引用格式：朱若琳，蓝善祯，朱紫星．视觉-语言多模态预训练模型前沿进展[J]．中国传媒大学学报（自然科学版），2023，30(01)：66-74.
文章编号：1673-4793(2023)01-0066-09

视觉-语言多模态预训练模型前沿进展

朱若琳*，蓝善祯，朱紫星
（中国传媒大学信息与通信工程学院，北京 100024）

摘要：近年来，多模态预训练学习在视觉-语言任务上蓬勃发展。大量研究表明，多个模态特征的表征学习预训练有利于视觉-语言下游任务的效果提升。多模态表征预训练旨在采用自监督的学习范式，包括对比学习，掩码自监督等，在大规模的图文相关性数据上进行训练，通过学习模态自身与模态间的知识先验，使模型获得通用的、泛化性较强的视觉表征能力。后BERT时代，本文介绍了视觉多模态领域基于Transformer的相关工作；对主流多模态学习方法的发展脉络进行梳理，分析了不同方法的优势和局限性；总结了多模态预训练的各种监督信号及其作用；概括了现阶段主流的大规模图像-文本数据集；最后简要介绍了几种相关的跨模态预训练下游任务。

关键词：多模态预训练；视觉-语言预训练；表征学习
中图分类号：TP391.4　**文献标识码**：A

A survey on Vision-Language multimodality pre-training

ZHU Ruolin*，LAN Shanzhen，ZHU Zixing
（Communication University of China，Beijing 100024, China）

Abstract: Multimodal pre-training has shown increased interest on vision-language tasks. Recent comprehensive studies have demonstrated that, multimodal representations training can benefit the Vision-Language downstream tasks. Multimodal pre-training requires a large-scale training data and self-supervised learning. This paper reviews some significant transformer-base researches about Vision-Language (VL) pre-training, which came out after BERT. Firstly, the application background and development significance of multimode pretraining are expounded. Secondly, this paper introduces the development of mainstream multimodal networks and analyzes the advantages and disadvantages of methods. Then, we explain cost functions used in multi-task pre-training. Next, We then illustrate the large-scale image-text database mentioned in recent studies. In the end, combining different VL downstream tasks, this paper describes the task objectives, datasets and training methods.

Keywords: multimodal pre-training; Vision-Language (VL) training; representation learning

1　引言

大规模预训练模型的泛化性表征可以迁移到各种下游任务，因此多模态预训陷入局部最优。近期，自然语言处理领域，出现BERT（Bidirectional Encoder Representation from Transformers）等一系列的大规模预训练工作，将语言掩码技术应用到Transformer的预

基金项目：国家重点研发计划（2018YFB1404103）
作者简介（*为通讯作者）：朱若琳（1994-09），女，博士研究生，主要从事视频理解研究。Email：zhuruolin@cuc.edu.cn；蓝善祯，中国传媒大学信息与通信工程学院，博士，副教授，主要从事数字视频技术和视觉信号处理研究；朱紫星，中国传媒大学信息与通信工程学院，硕士研究生，主要从事视频理解研究。

训练中，从而获得强泛化能力的特征。受到的启发，一系列的视觉-语言预训练工作应运而生，如CLIP[2]、VL-BERT[3]、ALING[4]。并且这些多模态特征的预训练在视觉-语言下游任务中也表现出较强的迁移能力。

图文检索[5][6][7]、视觉问答[8][9][10]、视觉推理任务[11][12][13]、视觉分割[14]、图文生成[15][16]中，多模态的预训练特征都超越旧有的单一模态框架和非预训练的研究方案。先前的工作一般直接在目标数据集上进行训练，获得的特征缺乏泛化性。不同的任务之间都需要重新进行特征训练。换言之，不同任务之间训练目标差异化较大，网络特征的迁移能力具有局限性。在训练阶段，先前的训练方法容易出现过拟合。多模态预训练相比旧有的方法，模型训练效果得到了提升。

目前多模态预训练面临几个问题：(1)为了保证深度神经网络充分训练以及模型的泛化能力，需要一个上亿体量的数据集。构建数据集本身存在一定难度，最常见的方式是网上爬取得到的图像文本对，不可避免的是如此获得的样本会存在大量的噪声。降低噪声对网络训练的影响成为提升预训练模型的关键之一；(2)视觉特征与文本特征不同，图像像素是连续性的变化。必须将视觉特征离散化处理输入到Transformer中，这里离散化处理的方式会影响到最终的模态融合；(3)多个模态之间特征对齐的监督信号，需从高层语义出发将两个模态进行对齐。促进模态融合的监督信号的设计也是多模态训练的难点所在；(4)最为重要的是训练资源。无论是海量多模态数据集，还是Transformer的训练都是十分耗费训练资源的。

2 多模态预训练相关进展

近两年，伴随着自然语言处理的Transformer预训练模型的发展，如ELMo[17]、BERT[18]、GPT[19]等，推动了自然语言处理任务的革新，基于Transformer的多模态领域发展也突飞猛进。视觉-语言预训练学习VLP(Vision-and-Language Pre-training)[3][6][15][20][21][22]是指基于海量图像-文本对数据训练跨模态的通用表征，得到的预训练模型可以直接微调适配下游视觉-语言任务。

