软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(11):5118−5133 [doi: 10.13328/j.cnki.jos.007321] [CSTR: 32375.14.jos.007321]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于多元实体对齐的视觉-语言多模态预训练*

李 登[1], 武阿明[2], 韩亚洪[1]

[1](天津大学 智能与计算学部, 天津 300350)
[2](西安电子科技大学 电子工程学院, 陕西 西安 710401)
通信作者: 韩亚洪, E-mail: yahong@tju.edu.cn

摘 要: 视觉-语言预训练 (visual-language pre-training, VLP) 旨在通过在大规模图像-文本多模态数据集上进行学习得到强大的多模态表示. 多模态特征融合、对齐是多模态模型训练的关键挑战. 现有的大多数视觉-语言预训练模型对于多模态特征融合、对齐问题主要方式是将提取的视觉特征和文本特征直接输入至 Transformer 模型中. 通过 Transformer 模型中的 attention 模块进行融合, 由于 attention 机制计算的是两两之间的相似度, 因而该方法难以实现多元实体间的对齐. 鉴于超图神经网络的超边具有连接多个实体、编码高阶实体相关性的特性, 进而实现多元实体间关系的建立. 提出基于超图神经网络的多元实体对齐的视觉-语言多模态模型预训练方法. 该方法在 Transformer 多模态融合编码器中引入超图神经网络学习模块学习多模态间多元实体的对齐关系以增强预训练模型中多模态融合编码器实体对齐能力. 在大规模图像-文本数据集上对所提视觉-语言预训练模型进行预训练并在视觉问答、图文检索、视觉定位以及自然语言视觉推理多个视觉-语言下游任务上进行微调实验, 实验结果表明所提方法相比于 baseline 方法在多个下游任务中性能均有提升, 其中在 NLVR^2 任务上相比 baseline 方法准确率提升 1.8%.

关键词: 视觉-语言预训练; 超图神经网络; 多元实体对齐; 注意力机制; 多模态理解
中图法分类号: TP18

中文引用格式: 李登, 武阿明, 韩亚洪. 基于多元实体对齐的视觉-语言多模态预训练. 软件学报, 2025, 36(11): 5118–5133. http://www.jos.org.cn/1000-9825/7321.htm
英文引用格式: Li D, Wu AM, Han YH. Visual-language Multimodal Pre-training Based on Multi-entity Alignment. Ruan Jian Xue Bao/Journal of Software, 2025, 36(11): 5118–5133 (in Chinese). http://www.jos.org.cn/1000-9825/7321.htm

Visual-language Multimodal Pre-training Based on Multi-entity Alignment

LI Deng[1], WU A-Ming[2], HAN Ya-Hong[1]

[1](College of Intelligence and Computing, Tianjin University, Tianjin 300350, China)
[2](School of Electronic Engineering, Xidian University, Xi’an 710401, China)

Abstract: Visual-language pre-training (VLP) aims to obtain a powerful multimodal representation by learning on a large-scale image-text multimodal dataset. Multimodal feature fusion and alignment is a key challenge in multimodal model training. In most of the existing visual-language pre-training models, for the multimodal feature fusion and alignment problem, the main approach is that the extracted visual features and text features are directly input into the Transformer model. Since the attention mechanism in the Transformer calculates the similarity between pairs, it is difficult to achieve the alignment among multiple entities. Considering that the hyperedges of hypergraph neural networks possess the characteristics of connecting multiple entities and encoding high-order entity correlations, thus enabling the establishment of relationships among multiple entities. In this study, a visual-language multimodal model pre-training method based on multi-entity alignment of hypergraph neural networks is proposed. In this method, the hypergraph neural network learning module is

* 基金项目: 国家自然科学基金 (62376186, 61932009)
收稿时间: 2023-07-07; 修改时间: 2024-01-25; 采用时间: 2024-11-07; jos 在线出版时间: 2025-02-26
CNKI 网络首发时间: 2025-02-27

introduced into the Transformer multi-modal fusion encoder to learn the alignment relationship of multi-modal entities, thereby enhancing the entity alignment ability of the multi-modal fusion encoder in the pre-training model. The proposed visual-language pre-training model is pre-trained on the large-scale image-text datasets and fine-tuned on multiple visual-language downstream tasks such as visual question answering, image-text retrieval, visual grounding, and natural language visual reasoning. The experimental results indicate that compared with the baseline method, the proposed method has performance improvements in multiple downstream tasks, among which the accuracy is improved by 1.8% on the NLVR^2 task.

Key words: visual-language pre-training (VLP); hypergraph neural network; multi-entity alignment; attention mechanism; multi-modal understanding

随着基于 Transformer 网络结构的自然语言处理预训练模型[1]在自然语言处理任务以及计算机视觉任务中的成功应用[2], 基于 Transformer 的视觉-语言多模态预训练这一研究也逐渐引起了广泛的研究学者兴趣. 视觉-语言预训练旨在从大规模的图像-文本数据集中学习得到具有强大的泛化能力视觉-语言联合特征表示, 然后将该泛化性能强的特征表示作用于下游的多模态任务进行微调后可显著提升下游任务性能[3].

视觉-语言联合特征表示学习不仅需要对视觉内容和文本语义进行理解, 同时多模态特征间的融合、对齐也是一大关键挑战[4]. 过去几年, 许多基于融合与对齐的大规模视觉-语言多模态预训练模型框架被提出[5−20]. 其训练方式可分为端到端训练和二阶段训练两种方式. 二阶段多模态预训练模型通过预训练的目标检测模型对图像中的目标进行提取编码后与文本特征一起输入多模态融合编码器进行融合对齐. 而端到端训练的多模态预训练模型则是直接输入图像和文本进行训练. 这些方法中的多模态融合编码器大多基于 Transformer[21]结构. Transformer 模型中的多头自注意力模块能够捕获长距离依赖的特征, 并且使用位置编码对输入数据进行预处理, 该方法不依赖于序列的先后关系, 使得模型能够并行计算, 因此, 其被广泛运用于预训练模型中. 此外, 也有研究学者通过对比学习的方法进行多模态特征间的对齐. Li 等人[19]和 Yang 等人[20]提出采用对比损失的方法去对齐的图文与文本之间的特征, 使得匹配的多模态特征尽量靠近, 非匹配的多模态特征之间距离尽量远离.

当前主流多模态预训练大多基于 Transformer 模块进行构建, 然而 Transformer 模块中的多头注意力机制中的 attention map 是通过计算两两特征之间相似度得到, 其捕获的是多模态数据中两两实体之间的对齐关系, 而难以捕获多个实体之间的对齐关系. 而在多模态分析场景任务中, 实体关系不仅存在两两对齐关系, 还存在着多个实体的对齐关系 (比如, 一台电脑可能包含着主机、显示器、键盘和鼠标等多个实体), 本文将这种关系定义为多元实体对齐关系. 捕获多元实体的对齐关系有利于编码实体的高阶关系, 构建完善的实体对齐关系从而增强多模态预训练模型特征表达能力.

针对上述问题, 本文提出基于超图神经网络的多元对齐的多模态预训练方法. 相比于一般的图神经网络中通过边以及注意力机制连接两个节点, 从而构建两两实体之间的关系. 超图中的超边可以连接多个节点, 从而对多元实体的对齐关系进行编码. 因此本文将超图神经网络引入至多模态特征融合编码器当中, 以实现多模态数据中多元实体对齐关系的构建, 从而提升预模型对多模态数据的特征表达能力. 图 1 所示为模型框架图, 通过超图神经网络分别对视觉模态和文本模态中的信息进行超图的构建进而分别对两种模态实现高阶实体的对齐关系进行编码, 然后通过注意力机制跨模态融合对齐模块对视觉模态及文本模态进行融合. 具体为, 首先, 分别由视觉模型 ViT[22]对图像特征进行提取以及由文本 BERT[1]模型对文本语义特征进行提取, 提取得到的特征可视为各模态的实体特征并且进行超图神经网络的构建, 其中, 通过聚类以及 KNN 近邻算法构建表示节点和超边之间关系的关联矩阵, 进而构建邻接矩阵. 在对超图神经网络中的邻接矩阵以及节点特征的表示进行学习后, 通过图卷积神经网络对超边的信息进行聚合从而实现多模态数据中多元实体的对齐关系构建, 而后通过 Transformer 中的 self-attention 对输出的多模态特征进行融合并输入至预训练任务 head 进行训练. 本文设定预训练任务为 masked language modeling (MLM) 以及 image-text matching (ITM) 从而实现视觉-语言模型的预训练. 在下游任务实验中, 本文选取了视觉问答任务 (visual question answering, VQA)、图文检索任务 (image-text retrieval)、视觉定位任务 (visual grounding) 以及自然语言视觉推理任务 (natural language for visual reasoning, NLVR) 进行了有效性的实验验证.

