基于跨模态引导和对齐的多模态预训练方法

才 华[1,2],易亚希[1],付 强[3],冉 越[1],孙俊喜[4]
(1. 长春理工大学电子信息工程学院,吉林长春 130022;2. 长春中国光学科学技术馆,吉林长春 130117;
3. 长春理工大学空间光电技术研究所,吉林长春 130022;4. 东北师范大学信息科学与技术学院,吉林长春 130117)

摘 要: 现有的视觉语言多模态预训练方法仅在图像和文本的全局语义上进行特征对齐,对模态间细粒度特征交互的探索不足. 针对这一问题,本文提出了一种基于跨模态引导和对齐的多模态预训练方法. 该方法在模态特征提取阶段,采用基于视觉序列压缩的双流特征提取网络,在视觉编码器中联合图像和文本信息逐层引导视觉序列压缩,缓解与文本无关的冗余视觉信息对模态间细粒度交互的干扰;在模态特征对齐阶段,对图像和文本特征进行细粒度关系推理,实现视觉标记与文本标记的局部特征对齐,增强对模态间细粒度对齐关系的理解. 实验结果表明,本文方法能够更好地对齐视觉文本的细粒度特征,在图文检索任务中,微调后的图像检索和文本检索的平均召回率分别达到了86.4% 和94.88%,且零样本图文检索的整体指标相较于经典图文检索算法CLIP(Contrastive Language-Image Pre-training)提升了5.36%,在视觉问答等分类任务中,准确率也优于目前主流多模态预训练方法.

关键词: 多模态预训练;跨模态引导;视觉序列压缩;双流特征提取;细粒度关系推理;局部特征对齐
基金项目: 国家自然科学基金(No.61890963,No.U2341226);吉林省人才专项(No.20240602015RC);西安市飞行器光学成像与测量技术重点实验室开放基金(No.2023-13)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2024)10-3368-14
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240271

Multimodal Pretraining with Cross-Modal Guidance and Alignment

CAI Hua[1,2], YI Ya-xi[1], FU Qiang[3], RAN Yue[1], SUN Jun-xi[4]
(1. *School of Electronic Information and Engineering*, *Changchun*, *Jilin* 130022, *China*;
2. *Changchun China Optics Science and Technology Museum*, *Changchun*, *Jilin* 130117, *China*;
3. *School of Opto-Electronic Engineer*, *Changchun University of Science and Technology*, *Changchun*, *Jilin* 130022, *China*;
4. *School of Information Science and Technology*, *Northeast Normal University*, *Changchun*, *Jilin* 130117, *China*)

Abstract: Current multimodal pre-training techniques for visual languages predominantly focus on aligning global semantic features between images and text, yet they inadequately explore the granular feature interactions between modalities. Addressing this gap, this paper proposes a novel multimodal pre-training strategy informed by cross-modal guidance and alignment. Our method employs a dual-stream feature extraction network designed for visual sequence compression, to facilitate modality feature extraction. During this phase, a synergistic image-text guidance is integrated within the visual encoder, orchestrating the compression of visual sequences layer by layer. This approach mitigates the obfuscation of modality-specific fine-grained interactions by irrelevant visual information. Subsequently, in the modality feature alignment phase, we implement fine-grained relational reasoning on the image and textual features to achieve localized feature alignment among visual tokens and textual tokens. This advancement bolsters the model´s comprehension of fine-grained alignment relationships. After fine-tuning, in the image-text retrieval tasks, our approach achieves an average recall rate of 86.4% for images and 94.88% for texts, which represents a significant 5.36% improvement in zero-shot image-text retrieval over the canonical CLIP (Contrastive Language-Image Pre-training) algorithm. Moreover, our method also surpasses existing mainstream multimodal pre-training methods in accuracy for classification tasks like visual question answering.