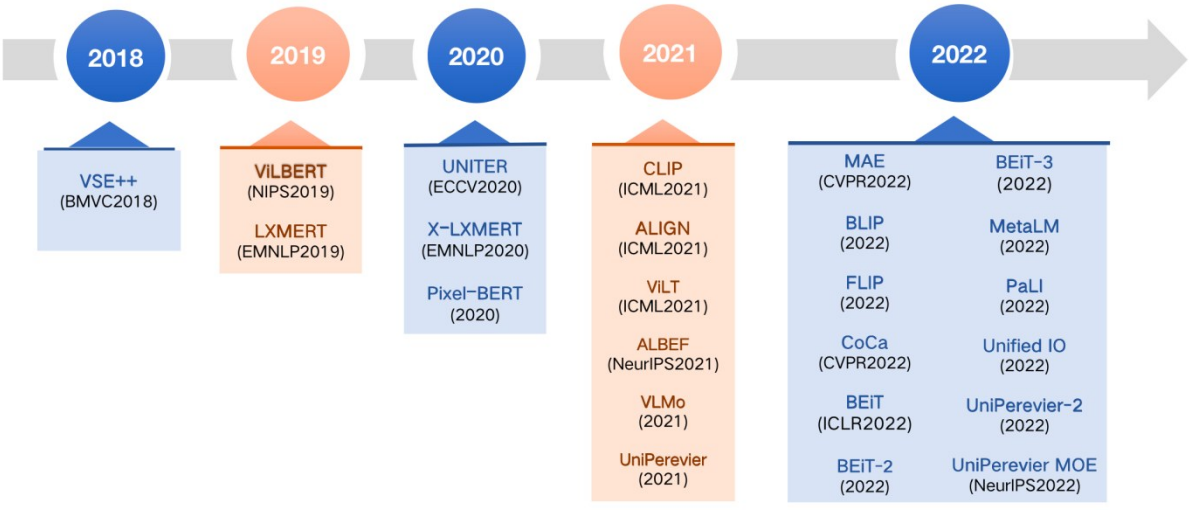

图1 视觉-语言多模态预训练模型发展历程

(1)双塔编码

双塔编码主要关注图像和文本的各自模态编码的表征对齐，采用最简单点乘融合特征。目前热点模型如CLIP[2]和ALIGN[4]等，这类方法使用了对比学习方式来训练大量网络噪声数据，采用余弦相似性来度量模态间的距离。这类方法的模态融合方式属于轻量型特征融合，如对于检索任务，底库特征提取可提前离线化存储，可以实现快速的多模态特征的检索。

早期工作有VSE0、VSE++[23]等任务使用图像检测技术将图像特征进行离散化处理，再联合自然语言处理领域成熟的文本编码器BERT[18]处理文本表征，最终通过轻量化融合实现特征表征。随后，OpenAI提出CLIP[2]，将视觉-语言预训练任务推向高潮，CLIP成为当下最常用的表征学习方法之一。在此基础上

的优化算法层出不穷,如Towhee技术团队的BLIP[16]是可清洗数据噪声的图像文本多任务预训练框架;近期Meta AI何凯明团队的FLIP[24]提出融合MAE[25]中的图像文本双掩码技术,训练阶段有效地提升了速度,并且模型精度远超CLIP。当然,CLIP也有一定局限性,由于特征融合机制太过简单,导致在复杂任务图像文本理解任务上表现一般。

(2)融合编码

融合编码框架使用Transformer进行跨模态融合。受到VSE++[23]双塔结构的启发,UNITER[26]在检测器输出区域特征编码和文本编码的基础上,使用Transformer进行多模态的融合编码。此前,Transformer通常用来编码单模态特征,UNITER提出跨模态的条件掩码技术和多任务学习,不同于以往双塔模型,联合使用语言掩码模型(MLM)、图像区域掩码(MRM)、图像文本匹配(ITM)和词汇区域对齐(WRA)四个监督任务,通过条件掩码监督强化模态间的特征对齐。ViLBERT[27]和LXMERT[28]提出使用三个Transformer分别进行图像编码、文本编码、特征融合。增加了融合阶段的网络深度后,混合编码模型框架在视觉-语言下游理解和生成类任务中都表现出优异的表征力。这类算法受限于网络训练和推理速度,并未得到工业界的广泛应用。限制这些框架应用的原因包括:第一,图像编码阶段都是对图像检测预训练模型Faster R-CNN[29]提取的区域特征进行Transformer编码,视觉特征的离散化处理需要更加简单的处理;第二,相较CLIP[2]工作,由于特征融合阶段不再是轻量化模型,所以在处理图像检索任务时,需要对图像和文本同时进行编码,推理速度受到影响。

