计算机应用
Journal of Computer Applications
ISSN 1001-9081,CN 51-1307/TP

《计算机应用》网络首发论文

题目： 多模态预训练模型综述
作者： 王惠茹，李秀红，李哲，马春明，任泽裕，杨丹
收稿日期： 2022-03-11
网络首发日期： 2022-06-15
引用格式： 王惠茹，李秀红，李哲，马春明，任泽裕，杨丹. 多模态预训练模型综述[J/OL]. 计算机应用. https://kns.cnki.net/kcms/detail/51.1307.TP.20220614.1841.010.html

Journal of Computer Applications ISSN 1001-9081
计算机应用 CODEN JYIIDU http://www.joca.cn

DOI:10.11772/j.issn.1001-9081.2022020296

多模态预训练模型综述

王惠茹[1]，李秀红[1*]，李哲[2]，马春明[1]，任泽裕[1]，杨丹[1]

(1.新疆大学 信息科学与工程学院，乌鲁木齐 830046;
2.香港理工大学 电子及资讯工程学系，香港 999077)

(*通信作者电子邮箱：xjulxh@xju.edu.cn)

摘 要：近年来，预训练模型(PTM)的出现，将人工智能带入了一个新时代。通过利用复杂的预训练目标和大量的模型参数，预训练模型可以有效地获得无标记数据中的丰富知识。在多模态中，预训练模型的发展还处于初期。依据具体模态的不同将目前大多数的多模态预训练模型分为图像-文本预训练模型和视频-文本预训练模型，依据数据融合方式的不同还可将多模态预训练模型分为单流模型和双流模型两类。首先总结了常见的预训练任务和验证实验所使用的下游任务；接着，梳理了目前多模态预训练领域的常见模型，并用表格列出各个模型的下游任务以及对模型的性能和实验数据进行比较；然后，介绍了 M6 模型、跨模态提示调优(CPT)模型、VideoBERT 模型和 Alicemind 模型在具体下游任务中的应用场景；最后，对多模态预训练模型相关工作面临的挑战以及未来可能的研究方向进行了总结。

关键词： 多模态；预训练模型；图像-文本预训练模型；视频-文本预训练模型；神经网络；单流模型；双流模型

中图分类号：TP391.1 **文献标识码：A**

Survey of multimodal pre-training models

WANG Huiru[1], LI Xiuhong[1*], LI Zhe[2],MA Chunming[1], REN Zeyu[1], YANG Dan[1]

(1. College of Information Science and Engineering, Xinjiang University, Urumqi Xinjiang 830046, China;
2. Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China)

Abstract: In recent years, the Pre-Training Model (PTM) has brought artificial intelligence into a new era. By using complex pre-training targets and a large number of model parameters, the pre-training model can effectively obtain acquire rich knowledge without annotation data. The development of the multimodal pre-training model is still in its infancy. According to the different modes, most of the current multimodal pre-training models are divided into the image-text pre-training models and video-text pre-training models. According to the different data fusion methods, the multimodal pre-training model can also be divided into two types: the single-stream model and the cross-stream model. First, we summarize common pre-training tasks and downstream tasks used in validation experiments. Secondly, we sorted out the common models in the area of multimodal pre-training, listed the downstream tasks of each model in tables and compared the performance of the models and experimental data. Then, we introduced the application scenarios of the M6 model, Cross-modal Prompt Tuning (CPT) model, VideoBERT model, and Alicemind model in specific downstream tasks. Finally, we introduced the challenges and future work related to the multimodal pre-training model and summarized the future research directions.

Key words: multimodal; Pre-Training Model (PTM); image-text Pre-Training Model; video-text Pre-Training Model; neural network; single-stream model; cross-stream model

收稿日期:2022-03-11； 修回日期: 2022-06-06； 录用日期: 2022-06-07。
基金项目：国家语委重点研发项目（ZDI135-96）。
作者简介: 王惠茹(1996—)，女，硕士研究生，新疆伊犁人，硕士研究生，主要研究方向：自然语言处理、图像处理；李秀红(1977—)，女，山东威海人，副教授，博士，主要研究方向：自然语言处理、图像处理；李哲(1992—)，男，山东泰安人，博士研究生，主要研究方向：多模态说话人识别、鲁棒性机器学习；马春明(1997—)，男，四川绵阳人，硕士研究生，主要研究方向：自然语言处理、事件抽取；任泽裕(1998—)，男，山西长治人，硕士研究生，主要研究方向：语音识别、图像处理；杨丹(1996—)，女，四川南充人，硕士研究生，主要研究方向：自然语言处理、图像处理。

0 引言

随着自监督的不断发展，预训练技术在学习视觉和语言表征信息方面发挥着重要的作用。预训练的作用是从大量的训练数据中提取出尽可能多的共性特征，使模型对特定任务的学习负担变轻。在大规模的未标记数据上对模型进行预训练，并使用特定于任务的标记数据对下游任务进行微调[1]。随着深度学习的发展，模型的参数越来越多。要完全训练模型参数、防止产生过拟合现象，就需要使用更大的数据集。针对这些问题，预训练模型慢慢出现。预训练简单来说就是指预先训练的一个模型或者指预先训练模型的过程。把一个已经训练好的图像分类[2]的模型的参数，应用到另一个类似任务上作为初始参数，类似这样训练模型的过程称作预训练。

多模态数据是指对于同一个描述对象，通过不同领域或视角获取到的数据，并且把描述这些数据的每一个领域或视角叫作一个模态[3]。多模态进行预训练期望学习到两种及多种模态间的关联关系，较单一模态进行预训练，可综合获取多种信息，使得预训练模型具有更好的泛化性。虽然视觉或语言等单一模态的理解在视觉或语言任务中不可或缺，但各个模态之间的相互关系也同样重要。若多模态模型无法将相关的视觉信息和语言单词进行联合表征，则经过预训练的单一模态的特征在许多任务中无法实现“微调即可用”。因此，在大规模无标注的多模态数据上学习到有利于下游任务关联、理解和推理的特征是非常重要的研究任务[4]。

多模态预训练模型虽然有所发展，但仍然面临着很多挑战:1)多模态数据的数据量规模远比自然语言语料库小很多。2)计算机视觉中预训练大多仅用于特征提取，将计算机视觉模型和自然语言处理模型共同训练的情况较少。目前没有好的联合训练算法，而且训练代价非常大。3)计算机视觉中的对象识别，目前的类别仅有 1000 类，对真实场景覆盖率低且识别精度不高，使得预训练的输入本身存在误差。4)多模态预训练模型目前大多采用 Transformer 机制，代价较大。这样是否最合适对图像文字、视频-文字建立关联并不确定。5)图片和视频的预训练模型不一样。视频播放是有时序的，所以对于视频的分割需要按照固定的时长进行，并且视频预训练的代价比图片和文字的预训练大很多。最近，人们开始对多模态任务进行自我监督学习，方法是对大型图像/视频和文本对进行预训练，然后对下游任务进行微调。例如，VideoBERT[5]应用 BERT(Bidirectional Encoder Representation from Transformers)从视频文本对中学习视频帧特征和语言标记的联合分布。ViLBERT(Vision and Language BERT)[4]和 LXMERT(Learning Cross-Modality Encoder Representations from Transformers)[6]引入了双流体系结构。在另一方面，B2T2(Bounding Boxes in Text Transformer)、VisualBERT[7]、Unicoder-VL(Universal encoder for Vision and Language by cross-modal pre-training)[8]和 VL-BERT(Visual-Linguistic BERT)[9]使用了单流体系结构，将单个 Transformer 应用于图像和文本。视觉-语言预训练（Vision-Language Pre-Training, VLP）将预训练的模型应用于图像字幕和视觉问答(Visual Question Answering, VQA)。并引入多任务学习和对抗训练用于进一步提高性能。

就以上问题撰写多模态预训练模型综述，对以上挑战以及相应解决方案进行总结阐述，以便为后续多模态预训练模型的研究者做简单参考。

1 相关工作

大规模预训练模型增强了对多模态数据的研究兴趣，例如图像-文本预训练或视频-文本预训练。考虑到图像和视频属于视觉，而文本和语音属于语言，因此将大多数多模态预训练模型归类为视觉-语言(Vision-Linguistic, V&L)预训练模型。V&L 任务根据具体模态不同可细分为图像-文本任务、视频-文本任务和视频-音频任务。以下对 V&L 多模态预训练的最近工作进行总结概述。

对于图像-文本预训练模型(Pre-Training Model, PTM)，当前的大多数工作都是基于视觉-语言 BERT 的架构。主要挑战在于统一语义空间中视觉和文本内容的对齐。因此发展出两种模型架构设计：双流和单流。

在双流模型方面，2019 年由 Lu 等提出 ViLBERT 模型[4]，首次将 BERT 结构扩展到多模态双流模型中，使用类似 BERT 的架构学习对图像-文本的联合表示，但由于视觉和语言都有单独的 Transformer 结构，导致参数量显著提高。同 ViLBERT 相似，LXMERT 模型[6]也是将两个 Transformer 应用于图像和文本，并通过第三个 Transformer 进行融合。2021 年，Radford 等提出的 CLIP(Contrastive Language-Image Pre-Training)模型[10]，用 4 亿个来自网络的图文数据对，将文本作为图像标签进行训练，使用两个编码器分别处理文本和图片，在图像-文本检索任务上取得了显著的性能，但在其他视觉-语言任务中表现不佳。针对以上问题，之后由 Li 等提出的 ALBEF(ALign BEfore Fuse)模型[11]，引入中间的图像-文本对比损失，首先将单模态图像表示与文本表示进行对齐，再与多模态编码器进行融合，引导视觉和语言表示学习，在多个下游任务中获得了更快的推理速度。

在单流模型中，Sun 等在 2019 年提出 VideoBERT[5]，作为单流模型，它在结构上使用堆叠的 Transformer 结构，使用聚类技术对视频帧和音频语言进行处理。VisualBERT[7]，与 VideoBERT 相比拥有更简单的架构，可以在无监督条件下建立语音、图像之间的联系，但还未将该模型应用于纯图像

任务中。随即，2020 年由 Li 等提出的 Unicoder-VL[8]作为图像-文本领域的预训练模型，继续采用堆叠的 Transformer 结构，相较于以上三个模型，使用大量的图像-文本对进行训练，可学习常见的跨模态知识并应用于更广泛的下游任务中，但无法从单个的图像模态中提取信息。

“紫东太初”[12]作为全球首个图文音三模态模型，通过在三个基本模态中加入多模态编码器网络，同时具备跨模态理解与跨模态生成能力，在预训练模型领域取得了突破性进展。“紫东太初”在图文跨模态理解与生成上，都能领先目前业界的 SOTA(State-Of-The-Art)模型，高效完成跨模态检测、视觉问答、语义描述等下游任务。