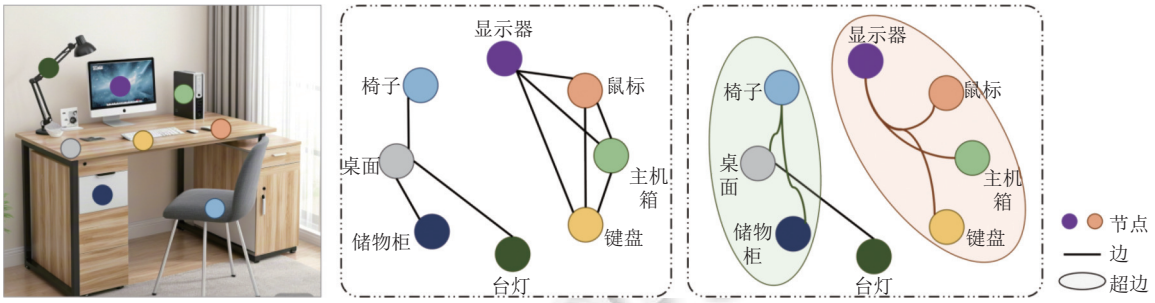
本文的主要贡献如下.

© 中国科学院软件研究所 http://www.jos.org.cn

(1) 针对基于 Transformer 的视觉-语言预训练模型中注意力机制实体对齐能力弱的问题, 本文提出了一种将图神经网络与 Transformer 结合的多模态实体对齐模块, 以增强多模态数据间实体的对齐关系.

(2) 借助超图中的超边可连接多个实体的特性, 本文提出一种基于超图神经网络的多元实体的对齐方法, 对多模态实体的高阶相关关系进行编码.

(3) 所提出的方法在多个视觉-语言下游任务中进行了实验, 相比于其他基于 Transformer 的视觉-语言预训练模型 baseline 方法, 本文方法均有提升, 相关实验结果验证了本文方法的有效性.



(a) 输入图像示例 (b) 二元对齐 (c) 多元对齐

图 1 多元实体对齐与二元实体对齐对比

本文第 1 节介绍视觉-语言多模态预训练相关方法以及研究现状. 第 2 节介绍本文所提出的基于图神经网络多元对齐的视觉-语言预训练方法. 第 3 节通过对比实验验证了所提模型的有效性. 第 4 节进行总结全文.

1 相关工作

目前, 关于视觉-语言预训练的研究可分为两阶段预训练和端到端预训练两种类型的模型框架, 两阶段预训练指的是: 第 1 阶段, 通过离线预训练好的目标检测模型将图像中的目标区域提取出来; 第 2 阶段, 将离线提取的视觉特征与文本一起输入至基于 Transformer 的预训练模型进行预训练. 端到端预训练则指的是直接将图像与文本输入至预训练模型中进行预训练, 同时对预训练模型中的视觉特征提取网络进行优化以及参数的更新.

1.1 两阶段视觉-语言多模态预训练

目前大多数两阶段视觉-语言多模态预训练方法还是采用卷积神经网络 (convolutional neural network, CNN) 目标检测模型提取图像目标特征, 然后通过类似于自然语言处理预训练模型 BERT 的自监督方式进行训练, 如 Lu 等人[18]提出的 ViLBERT、Tan 等人[5]、Su 等人[23]以及 Chen 等人[6] 通过预训练的 Faster R-CNN[24]模型对图像中目标候选区域进行特征提取, 然后构建生成图像 Embedding 与文本 Embedding 一起输入至基于 Transformer 的模块中进行多模态特征的交互融合并得到最后的视觉-语言联合表征. Yu 等人[12]提出将场景图的信息引入至预训练模型当中, 根据解析的场景图设计了场景图预测预训练任务, 使得模型能够更好地对图像和文本中的实体进行对齐. Gan 等人[25]提出将对抗训练引入到视觉-语言表达学习训练中从而提升特征表达的泛化能力. Li 等人[7]提出将图像检测得到的目标标签用于图像文本语义对齐学习. Li 等人[26]提出了一种统一的预训练模型框架, 能同时进行单模态和多模态的内容理解和生成任务, 并可利用大量开放域文本语料和图像来提高视觉和文本的理解能力. 由于两阶段多模态预训练视觉特征是通过离线方式获取, 因此在预训练以及下游任务微调过程中, 视觉特征提取网络均不会得到优化和参数的更新. 同时, 通过离线的目标检测模型提取的目标区域, 不可避免地会出现有区域的重叠情况, 从而导致图像特征之间产生混淆.

1.2 端到端视觉-语言多模态预训练

随着视觉 Transformer 模型的发展, 多模态预训练模型中的视觉内容特征提取模块也可通过 Transformer 模型进行构建从而实现端到端的训练. Huang 等人[14]较早提出了端到端的视觉-语言预训练模型框架, 并提出了基于视

觉字典的 masked visual modeling 预训练任务. Huang 等人[15]提出从图像的像素级别与文本语义进行对齐, 减少了视觉-语言模型在特定任务上视觉表征的限制. Xu 等人[27] 提出在预训练过程中将目标检测任务以及图像描述生成任务增加到预训练任务中从而去增强视觉内容特征的学习, 然而该方法需要大量的目标检测标签, 因而难于扩展到对于数据量需求大的大规模模型训练当中. Kim 等人[16]提出了一个纯 Transformer 结构的视觉-语言预训练模型, 使用预训练的 ViT 对多模态融合 Transformer 模块进行初始化, 使得可以直接通过多模态融合模块对视觉特征进行处理. Liu 等人[17]提出了一个目标感知的端到端预训练模型, 将从 CNN 模型提取得到的视觉网格特征输入至 Transformer 模型中以联合学习多模态表示, 同时该方法还将目标检测知识蒸馏引入预训练模型中, 以便学习不同语义层次的多模态对齐. Li 等人[19] 和 Yang 等人[20]通过引入了一种对比损失, 在多模态融合前对图像特征和文本特征进行对齐, 使得多模态融合编码器能够更好地进行多模态表示学习. X-Decoder[28]提出了一种通用的解码模型可以实现像素级别的粒度的视觉语言交互任务. PTP[29]构建了一种跨模态提示学习 (prompt learning) 的多模态预训练范式, 将物体的位置信息作为提示, 以引导增强预训练模型中的视觉定位能力. 目前, 多模态融合对齐方式主要基于 Transformer 和对比学习的方式, 而 Transformer 结构中 self-attention 以及对比学习中正负样本的判别均通过计算两两特征直接相似度得到, 这些方法主要考虑的两两实体之间的对齐而难以捕获多模态数据中多元实体的对齐关系, 因此, 本文基于超图神经网络汇聚超边的特性, 提出基于超图网络的多模态融合对齐模块实现多元实体对齐关系的构建, 从而增强预训练模型视觉-语言联合特征表示泛化能力.

1.3 超图神经网络

自 Schölkopf 等人[30]提出超图这一模型结构后, 超图的相关研究得到了快速的发展[31]. 受卷积神经网络启发, Feng 等人[32]提出了一种超图神经网络 (hypergraph neural network), 在该方法中, 设计了超边卷积运算来处理数据相关性的表示学习以对数据的高阶相关性进行建模. Ji 等人[33]针对协同过滤方法的建模不灵活以及高阶相关性建模不足的问题, 提出了一种双通道超图协同过滤 (dual channel hypergraph collaborative filtering) 框架. 也有研究学者针对数据中噪声产生的影响, 提出了 HyperGCN[34]的超图卷积网络模型, 通过对二元实体构建的边进行筛选, 从而对噪声信息进行过滤, 然而该方法可能会导致有效信息的丢失. Jiang 等人[35]提出了一种动态的超图卷积网络, 通过 K-means 聚类的方法构建超图网络, 并在训练过程中对聚类中心进行更新进而动态重构超图, 由于该方法每次迭代均需要重新聚类以及重构超图, 因此, 该方法效率较低, 不适用于大规模预训练模型. 受超图的高阶相关性建模启发, 本文将超图神经网络引入至 Transformer 的多头注意力模块中, 实现对视觉实体和文本实体的高阶相关性建模, 并通过多头注意力机制提取相关性强的实体特征信息, 进而增强视觉与文本的跨模态融合对齐.