Key words: multimodal pre-training; cross-modal guidance; visual sequence compression; dual-stream feature extraction; fine-grained relational reasoning; localized feature alignment
Foundation Item(s): National Natural Science Foundation of China (No.61890963, No.U2341226); Jilin Province

收稿日期:2024-03-28;修回日期:2024-07-28;责任编辑:覃怀银

Talent Development Special Fund (No.202406020015RC); Xi' an Key Laboratory of Aircraft Optical Imaging and Measurement Technology (No.2023-13)

1　引言

视觉语言多模态学习属于计算机视觉与自然语言处理的交叉领域,旨在通过同时处理视觉和语言两种模态信息,实现跨模态的联合学习与理解. 在视频理解[1]、生物医学[2,3]、自动驾驶[4]等领域应用广泛. 与单一模态信息处理不同,图像数据和文本数据在表达方式和信息密度上都存在较大差异,这种不可忽视的模态异构性和语义鸿沟给多模态模型学习视觉语言对齐关系带来了极大的挑战.

现有的多模态预训练模型在处理视觉语言模态的对齐问题上,主要采取两种方法. 一种方法是基于图像区域和文本单词的对齐[5~7],利用训练好的目标检测器识别图像中所有潜在对象,实现文本单词与具体视觉对象的细粒度特征对齐. 然而,这种方法对下游任务的迁移受到目标检测器的类别限制,且识别到的潜在对象也可能与文本描述毫无关联. 另一种方法基于图文对比学习ITC(Image-Text Contrastive)对齐[8~10],专注于在图像和文本的全局语义上进行粗粒度对齐,克服了目标检测器的限制. 但仅进行全局语义的跨模态特征对齐,容易导致多模态模型忽略视觉对象和文本单词之间的细粒度对齐关系,而这对于许多下游任务是至关重要的[11,12]. 为此,FILIP[13](Fine-grained Interactive Language-Image Pre-training)提出了更精细的标记级对比对齐,但没有做进一步的模态交互融合,在复杂的视觉语言理解任务上表现较差. ALBEF[14](ALign BEfore Fuse)提出先对齐再融合的多模态学习策略,在多模态融合编码器中学习模态间细粒度交互关系,在各类多模态下游任务中取得了不错的效果,但其将模态间的深层次交互完全依托于融合过程中的注意力机制实现,缺乏更为显式的细粒度对齐手段.

此外,由于在现有的多模态数据集中,图像通常包含比简短文本描述更丰富的信息,这种模态间的信息不对称性导致多模态模型在使用ViT[15](Vision Transformer)提取图像特征时,会不可避免地编码与文本内容无关的冗余背景信息,这不仅增加了网络模型的计算负担,同时也会阻碍文本和图像之间的细粒度交互[16]. 尽管在计算机视觉领域,研究者对于ViT中的长视觉序列问题,提出了许多序列修剪[17,18]、序列重组[19]以及序列压缩[20]等算法,以减少视觉序列长度,提升计算效率. 但这些方法仅使用单一的视觉信息指导视觉序列修剪,忽略了文本信息,因此并不适合直接应用于需要同时处理两种模态信息的视觉语言多模态模型.

针对上述问题,本文遵循先对齐再融合的预训练策略,提出了一种基于跨模态引导和对齐的多模态预训练方法(Multimodal pretraining with cross-modal Guidance and Alignment,MulGA),在视觉语言特征提取过程中,通过双模态信息的联合引导,逐步压缩视觉序列中与文本描述无关的冗余标记,缓解背景特征对模态交互的干扰;同时在模态对齐阶段进行细粒度关系推理,加强模型对视觉对象和文本单词之间深层关联的理解,实现更精准的特征对齐.

2　方法架构

本文提出的视觉语言多模态预训练模型MulGA整体网络框架如图1所示. 该框架由基于视觉序列压缩的双流特征提取网络、细粒度关系推理模块以及多模态交互融合模块三个核心组件构成.

首先,输入的图像-文本对经过双流特征提取网络提取特征向量,并在细粒度关系推理模块中进行初步的特征细粒度对齐处理;随后,将初步对齐的图像特征和文本特征馈送到多模态交互融合模块,通过执行图像文本匹配(Image-Text Matching,ITM)和掩蔽语言建模(Masked Language Modeling,MLM)两个预训练任务,实现进一步的模态细粒度融合;最终输出深度融合后的多模态表征. 具体流程如算法1所示.

2.1　基于视觉序列压缩的双流特征提取网络

本文采用双流特征提取架构,通过两个独立的单模态编码器分别提取输入的图像和文本特征,如图2所示.

