifying vision-and-language tasks via text generation[C]. International Conference on Machine Learning, 2021: 1931-1942.

[61] Sun C, Myers A, Vondrick C, Murphy K, Schmid C. Videobert: A joint model for video and language representation learning[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 7464-7473.

[62] Sun C, Baradel F, Murphy K, Schmid C. Learning video representations using contrastive bidirectional transformer[J]. arXiv preprint arXiv:1906.05743, 2019.

[63] Zhu L, Yang Y. Actbert: Learning global-local video-text representations[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020: 8746-8755.

[64] Li L, Chen Y-C, Cheng Y, Gan Z, Yu L, Liu J. Hero: Hierarchical encoder for video+ language omni-representation pre-training[J]. arXiv preprint arXiv:2005.00200, 2020.

[65] Akbari H, Yuan L, Qian R, Chuang W-H, Chang S-F, Cui Y, Gong B. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text[J]. Advances in Neural Information Processing Systems, 2021, 34.

[66] Tang Z, Lei J, Bansal M. Decembert: Learning from noisy instructional videos via dense captions and entropy minimization[C]. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021: 2415-2426.

[67] Lei J, Li L, Zhou L, Gan Z, Berg T L, Bansal M, Liu J. Less is more: Clipbert for video-and-language learning via sparse sampling[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 7331-7341.

[68] Xu H, Ghosh G, Huang P-Y, Arora P, Aminzadeh M, Feichtenhofer C, Metze F, Zettlemoyer L. VLM: Task-agnostic video-language model pre-training for video understanding[J]. arXiv preprint arXiv:2105.09996, 2021.

[69] Luo H, Ji L, Shi B, Huang H, Duan N, Li T, Li J, Bharti T, Zhou M. Univl: A unified video and language pre-training model for multimodal understanding and generation[J]. arXiv preprint arXiv:2002.06353, 2020.

[70] Jozefowicz R, Vinyals O, Schuster M, Shazeer N, Wu Y. Exploring the limits of language modeling[J]. arXiv preprint arXiv:1602.02410, 2016.

[71] Yang L, Tang K, Yang J, Li L-J. Dense captioning with joint inference and visual context[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 2193-2202.

[72] Hudson D A, Manning C D. Gqa: A new dataset for real-world visual reasoning and compositional question answering[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019: 6700-6709.

[73] Agrawal H, Desai K, Wang Y, Chen X, Jain R, Johnson M, Batra D, Parikh D, Lee S, Anderson P. Nocaps: Novel object captioning at scale[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 8948-8957.

[74] Reddy K K, Shah M. Recognizing 50 human action categories of web videos[J]. Machine vision and applications, 2013, 24(5): 971-981.

[75] Ordonez V, Kulkarni G, Berg T. Im2text: Describing images using 1 million captioned photographs[J]. Advances in neural information processing systems, 2011, 24.

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

[77] Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick C L. Microsoft coco: Common objects in context[C]. European conference on computer vision, 2014: 740-755.

[78] Plummer B A, Wang L, Cervantes C M, Caicedo J C, Hockenmaier J, Lazebnik S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models[C]. Proceedings of the IEEE international conference on computer vision, 2015: 2641-2649.

[79] Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li L-J, Shamma D A. Visual genome: Connecting language and vision using crowdsourced dense image annotations[J]. International journal of computer vision, 2017, 123(1): 32-73.

[80] Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick C L, Parikh D. Vqa: Visual question answering[C]. Proceedings of the IEEE international conference on computer vision, 2015: 2425-2433.

[81] Chang A, Dai A, Funkhouser T, Halber M, Niessner M, Savva M, Song S, Zeng A, Zhang Y. Matterport3d: Learning from rgb-d data in indoor environments[J]. arXiv preprint arXiv:1709.06158, 2017.

[82] Rostamzadeh N, Hosseini S, Boquet T, Stokowiec W, Zhang Y, Jauvin C, Pal C. Fashion-gen: The generative fashion dataset and challenge[J]. arXiv preprint arXiv:1806.08317, 2018.