2 基于多元实体对齐的多模态预训练

一般的图神经网络的节点之间通过边对节点进行连接, 如图 1(b), 图神经网络的边构建的是两两节点之间的关系, 这种对齐为二元对齐. 不同于简单的图神经网络, 超图神经网络是一种具有超边的图神经网络, 如图 1(c), 超图中的超边可以连接两个或两个以上节点, 对高阶关系进行编码, 这种对齐方式为多元对齐. 定义超图为 $G = (V_g, \varepsilon, W)$, 其中 V 为包含 N 个节点的集合, ε 表示超边, W 表示超边的权重. 超图的关联矩阵可以通过 $H \in \mathbb{R}^{|V|\times|\varepsilon|}$ 表示, 其中 $H_{i,j} = 1$ 表示节点 V_{gi} 和超边 ε_j 相关联.

鉴于超图神经网络强大的关系构建能力和特征提取能力, 本文提出一种基于超图神经网络的多元实体对齐的视觉-语言预训练模型, 如图 2 所示, 该模型主要包含视觉编码器、文本编码器、对比损失模块以及基于图神经网络的多模态融合对齐模块.

首先, 对输入的图像和文本进行特征提取, 对于视觉模态部分, 本文采用 ViT[22]模型作为图像编码器对图像内容进行提取, 该 ViT 模型为 12 层的 Transformer 结构模型, 并由在 ImageNet-1K 上预训练的模型对模型参数进行初始化. 当输入一张图像, ViT 模型将其 (图片) 编码为一序列的包含 [CLS] 的视觉 Embedding 特征向量. 对于文本模态部分, 本文采用一个 6 层的 BERT 模型作为文本编码器对文本的语义信息进行提取, 该 BERT 模型由预训练

© 中国科学院软件研究所 http://www.jos.org.cn

好的 BERT-base[1]模型前 6 层参数进行初始化. 同样, 对于输入的文本, 文本编码器将其编码为一序列的包含 [CLS] 的文本 Embedding 特征向量.

其次, 图像和文本两种模态数据可以视为对同一事物进行描述的两个不同视角. 通过引入对比学习对提取的图像特征和文本特征进行匹配可以增强视觉编码器以及文本编码器的特征提取能力. 训练时, 将视觉 Embedding 特征向量中的 [CLS] token 以及文本 Embedding 特征向量中的 [CLS] token 输入至对比学习模块中, 通过最小化 InfoNCE 损失使得匹配的图像文本特征对在统一的特征空间尽量靠近, 而非匹配的图像文本特征对尽量远离.

接着, 在提取得到视觉 Embedding 特征向量以及文本 Embedding 特征向量后, 将这两个特征向量输入至基于图的多模态特征融合对齐模块基于图卷积神经网络的汇聚边信息的能力实现多模态特征的多元实体对齐关系的构建. 对视觉 Embedding 特征构建稠密图结构, 对图神经网络结构中的邻接矩阵以及节点特征进行学习, 然后通过卷积神经网络汇聚多边信息. 对稀疏的文本 Embedding 特征通过 Transformer 中的多头 self-attention 模块进行学习. 在将学习得到的视觉特征和文本特征进行对齐后, 最终由构建的掩蔽文本预测 (MLM) 以及图像-文本对齐 (ITM) 预训练任务进行预训练.

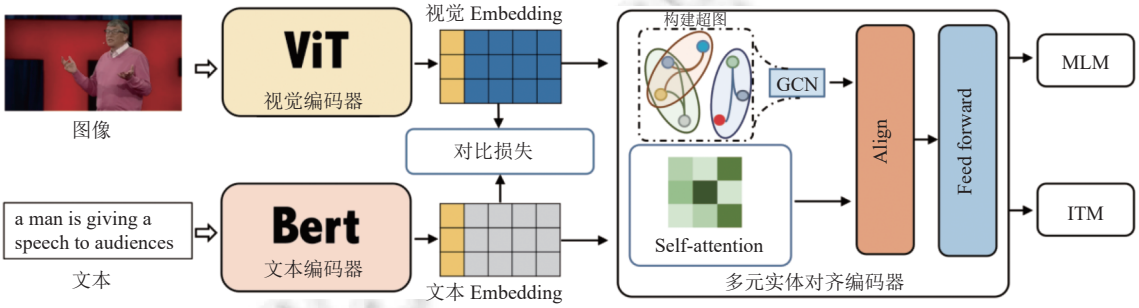


图 2 基于多元实体对齐的视觉-语言预训练模型框架图

2.1 视觉-语言特征对比学习

对比学习研究较早出现在视觉任务上[36], 其主要思想为对于给定的某一类别图像, 通过图像增强方式对齐进行变换得到该类别图像的另一视图, 即位正样本. 而其他类别的图像均为该类别图像的负样本. 本质上两种视图中的类别应该还属于同一类别, 因此其与正样本在统一的特征空间中的距离应该较近, 而与其他类别距离应该较远. 对比学习则是通过构造对比损失函数使得样本中类内之间距离尽量靠近, 而类间距离尽量远离, 从而增强网络的特征提取能力. 而在视觉-语言多模态的表示学习当中, 输入为图像和文本两个模态, 可以认为这两个模态是对同一事物的两个不同视图. 因此也可通过对比学习的方式增强视觉编码器以及文本编码器的特征提取能力. 这一过程可以通过最大化图像和文本之间的互信息进行, 而在具体实现中则可通过最小化 InfoNCE[37]损失函数对图像和文本的互信息下界进行最大化, 其表达式可表示为:

$$L_{\mathrm{NCE}} = -E_{p(a,b)}\left[\log \frac{\exp(s(a,b))}{\sum_{\hat{b}\in\hat{B}} \exp(s(a,\hat{b}))}\right] \tag{1}$$

其中, $s(a,b)$ 为计算 a 与 b 之间的相似度, $\hat{B}$ 为存储对比学习的样本的集合, 其中包含 1 个正样本以及 N–1 个负样本.

对于图像-文本对的对比损失 L_{itc} 可以表示为:

$$L_{\mathrm{itc}} = -\frac{1}{2}E_{p(I,T)}\left[\log \frac{\exp(s(I,T)/\tau)}{\sum_{m=1}^{M} \exp(s(I,T_m)/\tau)} + \log \frac{\exp(s(T,I)/\tau)}{\sum_{m=1}^{M} \exp(s(T,I_m)/\tau)}\right] \tag{2}$$

其中, I 表示输入的视觉 Embedding 特征向量, T 表示输入的文本 Embedding 特征向量, 最小化图像-文本对的对比损失函数可以理解为最大化互信息的对称问题, 使得匹配的图像-文本对距离靠近, 而非匹配的图像-文本对距离

© 中国科学院软件研究所 http://www.jos.org.cn

远离, 从而有利于多模态间特征的联合表示学习.

2.2 基于超图 Transformer 的多模态特征融合

为了更好地对多模态数据中的实体关系进行构建, 我们对基于 Transformer 的多模态融合编码器引入超图神经网络, 通过超图神经网络的高阶相关性建模特性对超边信息聚合实现多元实体的对齐, 从而增强多模态融合编码器的实体关系构建的能力. 图 3 所示为将超图神经网络引入至 Transformer 模块中的多模态特征融合对齐编码器框架图.