文本编码器由6层Transformer[21]块组成,每层包含一个多头自注意力(Multi-head Self-Attention,MSA)和一个前馈神经网络(Feedforward Neural Network,FNN),层归一化和残差连接用于防止梯度消失. 接收原始文本和掩蔽文本输入,将输入的文本单词转换为词嵌入向量$\boldsymbol{t}_i$,并与位置嵌入$\boldsymbol{T}_{\text{pos}}$求和构造输入表示. 需要注意的是,掩蔽文本输入仅作为后续MLM预训练任务使用,不参与文本编码器的梯度更新.

文本编码器计算过程,见式(1)~(3).

$$\boldsymbol{W}_0=(\boldsymbol{t}_{\text{cls}},\boldsymbol{t}_1,\boldsymbol{t}_2,\boldsymbol{t}_3,\cdots,\boldsymbol{t}_N)+\boldsymbol{T}_{\text{pos}} \tag{1}$$

$$\boldsymbol{W}_l'=\text{LN}(\text{MSA}(\boldsymbol{W}_{l-1})+\boldsymbol{W}_{l-1}) \tag{2}$$

$$\boldsymbol{W}_l=\text{LN}(\text{MLP}(\boldsymbol{W}_l')+\boldsymbol{W}_l') \tag{3}$$

其中,$\boldsymbol{W}_l$表示第l层编码器中序列化的文本特征矩阵.

视觉编码器由12层的ViT块组成,将输入图像$I\in\mathbf{R}^{H\times W\times C}$拆分成$N$个$P\times P$大小的图像块,其中$N=H\times W/P^2$. 每个图像块被展平并通过一个可训练的线性投影映射成D维的图像块嵌入向量$\boldsymbol{v}_i$,并与位置嵌入