在多模态预训练中，数据资源也尤为重要。最广泛使用的语料库是从网络收集的图像-文本对，包括概念字幕[13]、SBU (Stony Brook University) Captions[14]或为特定任务设计的 V&L 数据集，包括 COCO(Common Objects in Context)[15]、Flickr30K[16]、GQA[17]、VQA[18]和 Visual Genome[19]。UNITER(UNiversal Image-TExt Representation)[20]结合了上述几个数据集，产生了 560 万个用于训练的图像-文本对。与 UNITER 在架构和预训练任务上类似，ImageBERT[21]进一步构建一个包含 1000 万网页图像-文本对的数据集，并将其用作预训练数据集，从而在图像-文本检索任务上获得比 UNITER 更好的性能。除了并行的图像-文本数据，VL-BERT[9]发现，合并额外的纯文本语料库(如 BookCorpus[22])和维基百科(Wikipedia)有助于文本理解，特别是像视觉常识推理这样的长而复杂的句子任务。与只使用容易得到的数据(如图像文本对或文本语料库)的工作不同，Lu 等[23]通过对几乎所有的 V&L 任务进行联合多任务训练来验证专用数据集。

多模态以其表示信息的多样性，受到研究者的青睐。最近，多模态预训练模型也相继问世，引领了模型训练的新方向。目前多模态预训练模型还处于初期阶段，遵循大多数自然语言处理(Natural Language Processing, NLP)预训练模型，从有限的图像-语言数据集、视频-语言数据集中学习联合表示，计算图像或视频片段和文字描述的距离，并实现图像-文字、视频-文字之间的转换。虽然还不成熟，但已经展现出一定的前景。

2 多模态预训练任务

多模态预训练任务较多，这里主要在 UNITER 模型[20]的基础上详细介绍 4 个常见的多模态预训练任务。

2.1 MLM 预训练任务

掩码语言建模预训练任务(Masked Language Modeling, MLM)[24]是指在句子表示中随机掩盖一些字词，然后模型基于其他的文本标记和所有的图像标记来预测这些被掩盖的标记的一种任务。UNITER 模型的输入是图片、文本对，随机的删除一些位置的词语，目标是在这些删除位置让 UNITER 去还原原本的词语，MLM 任务图如图 1 所示。用到的损失函数是负对数似然函数(negative log-likelihood)。

Masked Language Modeling(MLM)

图 1 MLM 任务图

Fig. 1 Masked Language Modeling task map

图像区域和输入词表示如式(1)~(2):

$$\boldsymbol{v}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_K] \tag{1}$$

$$\boldsymbol{w}=[\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_T] \tag{2}$$

最小化负对数似然函数为式(3):

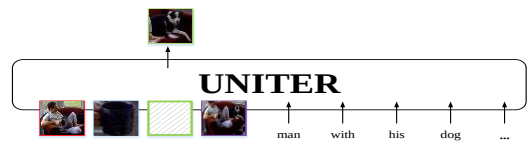
$$L_{\mathrm{MLM}}(\theta)=-\mathrm{E}_{(\boldsymbol{w},\boldsymbol{v})\sim D}\log P_\theta(\boldsymbol{w}_m|\boldsymbol{w}_{\backslash m},\boldsymbol{v}) \tag{3}$$

2.2 MRM 预训练任务

2.2.1 MRC 预训练任务

掩码区域分类预训练任务(Masked Region Classification, MRC)[25]，需要在区域标记中随机掩盖一些字词，然后根据其他的图片标记和所有的文本词语来预测这些被掩盖的字词。具体来说，经过 Faster R-CNN（Faster Region-based Convolutional Neural Network）算法[26]每个区域会得到一个标记，模型需要预测被掩盖字 词的类别，使之和 Faster R-CNN 的标记相同。

掩码区域建模(Masked Region Modeling, MRM)[27]，类似于 MLM，同样可以采样图像区域，并以 15%的概率掩盖它们的视觉特征。在给定剩余区域和所有单词的情况下，训练该模型重建掩码区域。被掩盖区域的视觉特征被零代替。但是与用离散标签表示的文本标记不同，视觉特征是高维连续的。MRM 任务图如图 2 所示。



Masked Region Modeling（MRM）

图 2 MRM 任务图

Fig. 2 Masked Region Modeling task map

图像区域和输入词表示如式(1)~(2)。

MRM 的最小化负似然函数为式(4):

$$L_{\mathrm{MRM}}(\theta)=E_{(\boldsymbol{w},\boldsymbol{v})\sim D}\log f_\theta(\boldsymbol{v}_m|\boldsymbol{v}_{\backslash m},\boldsymbol{w}) \tag{4}$$

最终目标是最小化交叉熵损失(Cross Entropy Loss, CE)，如式(5):

$$f_\theta(\boldsymbol{v}_m|\boldsymbol{v}_{\setminus m},\boldsymbol{w})=\sum_{i=1}^{M}CE(c(\boldsymbol{v}_m^{(i)}),g_\theta(\boldsymbol{v}_m^{(i)})) \tag{5}$$

2.2.2 MRC-KL 预训练任务

掩码区域分类 KL 散度(Kullback-Leibler divergence)预训练任务(Masked Region Classification with KL-Divergence, MRC-KL)[28]，同样是随机掩盖区域标记，但是不同的是，这里不是做分类任务，而是需要计算 Faster R-CNN 特征和掩盖区域的分布差异，使得掩盖区域的分布和 Faster R-CNN 特征的分布尽可能相似，所以损失函数用的是 KL 散度。

掩码区域分类 MRC 的任务是预测一个被掩盖的视觉区域的对象类别，激活未掩盖的视觉上下文和标记。MRC-KL 变量[7]测量预测分布的 KL 散度，而不是针对单个对象类的交叉熵。

MRM 的最小化负似然函数为式(4)。

最小化两个分布之间的 KL 差距如式(6):

$$f_\theta(\boldsymbol{v}_m|\boldsymbol{v}_{\setminus m},\boldsymbol{w})=\sum_{i=1}^{M}D_{KL}(\tilde{c}(\boldsymbol{v}_m^{(i)})||g_\theta(\boldsymbol{v}_m^{(i)})) \tag{6}$$

2.2.3 MRFR 预训练任务

MRC 可以被视为视觉 MLM，需要 V&L 模型来预测蒙面对象的类别。掩码区域特征回归(Masked Region Feature Regression, MRFR)[20]进一步要求 V&L 模型恢复掩盖对象区域的视觉特征。MRFR 预测被掩盖掉的 RoI(Region of Interest)特征。随机掩盖掉 15%的 RoI(全部替换为零向量)，损失为输出 RoI 特征与特征抽取模型的 RoI 特征间的 L2 距离。

MRFR 类似于 MLM 任务，是一个流行的图像预训练任务。在短语上，MRFR 掩盖对象特征，模型根据文本侧类别标签和对象周围的信息预测原始对象级特征。在句子中，MRFR 掩盖了图像的对象特征，模型基于文本侧句子层面的整体信息和周围对象信息预测原始对象。

2.3 ITM 预训练任务

在图像-文本配对任务(Image-Text Matching, ITM)[29]中，需要对输入的图像-文本配对，随机替换其中的图片或者文本，最后预测输入的图像和文本是否有对应关系，属于一个二分类的问题。具体任务图如图 3 所示。

ITM 的最小化负似然函数为式(7):

$$L_{\mathrm{ITM}}(\theta)=-E_{(\boldsymbol{w},\boldsymbol{v})\sim D}[y\log s_\theta(\boldsymbol{w},\boldsymbol{v})+(1-y)\log(1-s_\theta(\boldsymbol{w},\boldsymbol{v}))] \tag{7}$$

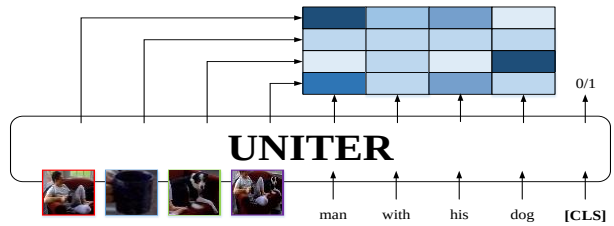


Image-Text Matching(ITM)

图 3 ITM 任务图

Fig. 3 Image-Text Matching task map

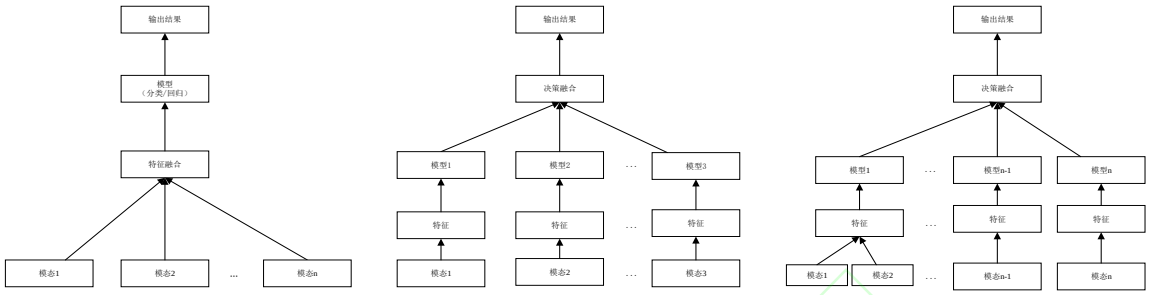
2.4 多模态数据融合

多模态数据融合是多模态预训练模型的重要的一部分，针对融合时期、融合程度和融合方式的不同，可将多模态数据融合分为早期融合、晚期融合和混合融合三种[30]。

为了解决各个模态之间的原始数据的不一致性，可以从每个模态中分别提取各自的特征表现形式，接着在特征级别上进行融合，称之为特征融合[31]。在深度学习中有时涉及从原始数据中学习特征的表示形式，导致有时在特征提取之前就进行数据融合，因此数据层和特征层的融合均称为早期融合。早期融合图如图 4(a)所示。在特征融合中，首先提取单一模态的特征，然后合并提取到的模态到融合特征中，再将该特征输入到指定模型中，输出预测结果。在该融合方法中，各模态特征经过转换、缩放等操作后得到的融合特征具有很高的维度，可使用线性判别分析(Linear Discriminant Analysis, LDA)[32]对融合特征进行降维处理。常常用早期融合结合语音识别中的音频和视频特征[33]。由于各种模态本身的差异，只进行简单的属性相接可能会忽略模态之间相关性和独有的特点，并可能产生数据之间的冗余和数据依赖[34]。并且在融合动作进行之前，要保证特征以相同的格式进行输入。