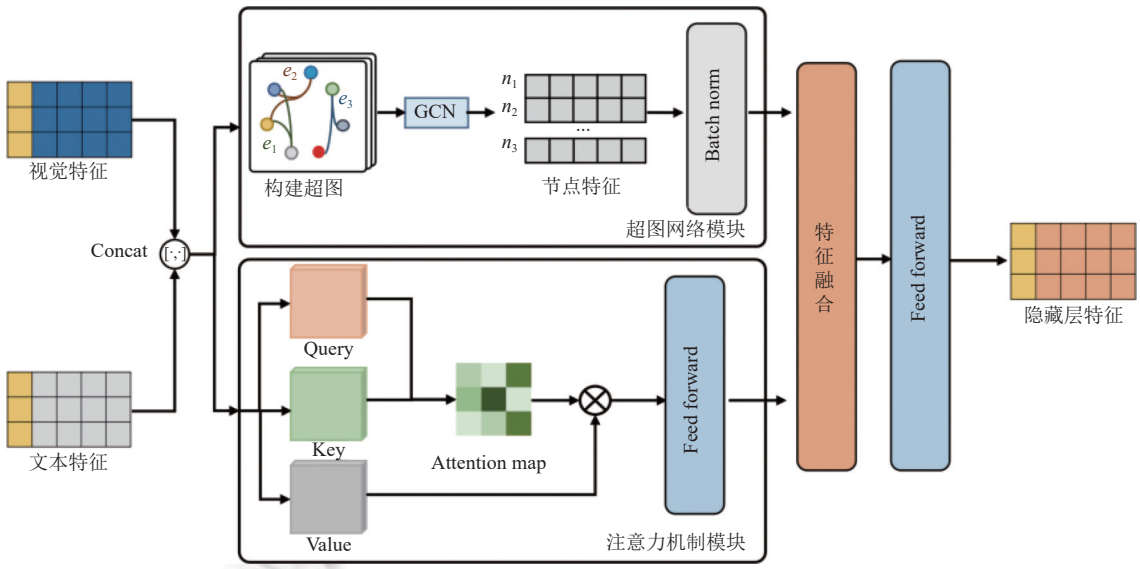


图 3 多模态融合对齐编码器框架图

视觉特征由视觉编码器 (ViT) 对输入图像提取得到, 文本特征由文本编码器 (BERT) 对于输入文本信息进行提取得到. 接着将视觉特征和文本特征拼接后分别输入至 Transformer 的多头注意力机制模块以及超图网络模块, 进行多模态实体的低阶与高阶相关性建模.

输入至注意力机制模块的特征首先经过非线性变换分别映射为 Query、Key 以及 Value 这 3 种值, 由 Query 和 Key 值计算相似度得到 attention map, 然后与 Value 计算得到注意力特征, 再经过由多层感知机 (multilayer perceptron, MLP) 构成的前向传递层 (feed forward 层) 得到输出特征 F_A. 该模块通过两两实体之间的相似度进行连接, 因此该模块构建的是二元实体连接表示. 注意力机制模块计算可表示如下 (为便于理解, 公式中省去残差模块的特征之和部分):

$$Q = F_q([I,T]) \tag{3}$$

$$K = F_k([I,T]) \tag{4}$$

$$V = F_v([I,T]) \tag{5}$$

$$F_A = MLP(LN(MSA(Q,K,V))) \tag{6}$$

其中, I 表示输入的视觉 Embedding 特征向量, T 表示输入的文本 Embedding 特征向量, $[\cdot,\cdot]$ 表示特征拼接操作, F_x 表示对应的非线形变换层. MSA 为多头自注意力函数 (multi-head self-attention), LN 为层归一化操作 (layer normalization), MLP 为多层感知机函数 (multilayer perceptron).

输入至超图网络模块的特征首先构建超图, 通过聚类构建多模态实体超图网络模块的超边 ϵ, 视觉与文本实体之间的关系是通过超边 ϵ 估计, 如果实体在同一条超边上同时出现, 则视为已连接, 由于一条超边可同时连接多个实体, 因此该模块可对高阶实体关系进行建模, 即实现多元实体的连接表示. 如此得到包含节点关系的超图关联

矩阵 H 以及相应的节点特征表示 V_g, 然后通过图卷积神经网络 (graph convolution network, GCN) 对超边的信息进行汇聚. 在超图网络模块中, 描述实体之间的关系的邻接矩阵可表示为:

$$A_{\text{hypergraph}} = D_v^{1/2} H W D_e^{-1} H^{\mathrm{T}} D_v^{1/2} \tag{7}$$

其中, W 表示超边的权重, H 表示超图关联矩阵, D_v 表示节点的度矩阵以及 D_e 表示超边的度矩阵. 最终超图网络模块输出特征 F_{hg} 可表示:

$$F_{\text{hg}} = MLP(LN(A_{\text{hypergraph}}[I,T]\Theta)) = MLP\left(LN\left(D_v^{1/2} H W D_e^{-1} H^{\mathrm{T}} D_v^{1/2}[I,T]\Theta\right)\right) \tag{8}$$

其中, Θ 为可学习参数. $D_v^{1/2} H W D_e^{-1} H^{\mathrm{T}} D_v^{1/2}[I,T]\Theta$ 表示视觉与文本节点信息聚合至超边节点上, $D_v^{1/2} H W$ 表示将汇聚后的边节点信息传输至原节点上. LN 为层归一化操作, MLP 为多层感知机函数. 将注意力机制模块的输出 F_A 以及超图网络模块的输出 F_{hg} 通过进行注意力机制融合后输入至前向传递层 (feed forward 层), 其输出结果则是当前多模态特征融合对齐编码器模块的隐藏层特征 F_{hidden}, 并作为下一层多模态融合对齐编码器的隐藏层特征输入.

$$F_{\text{hidden}} = Fusion(F_A, F_{\text{hg}}) \tag{9}$$

2.3 视觉-语言预训练任务

本文预训练任务包含有掩蔽文本预测 (MLM) 以及图像-文本对齐 (ITM) 两个预训练任务.

(1) 掩蔽文本预测任务 (MLM)

本文的 MLM 预训练任务与自然语言处理预训练模型 BERT[1]中的 MLM 预训练任务类似, 主要不同之处在于 BERT 模型中该任务利用的信息为纯文本信息, 而本文中同时利用了图像内容以及文本语义信息对掩蔽的单词进行预测. 通过随机概率从输入的所有单词 token 中随机选择一个 token 替换为 [MASK] token 作为掩蔽的单词. 训练时通过最小化交叉熵损失进行优化, 损失函数可表示为:

$$L_{\text{mlm}} = -E_{(I,\hat{T})\sim D} H\left(y^{\text{mask}}, p^{\text{mask}}\left(I,\hat{T}\right)\right) \tag{10}$$

其中, $H(\cdot,\cdot)$ 为交叉熵函数, $p^{\text{mask}}(I,\hat{T})$ 为预测概率值, y^{mask} 为真实值.

(2) 图像-文本对齐任务 (ITM)

图像-文本对齐任务即为对给定的图像-文本对预测判断是否是匹配. 在预训练过程中, 选择多模态融合对齐编码器输出的 [CLS] token 作为图像-文本对的联合特征表征, 而后输入至一层全连接网络层进行二分类得到预测概率为 p^{itm}. 预训练 ITM 任务的损失函数可表示为:

$$L_{\text{itm}} = -E_{(I,T)\sim D} H\left(y^{\text{itm}}, p^{\text{itm}}(I,T)\right) \tag{11}$$

其中, $p^{\text{itm}}(I,T)$ 为预测概率值, y^{itm} 为真实值.

综上, 整个预训练过程中共包含 3 部分损失, 分别为图像文本对比学习损失 L_{itc}、掩蔽文本预测损失 L_{mlm} 以及图像-文本对齐损失 L_{itm}. 其表达式可表示为:

$$L_{\text{total}} = L_{\text{itc}} + L_{\text{mlm}} + L_{\text{itm}} \tag{12}$$

3 实验分析

3.1 实验数据

实验数据集包括模型预训练数据集以及多模态下游任务数据集. 在多模态模型预训练阶段采用的数据集为 Conceptual Captions[38]、 SBU Caption[39]、COCO Captions[40]以及 Visual Genome[41]多模态数据集. 共包含 400 万张图像以及 510 万条图文对.

我们在视觉问答、下游任务数据集包括 VQA-v2.0[42]、Flickr30k[43]、NLVR^2[44]、RefCOCO+[45]. 其中视觉问答数据集 VQA-v2.0 共包含 204 721 张图像、1 105 904 对问答, 图像来自真实场景以及 COCO 数据集, 问题类型包括开放型、非型以及多个选项型. Flickr30k 是常用的图像描述数据集, 其包含 31 783 图像以及每张图像包含 5 条相关的描述文本. NLVR^2 是关于自然语言和图像联合推理的一个数据集, 该数据集包含 107 292 条英文句子文

© 中国科学院软件研究所 http://www.jos.org.cn

本与网络图像对, 其中每条文本描述对应两张分别为匹配与不匹配的图像. RefCOCO+是一个视觉定位数据集, 该数据集包含 141 564 条描述句, 对应包含着 49 856 个目标以及 19 992 张图像. 表 1 给出了预训练数据集以及下游任务数据集所对应的详细信息.