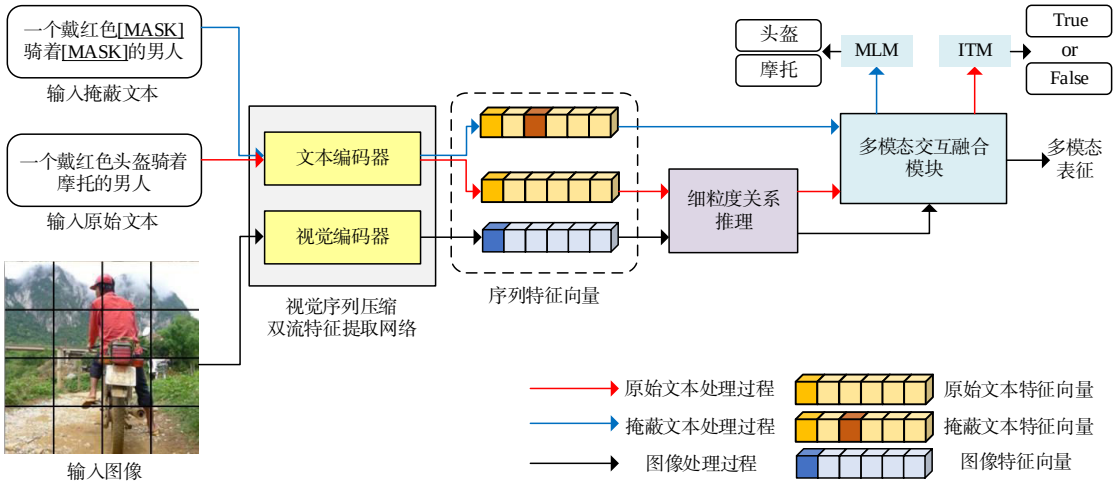


图 1 MulGA 模型整体框架

算法 1 MulGA

输入:图像 I,原始文本 T,掩蔽文本 T_{mask},视觉标记保留比例 λ

输出:多模态表征 $\boldsymbol{F}_{\text{fusion}}$

Function MulGA(I, T, T_{mask}, λ)
  //基于视觉序列压缩的双流特征提取
  $\boldsymbol{W}, \boldsymbol{W}_{\text{mask}} \leftarrow \text{BERT}(T, T_{\text{mask}})$;
  for $i=1$ to 12 do
    if $i \leqslant 6$ then
      $\boldsymbol{Z}_i \leftarrow \text{ViT}(I)$; //前六层保持不变
    else $i \leqslant 9$ then
      $\boldsymbol{m}_{\text{cls}} \leftarrow \boldsymbol{t}_{\text{cls}} + \boldsymbol{v}_{\text{cls}}$; //合并 CLS 标记
      $\boldsymbol{a}_{\text{cls}} \leftarrow \text{MSA}(\boldsymbol{m}_{\text{cls}}, \boldsymbol{Z}_i)$;
      划分相关序列 $\boldsymbol{Z}^{\text{R}}$ 和非相关序列 $\boldsymbol{Z}^{\text{NR}}$;
      $s_{x,y} \leftarrow \text{Similarity}(\boldsymbol{v}_x \in \boldsymbol{Z}^{\text{NR}}, \boldsymbol{v}_y \in \boldsymbol{Z}^{\text{R}})$;
      合并融合,得到压缩视觉序列 $\boldsymbol{Z}_{\text{out}}$;
    else
      $\boldsymbol{Z}_i \leftarrow \text{ViT}(\boldsymbol{Z}_{i-1})$; //后三层保持不变
  //细粒度关系推理
  for $k=1$ to M do
    //计算图像和文本特征的标记级相似度
    $S_{\text{I}}, S_{\text{T}} \leftarrow \text{Token\_Similarity}(\mathbf{Z}^k, \mathbf{W}^k)$;
  计算对比损失 $L_{\text{itc}} \leftarrow \text{ITC}(S_{\text{I}}, S_{\text{T}})$;
  //多模态交互融合
  $\boldsymbol{F}_{\text{fusion}}, \boldsymbol{F}_{\text{fusion}}^{\text{mask}} \leftarrow \text{MCA}(\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{W}_{\text{mask}})$;
  执行预训练任务 ITM,MLM($\boldsymbol{F}_{\text{fusion}}, \boldsymbol{F}_{\text{fusion}}^{\text{mask}}$);
  计算总损失 $L_{\text{total}} \leftarrow L_{\text{itc}} + L_{\text{itm}} + L_{\text{mlm}}$;
  return

$\boldsymbol{V}_{\text{pos}}$ 求和构造序列化的视觉特征矩阵 $\boldsymbol{Z}_0$,如式(4)所示:

$$\boldsymbol{Z}_0 = (\boldsymbol{v}_{\text{cls}}, \boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \cdots, \boldsymbol{v}_N) + \boldsymbol{V}_{\text{pos}} \tag{4}$$

在输入数据中,并不是所有的图像内容都有与之相匹配的文本描述,有必要对视觉序列进行修剪操作,减少与文本无关的冗余视觉信息. 但现有的主流视觉序列修剪方法只利用了单一的视觉信息,并不适合需要处理两种模态信息的多模态模型. 为此,本文提出图像文本联合引导的视觉序列压缩(image-text coordinated Visual Sequence Compression,VSC)方法,如图 3 所示.

保持视觉编码器中的前六层和后三层不变,仅在中间三层逐层选择并压缩与文本描述不相关的视觉标记,避免因压缩过多视觉标记导致图像结构信息丢失.

对于第 l 层视觉序列矩阵 $\boldsymbol{Z}_l \in \mathbf{R}^{(n+1)\times D}$,首先将序列中图像 CLS 特征向量 $\boldsymbol{v}_{\text{cls}}$ 和文本编码器输出的文本 CLS 特征向量 $\boldsymbol{t}_{\text{cls}}$ 合并,获得具有多模态语义的联合特征向量 $\boldsymbol{m}_{\text{cls}}$,如式(5)所示:

$$\boldsymbol{m}_{\text{cls}} = \frac{1}{2}(\boldsymbol{v}_{\text{cls}} + \boldsymbol{t}_{\text{cls}}) \tag{5}$$

然后将联合标记与视觉序列级联送入多头自注意力层,计算联合标记对每个视觉标记的注意力得分 $\boldsymbol{a}_{\text{cls}}$,作为视觉标记与文本描述相关程度的评估标准,如式(6)所示:

$$\boldsymbol{a}_{\text{cls}} = \text{softmax}\left(\frac{\boldsymbol{m}_{\text{cls}} \cdot \boldsymbol{Z}_l[1{:}]^{\text{T}}}{\sqrt{d}}\right) \tag{6}$$

依据 $\boldsymbol{a}_{\text{cls}} = (a_1, a_2, \cdots, a_n)$ 和预先定义的视觉标记保留比例 λ 对特征序列进行分组. 将需要保留的高得分标记划分为相关序列 $\boldsymbol{Z}^{\text{R}}$,可以忽视的低得分标记划分为非相关序列 $\boldsymbol{Z}^{\text{NR}}$.