为了优化第一点限制,研究者相继提出了X-LXMERT[30]和Pixel-BERT[31]等,移除检测预训练模型,使用卷积神经网络CNN的网格化特征输入Transformer进一步编码。然而,卷积神经网络CNN带来一定计算量,影响推理速度。ViLT[32]针对推理速度问题进行了优化:取代检测模型和卷积神经网络CNN,提出极简化的网络设计,除模态融合网络外没有采用额外的特征提取网络,参考ViT[33],将图像直接进行分块投影后输出到Transformer中,实现了极简化的图像处理。实验显示,该方法在参数量和运行时间上都能明显降低,模型效果明显优于Pixel-BERT等融合编码框架,但是较CLIP[2]双塔框架还是有一定差距。

直观上讲视觉编码要比文本编码更加复杂,ViLT[32]在视觉编码的处理上过于简化。Salesforce Research提出ALBEF[22],采用比语言编码器更深的视觉编码器处理视觉特征,并提出使用图像文本对比学习的监督信号,在单模态特征输入融合模块前进行特征对齐,针对噪声数据提出采用动量蒸馏的方法进行自监督训练。ALBEF[22]在图像检索领域反超CLIP[2]和ALIGN[4],并在多模态预训练任务上一定程度上缓解了物理训练成本的问题。在视觉问答VQA和视觉推理NLVR的任务上达到了SOTA效果。

后续一系列工作都是在ALBEF[22]基础上展开的,如CoCa[34]、VLMo[35]、BLIP[16]等。为了在不同的任务上提升图像特征的可学习性,Google的CoCa使用了注意力池化层作为可学习参数应用到图像特征处理上。CoCa还针对ALBEF、VLMo[35]等网络存在训练效率问题进一步优化,去除图文对比监督,文本端都进行掩码,文本侧Transformer只需要一次推理。

(3)其它编码

VLMo[35]兼顾了前两种框架的优点,灵活切换双塔编码和融合编码结构,提出将原始Transformer子模块替换为多专家模块MoME。原始的单个Transformer子模块中仅包含单个前向注意力层FFN,MoME则是包含三个前向注意力层,分别关注到视觉、语言和融合模态。这样通过采用控制前向注意力层,可以实现灵活地切换双塔编码和融合编码框架,兼顾了两个编码框架的优点。同时提出了分阶段训练的思路,一定程度上解决了图像-文本对标注数据补充的问题。

MoME的核心思想是共享不同模态间的自注意力层参数,后续的BLIP[16]模型上也保留了这一思想。BLIP的网络设计中大量共享了自注意力和交叉注意力层的参数。即便在增加了生成器的情况下,模型也没有增加大量的参数。

在CoCa[34]提出精简监督信号思想之后,微软相继提出BEiT-2[36]、BEiT-3[37]。BEiT-3提出将图像看成一种"语言",文本、图像和图像-文本融合模态都使用Multiway Transformer[35]学习,Multiway Transformer由VLMo中的MoME组成。不同于以往的视觉-语言预训练模型都会采用多任务训练,BEiT-3只训练统一的生成任务,因而仅使用图像和文本的掩码建模。实验证明,单一监督信号也能实现了SOTA的迁移能力。

近期研究工作出现一系列基于语言指导网络学习"Image-and-text to Text"的框架,例如微软的MetaLM[38]和Google的PaLI[39][40],本质上是一种文本生成类框架。通过Prompt控制网络进行不同训练任务,

给出相应的文本输出，以跨任务的迁移能力来进行训练成本。

此外，以Unified IO为代表的通用模型也相继提出，UniPerevier[41]、UniPerevier-2[42]、UniPerevier MOE[43]等。Unified IO[44]执行多种AI任务，包含图像生成、目标检测、深度估计、姿态估计等机器视觉任务，也能执行自然语言处理的问答和推理任务。在下游任务迁移时不需要添加额外的结构，即可对预训练阶段没有使用过的数据或者任务进行零样本推理。然而，UniPerevier MOE[43]还不能达到SOTA效果。

3 模态融合监督信号

在多模态预训练任务中，为了提升表征的迁移能力大多采用多任务训练方式。目前主流的方法中提及的多任务监督信号包含如下几种：

（1）图文对比监督（ITC）

图文对比监督（ITC）目标是学习更佳的单一模态间的特征表达，一般作用于单模态特征编码阶段。在双塔模型里是最重要的多模态监督信号，用于监督对齐两个模态的信息。网络训练阶段，给出N个图像样本对，预测N×N个可能的样本对之间图像文本是否匹配。ITC主要学习一个相似度量，对于每对图像文本，分别计算图像到文本的相似度和文本到图像的相似度，并进行softmax归一化处理，计算公式如下：

$$s_{i,j}^{I2T}=\hat{h}_i^{v\top}\hat{h}_j^{w},\ s_{i,j}^{T2I}=\hat{h}_i^{w\top}\hat{h}_j^{v} \tag{1}$$

$$p_i^{I2T}(I)=\frac{exp(s_{i,i}^{I2T}/\sigma)}{\sum_{j=1}^{M}exp(s_{i,j}^{I2T}/\sigma)} \tag{2}$$

$$p_i^{T2I}(I)=\frac{exp(s_{i,i}^{T2I}/\sigma)}{\sum_{j=1}^{M}exp(s_{i,j}^{T2I}/\sigma)} \tag{3}$$