表 1 预训练及下游任务实验数据集

分组	数据集	图像数量	图像文本对数量
预训练	Conceptual Captions	3.0M	3.0M
	SBU Caption	1.0M	1.0M
	COCO Captions	110k	555k
	Visual Genome	103k	5M
下游任务	VQA-v2.0	204k	1.1M
	Flickr30k	32k	160k
	$NLVR^2$	214k	107k
	RefCOCO+	20k	142k

3.2 实验设定

本文所有实验均在 8 卡 NVIDIA V100 上进行, 通过深度学习框架 PyTorch[40]进行分布式训练. 本文将 ALBEF 模型作为 baseline 方法, 并基于该方法构建本文基于多元实体对齐的多模态预训练模型. 预训练模型中的视觉编码器模块采用 12 层参数量为 85.8M 的 ViT-Base 模型, 文本编码器以及多模态融合对齐编码器中的 Transformer 结构则均为 6 层网络结构, 分别采用 BERT 的前 6 层和后 6 层参数进行初始化. 预训练过程中 batch size 设定为 256, 共训练 30 个 epoch. 采用权重衰减为 0.02 的 AdamW[36]优化器进行优化. 初始学习率设定为 1E–4 并根据 cosine 函数进行逐步衰减至 1E–5. 预训练时输入图像为随机裁剪至大小为 256×256 尺寸图像, 而对下游任务进行微调时则根据 ViT 中插值方法增大至 384×384.

本文将预训练模型在 4 个视觉-语言下游任务上进行了微调实验. 分别为视觉问答、图文检索、自然语言视觉推理、视觉定位 visual grounding. 下面对各下游任务实验设定进行介绍.

视觉问答 (visual question answering, VQA) 任务旨在对于给定的图像和文本问题, 通过对视觉内容和文本语义的理解预测出正确答案. 实验数据采用 VQA-v2.0 数据集. 本文中, 该下游任务实验设定与 ALBEF 方法一致. 将视觉问答预测任务视为生成任务由 6 层 Transformer 结构解码器生成答案. 将多模态融合对齐编码器的输出输入至解码器当中进行解码生存答案. 实验时, 为对比公平, 本文仅对相同的 3 192 个问题进行测试生成对应答案. 实验时, 设定优化器为权重衰减系数为 0.02 的 AdamW 优化器, 初始学习率为 1E–5, 经过 4 个 epoch 的 warm up 学习策略达到 2E–5 而后以 cosine 函数衰减策略最终衰减至 1E–6. 设定模型训练 epoch 数量为 8.

图文检索 (image-text retrieval) 任务包含有根据输入图像检索文本的任务以及根据输入文本检索图像的任务. 本文由预训练得到的模型在 COCO 数据集上对两种类型的检索任务分别进行了实验. 实验时, 设定优化器为权重衰减系数为 0.02 的 AdamW 优化器, 初始学习率为 1E–5, 学习策略达到 2E–5 而后以 cosine 函数衰减策略最终衰减至 1E–6. 设定模型训练 epoch 数量为 8.

自然语言视觉推理 (natural language for visual reasoning, NLVR) 任务旨在对给定的图像-文本对进行判断是否与对应的文本相匹配. 本文在 $NLVR^2$ 数据集上对 107 292 条文本语句匹配的图像进行实验评估. 对于该任务由于需要对图像文本对进行新的编码因此还需要进行额外的预训练. 与 ALBEF 方法一致, 本文设计了一个文本分配任务, 即对于输入的两张图像和一条文本, 模型对文本分配进行分类判断是与哪张图像相匹配. 实验时, 对于增加的额外预训练, 设定优化器为权重衰减系数为 0.02 的 AdamW 优化器, 初始学习率为 1E–5, 学习策略达到 2E–5 而后以 cosine 函数衰减策略最终衰减至 1E–5. 设定模型训练 epoch 数量为 1. 对于下游任务预测训练, 设定初始学习率为 1E–5, 学习策略达到 2E–5 而后以 cosine 函数衰减策略最终衰减至 1E–6, 设定训练 epoch 数量为 10. 由于该任务输入是一条文本对应着两张图像, 因此参考 ALBEF 方法对视觉编码器进行了扩展为 2 个参数共享的视觉编码器以对输入两张图像进行处理.

© 中国科学院软件研究所 http://www.jos.org.cn

视觉定位 (visual grounding) 任务旨在由输入的图像和对应的特定文本描述, 定位出图像中相对应的区域. 本文实验中的实验设定为弱监督视觉定位, 即给定的训练集不包含目标框的标注. 本文由预训练得到的模型在 RefCOCO+视觉定位 Benchmark 数据集上进行了微调实验. 实验设定初始学习率为 1E–5, 训练 epoch 数量为 5.

本文与多个视觉-语言多模态预训练模型进行了对比. 其中, VisualBERT[9]和 VL-BERT[23]均为通过 Transformer 的自注意力机制实现图像区域的文本的对齐. LXMERT[5]构建了关系编码器用于学习视觉和语言特征之前的关系. 12-in-1[8]通过构建干净的视觉-语言多任务, 使得在训练过程中, 确保不出现信息泄漏. OSCAR[7]将图像中检测的物体标签作为锚点输入. UNITER[6]构建了一种通用的图像文本表征空间. ViLT[16]使用简单的线性映射大大减少了视觉编码器的参数量. VILLA[25]通过对抗训练的方式增强模型的泛化能力. UNIMO[26]、ALBEF[19]和 FLAVA[46]均为构建跨模态的对比学习方法将文本和图像映射到统一空间中, 从而提升视觉和文本的理解能力的学习方法. METER-SwinB[47]引入交叉注意力促进多模态的融合. X-Decoder[28]构建了一种像素级和图像级语义的视觉语言对齐通用的编码框架. PTP[29]引入了一种新颖的位置引导文本提示范例, 以增强视觉语言多模态预训练模型中的视觉定位能力. 与上述方法不同, 本文引入超图的多元实体链接特性, 构建了一种多元实体对齐的视觉语言多模态模型框架.

3.3 实验结果与分析

(1) VQA 以及 NLVR^2 下游任务实验结果分析

表 2 所示为本文提出的视觉-语言预训练模型在 VQA 以及 NLVR^2 下游任务进行微调训练的结果. 本文对比了目前主流的视觉-语言预训练模型, 包括两阶段训练以及端到端训练模型. 预训练模型的图像数量均约为 400 万张. 对于视觉问答 VQA 任务, 分别在 VQA-v2.0 的 test-dev 以及 test-std 数据集上进行了对比评估, 其中在 test-std 上比 ALBEF[19]模型提升约 0.06%. 对于自然语言视觉推理 NLVR 任务, 分别在 NLVR^2 的 dev 以及 test 数据集上进行了对比评估, 其中在 dev 测试集上比 baseline 模型 ALBEF[19]提升约 1.0%, 在 test 上比 ALBEF[19]模型提升约 1.8%. 实验结果表明本文提出的预训练方法在视觉问答以及自然语言视觉推理下游任务上均能提升性能且对于自然语言视觉推理任务上具有较大幅度的提升, 这说明提出基于图卷积神经网络的多模态特征多元对齐预训练方法能够增强多模态特征间的对齐关系, 从而提升模型视觉-语言特征泛化能力.