对于非相关序列中的视觉标记,直接删除会导致信息丢失,简单融合又会生成额外的标记,破坏视觉信息的整体性[20]. 为了避免这些问题,本文采用一种序列压缩策略. 首先,对非相关序列中的每一个视觉标记 $\boldsymbol{v}_x = \boldsymbol{Z}^{\text{NR}}[:, x]$,在相关标记 $\boldsymbol{v}_y = \boldsymbol{Z}^{\text{R}}[:, y]$ 中找到其最近邻的视觉标记 $\boldsymbol{v}_{\text{near}}$.

最近邻标记计算过程见式(7)和式(8).

$$s_{x,y} = \frac{\boldsymbol{v}_x^{\text{T}} \boldsymbol{v}_y}{\|\boldsymbol{v}_x\| \|\boldsymbol{v}_y\|} \tag{7}$$

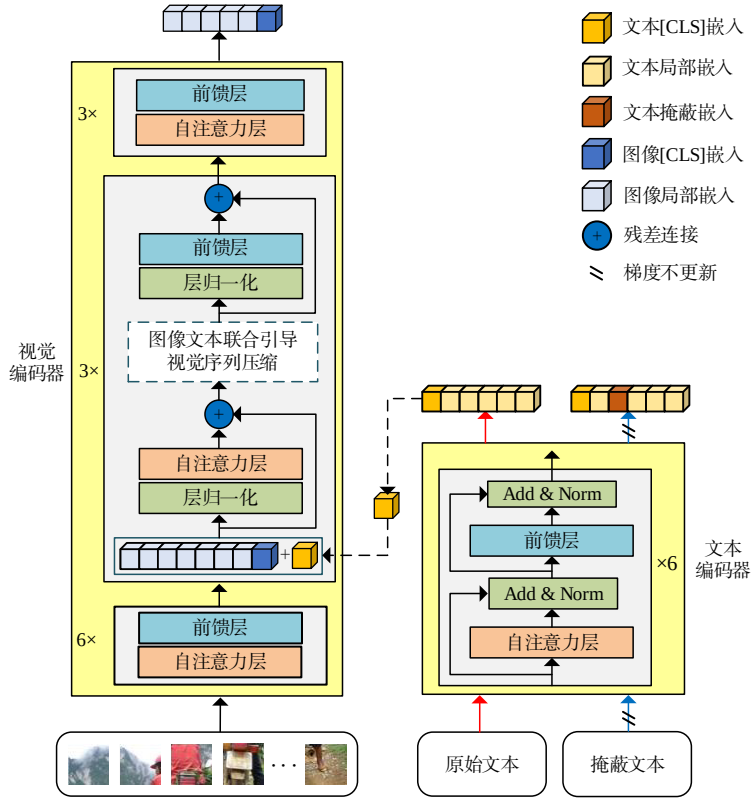


图2　双流特征提取架构示意图

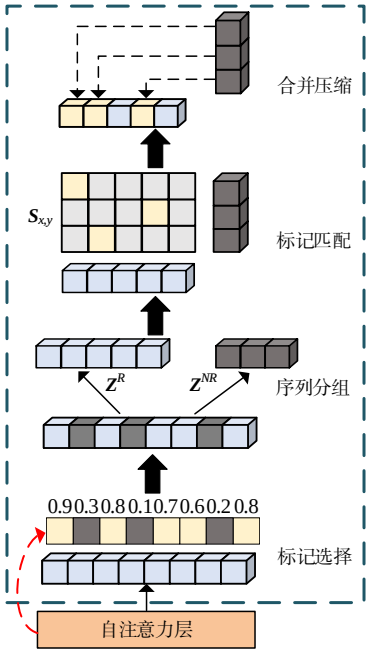


图3　图像文本联合引导视觉序列压缩示意图

$$\boldsymbol{v}_{\text{near}} = \underset{\boldsymbol{v}_y = \boldsymbol{Z}^{\text{R}}[:,y]}{\arg\max}\, s_{x,y} \tag{8}$$