晚期融合方法也称决策级融合方法，先用对应的模型对各个模态进行建模训练，再融合多个模型输出的结果。这种方法主要采用最大值结合、平均值结合、贝叶斯规则等结合规则来确定不同模型输出结果的结合策略，较早期融合方法解决了一定的数据异步性。这样融合的好处是融合模型的错误来自不同的分类器，不会造成错误的累加效应。针对不同的模态可以选择各自最适合的分析方法，如音频使用隐马尔可夫模型[35]图像使用可支持向量机(Support Vector Machine, SVM)[36]，但这加大了融合的难度。晚期融合如图 4(b)所示。

混合融合方法在综合了早期融合和晚期融合优点的同时，也增加了训练的难度。在深度学习中，各模型灵活性和不确定性较大，大多使用混合融合方法。模型图如图 4(c)所示。



(a)早期融合图　(b)晚期融合图　(c)混合融合图

图 4 三种融合方式流程图

Fig. 4 Flow chart of three fusion method

研究表明，每个融合方式并无确定的优劣之分，在不同的实验条件下，可以尝试不同的融合方式来达到预期的效果。三种融合方式的具体参数比较如表 1 所示。

表 1 三种融合方式比较

Tab. 1 Comparison of three fusion methods

融合方式	信息损失	融合难度	容错性	输出	时序模型
早期融合	中	难	差	分类	否
后期融合	大	中	中	回归	是
混合融合	小	易	好	分类	否

3 多模态预训练模型

多模态预训练模型一般分为单流结构和双流结构。单流式预训练模型在早期就将不同模态的信息进行融合，通过注意力机制进行多模态交互。来自不同模态的特征要被结合在一起输入模型进行交互[37]。最简单的方法就是直接拼接。将提取出的视觉模型和文本特征映射到相同的维度后，直接将视觉特征的序列和文本拼接在一起。此外，一些特殊的数据类型提供了文本词语到视觉区域的映射关系，可以直接将视觉特征作为词语插入到文本中。

双流模型则先用不同的结构分别对两个模态进行编码，再通过互注意力(co-attention)机制[38]进行跨模态融合。视觉预训练模型的网络层数较多，因而在进行多模态交互之前所需的处理也少一些。而文本特征则没有经过较深模型的处理。双流模型结构灵活，可以根据具体情况决定在交互前对不同模态进行不同的处理。

单流与双流两种模型结构各有各的优点，并没有好坏之分。或者二者的性能本来就是相近的，神经网络可以自己学到合适的交互方式。以下对常见的一些单流和双流模型进行总结分析。

3.1 单流式多模态预训练模型

3.1.1 VideoBERT 模型

Sun 等提出的 VideoBERT 模型[5]，是一个在 BERT 模型的基础上捕捉语言和视觉领域的结构，是一个视频和语言表征学习的联合模型。由于大多数视频包含同步的音频和视频信号，这两种模态可以相互监督，以学习强自监督的视频表示[39]。模型结构如图 5 所示。

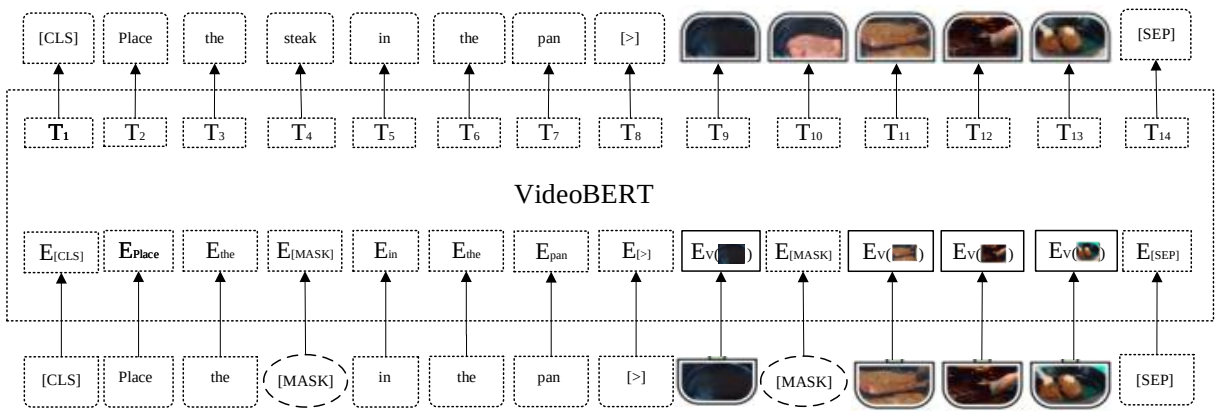


图 5 VideoBERT 模型[5]

Fig. 5 VideoBERT model

BERT 通过使用“掩码语言模型”训练目标来学习语言表示。联合概率分布如式(8):

$$\mathrm{p}(\boldsymbol{x}|\theta)=\frac{1}{Z(\theta)}\prod_{l=1}^{L}\phi_l(\boldsymbol{x}|\theta)\propto exp\left(\sum_{l=1}^{L}\log\phi_l(\boldsymbol{x}|\theta)\right) \tag{8}$$

其中：$\boldsymbol{x}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_L]$是一组离散标记；$\phi_l(\boldsymbol{x})$是第$l$个势函数；参数$\theta$，$Z$为配分函数。

学习每个单词标记以及这些标记的嵌入，然后对嵌入向量求和以获得每个标记的连续表示。$f(\boldsymbol{x}_{\backslash l})$是一个多层双向 Transformer 模型[40]。

每个位置的对数势能函数由式(9)~(10)定义：

$$\log\phi_l(\boldsymbol{x}|\theta)=\boldsymbol{x}_l^{\mathrm{T}}f_\theta(\boldsymbol{x}_{\backslash l}) \tag{9}$$

$$L(\theta)=E_{\boldsymbol{x}\sim D}\sum_{i=1}^{L}\log p(\boldsymbol{x}_l|\boldsymbol{x}_{\backslash l};\theta) \tag{10}$$

$\boldsymbol{x}_l$是第l个标记的独热(one-hot)向量，函数$f(\boldsymbol{x}_{\backslash l})$是一个多层双向 Transformer，它采用一个$L\times D_1$大小的张量，包含对应于$\boldsymbol{x}_{\backslash l}$的$D_1$维嵌入向量，并返回一个$L\times D_2$张量，其中$D_2$是每个 Transformer 节点的输出大小，训练模型以近似最大化对数似然估计。在实际应用中，可以通过采样位置和训练句子来对损失函数的对数形式随机优化。

联合分布概率与对数势能函数主要是将 BERT 扩展到视频和语言数据进行联合建模，而在 VideoBERT 中不仅需要简单地做扩展序列建模，而且对视频帧、语言的处理以及它们之间的关系也更为关注。

为了将 BERT 扩展到视频，利用预训练的语言模型和可扩展的实现来进行推理和学习，并将原始的视觉数据转换成离散的标记序列。为此，通过使用预训练模型将分级矢量量化应用于从视频中导出的特征来生成“视觉单词”序列，这种方法鼓励模型关注视频中的高级语义和更长时间范围的动态信息。

3.1.2 HERO 模型

Li 等提出的 HERO(Hierarchical EncodeR for Omnirepresentation learning)模型[41]在分层结构中编码多模态输入，其中视频帧的局部上下文由跨模态 Transformer 通过多模态融合捕获，而全局视频上下文由时间 Transformer 捕获。它将视频剪辑的帧和字幕句子的文本标记作为输入，再将它们送到视频嵌入器和文本嵌入器，以提取初始表示。HERO 计算分层过程中的上下文化视频嵌入。首先，每个视觉帧的本地文本上下文被交叉模态 Transformer 捕获，计算字幕句子和其相关视觉帧之间的上下文化多模态嵌入。然后，整个视频剪辑的编码帧嵌入被馈送到时间 Transformer，以学习全局视频上下文并获得最终的上下文视频嵌入。

如图 6 所示，MLM 是掩码语言建模任务，掩码帧建模(Masked Frame Modeling, MFM)任务，视频字幕匹配(Video-Subtitle Matching, VSM)任务。

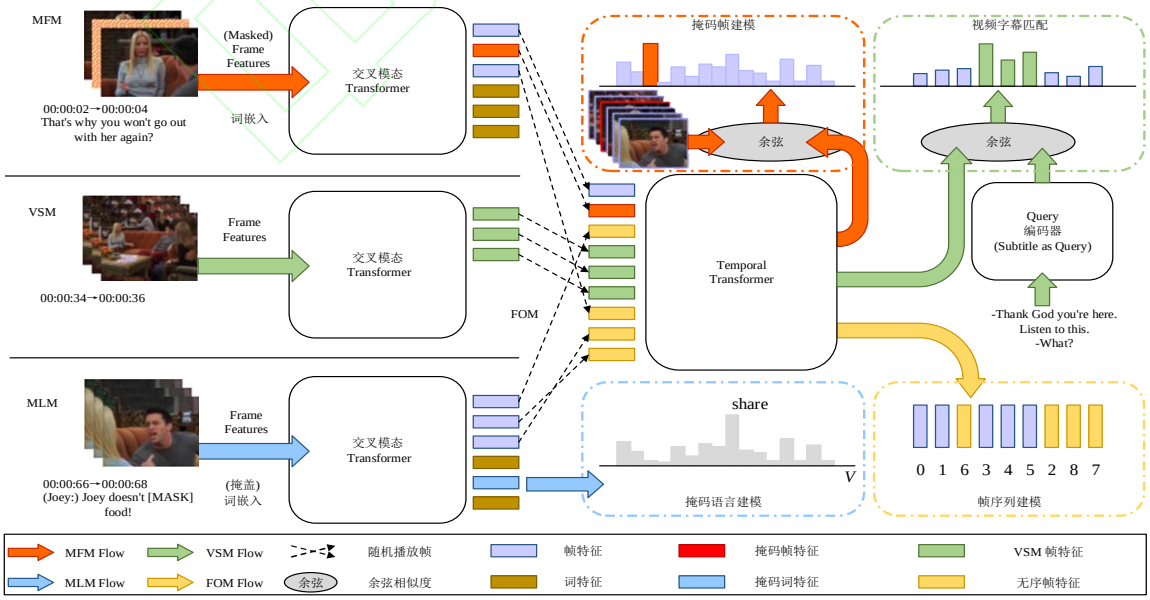


图 6 HERO 模型[41]