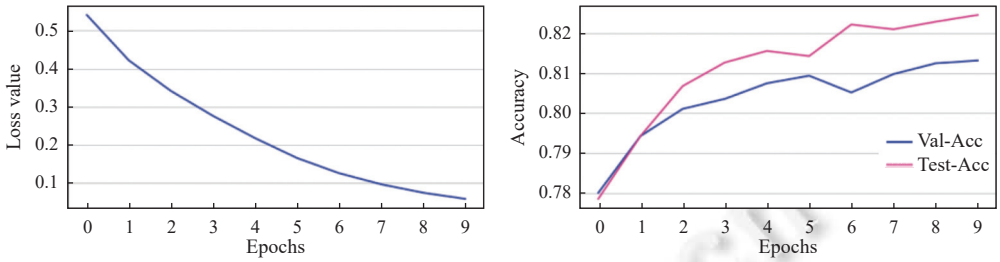
表 2 VQA 以及 NLVR^2 下游任务性能比较

方法	预训练图像数量 (M)	VQA (%)		NLVR^2 (%)	
		test-dev	test-std	dev	test
VisualBERT[9]	4	70.80	71.00	67.40	—
VL-BERT[23]	4	71.16	—	—	—
LXMERT[5]	4	72.42	72.54	74.90	—
12-in-1[8]	4	73.15	—	—	76.95
OSCAR[7]	4	73.16	73.44	78.07	—
UNITER[6]	4	72.70	72.91	77.18	78.28
ViLT[16]	4	70.94	—	75.70	—
UNIMO[26]	4	73.29	—	—	79.10
VILLA[25]	4	73.59	73.67	78.39	79.03
ALBEF[19]	4	**74.54**	74.70	80.24	80.50
FLAVA[46]	70	72.8	—	—	—
X-Decoder[28]	4	74.1	74.2	—	—
PTP[29]	4	73.44	**76.16**	77.31	78.50
Ours	4	74.43	74.76	**81.25**	**82.29**

图 4 所示为经过预训练的视觉-语言模型在自然语言视觉推理下游任务上的训练 10 个 epoch 的损失曲线图 (a) 以及验证集与测试集准确率的曲线图 (b). 从图 4(a) 中可以看出在经过预训练之后, 对下游任务微调训练过程中,

© 中国科学院软件研究所 http://www.jos.org.cn

训练损失能够稳步地下降, 说明下游任务的模型能够比较稳定地进行训练. 从图 4(b) 中可以看出下游任务训练的第 1 个 epoch 后验证集与测试集的准确率便可达到 0.78, 表明视觉-语言预训练模型提取得到的泛化表征能够很好地对多模态特征进行表示并且提升下游任务的初步训练的准确率.



(a) 训练损失 (b) 验证集和测试集准确率

图 4 自然语言视觉推理 (NLVR^2) 下游任务训练损失以及验证集和测试集准确率曲线图

(2) 图文检索下游任务实验结果分析

对于图文检索下游任务, 本文分别对根据图像检索文本以及根据文本检索图像这两个子任务进行了实验. 表 3 所示为预训练模型在图文检索下游任务进行微调训练的结果, 实验数据集为 COCO. 分别对与 ImageBERT[10]、OSCAR[7]、UNITER[6]、ViLT[16]、ALBEF[19]、METER-SwinB[47]、X-Decoder[28]以及 PTP[29]这些视觉-语言预训练模型进行了对比. 其中 ImageBERT 预训练模型的图像数量为 600 万张其他模型均约 400 万张. 本文对比了检索结果 Top1、Top5 以及 Top10 中召回结果. 可以看出, 与性能较好的模型 ALBEF 相比, 对于图像检索文本任务, 本文提出的方法在 R@1、R@5 以及 R@10 分别提升约 1.1%、1.2% 以及 0.7%. 对于文本检索图像任务, 本文方法在 R@1、R@5 以及 R@10 分别提升约 0.3%、0.6% 以及 0.3%. 这一实验结果表明本文提出的方法能够通过增强多模态实体间对齐关系从而提升图文检索下游任务性能.

表 3 图文检索下游任务性能比较

方法	预训练图像数量 (M)	Text Retrieval (%)			Image Retrieval (%)		
		R@1	R@5	R@10	R@1	R@5	R@10
ImageBERT[10]	6	66.4	89.8	94.4	50.5	78.7	87.1
OSCAR[7]	4	70.0	91.1	95.5	54.0	80.8	88.5
UNITER[6]	4	65.7	88.6	93.8	52.9	79.9	88.0
ViLT[16]	4	61.5	86.3	92.7	42.7	72.9	83.1
ALBEF[19]	4	73.1	91.4	96.0	56.8	81.5	89.2
METER-SwinB[47]	4	72.96	92.02	96.26	54.85	81.41	89.31
X-Decoder[28]	4	71.2	—	—	54.5	—	—
PTP[29]	4	68.3	91.2	94.7	45.6	80.6	89.2
Ours	4	**74.2**	**92.6**	**96.7**	**57.1**	**82.1**	**89.5**

(3) 视觉定位下游任务实验结果分析

对于视觉定位下游任务, 本文实验设定为弱监督方式训练的任务, 即训练数据集中不提供目标区域的标注框. 实验数据集为 RefCOCO+数据集, 训练完成后, 分别在验证集、测试集 A 以及测试集 B 上进行了测试. 表 4 所示为预训练模型在视觉定位下游任务进行微调训练得到的测试结果, 分别与 ARN[48]、CCL[49]以及 ALBEF[19]这些模型的结果进行了对比. 可以看出, 与性能较好的模型 ALBEF 相比, 本文提出的方法在验证集上略低, 而在测试集 A 以及测试集 B 分别提升约 0.1% 以及 0.7%. 在 RefCOCO+数据集中, 由于在 RefCOCO+数据集的测试集 Test A 中图像主要物体为人, 而测试集 Test B 中的图像则包含所有其他不同的物体. 划分的这两个测试集存在着不同的数据分布, 且测试集 Test A 包含的人的样本相对较容易识别, 而测试集 Test B 中可能包含更具挑战性的样本. 从

© 中国科学院软件研究所 http://www.jos.org.cn

而导致对模型在测试集 Test B 上的性能与测试集 Test A 的性能存在一个明显差距. 从平均值来看, 本文提出的方法性能优于 baseline 模型 ALBEF. 实验结果表明, 本文提出的基于图神经网络的多模态数据多元实体对齐的方法能够增强视觉与文本信息的对齐关系.

表 4 弱监督视觉定位下游任务性能比较 (%)

方法	Val	Test A	Test B	Avg.
ARN[49]	32.78	34.45	32.13	33.12
CCL[50]	34.29	36.91	33.56	34.92
ALBEF[19]	**58.46**	65.89	46.25	56.87
Ours	58.11	**66.03**	**46.94**	**57.02**

(4) 可视化结果分析

Grad-CAM[50]方法是一种对神经网络做出决策进行解释性分析的一种方法, 通过该方法可以对激活特征图进行可视化分析. 本文通过 Grad-CAM 方法, 分别对视觉问答模型多模态编码器中跨模态对齐模块、视觉定位模型多模态编码器中跨模态对齐模块以及多模态编码器中第 3 层跨模态对齐模块注意力图进行了可视化分析. 如图 5 为视觉问答 VQA 模型多模态编码器中跨模态对齐模块注意力图可视化结果, 从图 5 中可以看出, 本文提出的基于图神经网络的多模态编码器能够很好地捕获到视觉和文本对应的关键特征, 同时对于相同的输入图像内容也可以根据输入的不同的问题文本句子提取得到相关度高的特征进而对最终的预测做出决策. 图 6 所示为视觉定位模型中单个单词对应跨模态对齐模块注意力图可视化结果, 从图 6 中可以看出对于给定的句子中的不同目标实体, 注意力图中激活区域能够集中对应到相应的图像特征区域. 图 7 所示为多模态编码器中第 3 层跨模态对齐模块注意力图可视化结果, 结果表明第 3 层跨模态融合对齐模块能够根据输入图像与文本信息很好地对相应的实体进行对齐. 上述结果表明本文提出的基于图卷积神经网络多元实体对齐的视觉-语言多模态预训练模型能够很好地对多模态数据中的实体构建对齐关系, 提升视觉-语言预训练模型的泛化能力, 并在对视觉-语言下游任务进行微调后能够提升下游任务中的性能.

图 5 单个单词对应跨模态对齐模块注意力图 Grad-CAM 可视化结果

图 8 所示为自然语言视觉推理下游任务测试可视化结果, 可以看到将本文提出的基于图卷积神经网络的多元实体对齐的视觉-语言多模态预训练微调至自然语言视觉推理下游任务进行训练后, 模型能够很好地对输入的两张图像和文本中实体的关系进行构建, 从而对输入的两张图像内容进行较好的预测.