其中，$\left\{\hat{h}_i^{v}\right\}_{i=1}^{N}$和$\left\{\hat{h}_i^{w}\right\}_{i=1}^{N}$分别表示图像和文本特征向量。$s_{i,j}^{I2T}$代表第$i$张图像特征$\hat{h}_i^{v}\epsilon\mathbb{R}^{D}$到第$j$个文本特征之间的相似度，$s_{i,j}^{T2I}$代表文本到图像的相似度。$\sigma$是一个可学习的参数。公式（2）和（3）是softmax归一化处理。

最终，ITC的损失函数定义为：

$$\mathscr{L}_{ITC}=\frac{1}{2}E_{(I,T)\sim D}[H(y^{I2T}(I),p^{I2T}(I))+H(y^{T2I}(I),p^{T2I}(I))] \tag{4}$$

$y^{I2T}(I)$和$y^{T2I}(I)$是对真值标签的one-hot编码，其中负样本对的概率为0，正样本对概率为1。$H(\cdot)$表示交叉熵损失函数。

（2）图文匹配监督（ITM）

图像文本匹配训练的目标是预测图像和文本是否描述一致，可以看作一个二分类问题。通常采用Transformer输出层的融合特征向量输入分类器，使用交叉熵损失函数监督。

$$\mathscr{L}_{ITM}=E_{(I,T)\sim D}H(y^{itm},p^{itm}(I,T)) \tag{5}$$

y^{itm}是二维的one-hot编码，代表正负（匹配和不匹配）两个标签。图文匹配在训练阶段相较其它任务的训练目标更简单，收敛更快。所以，在很多算法中都进行了困难负样本对的挖掘[22]，挖掘一些与图像描述相近的负样本进行进一步的预训练。这些负样本虽然同真实标签语义上相似，但是仍存在一些细粒度上的差别。反之，也可以挖掘一些与文本描述相近的负样本图像，这样能保证在多任务训练阶段图像文本监督信号得到充分的训练。

（3）语言掩码监督（MLM）

参照BERT[18]，语言掩码随机选择文本描述中部分单词，将其用特殊标志符［MASK］替换。模型在训练过程中会读取图像信息和文本中剩余词汇，预测出掩码区域对应的单词。语言掩码监督（MLM）为最小化交叉熵损失函数，可以进一步促进文本学习视觉模态的信息，从而完成多模态间语义对齐。

$$\mathscr{L}_{MLM}=E_{(I,\hat{T})\sim D}H(y^{msk},p^{msk}(I,\hat{T})) \tag{6}$$

其中，y^{msk}指one-hot编码的分布，对应正确词汇标签位置的概率值为1。这里的掩码对象按照一定掩码概率对整个句子随机选择，其中掩码概率为可调参数。对于掩码对象，不仅可以对单词进行掩码，也可以直接对Transformer输入单元掩码。此外，在多模态训练中语言掩码一般同图文对比监督（ITC）和图文匹配监督（ITM）同时使用，但是语言掩码监督是对掩码后的文本提取特征，而不是对原始文本。因此，同时使用语言掩码和其它两种监督信号时需要做两次甚至更多次前向推理[22]，这也给预训练带来了一些资源开销。

（4）文本词汇图像区域对齐（WRA）

文本词汇图像区域对齐（WRA）使用了最优传输理论（Optimal Transport，OT）[26]，用于显式地对齐细粒度的文本和图像区域。特别说明，此处提及的图像区域指的是图像检测中定位网络RPN的输出。首先，将文本T和图像区域I转换成两个离散分布(u,v)，这两个分布满足如下形式：

$$u=\sum_{i=1}^{N}a_i\delta_{T_i} \tag{7}$$

$$v = \sum_{j=1}^{K} b_j \delta_{I_j} \tag{8}$$

$$\sum_{i=1}^{N} a_i = \sum_{j=1}^{K} b_j = 1 \tag{9}$$

其中δ_{T_i}是T_i的狄拉克函数中心，$a = \{a_i\}_{i=1}^{N} \in \Delta_N$和$b = \{b_j\}_{j=1}^{K} \in \Delta_K$分别是N和K维的权重矩阵，$u$和$v$是概率分布。

$$\Pi(a,b) = \{P \in \mathbb{R}_+^{N \times K} | P1_m = a,\ P^{\top}1_n = b\} \tag{10}$$