Fig. 6 HERO model

为了利用字幕和视频帧之间的内在一致性，对于每个字幕句子$\boldsymbol{s_i}$，通过跨模态注意学习相应标记$\boldsymbol{w}_{\boldsymbol{s}_i}$与其相关视觉框架$\boldsymbol{v}_{\boldsymbol{s}_i}$之间的上下文嵌入信息，并使用多层 Transformer 结构。跨模式 Transformer 的输出是每个字幕标记和每个视频帧的上下文嵌入序列，由式(11)表示：

$$\boldsymbol{V}_{\boldsymbol{s_i}}^{\mathrm{cross}},\boldsymbol{W}_{\boldsymbol{s_i}}^{\mathrm{cross}}=f_{\mathrm{cross}}(\boldsymbol{V}_{\boldsymbol{s_i}}^{\mathrm{emb}},\boldsymbol{W}_{\boldsymbol{s_i}}^{\mathrm{emb}}) \tag{11}$$

在从跨模态 Transformer 的输出中收集了所有视觉帧嵌入 $\boldsymbol{V}^{\text{cross}}$ 之后，使用另一个 Transformer 作为时间注意力，从视频剪辑的全局上下文中学习上下文的视频嵌入。

为了避免丢失位置信息，使用残差连接[2]将 $\boldsymbol{V}^{\text{emb}}$ 加入。最终的上下文视觉嵌入计算如式(12):

$$\boldsymbol{V}^{\text{temp}}=f_{\text{temp}}(\boldsymbol{V}^{\text{emb}}+\boldsymbol{V}^{\text{cross}}) \tag{12}$$

其中，$f_{\text{temp}}(\cdot)$ 表示时间 Transformer,$\boldsymbol{V}^{temp}\in\mathbf{R}^{N_V\times d}$。

该模型提出了一种用于视频、语言全表示预训练的分层编码器，提出了一个层次结构，由跨模态 Transformer 和时间 Transformer 组成，用于多模态融合。提出了新颖的预训练任务来捕获局部和全局的时间对齐。在两个大规模视频数据集上进行预处理后，当转移到多个视频和语言任务时，HERO 超越了其他模型。

3.1.3 VL-BERT

VL-BERT 模型基于 Transformer 模型，以视觉和语言的嵌入特征作为输入，对于每个输入元素，主要由 4 个嵌入层组成，标记嵌入、视觉特征嵌入、片段嵌入以及序列位置嵌入。标记嵌入层主要是为每个特殊的元素分配特殊的标记，为每个视觉元素分配一个[IMG]标记。第二层的视觉特征嵌入层，是为了嵌入视觉信息新添加的层，由视觉外部特征以及几何特征拼接而成。对于非视觉部分输入的是整个图像提取到的特征，对于视觉部分输入为图像经过预训练之后的 Faster R-CNN 提取到的 RoI 区域图像的相应视觉特征[9]，每个 RoI 由一个 4 维向量表示：

$$(\frac{x_{LT}}{W},\frac{y_{LT}}{H},\frac{x_{RB}}{W},\frac{h_{RB}}{H}) \tag{13}$$

其中(x_{LT},y_{LT})和(x_{RB},y_{RB})分别表示左上角和右下角的坐标，W,H表示输入图像的宽度和高度，视觉特征嵌入附加到每个输入元素，是将视觉外观特征和视觉几何嵌入的串联作为输入。其他三个嵌入层遵循 BERT 模型的设计。

用于预训练 VL-BERT 的架构如图 7 所示，通过引入的 MLM 任务，使得 BERT 模型经过预训练后，可以从所有未被掩盖的词语信息中预测掩盖词[42],VL-BERT 模型中具有视觉信息的掩码语言建模任务与 MLM 类似，不同在于添加视觉嵌入层后，可以得到视觉和语言内容之间的依赖关系，根据未掩盖的单词与视觉特征，对模型训练以预测掩盖的单词，进一步使得视觉和语言内容保持一致。在预训练期间，将与掩盖的 RoI 相对应的最终输出特征加入具有 softmax 交叉熵损失的分类器中，用于对象类别分类[43-44]。VL-BERT 是一种可广泛应用于视觉-语言任务的预训练模型，在 BERT 模型基础上将捕获视觉信息作为一个新的输入，VL-BERT 预训练模型通过对视觉语言信息进行调整可应用于更多的下游任务，提高了三个下游任务的性能，其中在视觉常识推理(Visual Commonsense Reasoning, VCR)任务上预训练效果较为明显。

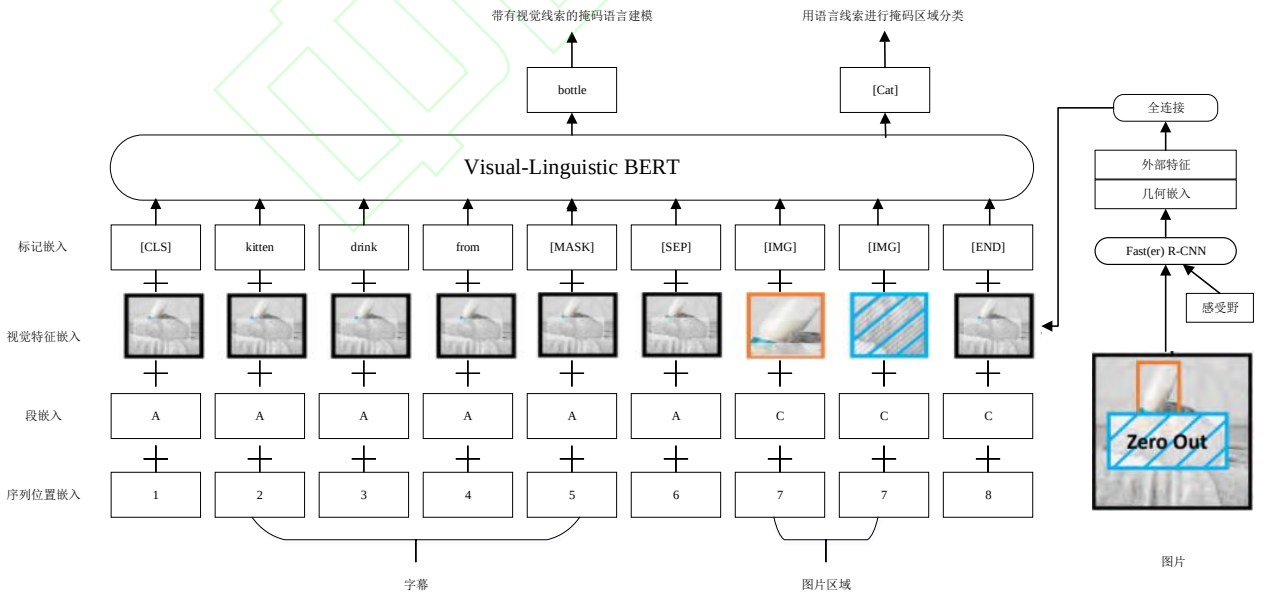


图 7 VL-BERT 模型[9]
Fig. 7 VL-BERT model

3.1.4 ImageBERT

ImageBERT 用于图像-文本联合嵌入，基于 Transformer 模型，使用图像视觉标记和文本标记作为输入，并对它们之间的关系进行建模。其中图像视觉标记是从 Faster R-CNN 模型[26]中提取的 RoI 特征，将这些嵌入信息输入到一个多层双向自注意力 Transformer 中，学习一个跨模态 Transformer 来模拟视觉区域和语言标记之间的关系[21]。

通过一层嵌入层将文本和图像编码成不同的嵌入信息。

语言嵌入中将输入的句子标记为 n 个子词标记[45]，每个子词标记的最终嵌入是通过结合词嵌入、片段嵌入和序列位置嵌入生成的。图像嵌入从视觉输入中生成，其中图像检测到的对象可以为语言部分提供整个图像的视觉上下文信息，通过将对象相较于全局图像的位置可以编码成 5 维向量，表示如下：

$$\boldsymbol{c}^{(i)}=\left(\frac{x_{tl}}{W},\frac{y_{tl}}{H},\frac{x_{br}}{W},\frac{y_{br}}{H},\frac{(x_{br}-x_{tl})(y_{br}-y_{tl})}{WH}\right) \tag{14}$$

(x_{tl},y_{tl}) 以及(x_{br},y_{br}) 分别代表边界框的左上角和右下角坐标，5 维向量的第五个分向量$\frac{(x_{br}-x_{tl})(y_{br}-y_{tl})}{WH}$是相较于整个图像的比例面积，最后还有序列嵌入以及段嵌入。

ImageBERT 模型的整体架构图 8 所示，该模型的预训练是同时在掩码语言建模、掩码对象分类、掩码区域特征回归和图像文本匹配四个任务上进行。

经过预训练后，得到关于视觉语言联合表示的预训练模型。ImageBERT 模型为更好利用不同种类的预训练数据，将多个预训练阶段应用于一个统一的网络结构，提出一个多阶段预训练框架，通过使用大规模的域外数据再使用小规模的域内数据对预训练模型进行训练，使得模型可以更好地收敛于最终任务[46]，让下游任务受益。在两个大规模数据集上进行文本到图像、图像到文本的检索任务都取得了较好的结果。

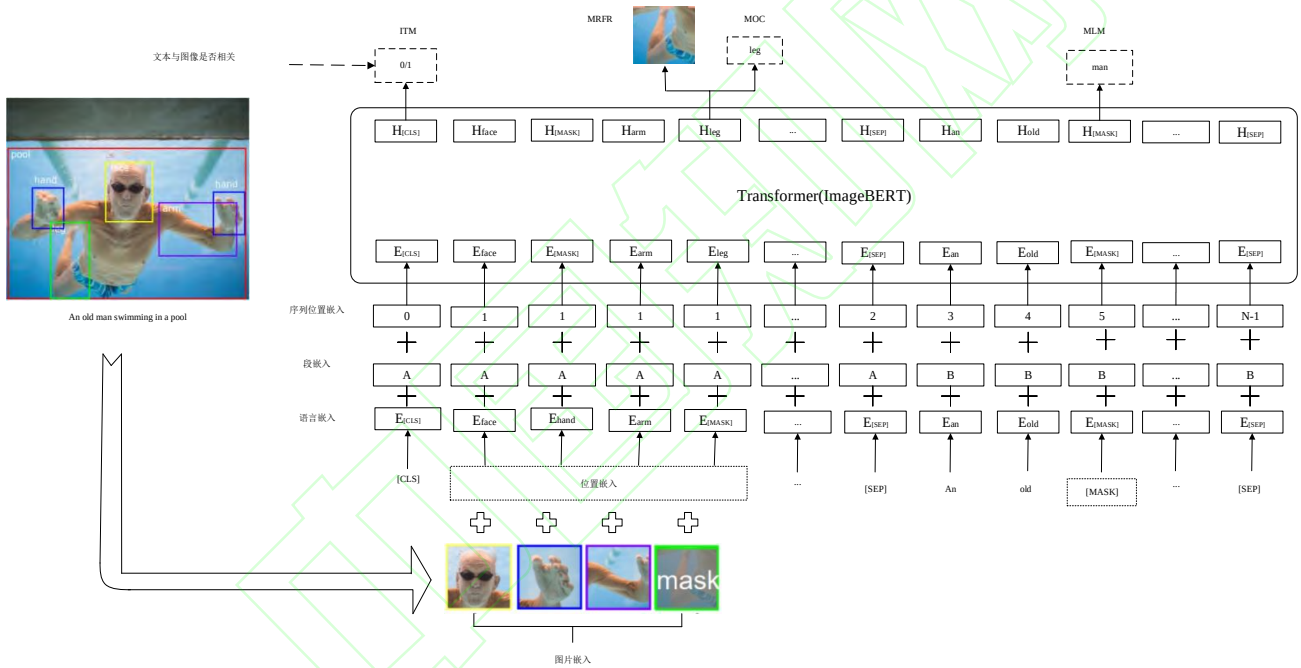


图 8 ImageBERT 模型

Fig. 8 ImageBERT model

3.2 双流式多模态预训练模型

3.2.1 UniVL 模型

由 Luo 等提出的 UniVL 模型[1]，主要结构有四个组件，包括两个单模编码器、一个交叉编码器和一个解码器。该模型使用各种特征提取器提取输入文本标记和视频帧序列的表示。然后，文本编码器采用 BERT 模型嵌入文本，视频编码器利用 Transformer 编码器嵌入视频帧最后，使用 Transformer 解码器来重构输入文本。模型结果如图 9 所示。