随后,计算非相关标记与最近邻相关标记的融合权值,将非相关标记加权合并到其最近邻标记中,获得更新后的令牌 $\boldsymbol{v}_{\text{update}}$. 在不损失信息也不增加额外标记的情况下,实现视觉序列压缩,如式(9)~(11)所示:

$$\theta_x = \frac{\exp(s_{x,y})}{\sum\limits_{\boldsymbol{v}_x = \boldsymbol{Z}^{\text{NR}}[:,x]} \exp(s_{x,y}) + \text{e}} \tag{9}$$

$$\theta_y = \frac{\text{e}}{\sum\limits_{\boldsymbol{v}_x = \boldsymbol{Z}^{\text{NR}}[:,x]} \exp(s_{x,y}) + \text{e}} \tag{10}$$

$$\boldsymbol{v}_{\text{updata}} = \theta_y \boldsymbol{v}_y + \sum_{\boldsymbol{v}_x = \boldsymbol{Z}^{\text{NR}}[:,x]} \theta_x \boldsymbol{v}_x \tag{11}$$

其中,θ_x 为每个非相关标记的融合权值,θ_y 为相关标记本身的融合权值.

最后,将当前层的视觉序列重构为 $\boldsymbol{Z}_l \in \mathbf{R}^{(k+1)\times D}$,并馈送到多层感知机模块进一步处理,其中 $k = \lambda n$.

2.2　基于对比学习的细粒度关系推理

细粒度关系推理模块(Fine-grained Relationship Inference Module,FRIM)利用视觉和语言表征的标记级相似度综合表示图像与文本以及文本与图像的相似度,并使用对比损失最大化正样本对的相似度、最小化负样本对的相似度,实现更细粒度和更具可解释性的跨模态对齐,具体流程如图4所示.

假设在一个训练批次中有 M 个图像-文本对 $\{I_k, T_k\}_{k=1}^{M}$,其中 $\{I_k, T_k\}$ 为正样本对,而 I_k 与其他文本组合 $\{I_k, T_m\}_{m\neq k}^{M}$ 为负样本对. 对于该批次中的第 k 个图像文本对 $\{I_k, T_k\}$,将其视觉特征 $\boldsymbol{Z}^k = (\boldsymbol{v}_{\text{cls}}, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_{n_1})$ 和文

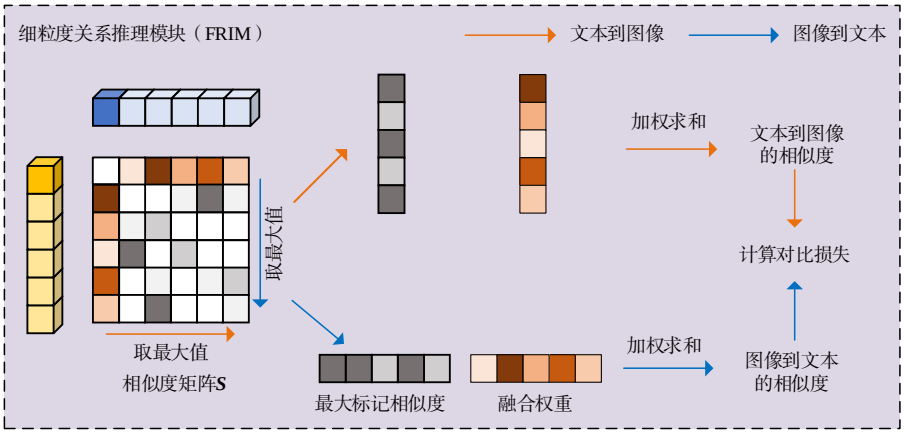


图4 细粒度关系推理模块

本特征 $\boldsymbol{W}^k=(\boldsymbol{t}_{\text{cls}},\boldsymbol{t}_1,\cdots,\boldsymbol{t}_{n_2})$ 经过单层感知机映射到公共嵌入空间，用 $g_{\text{v}}(\boldsymbol{Z}^k)$ 和 $g_{\text{t}}(\boldsymbol{W}^k)$ 表示，其中 n_1 和 n_2 为图像和文本的非填充标记数量. 为了实现图像与文本的细粒度对齐，细粒度关系推理模块需要对每个特征标记都计算相似度，最终获得一个 $(n_1+1)\times(n_2+1)$ 的标记级相似性矩阵 $\boldsymbol{S}$，如式(12)所示：

$$\boldsymbol{S}=g_{\text{v}}(\boldsymbol{Z}^k)^{\text{T}}g_{\text{t}}(\boldsymbol{W}^k) \tag{12}$$