其中，1_n是一个n维的单位阵，矩阵$P \in \mathbb{R}^{N \times K}$是两个模态$u$和$v$之间的转移矩阵。想要精准的计算出两个模态的转移关系是十分困难的，所以一般采用IPOT算法[45]进行近似最优化路径距离。近似计算得到转移矩阵，因此有WRA监督信号如下：

$$\mathscr{L}_{WRA}(\theta) = E_{(I,T)\sim D} H(y^{wra}, p^{wra}(I,T)) = \min_{P \in \Pi(a,b)} \sum_{i=1}^{N}\sum_{j=1}^{K} P_{ij} \cdot c(T_i, I_j) \tag{11}$$

其中$c(T_i, I_j)$是余弦距离度量T_i和I_j之间的距离。伴随着多模态预训练的发展，在ViLT[32]中将文本词汇图像区域对齐(WRA)优化为文本词汇图像子块对齐(WPA)，计算文本子集和图像块子集的对齐得分。

(5)图像区域掩码(MRM)

参考语言掩码监督(MLM)，有研究者提出在视觉模态也可以进行掩码处理[2, 12, 13]。以UNITER[26]提出的图像区域掩码(MRM)监督信号设计为例，文本掩码采用特殊字符[MASK]替换被掩码的字符，图像掩码则采用对掩码区域的值置零[26]。不同于文本掩码中是采用词汇类别直接进行监督，图像是更高维的连续信号，需要采用更加复杂的监督信息。三种图像区域掩码(MRM)的变体如下：

$$\mathscr{L}_{MRM} = E_{(\hat{I},T)\sim D} H(y^{mrm}, p^{mrm}(\hat{I},T)) \tag{12}$$

a)图像掩码区域特征回归(MRFR)

采用回归的思想，掩码后的特征$h_\theta(\hat{v}_i)$和输入的感兴趣区域的特征$r(\hat{v}_i)$进行L2回归。

$$H(y^{mrm}, p^{mrm}(\hat{I},T)) = \sum_{i=1}^{M} \left\| h_\theta(\hat{v}_i), r(\hat{v}_i) \right\|_2^2 \tag{13}$$

b)图像掩码区域分类(MRC)

顾名思义，就是对被掩码区域进行语义分类，将Transformer输出的特征输入到分类器中得到一个K类的预测，归一化的分布为$g_\theta(\hat{v}_i) \in \mathbb{R}^K$。这里的真值标签采用Faster R-CNN[29]预测结果中最高置信度的类别，one-hot编码后为$c(\hat{v}_i) \in \mathbb{R}^K$。这是因为区域定位网络RPN的输出并没真实标注信息。最终，即可以使用分类中常见的交叉熵损失CE (cross-entropy loss)。

$$H(y^{mrm}, p^{mrm}(\hat{I},T)) = \sum_{i=1}^{M} CE(c(\hat{v}_i), g_\theta(\hat{v}_i)) \tag{14}$$

c)基于KL散度的图像掩码区域分类(MRC-kl)

参照MRC，MRC-kl将单一的分类标签转化为软标签，即直接使用Faster R-CNN的预测结果作为标签$\tilde{c}(\hat{v}_i) \in \mathbb{R}^K$。通过最小化两个分布间的KL散度，实现提取软标签中蕴含的知识。

$$H(y^{mrm}, p^{mrm}(\hat{I},T)) = \sum_{i=1}^{M} D_{KL}(\tilde{c}(\hat{v}_i) \| g_\theta(\hat{v}_i))) \tag{15}$$

伴随着多模态网络的发展，图像掩码形式呈现多样性。例如UNITER[26]的图像编码器使用了检测网络结构，对区域检测器的输出进行掩码。ViLBERT[27]和LXMERT[28]中是对卷积网络输出的图像特征进行掩码。Pixel-BERT[31]则是直接对图像像素进行随机掩码。后续的MAE[25]和FLIP[24]都对像素掩码的几个变体进行了消融实验，结果表明对图像分片后进行随机掩码效果最佳。

4 预训练数据集

(1)组合数据集

视觉-语言预训练模型研究初期，如UNITER[26]以及更早的工作都采用组合多个数据集，包含Conceptual Captions[46]、SBU Captions(1m)[47]、COCO[48]、Visual Genome(VG)[49]四个数据集，总计400万张图像，500万个图像文本对，其中COCO和VG中单张图像可能对应多个文本。这些数据集均通过人工注释生成，如COCO Captions在COCO图片数据基础上，由人工标注图片描述得到。Visual Genome是李飞飞2016年发布的大规模图片语义理解数据集，含图像和问答数据，标注密集，语义多样。这两个数据集主要用于图像生成描述，然而由于图片数量较少，仅有330k和5M对，模型发展受到限制。

(2)网络噪声数据

CLIP[2]、ALIGN[3]等图文多模态预训练方法证实，对于大规模多模态数据，甚至不需要进行人工标注，自监督或弱监督训练模型也能超越有监督训练。除了本身的模型优化之外，目前的进展多依赖底层的上亿对图文数据。以CLIP预训练数据WIT为例，现有数据集规模不够大是导致目前多模态预训练方法不能得到充分训练的原因。因此，WIT网上爬取了4亿个图像文本对用于训练。由于网络爬取的数据集并未进行政治宗教的过滤，所以WIT数据集并未对外