[83] Sharma P, Ding N, Goodman S, Soricut R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning[C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018: 2556-2565.

[84] Changpinyo S, Sharma P, Ding N, Soricut R. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 3558-3568.

[85] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images[J], 2009.

[86] Torralba A, Fergus R, Freeman W T. 80 million tiny images: A large data set for nonparametric object and scene recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2008, 30(11): 1958-1970.

[87] Miller G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39-41.

[88] Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T. HMDB: a large video database for human motion recognition[C]. 2011 International conference on computer vision, 2011: 2556-2563.

[89] Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. arXiv preprint arXiv:1212.0402, 2012.

[90] Rohrbach M, Rohrbach A, Regneri M, Amin S, Andriluka M, Pinkal M, Schiele B. Recognizing fine-grained and composite activities using hand-centric features and script data[J]. International Journal of Computer Vision, 2016, 119(3): 346-373.

[91] Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P. The kinetics human action video dataset[J]. arXiv preprint arXiv:1705.06950, 2017.

[92] Gu C, Sun C, Ross D A, Vondrick C, Pantofaru C, Li Y, Vijayanarasimhan S, Toderici G, Ricco S, Sukthankar R. Ava: A video dataset of spatio-temporally localized atomic visual actions[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 6047-6056.

[93] Miech A, Zhukov D, Alayrac J-B, Tapaswi M, Laptev I, Sivic J. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019:

2630-2640.

[94] Pan Y, Li Y, Luo J, Xu J, Yao T, Mei T. Auto-captions on GIF: A Large-scale Video-sentence Dataset for Vision-language Pre-training[J]. arXiv preprint arXiv:2007.02375, 2020.

[95] Caba Heilbron F, Escorcia V, Ghanem B, Carlos Niebles J. Activitynet: A large-scale video benchmark for human activity understanding[C]. Proceedings of the ieee conference on computer vision and pattern recognition, 2015: 961-970.

[96] Sigurdsson G A, Varol G, Wang X, Farhadi A, Laptev I, Gupta A. Hollywood in homes: Crowdsourcing data collection for activity understanding[C]. European Conference on Computer Vision, 2016: 510-526.

[97] Li Y, Song Y, Cao L, Tetreault J, Goldberg L, Jaimes A, Luo J. TGIF: A new dataset and benchmark on animated GIF description[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 4641-4650.

[98] Zhou L, Xu C, Corso J J. Towards automatic learning of procedures from web instructional videos[C]. Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[99] Xu J, Mei T, Yao T, Rui Y. Msr-vtt: A large video description dataset for bridging video and language[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 5288-5296.

[100] Anne Hendricks L, Wang O, Shechtman E, Sivic J, Darrell T, Russell B. Localizing moments in video with natural language[C]. Proceedings of the IEEE international conference on computer vision, 2017: 5803-5812.

[101] Rohrbach A, Rohrbach M, Tandon N, Schiele B. A dataset for movie description[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015: 3202-3212.

[102] Sanabria R, Caglayan O, Palaskar S, Elliott D, Barrault L, Specia L, Metze F. How2: a large-scale dataset for multimodal language understanding[J]. arXiv preprint arXiv:1811.00347, 2018.

[103] Lei J, Yu L, Berg T L, Bansal M. Tvr: A large-scale dataset for video-subtitle moment retrieval[C]. European Conference on Computer Vision, 2020: 447-463.

[104] Liu J, Chen W, Cheng Y, Gan Z, Yu L, Yang Y, Liu J. Violin: A large-scale dataset for video-and-language inference[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10900-10910.

[105] Lei J, Yu L, Bansal M, Berg T L. Tvqa: Localized, compositional video question answering[J]. arXiv preprint arXiv:1809.01696, 2018.

[106] Tang Y, Ding D, Rao Y, Zheng Y, Zhang D, Zhao L, Lu J, Zhou J. Coin: A large-scale dataset for comprehensive instructional video analysis[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 1207-1216.

[107] Zhukov D, Alayrac J-B, Cinbis R G, Fouhey D, Laptev I, Sivic J. Cross-task weakly supervised learning from instructional videos[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 3537-3545.