对于图像的第 i 个视觉标记 $\boldsymbol{v}_i$，取其中除全局语义表示外 n_2 个相似度中最大的一个作为该视觉标记与整段文本的相似度，如式(13)所示：

$$\max_{0\leqslant j<n_2} g_{\text{v}}(\boldsymbol{v}_i)^{\text{T}}g_{\text{t}}(\boldsymbol{t}_j) \tag{13}$$

这样可以得到 n_1 个视觉标记与整段文本的相似度，同理也能获得 n_2 个文本标记与整幅图像的相似度. 为了进一步量化 I_k 和 T_k 之间的匹配程度，通过对全局语义表示与各标记之间的相似性进行 softmax 归一化处理，评估各个标记在匹配过程中的贡献程度，获得对应的融合权重 ω，进而加权融合得到图像到文本的相似度 $S_{\text{I}}(\boldsymbol{Z}^k,\boldsymbol{W}^k)$ 与文本到图像的相似度 $S_{\text{T}}(\boldsymbol{W}^k,\boldsymbol{Z}^k)$.

$S_{\text{I}}(\boldsymbol{Z}^k,\boldsymbol{W}^k)$ 和 $S_{\text{T}}(\boldsymbol{W}^k,\boldsymbol{Z}^k)$ 的计算过程见式(14)和式(15).

$$S_{\text{I}}(\boldsymbol{Z}^k,\boldsymbol{W}^k)=\begin{cases}\displaystyle\sum_{i=1}^{n_1}\omega_i g_{\text{v}}(\boldsymbol{v}_i)^{\text{T}}g_{\text{t}}(\boldsymbol{t}_p), & \text{if } p=\underset{0\leqslant j<n_2}{\arg\max}\, g_{\text{v}}(\boldsymbol{v}_i)^{\text{T}}g_{\text{t}}(\boldsymbol{t}_j)\\ 0, & \text{if } p\neq\underset{0\leqslant j<n_2}{\arg\max}\, g_{\text{v}}(\boldsymbol{v}_i)^{\text{T}}g_{\text{t}}(\boldsymbol{t}_j)\end{cases} \tag{14}$$

$$S_{\text{T}}(\boldsymbol{W}^k,\boldsymbol{Z}^k)=\begin{cases}\displaystyle\sum_{j=1}^{n_2}\omega_j g_{\text{t}}(\boldsymbol{t}_j)^{\text{T}}g_{\text{v}}(\boldsymbol{v}_p), & \text{if } p=\underset{0\leqslant i<n_1}{\arg\max}\, g_{\text{t}}(\boldsymbol{t}_j)^{\text{T}}g_{\text{v}}(\boldsymbol{v}_i)\\ 0, & \text{if } p\neq\underset{0\leqslant i<n_1}{\arg\max}\, g_{\text{t}}(\boldsymbol{t}_j)^{\text{T}}g_{\text{v}}(\boldsymbol{v}_i)\end{cases} \tag{15}$$

需要注意的是，与图文对比学习不同，此时的 $S_{\text{I}}(\boldsymbol{Z}^k,\boldsymbol{W}^k)$ 不一定等于 $S_{\text{T}}(\boldsymbol{W}^k,\boldsymbol{Z}^k)$.