公开。

网络爬取的数据包含大量噪声，这一点在后续多模态预训练工作中都有提及。网上爬取得到的图像文本对，文本描述不能很好的表征图像全部元素，称之为替代文本Alt Text[22][16]。这是由于搜索引擎更加关注的是关键词和有商业价值的词汇，而不是文本是否能全面描述图像。噪声数据的处理方法有很多种，ALBEF[22]中使用动量蒸馏的方法来降低噪声对训练的影响。BLIP[16]中使用在COCO数据集上微调过的编码器来清洗噪声数据，并使用生成器生成一个新的文本描述来扩充图像文本描述；巧妙地应用了分阶段训练思想，使用有噪声的数据集训练预训练模型，然后通过有标注的数据集微调后的模型清洗和扩充数据。

（3）公开数据集

LAION团队提出了LAION 400 million[50]/LAION 5 billion数据集。官方提供了在该数据集上预训练的模型，并附上预先计算的向量和搜索功能。LAION-5B通过Common Crawl爬虫工具获取文本和图片，使用CLIP[2]计算获取图像和文本的相似性，并删除相似度低于设定阈值的图文对（英文阈值0.28，其余阈值0.26），500亿图片保留了不到60亿，最后形成58.5亿个图文对，包括23.2亿的英语，22.6亿的100+语言及12.7亿的未知语言。

后续，LAION-5B的基础上应用了BLIP生成能力推出的LAION COCO。首先，对单张图像生成40个文本描述，并使用CLIP对文本描述排序；使用OpenAI的Vision Transformer Large选出最好的5个文本描述；最后，用OpenAI的Resnet50预训练模型选出最佳描述。

5 下游任务的迁移

下游任务的迁移主要包括微调和线性预测。通过微调策略修改预训练模型执行全新的视觉-语言下游任务，在下游任务训练中仅仅做了微调，甚至有些任务只是调整了分类器。无论是哪一种方法，都尽可能少的调整模型表征部分，保留了预训练得到的表征的泛化性，以免模型陷入局部最优解。

（1）视觉问答（VQA）

视觉问答任务[5][8][9][10]需要观察图像回答一个自然语言的问题。视觉问答任务分为闭集和开集两种：闭集就是给定一个回答的可能集合，从中选择正确的回答，可以看作多分类问题[51]；开集视觉问答则是通过网络生成回答的自然语言句子，本质上是生成问题，相对闭集问答更加复杂。通常用多模态预训练网络初始化编码器，并使用语言模型中损失函数监督微调网络参数。

以VQA2.0[52]数据集为例，总共包含110万个问题，每个问题对应10个回答集合。每个问题都与其语义的结构化表示相关联，并且约束应答者必须采用特定的推理步骤完成回答。许多VQA问题涉及空间理解和多步推理等，具有一定挑战性。数据集严控数据的平衡性和不同问题组的答案分布，以免使用语言和先验信息进行猜测。下游任务微调时，需要训练两个多层感知器，分别处理文本和图像，最终输出尽可能涵盖3129种可能的回答。参考文献[51]中，将VQA问题转化成一个多标签分类任务。

（2）视觉推理（NLVR）

视觉推理任务[11][12][13]用于描述一句文本是否能正确描述一对图像，为二分类问题。相比传统的VQA问题，视觉推理是要让问题难度提升，必须经过推理才能回答。OSCAR[21]中提及可采用将三元组拆分成两个图像文本对的形式，每一个都包含一个文本和一张图像，利用预训练模型提取的多模态特征直接送入分类器进行分类预测。

（3）视觉蕴含（SNLI-VE）

视觉蕴含[5][53][54]是一个细粒度的视觉推理任务，给定一个假设或者前提，判断推理出其属于蕴含（假设为真）、中立（假设可能为真或假）和矛盾（假设为假）三种状态中的哪一种，本质上是三分类问题。视觉蕴含任务属于细粒度的多模态推理，必须在假设中找出至少一点视觉证据，证明假设与图像冲突。

参考文献[5]中，使用参数微调策略验证了CLIP在视觉蕴涵任务和视觉问答VQA中的零样本跨模态迁移能力。通用的方法是将预训练模型的多模态特征输入到多层感知器MLP中预测三类的分类得分。

（4）图像文本检索

图像文本检索[6][7]包含两个子任务：图像检索文本（TR）和文本检索图像（IR）。图像文本检索是跨模态检索的主要任务，其难点就在不同模态间的特征空间之间无法直接度量二者的相似性。CLIP[2]这类双塔模型最擅长此类问题，这是因为该任务目标与对比学习训练目标一致。

该类任务迁移阶段仅需要使用预训练表征，将目标图像/文本的表征与库内文本/图像特征对进行比对，按照余弦相似度给出评分。

6 结论

涉及的关键技术包括多任务训练、特征语义对齐、自监督学习、掩码等。由于自然语言处理领域的BERT[18]算法相对成熟,所以视觉-语言预训练任务中主要关注视觉模态,以及视觉-语言两种模态的融合方式。