(5) 消融实验

本文分析 6 层多模态特征融合编码器中图卷积神经网络的影响, 对多模态特征融合编码器引入图卷积神经网络的所在层进行了消融实验. 该消融实验在弱监督的视觉定位任务上进行, 分别测试了 RefCOCO+验证集、测试

© 中国科学院软件研究所 http://www.jos.org.cn

集 A 以及测试集 B 上的结果如表 5 所示, 当第 1 层引入图卷积神经网络, 在测试集 A 以及测试集 B 上优于不引入情况. 当所有层都引入图卷积神经网络时, 在测试集 A 以及测试集 B 上效果最好. 从而说明当所有层都引入图卷积神经网络时, 能够更好地提升图像和文本中实体对齐关系提取能力.

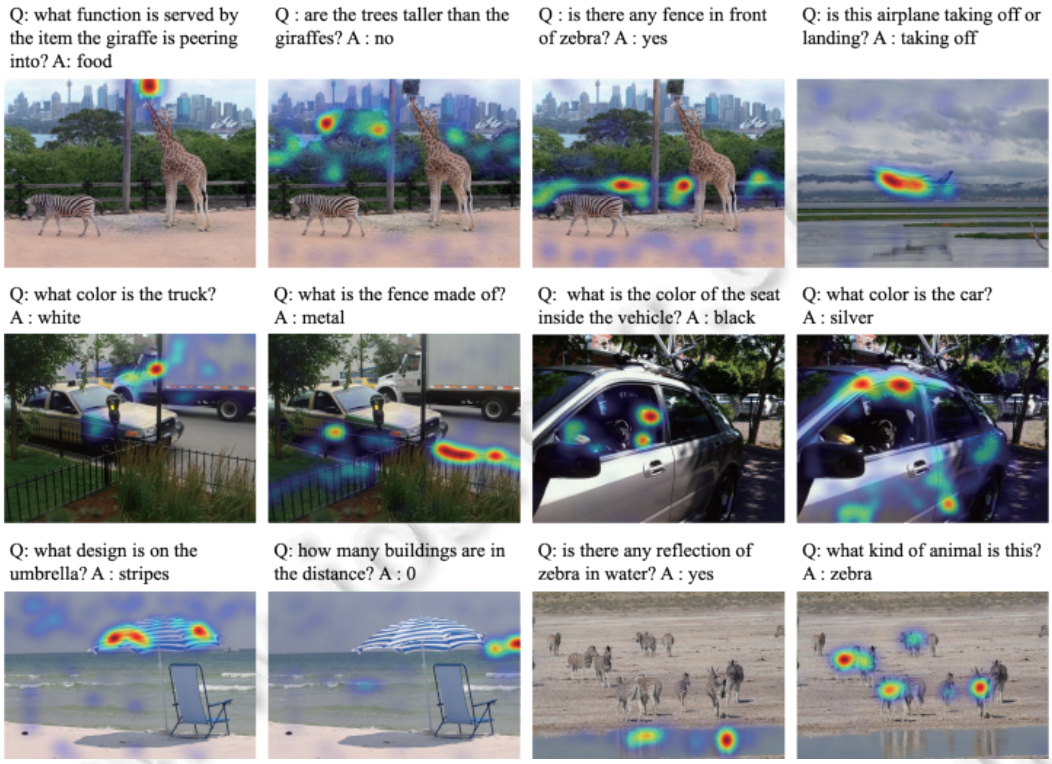


图 6　VQA 模型多模态编码器中跨模态对齐模块注意力图 Grad-CAM 可视化结果

图 7　多模态编码器中第 3 层跨模态对齐模块注意力图 Grad-CAM 可视化结果

图 8 自然语言视觉推理下游任务测试可视化结果

表 5 超图网络位置消融实验 (%)

所在层	Val	Test A	Test B
—	**58.46**	65.89	46.25
[1]	58.12	65.93	46.84
[1,3]	58.20	65.57	46.82
[1,3,5]	58.03	65.92	46.88
[1,3,4,5]	58.14	66.01	46.67
[1,2,3,4,5]	58.02	66.00	46.75
[1,2,3,4,5,6]	58.11	**66.03**	**46.94**

4 总 结

针对视觉-语言多模态预训练中多模态数据实体关系对齐欠缺的问题, 本文提出基于超图神经网络多元实体对齐的多模态预训练方法, 通过在 Transformer 多模态融合编码器中引入超图网络, 借助超图的超边的编码高阶相关性能力, 聚合学习得到多元实体的对齐关系从而增强多模态融合编码器的实体对齐构建能力, 进而提升视觉-语言多模态模型的泛化能力. 本文提出的方法在包含 400 万图像的大规模图像-文本数据集上进行了端到端的预训练, 并将预训练模型微调至 4 种视觉-语言下游任务中. 实验对比结果以及可视化分析结果验证了本文所提出的预训练模型方法能够有效地学习到多模态实体间的对齐关系, 并在多个视觉-语言下游任务性能均有所提升, 其中在 $NLVR^2$ 任务上相比于 baseline 方法 ALBEF 准确率提升 1.8%.

References:

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

[2] Wang NY, Ye YX, Liu L, Feng LZ, Bao T, Peng T. Language models based on deep learning: A review. Ruan Jian Xue Bao/Journal of Software, 2021, 32(4): 1082–1115 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6169.htm [doi: 10.13328/j.cnki.jos.006169]

[3] Uppal S, Bhagat S, Hazarika D, Majumder N, Poria S, Zimmermann R, Zadeh A. Multimodal research in vision and language: A review of current and emerging trends. Information Fusion, 2022, 77: 149–171. [doi: 10.1016/j.inffus.2021.07.009]

[4] Yang Y, Zhan DC, Jiang Y, Xiong H. Reliable multi-modal learning: A survey. Ruan Jian Xue Bao/Journal of Software, 2021, 32(4): 1067–1081 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6167.htm [doi: 10.13328/j.cnki.jos.006167]

© 中国科学院软件研究所 http://www.jos.org.cn

[5] Tan H, Bansal M. LXMERT: Learning cross-modality encoder representations from Transformers. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019. 5100–5111. [doi: 10.18653/v1/D19-1514]

[6] Chen YC, Li LJ, Yu LC, El Kholy A, Ahmed F, Gan Z, Cheng Y, Liu JJ. UNITER: Universal image-text representation learning. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 104–120. [doi: 10.1007/978-3-030-58577-8_7]

[7] Li XJ, Yin X, Li CY, Zhang PC, Hu XW, Zhang L, Wang ZJ, Hu HD, Dong L, Wei FR, Choi Y, Gao JF. OSCAR: Object-semantics aligned pre-training for vision-language tasks. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 121–137. [doi: 10.1007/978-3-030-58577-8_8]

[8] Lu JS, Goswami V, Rohrbach M, Parikh D, Lee S. 12-in-1: Multi-task vision and language representation learning. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10434–10443. [doi: 10.1109/CVPR42600.2020.01045]

[9] Li LH, Yatskar M, Yin D, Hsieh CJ, Chang KW. VisualBERT: A simple and performant baseline for vision and language. arXiv:1908.03557, 2019.

[10] 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. arXiv:2001.07966, 2020.

[11] Li G, Duan N, Fang YJ, Gong M, Jiang DX. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 11336–11344. [doi: 10.1609/aaai.v34i07.6795]

[12] Yu F, Tang JJ, Yin WC, Sun Y, Tian H, Wu H, Wang HF. ERNIE-ViL: Knowledge enhanced vision-language representations through scene graphs. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. Vancouver: AAAI, 2021. 3208–3216. [doi: 10.1609/aaai.v35i4.16431]

[13] Zhang PC, Li XJ, Hu XW, Yang JW, Zhang L, Wang LJ, Choi Y, Gao JF. VinVL: Revisiting visual representations in vision-language models. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 5575–5584. [doi: 10.1109/CVPR46437.2021.00553]

[14] Huang ZC, Zeng ZY, Huang YP, Liu B, Fu DM, Fu JL. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 12971–12980. [doi: 10.1109/CVPR46437.2021.01278]

[15] Huang ZC, Zeng ZY, Liu B, Fu DM, Fu JL. Pixel-BERT: Aligning image pixels with text by deep multi-modal Transformers. arXiv:2004.00849, 2020.

[16] Kim W, Son B, Kim I. ViLT: Vision-and-language Transformer without convolution or region supervision. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 5583–5594.