文本编码器的结构和 BERT 模型相同，也使用 BERT-base 模型进行参数的初始化。视频编码器部分使用 S3D(Separable 3D convolutional neural network)模型[47]提取图片特征，再使用 6 层的 Transformer 对图像序列的特征进行进一步的处理，如式(15)~(16):

$$\boldsymbol{T}=BERT(\boldsymbol{t}) \tag{15}$$

$$\boldsymbol{V}=Transformer(\boldsymbol{F}_v) \tag{16}$$

文本编码器和视频编码器主要关注单模态。为了使文本和视频充分交互，设计了交叉编码器，它将文本和视频模态特征都作为输入，如式(17):

$$\boldsymbol{M}=Transformer([\boldsymbol{T};\boldsymbol{V}]) \tag{17}$$

其中[;]表示按照顺序的维度进行组合操作，用 Transformer 获得解码后的特征 $\boldsymbol{D}$，如式(18):

$$\boldsymbol{D}=Transformer(\boldsymbol{M}) \tag{18}$$

UniVL 模型，在视频语言理解和生成任务中可灵活使用，能够学习到较强的视频文本表示。和该模型一起提出的逐步预训练和增强视频表示预训练策略，使得模型的训练效果更好。

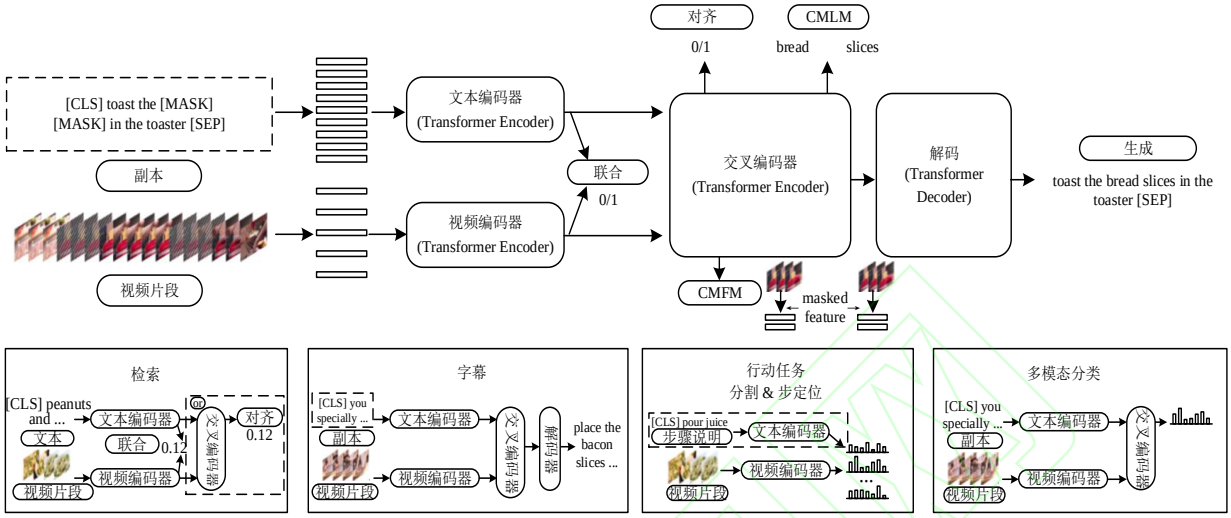


图 9 UniVL 模型图[1]

Fig. 9 UniVL model

3.2.2 ERNIE-ViL 模型

ERNIE(Enhanced Representation through knowledge Integration)[48-49]是由百度提出用于对文本进行建模的模型，为了对图文信息进行多模态建模，百度在后续还提出了 ERNIE-ViL 模型[50]。

如图 10 所示，左边是网络输入输出示例，而右图是对文本进行场景图解析后的各个文本元素。图 10 中的视觉元素能找到对应的文本元素。文本具有一定的结构化信息，在多模态信息融合过程中考虑这些结构化信息有助于模型理解图中物体之间的结构关系、属性、类别等。左图的输入分别是对一张图进行 RoI 检测之后的 RoI 区域图片特征，另一个输入是文本的嵌入特征。此处的文本 W 需要利用场景图解析(Scene Graph Parsing)方法解析出各种文本元素，解析结果可以表示为式(19):

$$G(\boldsymbol{w}) =< O(\boldsymbol{w}), E(\boldsymbol{w}), K(\boldsymbol{w}) > \quad (19)$$

视觉元素输入是一系列显著性 RoI 区域的特征(以及其检测框)，文本元素是经过场景图解析后的诸多文本结构信息，那么对掩码语言模型进行扩展，可以掩盖某个模态的信息，然后期望模型可以根据另一个模态的信息“恢复”出被掩盖的信息，这个恢复过程既考虑了同一个模态的上下文信息，也考虑到了跨模态的建模。

对象预测，可以表示为式(20):

$$L_{\text{obj}}(\theta) = -E_{(\boldsymbol{w},\boldsymbol{v})\sim D}\log(P(\boldsymbol{w}_{oi}|\boldsymbol{w}_{\setminus w_{oi}}, \boldsymbol{v})) \quad (20)$$

属性预测可以表示为式(21):

$$L_{\text{attr}}(\theta) = -E_{(\boldsymbol{w},\boldsymbol{v})\sim D}\log(P(\boldsymbol{w}_{ai}|\boldsymbol{w}_{w_{oi}}, \boldsymbol{w}_{\setminus w_{ai}}, \boldsymbol{v})) \quad (21)$$

关系预测则可以表示为式(22):

$$L_{\text{rel}}(\theta) = -E_{(\boldsymbol{w},\boldsymbol{v})\sim D}\log(P(\boldsymbol{w}_{ri}|\boldsymbol{w}_{oi1}, \boldsymbol{w}_{oi2}, \boldsymbol{w}_{\setminus w_{ri}}, \boldsymbol{v})) \quad (22)$$

该模型首次利用视觉结构化知识(场景图)，取得了良好的效果。

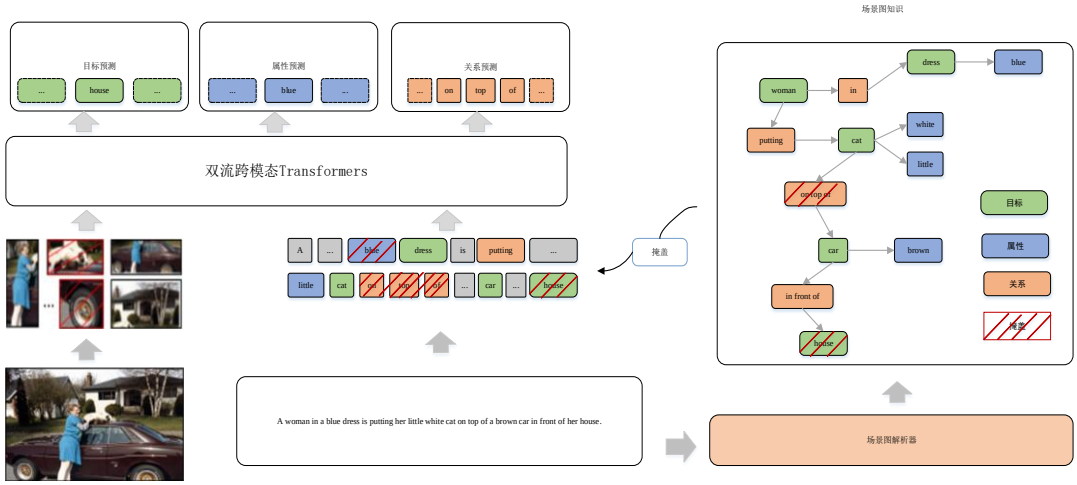


图 10 ERNIE-ViL 模型结构图[50]

Fig. 10 ERNIE-ViL model structure diagram

3.2.3 VilBERT 模型

Lu 等提出的 ViLBERT 模型[4]，将流行的 BERT 架构扩展到多模态双流模型中，通过共同注意力 Transformer 层单独对视觉和文本输入信息进行预训练。输入的文本经过文本嵌入层后，再输入到文本的单模态 Transformer 编码器中提取上下文信息。使用预训练 Faster R-CNN 对于图片生成候选区域提取特征并送入图像嵌入层。然后将获取好的文本和图像的特征通过共同注意力 transformer 模块进行交互融合，得到最后的表征。将该模型应用到视觉问答[18]、视觉常识推理[51]、指示表达定位(Grounding Referring Expressions, GRE)[52]、图片图像检索[53]等下游任务上，取得了较好的结果。

3.2.4 ActBERT 模型

为了同时进行文字与动作和局部区域的视觉输入关联，ActBERT[54]提出了一个简单的思路，就是在输入层同时加入全局动作特征与局部区域特征。为了得到动作特征，首先从源数据集文本中提取动词，并构建出一个新的动词词汇表。为了得到局部特征，ActBERT 使用在 COCO[12]上预训练的 Faster R-CNN 在视频帧上产生物体候选框，每帧仅保留部分候选框以减少冗余，这些候选框对应的特征将作为网络输入，而候选框在 COCO 上的类别分布将作为 ActBERT 预测类别。