伴随着网络的逐步优化,视觉信息的离散化可总结为三种方式:a)预训练目标检测器提取区域特征;b)使用卷积神经网络输出的图像特征网格化;c)直接对图像分块映射输入到Transformer网络中。图像分块映射方法表现更优异,推理速度最快,这种方式尽可能地保留了图像原始信息,交由Transformer网络提取特征,获得的特征与文本特征更容易进行特征对齐。如采用卷积网络提取特征,会受到感受野的限制;采用检测网络区域特征的缺点就更为明显,区域特征更多是对应文本和图像中的名词信息,损失了文本中空间位置信息和物体间的动作关系。

视觉和文本单模态的特征提取任务复杂度不同,从ALBEF[22]可以得出,视觉任务需要更深的网络来处理,相比之下文本信息已经有了比较好的预训练模型,大部分的算法都是直接使用现有模型。如果不对图像的单模态进行额外处理,例如ViLT[32]弱化图像模态特征处理,效果上达不到理想的效果。因此,融合模块前的单模态特征处理确实能有效提升预训练效果。

融合模块的设计也十分必要,双塔模型CLIP[2]的点乘融合确实有利于检索任务对时效性的要求,并且效果明显优于单模态。如ALBEF[22],ViLT[32]系列的研究都使用了Transformer进行融合,在视觉-语言理解类的下游任务中表现出极大的优势。当然,也有更好融合框架VLMo[35]继承两类框架的优势,但是也带了一定训练资源的开销,是限制多模态预训练发展的原因之一。

参考文献(References):

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database [C]. In CVPR, 2009.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision[C]//In International Conference on Machine Learning, PMLR, 2021:8748-8763.

[3] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: pretraining of generic visual-linguistic representations. In 8th International Conference on Learning Representations, ICLR, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020.

[4] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, YunHsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision [C]//In International Conference on Machine Learning. PMLR, 2021: 4904-4916.

[5] Song H, Dong L, Zhang W, Liu T, Wei F.CLIP Models are Few-Shot Learners: Empirical Studies on VQA and Visual Entailment[C]. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022:6088-6100.

[6] Chen H, Ding G, Liu X, Lin Z, Liu J, Han J. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval[C]. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020:12655-12663.

[7] Chen F, Chen X, Shi J, Zhang D, Chang J, Tian Q. HiVLP: Hierarchical Vision-Language Pre-Training for Fast Image-Text Retrieval[C]. arXiv preprint arXiv:2205.12105. 2020.

[8] Wu Q, Teney D, Wang P, Shen C, Dick A, Van Den Hengel A. Visual question answering: A survey of methods and datasets [C]. Computer Vision and Image Understanding 163, 21-40 2020.

[9] Kafle K, Kanan C. Visual question answering: Datasets, algorithms, and future challenges[C]. Computer Vision and Image Understanding 163, 3-20, 2017.

[10] Kafle K, Kanan C. An analysis of visual question answering algorithms[C]. In: Proceedings of the IEEE International Conference on Computer Vision, 2017:1965-1973.

[11] Li K, Y Zhang, K Li, et al. Visual semantic reasoning for image-text matching[C]. In ICCV, 2019: 4653-4661.

[12] Suhr A, Lewis M, et al. A corpus of natural language for visual reasoning[C]. In: ACL, 2017: 217-223.

[13] Marasovi´c A, Bhagavatula C, sung Park J, Le Bras R, Smith N A, Choi Y. Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs[C]. In: Findings of the Association for Computational Linguistics: EMNLP 2020,. 2020:2810-2829.

[14] Li B, Weinberger K Q, Belongie S, Koltun V, & Ranftl R. Language-driven Semantic Segmentation [C]. arXiv: 2201.03546. 2022.

[15] Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT pre-training of image Transformers[C]. arXiv:2106.08254, 2021.

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping languageimage pre-training for unified vision-language understanding and generation[C]. arXiv preprint arXiv:2201.12086, 2022.

[17] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations[C]. In NACCL, 2018.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding[C]. arXiv preprint arXiv: 1810.04805, 2018.

[19] Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning[R]. Technical report, OpenAI, 2018.

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision[C]//In Marina Meila and Tong Zhang. Proceedings of the 38th International Conference on Machine Learning, ICML, 18-24 July 2021, Virtual Event, volume 139 of Proceedings of Machine Learning Research, PMLR, 2021: 8748-8763.

[21] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks[C]. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm. Computer Vision -ECCV 2020 -16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX, volume 12375 of Lecture Notes in Computer Science, pages 121-137. Springer, 2020. DOI: 10.1007/978-3-030-58577-8_8.

[22] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation[C]. Advances in Neural Information Processing Systems, 2021:34.

[23] Faghri F, D J Fleet, J R Kiros, et al. VSE++: improving visual-semantic embeddings with hard negatives. In BMVC, 2018:12.

[24] Li Y, Fan H, Hu R, Feichtenhofer C, & He K. Scaling Language-Image Pre-training via Masking[C]. arXiv: 2212.00794, 2022.