[17] Liu YF, Wu CF, Tseng SY, Lal V, He XM, Duan N. KD-VLP: Improving end-to-end vision-and-language pretraining with object knowledge distillation. In: Proc. of the 2022 Findings of the Association for Computational Linguistics. Seattle: Association for Computational Linguistics, 2022. 1589–1600. [doi: 10.18653/v1/2022.findings-naacl.119]

[18] Lu JS, Batra D, Parikh D, Lee S. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 2.

[19] Li JN, Selvaraju RR, Gotmare AD, Joty S, Xiong CM, Hoi SCH. Align before fuse: Vision and language representation learning with momentum distillation. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems. Curran Associates Inc., 2021. 742.

[20] Yang JY, Duan JL, Tran S, Xu Y, Chanda S, Chen LQ, Zeng BLD, Chilimbi T, Huang JZ. Vision-language pre-training with triple contrastive learning. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 15650–15659. [doi: 10.1109/CVPR52688.2022.01522]

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

[22] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. of the 9th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2021.

[23] Su WJ, Zhu XZ, Cao Y, Li B, Lu LW, Wei FR, Dai JF. VL-BERT: Pre-training of generic visual-linguistic representations. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[24] Ren SQ, He KM, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]

[25] Gan Z, Chen YC, Li LJ, Zhu C, Cheng Y, Liu JJ. Large-scale adversarial training for vision-and-language representation learning. In:

© 中国科学院软件研究所 http://www.jos.org.cn

Proc. of the 34rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 555.

[26] Li W, Gao C, Niu GC, Xiao XY, Liu H, Liu JC, Wu H, Wang HF. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing. Association for Computational Linguistics, 2021. 2592–2607. [doi: 10.18653/v1/2021.acl-long.202]

[27] Xu HY, Yan M, Li CL, Bi B, Huang SF, Xiao WM, Huang F. E2E-VLP: End-to-end vision-language pre-training enhanced by visual learning. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing. Association for Computational Linguistics, 2021. 503–513.

[28] Zou XY, Dou ZY, Yang JW, Gan Z, Li LJ, Li CY, Dai XY, Behl H, Wang JF, Yuan L, Peng NY, Wang LJ, Lee YJ, Gao JF. Generalized decoding for pixel, image, and language. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 15116–15127. [doi: 10.1109/CVPR52729.2023.01451]

[29] Wang AJ, Zhou P, Shou MZ, Yan SC. Enhancing visual grounding in vision-language pre-training with position-guided text prompts. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2024, 46(5): 3406–3421. [doi: 10.1109/TPAMI.2023.3343736]

[30] Schölkopf B, Platt J, Hofmann T. Learning with hypergraphs: Clustering, classification, and embedding. In: Proc. of the 20th Int'l Conf. on Neural Information Processing Systems. Vancouver: MIT Press, 2007. 1601–1608.

[31] Hu BD, Wang XG, Wang XY, Song ML, Chen C. Survey on hypergraph learning: Algorithm classification and application analysis. Ruan Jian Xue Bao/Journal of Software, 2022, 33(2): 498–523 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6353.htm [doi: 10.13328/j.cnki.jos.006353]

[32] Feng YF, You HX, Zhang ZZ, Ji RR, Gao Y. Hypergraph neural networks. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI, 2019. 3558–3565. [doi: 10.1609/aaai.v33i01.33013558]

[33] Ji SY, Feng YF, Ji RR, Zhao XB, Tang WW, Gao Y. Dual channel hypergraph collaborative filtering. In: Proc. of the 26th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. ACM, 2020. 2020–2029. [doi: 10.1145/3394486.3403253]

[34] Yadati N, Nimishakavi M, Yadav P, Nitin V, Louis A, Talukdar P. HyperGCN: A new method of training graph convolutional networks on hypergraphs. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 135.

[35] Jiang JW, Wei YX, Feng YF, Cao JX, Gao Y. Dynamic hypergraph neural networks. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: ijcai.org, 2019. 2635–2641. [doi: 10.24963/ijcai.2019/366]

[36] He KM, Fan HQ, Wu YX, Xie SN, Girshick R. Momentum contrast for unsupervised visual representation learning. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9726–9735. [doi: 10.1109/CVPR42600.2020.00975]

[37] van den Oord A, Li YZ, Vinyals O. Representation learning with contrastive predictive coding. arXiv:1807.03748v1, 2019.

[38] Sharma P, Ding N, Goodman S, Soricut R. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018. 2556–2565. [doi: 10.18653/v1/P18-1238]

[39] Ordonez V, Kulkarni G, Berg TL. Im2Text: Describing images using 1 million captioned photographs. In: Proc. of the 25th Int'l Conf. on Neural Information Processing Systems. Granada: Curran Associates Inc., 2011. 1143–1151.

[40] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Common objects in context. In: Proc. of the 13th European Conf. on Computer Vision. Zurich: Springer, 2014. 740–755. [doi: 10.1007/978-3-319-10602-1_48]

[41] Krishna R, Zhu YK, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li LJ, Shamma DA, Bernstein MS, Li FF. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. Int'l Journal of Computer Vision, 2017, 123(1): 32–73. [doi: 10.1007/s11263-016-0981-7]

[42] Antol S, Agrawal A, Lu JS, Mitchell M, Batra D, Zitnick CL, Parikh D. VQA: Visual question answering. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 2425–2433. [doi: 10.1109/ICCV.2015.279]

[43] Plummer BA, Wang LW, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 2641–2649. [doi: 10.1109/ICCV.2015.303]

[44] Suhr A, Zhou S, Zhang A, Zhang I, Bai HJ, Artzi Y. A corpus for reasoning about natural language grounded in photographs. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 6418–6428. [doi: 10.18653/v1/P19-1644]

[45] Yu LC, Poirson P, Yang S, Berg AC, Berg TL. Modeling context in referring expressions. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer. 2016. 69–85. [doi: 10.1007/978-3-319-46475-6_5]

© 中国科学院软件研究所 http://www.jos.org.cn

[46] Singh A, Hu RH, Goswami V, Couairon G, Galuba W, Rohrbach M, Kiela D. FLAVA: A foundational language and vision alignment model. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 15617–15629. [doi: 10.1109/CVPR52688.2022.01519]

[47] Dou ZY, Xu YC, Gan Z, Wang JF, Wang SH, Wang LJ, Zhu CG, Zhang PC, Yuan L, Peng NY, Liu ZC, Zeng M. An empirical study of training end-to-end vision-and-language Transformers. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 18145–18155. [doi: 10.1109/CVPR52688.2022.01763]

[48] Liu XJ, Li L, Wang SH, Zha ZJ, Meng DC, Huang QM. Adaptive reconstruction network for weakly supervised referring expression grounding. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 2611–2620. [doi: 10.1109/ICCV.2019.00270]

[49] Zhang Z, Zhao Z, Lin ZJ, Zhu JM, He XQ. Counterfactual contrastive learning for weakly-supervised vision-language grounding. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1521.

[50] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 618–626. [doi: 10.1109/ICCV.2017.74]

附中文参考文献:

[2] 王乃钰, 叶育鑫, 刘露, 凤丽洲, 包铁, 彭涛. 基于深度学习的语言模型研究进展. 软件学报, 2021, 32(4): 1082–1115. http://www.jos.org.cn/1000-9825/6169.htm [doi: 10.13328/j.cnki.jos.006169]

[4] 杨杨, 詹德川, 姜远, 熊辉. 可靠多模态学习综述. 软件学报, 2021, 32(4): 1067–1081. http://www.jos.org.cn/1000-9825/6167.htm [doi: 10.13328/j.cnki.jos.006167]

[31] 胡秉德, 王新根, 王新宇, 宋明黎, 陈纯. 超图学习综述: 算法分类与应用分析. 软件学报, 2022, 33(2): 498–523. http://www.jos.org.cn/1000-9825/6353.htm [doi: 10.13328/j.cnki.jos.006353]

李登(1992－), 男, 博士生, CCF 学生会员, 主要研究领域为多模态内容理解, 开放场景感知与理解.

韩亚洪(1977－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为多媒体内容理解, 人工智能安全.

武阿明(1987－), 男, 博士, 副教授, 博士生导师, CCF 专业会员, 主要研究领域为多模态内容理解, 开放场景感知与理解.

© 中国科学院软件研究所 http://www.jos.org.cn