针对以上的单流和双流模型，作出表 2 所示比较。

表 2 不同模型的比较

Tab. 2 Comparison of different models

模型名称	初始模型	预训练任务	视觉构成	下游任务	特征提取器	预训练数据集
Videobert[5]	BERT	1)根据文本预测视频，根据文本自动插图；2)根据视频预测文本，对视频自动生成摘要	视频帧	1)视频字幕；2)零样本动态分类	S3D[47]	Cooking312k[5]
HERO[41]	-	1)掩码语言建模；2)掩码帧建模；3)视频字幕匹配；4)帧顺序建模	视频帧	1)视频语言推断;2)视频字幕匹配	-	TV 数据集[55], Howto100M[56]
VL-BERT[9]	BERT	1)掩码文本预测；2)掩码图像类别预测	图像 RoI	1)视觉常识推理;2)视觉问答; 3)引用表达式理解	Fast R-CNN[57]	Conceptual captions[13]
ImageBERT[21]	BERT	1)掩码文本预测；2)掩码图像类别分类；3)掩码图像特征回归；4)图像文本对齐	图像 RoI	1)图像检索;2)文本检索	Fast R-CNN[57]	LAIT 数据集
UniVL[1]	Transformer	1)条件屏蔽语言模型；2)条件屏蔽帧模型；3)视频文本对齐；4)语言重建	视频帧	1)多模态视频字幕;2)动作分割;3)动作步骤定位;4)多模态情感分析;5)基于文本的视频检索	S3D[47] Resnet-152 和 Resnet-101[2]	HowTo100M[56]
ERNIE-ViL[50]	Transformer	1)视觉问答[18]；2)视觉常识推理[51]；3)图像&文本检索	图像 RoI	场景图预测(目标预测、属性预测、关系预测)	-	Conceptual Captions(CC)[13] SBU Captions[14]
ViLBERT[5]	BERT	1)掩码文本预测；2)掩码图像类别预测；3)图像文本对齐	-	1)图像检索;2)视觉常识推理;3)视觉问答;4)引用表达式理解	Fast R-CNN[57] 和 Resnet-101[2]	Visual Genome[19]
ActBERT[54]	BERT	1)掩码文本预测；2)掩码动作预测；3)掩码物体预测(4)视频文本对齐	视频帧	1)视频检索;2)视频问答;3)视频描述生成;4)行为分割;5)动作定位	Faster R-CNN[57]	HowTo100M[56]

在常见的 6 个下游任务(视觉问答[58]、视觉常识推理[51]、图像区域定位[52]、图像检索(Image Retrieval, IR)[59]、视频检索(Video Retrieval, VR)[60]和视频字幕(Video Captioning, VC)[61]生成)中，对相关模型的性能进行比较。

其中，视觉问答，是指对视觉图像的自然语言问答，连接视觉和语言，旨在通过对图像理解的基础上，根据具体的问题进行回答。视觉常识推理，旨在通过给定的图片、区域和问题后，从选项中选出答案并指明原因。主要评价指标包括是否正确选择了答案(Q->A)、是否对选择的答案给出合理的原因(QA->R)以及综合评价答案和原因(Q->AR)。图像区域定位，又称为看图识物，目标是根据文字信息在图像中标注出对应的物体。图像检索，主要是根据文字描述在图片库中搜索对应图片。视频检索与之类似，以召回率作为评价指标。视频字幕生成是一个视频图形序列到文本序列的序列到序列任务，旨在将视频翻译为自然语言。

如表 3 所示，是 VL-BERT、VilBERT 以及 ERNIE-Vil 三

个预训练模型在 VQA、VCR 以及 GRE 任务上的实验对比。

如表 4 所示，是现有部分模型在剩余三个下游任务上的实验数据对比。其中 HERO 模型在视频字幕时刻检索(Video-subtitle Moment Retrieval, TVR)和视频字幕描述(Video-subtitle caption description, TVC)任务中与 SOTA 基线模型(MMT（MultiModal Transformer）模型[62]和 XML(Cross-modal Moment Localization)模型[62])进行比较。

表 3 各模型在 VQA、VCR、GRE 任务上的实验对比　　单位:%

Tab. 3 Experimental comparison of each model on VQA,VCR and GRE tasks　　unit:%

模型	VQA（数据集 VQA2.0）		VCR（数据集 VCR）						GRE（数据集 RefCOCO+[63]）		
			Q->A		QA->R		Q->AR				
	test-dev	test-std	val	test	val	test	val	test	val	testA	testB
VL-BERT	71.79	72.22	75.5	75.8	77.9	78.4	58.9	59.7	72.59	78.57	62.30
ViLBERT	70.55	70.92	72.42	73.3	74.47	74.6	54.04	54.8	72.34	78.52	62.61
ERNIE-ViL	73.78	73.96	78.52	79.2	83.37	83.5	65.81	66.3	74.24	80.97	64.70

表 4 各模型在 IR、VR、VC 任务上的实验对比　　单位:%

Tab. 4 Experimental comparison of each model on IR, VR and VC tasks　　unit:%

模型	IR（数据集 Flickr30k[64]）			VR/TVR（数据集 YouCook2[65]）				VC/TVC（数据集 YouCook2）				
	R@1	R@5	R@10	R@1	R@5	R@10	Media R	B-3	B-4	METEOR	ROUNG-L	CIDEr
ViLBERT	58.2	84.9	91.52	-	-	-	-	-	-	-	-	-
ImageBERT	73.1	92.6	96.00	-	-	-	-	-	-	-	-	-
ActBERT	-	-	-	9.6	26.7	38.0	19	8.66	5.41	13.30	30.56	0.65
UniVL[1]	-	-	-	28.9	57.6	70.0	4	16.46	11.17	17.57	40.09	1.27
Videobert	-	-	-	-	-	-	-	6.80	4.04	11.01	27.50	0.49
ERNIE-ViL	75.1	93.42	96.26	-	-	-	-	-	-	-	-	-
HERO	-	-	-	6.21 (3.25)	-	19.34 (13.41)	-	12.35 (10.87)	-	17.64 (16.91)	34.16 (32.81)	49.98 (45.38)

3.3 基于 Prompt Turning 的多模态预训练模型

由于模型预训练和整合之间的客观形式存在显著差异，Yao 等[66]提出跨模态提示调优(Cross-modal Prompt Tuning, CPT)，在图像和文本中使用基于颜色的共同参照标记重新构建了视觉定位问题，使得视觉-语言预训练模型(Visual-Language PTM, VL-PTM)在少样本甚至零样本的情形下展现出强大的视觉预测能力。

CPT 由视觉子标识和文本子标识两部分组成。视觉子标识，目的是通过可分辨的标记区分每一个区域，比如颜色。这个子标识是直接加在原图中的，所以既没有改变模型结果，也没有改变参数。另一个子标识是文本子标识，目的是在图片和文本问题之间建立一个链接，这里使用的模板为式(23):

$$T(q) = [\mathrm{CLS}]\ q\ \text{is in}\ [\mathrm{MASK}]\ \text{color}\ [\mathrm{SEP}] \tag{23}$$

视觉-语言预训练模型(VL-PTM)通过这样的提示来决定哪个颜色的区域放在相应的位置最合适，如式(24):

$$P(\boldsymbol{v}^* = \boldsymbol{v}_i|\mathbf{R}, q) = P([\mathrm{MASK}] = \boldsymbol{c}_\omega^i(\bar{\psi}(\mathbf{R}; C), T(q))) = \frac{\exp\left(\boldsymbol{h}_{[\mathrm{MASK}]}^{\mathrm{T}}\boldsymbol{c}_\omega^i\right)}{\Sigma_{c_j \in C}\exp\left(\boldsymbol{h}_{[\mathrm{MASK}]}^{\mathrm{T}}\boldsymbol{c}_\omega^j\right)} \tag{24}$$

该过程未引入任何新的参数，也减轻了预训练和调优之间的差距，从而提高了调优 VL-PTM 的数据效率。在少样本、零样本任务中表现良好。在目标检测[67]、谓元分类[68]、场景图分类[69]等视觉任务中也有良好的表现。

尽管在一些任务中的性能很好，但存在以下两个局限性：1)颜色干扰。CPT 通过在图像和文本中添加基于颜色的提示，利用颜色来连接视觉和文本语义。基于颜色的提示可能会被原始图像和文本中的颜色所干扰。2)计算效率低。为了最大限度地避免颜色干扰，并考虑有限数量的候选颜色选择，采用了小的图像区域束大小。这意味着需要将一个数据实例多次输入模型以获得结果，降低了计算效率。

4 应用

4.1 M6 模型在文本图像生成中的应用

现有的多模态预训练方法都只关注具有视觉和语言输入的多模态任务，视觉与语义是相通的。在应用领域提出了一种跨模态预训练方法 M6(Multi-Modality to Multi-Modality

Multitask Mega-transformer)，将其对于单模态和多模态数据来统一预训练[70]。将模型规模扩大到 100 亿和 1000 亿个参数，可以使模型应用于一系列下游应用，展示了 M6 出色的性能。此外，还设计了一个文本图像生成的下游任务，并表明微调后的 M6 可以生成高分辨率和丰富细节的高质量图像。

文本图像生成方面的应用利用了文本图像生成的两阶段框架，包括离散表示学习[71]和语言建模。在离散表示学习中，集中在将图像转换成离散码序列。在下一阶段中，需要建立一个语言模型来学习生成文本和代码序列。在微调中，需要将代码嵌入层和输出层添加到预训练的 M6 中。将单词序列和上述生成的代码序列作为输入，并设定了训练的自回归语言建模目标。在推理阶段，输入文本序列，通过前 k 个抽样自回归生成代码。最后一步是从第一阶段用生成器将代码序列转换成图像。

多模态到文本的转换，是建立在图像到文本转换的基础上，此外还增加了隐藏的语言输入，因此模型需要学习，同时基于视觉信息和噪声语言信息生成目标文本。这个任务使得模型能够生成图 11 所展示的示例。可以发现，生成的图像质量高，生成的对象与真实对象相似。此外，在图 12 中，该模型能够根据查询军旅风迷彩高跟鞋来设计新款的物品样式，为现实工业场景中的创意设计提供了空间，如服装设计、鞋子设计等。

图 11 减震透气跑鞋生成图像

Fig. 11 Shock-absorbing and breathable running shoes generate images

图 12 军旅风迷彩高跟鞋生成图像

Fig. 12 Military style camouflage heels generated image

4.2 VideoBERT 在视频剪辑中的应用

视频剪辑首先转换为视频标记(每个示例有两个视频标记)，并使用它们的中心点来进行可视化。使用 VideoBERT[5] 预测视频剪辑中的名词和动词。VideoBERT 将视频转化为一系列“可视化单词”(visual words)。视频由一系列图片构成，一幅图片对应一帧，将 n 个连续的帧构成一个片段 clip，使用计算机视觉领域的模型进行特征提取，最终抽取了特征向量，随后对所有特征向量做分层矢量量化(hierarchical vector quantization)[72]，即聚类，得到 20736 个类，每个视频都有属于自己的一个类，这个类就是文本处理时的标记(visual token)。看图猜词，如图 13，将一些关键动词和名词掩盖，来猜测它们的具体内容。