[25] Alan Baade, Puyuan Peng, and David Harwath. MAEAST: Masked autoencoding audio spectrogram transformer[C]. arXiv:2203.16691, 2022.

[26] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning[C]. In ECCV, 2020.

[27] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[C]//In Hanna M Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d′Alché-Buc, Emily B Fox, and Roman Garnett. Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, 2019: 13-23.

[28] Tan H, M Bansal. LXMERT: learning cross-modality encoder representations from transformers[C]//In K Inui, J Jiang, V Ng, X Wan. EMNLP, 2019: 5099-5110.

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks[C]. In NuerIPS, 2015: 91-99.

[30] Cho J, Lu J, Schwenk D, Hajishirzi H, and Kembhavi A. X-lxmert: Paint, caption and answer questions with multi-modal transformers[C]. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020: 8785-8805.

[31] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers[C]. CoRR, abs/2004.00849, 2020.

[32] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision[C]. 2021.

[33] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C]. preprint arXiv:2010.11929, 2020.

[34] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive captioners are image-text foundation models[C]. arXiv preprint arXiv: 2205.01917, 2022.

[35] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pretraining with mixture-of-modality-experts[C]. arXiv preprint arXiv: 2111.02358, 2021.

[36] Peng Z, Dong L, Bao H, Ye Q, & Wei F. BEiT v2: Masked

Image Modeling with Vector-Quantized Visual Tokenizers[C]. arXiv 2208.06366, 2022.

[37] Wang W, Bao H, Dong L, Bjorck J, Peng Z, Liu Q, Aggarwal K, Mohammed O K, Singhal S, Som S, & Wei F. Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks[C]. arXiv2208.10442, 2022.

[38] Hao Y, Song H, Dong L, Huang S, Chi Z, Wang W, Ma S, & Wei F. Language Models are General-Purpose Interfaces[C]. arXiv:2206.06336, 2022.

[39] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer[C]. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021: 483-498.

[40] Chen X, Wang X, Changpinyo S, Piergiovanni A J, Padlewski P, Salz D, Goodman S, Grycner A, Mustafa B, Beyer L, Kolesnikov A, Puigcerver J, Ding N, Rong K, Akbari H, Mishra G, Xue L, Thapliyal A, Bradbury J, Kuo W, Seyedhosseini M, Jia C, Ayan B K, Riquelme C, Steiner A, Angelova A, Zhai X, Houlsby N, Soricut R. PaLI: A Jointly-Scaled Multilingual Language-Image Model[C]. arXiv:2209.06794, 2022.

[41] Zhu X, Zhu J, Li H, Wu X, Wang X, Li H, Wang X, & Dai J. Uni-Perceiver: Pre-training Unified Architecture for Generic Perception for Zero-shot and Few-shot Tasks[C]. arXiv: 2112.01522, 2021.

[42] Li H, Zhu J, Jiang X, Zhu X, Li H, Yuan C, Wang X, Qiao Y, Wang X, Wang W, & Dai J. Uni-Perceiver v2: A Generalist Model for Large-Scale Vision and Vision-Language Tasks[C]. arXiv 2211.09808, 2022.

[43] Zhu J, Zhu X, Wang W, Wang X, Li H, Wang X, & Dai J. Uni-Perceiver-MoE: Learning Sparse Generalist Models with Conditional MoEs[C]. arXiv. 2206.04674, 2022.

[44] Lu J, Clark C, Zellers R, Mottaghi R, & Kembhavi A. Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks[C]. arXiv2206.08916, 2022.

[45] Xie Y, Wang X, Wang R, Zha H. A fast proximal point method for Wasserstein distance[C]. In: arXiv:1802.04307, 2018.

[46] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning[C]. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 2556-2565.

[47] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs[C]//John Shawe-Taylor, Richard S Zemel, Peter L Bartlett, Fernando C N Pereira, and Kilian Q Weinberger. Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain, 2011: 1143-1151.

[48] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO Captions: Data collection and evaluation server[C]. arXiv preprint arXiv:1504.00325, 2015.

[49] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations[C]. International journal of computer vision, 2017, 123 (1):32-73..

[50] Schuhmann C, Vencu R, Beaumont R, Kaczmarczyk R, Mullis C, Katta A, Coombes T, Jitsev J, and Komatsuzaki A. Laion-400m: Open dataset of clipfiltered 400 million image-text pairs[C]. arXiv preprint arXiv:2111.02114, 2021.

[51] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering[C]. In CVPR, 2018: 6077-6086,.

[52] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering[C]. In ICCV, 2015.

[53] Xie N, Lai F, et al. Visual entailment: A novel task for fine-grained image understanding[C]. arXiv preprint arXiv: 1901.06706, 2019.

[54] Xie N, Lai F, Doran D, Kadav A. Visual entailment task for visually-grounded language learning[C]. arXiv preprint arXiv: 1811.10582, 2018.

编辑:龙学锋