图 13 使用 VideoBERT 预测视频剪辑中的名词和动词

Fig. 13 Predict nouns and verbs in video clips with VideoBERT

4.3 CPT 模型在视觉-语言预训练模型中的应用

CPT 使 VL-PTM 仅使用少量训练实例就能区分被同一类型对象分散的目标对象，而微调方法很难成功。在复杂条件下，CPT 也具有优秀的性能表现。例如，与目标相似的物体需要复杂的推理才能识别成功，但通常会产生合理的预测。例如在图 14(b)中，CPT 预测一个附近的苹果，而微调基线错误预测为一个碗。原因是 CPT 最大限度地使用了 VL-PTM 的预训练参数，这避免在少量微调后出现错误的预测。然而，发现 CPT 还会受到原始图像区域和文本颜色的干扰。例如，当候选区域由红色方块着色时，模型很难识别红色碗(图 14（c）)。

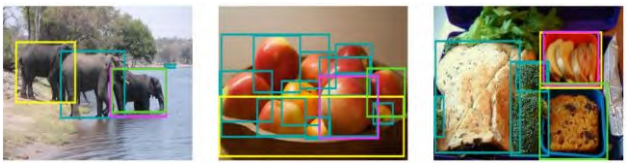


图 14 边界框给出的图像区域建议[48]

Fig. 14 Suggestions for image areas given by the border frame[48]

4.4 Alicemind 在电商商品理解和图文搜索上的应用

多模态预训练模型体系 AliceMind(ALIbaba's Collection of Encoder-decoders from Mind)[73]具有问答、搜索、摘要生成、对话等多种能力，目前已经在电商、客服、广告等数十个核心业务应用落地。如图 15 是 Alicemind 有关于视觉问答方面的各个类别示例，可以看出 Alicemind 熟悉日常类别知识、在目标物体识别与记忆以理解和推理图像中物体间的关系等，但是在逻辑推理问题上效果不佳，应用场景之一就是多模态商品理解，商品是电商的一个基本元素，电商的实质是将商品和人联系起来，然而在商品理解中，依靠单一模态无法满足电商应用中更复杂的语义关系[74]，多模态理解可以更好地获取对商品的三个认知：品牌类别识别、同款识别及需要从复杂场景里根据商品标题去选择正确的商品主体。下游应用展示了在网购平台进行商品发布，通过图片和文本展示商品信息以及推送更多该商品的同款信息来了解价格等，在图文搜索场景中，多模态可以帮助给到用户更准确的商品结果，也可以推荐相似的商品去吸引更多的用户购买。

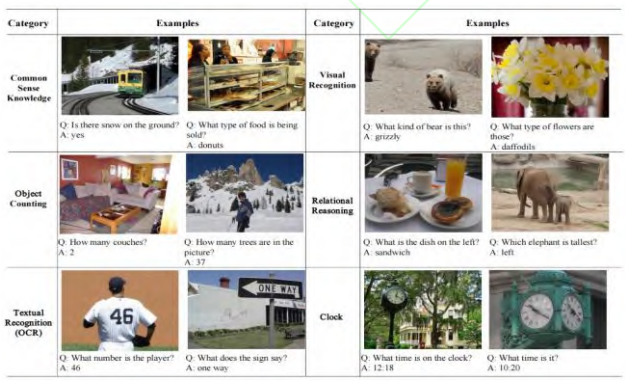


图 15 视觉问答各个类别示例

Fig. 15 Visual Q&A each category example

4.5 文本视频转换

视频可以理解为一组快速播放的图片，VideoBERT 模型可以根据文本自动插图生成视频，还可以依据视频预测文本，对视频自动生成摘要[75]。如图 16 可以根据给定的一些分成句子的食谱文本，通过 VideoBERT 生成视频标记序列，然后生成了一个有关于食谱的视频，更形象的说明食谱中的文字，视频字幕生成目标如图 17 是根据不同复杂视频内容给出一句或多句文字描述，VideoBERT 模型将视频数据融入 BERT，可以对一个复杂视频生成字幕，生成的字幕内容生动具体，实际应用上可以用于后期的视频搜索或检索，也可以进行人机交互或者帮助有视觉障碍的人理解现实情况。

图 16 文本生成视频

Fig. 16 Text generation video

图 17 视频字幕生成

Fig. 17 Video subtitle generation

5 总结与展望

目前的多模态预训练模型相关工作已经取得了一定的进展，在多个下游任务上有了不俗的表现。多模态预训练模型不仅可以应用于图像和文本，在视频和音频中也得到广泛应用。多模态预训练还可以应用于图像-文本检索、图像-文本生成、文本-图像生成等下游任务。然而，为多模态预训练找到一个“真实”的应用场景仍然是一个挑战，相对“真实”的应用场景可以反映出多模态预训练的训练结果，这里可以尝试利用许多成本很低且有效的工程技巧来解决，以及如何通过多模态多语言的 PTM 将源语言音频直接生成为目标语言的文本或目标语言的音频也是值得探索的[76]。

未来的工作可能从以下几个方向取得进一步的进展:第一是单模态下游任务上能否取得提升。现在大部分多模态预训练模型都是在多模态的下游任务上进行测试，少有工作在单模态任务如自然语言处理任务与单模态预训练模型进行全面的比较。如果认为模型在多模态数据上通过预训练能够更加充分地理解语义，那么直觉上看多模态预训练模型与单模态模型在相近的实验设置下(如语料规模相似)应当取得更好

的成绩。第二是更精细的挖掘不同模态数据间的相关信息并设计更巧妙的预训练任务。比如挖掘图像-文本之间，名词与物体对象之间的相关性，使得模型建立词语与物体对象之间的相关性。第三是设计更高效的模型架构以及挖掘更大规模的高质量多模态数据。

参考文献

[1] LUO H, JI L, SHI B, et al. UniVL: A unified video and language pre-training model for multimodal understanding and generation [EB/OL].[2020-09-15].https://arxiv.org/pdf/2002.06353.pdf .

[2] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE conference on computer vision and pattern recognition.Piscataway:IEEE, 2016: 770-778.

[3] 赵亮.多模态数据融合算法研究[D].辽宁:大连理工大学,2018:1-2.(ZHAO L. Research on multimodal data fusion methods[D]. Liaoning:Dalian University of Technology, 2018:1-2).

[4] LU J, BATRA D, PARIKH D, et al. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks [EB/OL]. [2019-08-06]. https://arxiv.org/pdf/1908.02265.pdf.

[5] SUN C, MYERS A, VONDRICK C, et al. VideoBERT: A joint model for video and language representation learning [C]// Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Piscataway:IEEE 2019: 7464-7473.

[6] TAN H, BANSAL M. LXMERT: Learning cross-modality encoder representations from transformers [C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroundsburg, PA:Association for Computational Linguistics, 2019: 5100-5111.

[7] LI L H, YATSKAR M, YIN D, et al. VisualBERT: a simple and performant baseline for vision and language[EB/OL].[2019-08-09] . https://arxiv.org/pdf/1908.03557.pdf.

[8] LI G, DUAN N, FANG Y, et al. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training[C]//Proceedings of the 2020 AAAI Conference on Artificial Intelligence. Menlo Park,CA:AAAI Press, 2020: 11336-11344.

[9] SU W, ZHU X, CAO Y, et al. VL-BERT: Pre-training of generic visual-linguistic representations [EB/OL].[2020-02-18].https://arxiv.org/pdf/1908.08530.pdf.

[10] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision [C]// Proceedings of the 38th International Conference on Machine Learning. New York :PMLR, 2021: 8748-8763.

[11] LI J, LIN X, HAN S, et al. Align before fuse: vision and language representation learning with[C]// Proceedings of the 2021 Conference on Neural Information Processing Systems.Stroundsburg, PA:Association for Computational Linguistics, 2021: 1978-1992.

[12] 赵广立.跨模态通用 AI 平台“紫东太初”发布[N].中国科学报，2021-07-12(4).(ZHAO G L. Cross-modal universal AI platform "Purple east taichu" was released[N]. Chinese Journal of Science，2021-07-12(4).)

[13] SHARMA P, DING N, GOODMAN S, et al. 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). Stroundsburg, PA:Association for Computational Linguistics, 2018: 2556-2565.

[14] ORDONEZ V, KULKARNI G, BERG T L. Im2Text: describing images using 1 million captioned photographs[C]//Proceedings of the 24th International Conference on Neural Information Processing Systems. New York : Curran Associates Inc, 2011: 1143-1151.

[15] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]// Proceedings of the 2014 European Conference on computer vision(ECCV). Cham: Springer, 2014: 740-755 .

[16] PLUMMER B A, WANG L, CERVANTES C M, et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models[C]//Proceedings of the 2015 IEEE international conference on computer vision. Piscataway:IEEE, 2015: 2641-2649.

[17] HUDSON D A, MANNING C D. GQA: a new dataset for real-world visual reasoning and compositional question answering[C]//Proceedings of the 2019 IEEE/CVF conference on computer vision and pattern recognition. Piscataway:IEEE , 2019: 6700-6709.

[18] ANTOL S, AGRAWAL A, LU J, et al. VQA: Visual question answering[C]//Proceedings of the 2015 IEEE international conference on computer vision. Piscataway:IEEE, 2015: 2425-2433.

[19] KRISHNA R, ZHU Y, GROTH O, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations[J]. International journal of computer vision, 2017, 123(1): 32-73.

[20] CHEN Y C, LI L, YU L, et al. Uniter: Universal image-text representation learning[C]//Proceedings of the 2020 European conference on computer vision(ECCV). Cham: Springer, 2020: 104-120.

[21] QI D, SU L, SONG J, et al. ImageBERT: Cross-modal pre-training with large-scale weak-supervised image-text data[EB/OL].[2020-01-23]. https://arxiv.org/pdf/2001.07966.pdf.

[22] ZHU Y, KIROS R, ZEMEL R, et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books[C]//Proceedings of the 2015 IEEE international conference on computer vision. Piscataway:IEEE, 2015: 19-27.

[23] LU J, GOSWAMI V, ROHRBACH M, et al. 12-in-1: multi-task vision and language representation learning[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE, 2020: 10437-10446.

[24] SINHA K, JIA R, HUPKES D, et al. Masked language modeling and the distributional hypothesis: order word matters pre-training for little[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroundsburg,PA:Association for Computational Linguistics, 2021: 2888-2913.

[25] ZHOU L, PALANGI H, ZHANG L, et al. Unified vision-language pre-training for image captioning and vqa[C]//Proceedings of the 2020 AAAI Conference on Artificial Intelligence.Menlo Park, CA:AAAI Press, 2020: 13041-13049.

[26] MENG R, RICE S G, WANG J, et al. A fusion steganographic algorithm based on faster R-CNN[J]. Computers, Materials & Continua, 2018, 55(1): 1-16.

[27] HUANG Z, ZENG Z, LIU B, et al. Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers[EB/OL].[2020-06-22]. https://arxiv.org/pdf/2004.00849.pdf.

[28] FRANK S, BUGLIARELLO E, ELLIOTT D. Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroundsburg, PA:Association for Computational Linguistics, 2021: 9847-9857.

[29] LEE K H, CHEN X, HUA G, et al. Stacked cross attention for image-text matching[C]//Proceedings of the 2018 European Conference on Computer Vision (ECCV). Cham: Springer 2018: 201-216.

[30] LAHAT D, ADALI T, JUTTEN C. Multimodal data fusion: an overview of methods, challenges, and prospects[J]. Proceedings of the IEEE, 2015,

103(9): 1449-1477.
[31] GAO J, LI P, CHEN Z, et al. A survey on deep learning for multimodal data fusion[J]. Neural Computation, 2020, 32(5): 829-864.
[32] BALAKRISHNAMA S, GANAPATHIRAJU A. Linear discriminant analysis-a brief tutorial[J]. Institute for Signal and information Processing, 1998, 18(1998): 1-8.
[33] NEFIAN A V, LIANG L, PI X, et al. Dynamic Bayesian networks for audio-visual speech recognition[J]. EURASIP Journal on Advances in Signal Processing, 2002, 2002(11): 1-15.
[34] MARTÍNEZ H P, YANNAKAKIS G N.Deep multimodal fusion: Combining d-iscrete events and continuous signals[C]//Proceedings of the 16th Inter-national conference on multimodal interaction.New York:ACM, 2014: 34-41.
[35] EDDY S R. What is a hidden Markov model?[J]. Nature biotechnology, 2004, 22(10): 1315-1316.
[36] HEARST M A, DUMAIS S T, OSUNA E, et al.Support vector machines[J]. IEEE Intelligent Systems and their applications, 1998, 13(4): 18-28.
[37] TURK M. Multimodal interaction: a review[J].Pattern recognition letters, 2014, 36: 189-195.
[38] LI X, SONG J, GAO L, et al. Beyond RNNs: positional self-attention with co-attention for video question answering[C]//Proceedings of the 2019 AAAI Conference on Artificial Intelligence. Menlo Park,CA:AAAI Press, 2019: 8658-8665.
[39] OWENS A, ISOLA P, MCDERMOTT J, et al. Visually indicated sounds[C]//Proceedings of the 2016 IEEE conference on computer vision and pattern recognition. Piscataway:IEEE, 2016: 2405-2413
[40] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems.New York : Curran Associates Inc, 2017: 6000-6010.
[41] LI L, CHEN Y C, CHENG Y, et al. HERO: Hierarchical Encoder for Video+ Language Omni-representation Pre-training[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroundsburg,PA:Association for Computational Linguistics, 2020: 2046-2065.
[42] YAMADA K, SASANO R, TAKEDA K. Semantic frame induction using masked word embeddings and two-step clustering[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Stroundsburg, PA:Association for Computational Linguistics, 2021: 811-816.
[43] KAMIGAITO H, HAYASHI K. Unified interpretation of softmax cross-entropy and negative sampling: with case study for knowledge graph embedding[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroundsburg, PA:Association for Computational Linguistics, 2021: 5517-5531.
[44] JANG E, GU S, POOLE B. Categorical reparameterization with gumbel-softmax[EB/OL].[2017-08-05]. https://arxiv.org/pdf/1611.01144.pdf.
[45] WU Y, SCHUSTER M, CHEN Z, et al. Google's Neural machine translation system: bridging the gap between human and machine translation[EB/OL].[2016-09-26]. https://arxiv.org/pdf/1609.08144.pdf.
[46] LI Z, FAN Z, TOU H, et al. MVP: Multi-stage vision-language pre-training via multi-level semantic alignment[EB/OL].[2022-02-28]. https://arxiv.org/pdf/2201.12596.pdf.
[47] XIE S, SUN C, HUANG J, et al. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification[C]//Proceedings of the 2018 European conference on computer vision (ECCV). Cham: Springer, 2018: 305-321.
[48] SUN Y, WANG S, LI Y, et al. ERNIE: Enhanced representation through knowledge integration[EB/OL].[2019-04-19]. https://arxiv.org/pdf/1904.09223.pdf.
[49] SUN Y, WANG S, LI Y, et al. ERNIE 2.0: A continual pre-training framework for language understanding[C]//Proceedings of the 2020 AAAI Conference on Artificial Intelligence. Menlo Park,CA:AAAI Press, 2020, 34(05): 8968-8975.
[50] YU F, TANG J, YIN W, et al. ERNIE-ViL: Knowledge enhanced vision-language representations through scene graphs[C]//Proceedings of the 2021 AAAI Conference on Artificial Intelligence. Menlo Park,CA:AAAI Press, 2021, 35(4): 3208-3216.
[51] ZELLERS R, BISK Y, FARHADI A, et al. From recognition to cognition: Visual commonsense reasoning[C]//Proceedings of the 2019 IEEE/CVF conference on computer vision and pattern recognition. Piscataway:IEEE, 2019: 6720-6731.
[52] YANG S, LI G, YU Y. Cross-modal relationship inference for grounding referring expressions[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE, 2019: 4145-4154.
[53] ZHANG Q, LEI Z, ZHANG Z, et al. Context-aware attention network for image-text retrieval[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE, 2020: 3536-3545.
[54] ZHU L, YANG Y. ActBERT: Learning global-local video-text representations[C]//Proceedings of the 2020 IEEE/CVF conference on computer vision and pattern recognition. Piscataway:IEEE, 2020: 8746-8755.
[55] LEI J, YU L, BANSAL M, et al. TVQA: Localized, compositional video question answering[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroundsburg, PA:Association for Computational Linguistics, 2018: 1369-1379.
[56] MIECH A, ZHUKOV D, ALAYRAC J B, et al. Howto100m: learning a text-video embedding by watching hundred million narrated video clips[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Piscataway:IEEE, 2019: 2630-2640.
[57] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(6): 1137-1149.
[58] YU Z, YU J, CUI Y, et al. Deep modular co-attention networks for visual question answering[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE, 2019: 6281-6290.
[59] ZENGGANG X, ZHIWEN T, XIAOWEN C, et al. Research on image retrieval algorithm based on combination of color and shape features[J]. Journal of Signal Processing Systems, 2021, 93(2): 139-146.
[60] GABEUR V, SUN C, ALAHARI K, et al. Multi-modal transformer for video retrieval[C]// Proceedings of the 2020 European Conference on Computer Vision (ECCV). Cham: Springer, 2020: 214-229.
[61] WANG B, MA L, ZHANG W, et al. Reconstruction network for video captioning[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE, 2018: 7622-7631.
[62] LEI J, YU L, BERG T L, et al. TVR: a large-scale dataset for video-subtitle moment retrieval[C]// Proceedings of the 2020 European Conference on Computer Vision(ECCV). Cham: Springer, 2020: 447-463.
[63] KAZEMZADEH S, ORDONEZ V, MATTEN M, et al. Referitgame: Referring to objects in photographs of natural scenes[C]//Proceedings of

the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroundsburg, PA:Association for Computational Linguistics, 2014: 787-798.

[64] YOUNG P, LAI A, HODOSH M, et al. 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.

[65] ZHOU L, XU C, CORSO J J. Towards automatic learning of procedures from web instructional videos[C]//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. Menlo Park,CA:AAAI Press, 2018: 7590-7598.

[66] YAO Y, ZHANG A, ZHANG Z, et al. CPT: Colorful prompt tuning for pre-trained vision-language models[EB/OL].[2021-10-08]. https://arxiv.org/pdf/2109.11797.pdf.

[67] ZHAO Z Q, ZHENG P, XU S, et al. Object detection with deep learning: A review[J]. IEEE transactions on neural networks and learning systems, 2019, 30(11): 3212-3232.

[68] HONG G, KIM Y, CHOI Y, et al. BioPREP: Deep Learning-based Predicate Classification with SemMedDB[J]. Journal of Biomedical Informatics, 2021, 122: 103888-103888 .

[69] LI J, LIN D, WANG Y, et al. Deep discriminative representation learning with attention map for scene classification[J]. Remote Sensing, 2020, 12(9): 1366.

[70] LIN J, MEN R, YANG A, et al. M6: A Chinese Multimodal Pretrainer[EB/OL].[2021-05-29]. https://arxiv.org/pdf/2103.00823.pdf.

[71] van DEN OORD A, VINYALS O, KAVUKCUOGLU K. Neural discrete representation learning [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. New York : Curran Associates Inc, 2017: 6309-6318.

[72] KASBAN H, HASHIMA S. Adaptive radiographic image compression technique using hierarchical vector quantization and Huffman encoding[J]. Journal of Ambient Intelligence and Humanized Computing, 2019, 10(7): 2855-2867.

[73] YAN M, XU H, LI C, et al. Achieving Human Parity on Visual Question Answering[EB/OL].[2021-11-19]. https://arxiv.org/pdf/2111.08896.pdf.

[74] CAI W, SONG Y, WEI Z. Multimodal data guided spatial feature fusion and grouping strategy for e-commerce commodity demand forecasting[EB/OL].[2021-12-06]. https://downloads.hindawi.com/journals/misy/2021/5568208.pdf.

[75] DESHPANDE A M, KALBHOR S R.Video-based Marathi Sign Language Recognition and Text Conversion Using Convolutional Neural Network[M]. Singapore:Springer, 2020: 761-773.

[76] HAN X, ZHANG Z, DING N, et al. Pre-trained models: Past, present and future[J]. AI Open, 2021, 2: 225-250 .

This work is partially supported by National Language Commission Key Project (ZDI135-96).

WANG Huiru born in 1996. M. S. candidate. Her research interests include natural language processing, image processing.

LI Xiuhong born in 1977. Ph. D., associate professor. Her research interests include Natural language processing, image processing.

LI Zhe born in 1992. Ph. D. candidate. His research interests include natural language processing, public opinion analysis.

MA Chunming born in 1997. M. S. candidate. His research interests include natural language processing, event extraction.

REN Zeyu born in 1998. M. S. candidate. His research interests include speech recognition, image processing.

YANG Dan born in 1996. M. S. candidate. Her research interests include natural language processing, image processing.