最后计算图文对比损失，实现图像和文本的细粒度交互和对齐，如式(16)～(18)所示：

$$L_k^{\text{i2t}}=-\frac{1}{M}\log\frac{\exp(S_{\text{I}}(\boldsymbol{Z}^k,\boldsymbol{W}^k)/\tau)}{\sum_{m=1}^{M}\exp(S_{\text{I}}(\boldsymbol{Z}^k,\boldsymbol{W}^m)/\tau)} \tag{16}$$

$$L_k^{\text{t2i}}=-\frac{1}{M}\log\frac{\exp(S_{\text{T}}(\boldsymbol{W}^k,\boldsymbol{Z}^k)/\tau)}{\sum_{m=1}^{M}\exp(S_{\text{T}}(\boldsymbol{W}^k,\boldsymbol{Z}^m)/\tau)} \tag{17}$$

$$L_{\text{itc}}=\frac{1}{2}\sum_{k=1}^{M}(L_k^{\text{i2t}}+L_k^{\text{t2i}}) \tag{18}$$

其中，τ 是可学习的温度参数，L_k^{i2t} 为图像 I_k 到文本的对比损失，L_k^{t2i} 为文本 T_k 到图像的对比损失，L_{itc} 为该训练批次的总对比损失.

2.3 多模态交互融合模块

多模态交互融合模块如图5所示，输入对齐后图像特征、原始文本特征以及掩蔽文本特征，其中掩蔽文本特征和原始文本特征经过多头自注意力层作为查询向量，图像特征作为键向量和值向量，在多头交叉注意力层(Multi-head Cross-Attention，MCA)进行融合. 通过图像文本匹配(ITM)和掩蔽语言建模(MLM)两个预训练任务训练融合编码器隐式挖掘图像和文本的细粒度交互关系，如式(19)～(22)所示：

$$\boldsymbol{M}_0=\boldsymbol{W} \tag{19}$$

$$\boldsymbol{M}_l''=\text{LN}\left(\text{MSA}\left(\boldsymbol{M}_{l-1}\right)\right)+\boldsymbol{M}_{l-1} \tag{20}$$

$$\boldsymbol{M}_l'=\text{LN}\left(\text{MCA}\left(\boldsymbol{M}_l'',\boldsymbol{Z}\right)\right)+\boldsymbol{M}_l'' \tag{21}$$

$$\boldsymbol{M}_l=\text{LN}\left(\text{MLP}(\boldsymbol{M}_l')\right)+\boldsymbol{M}_l' \tag{22}$$

其中，$\boldsymbol{W}$ 是文本编码器的输出，$\boldsymbol{Z}$ 是视觉编码器的输出，$\boldsymbol{M}_l$ 是第 l 层编码器中序列化的多模态特征矩阵.

图像文本匹配(ITM)任务将融合编码器输出的文本 CLS 嵌入作为多模态全局语义表示，经过一个全连

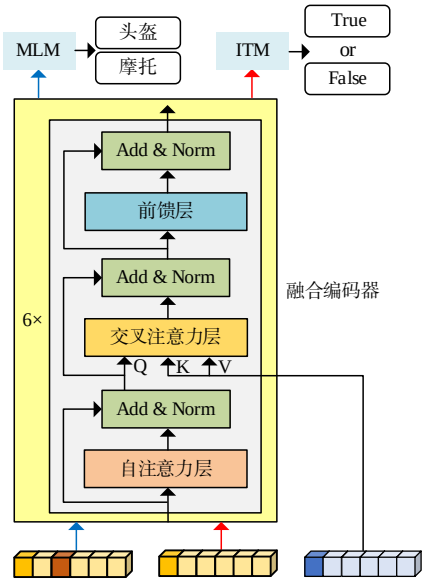


图5　多模态交互融合模块

接层和Softmax来预测图像和文本是否匹配. ITM损失计算如式(23)所示:

$$L_{\text{itm}}=\mathrm{E}_{(I,T)\sim M}(\mathrm{H}(y^{\text{itm}},p^{\text{itm}}(I,T))) \tag{23}$$

其中,p^{itm}为预测两类的概率,y^{itm}表示真值标签,y^{itm}当图像和文本相匹配时为1,不匹配时为0.

掩蔽语言建模(MLM)任务利用图像信息和文本上下文信息预测被随机掩蔽的文本标记. 由于提高MLM任务中文本的掩蔽比例,能够帮助模型更有效地利用视觉特征[22,23]. 因此本文以30%的概率随机掩蔽原始文本. MLM损失计算如式(24)所示:

$$L_{\text{mlm}}=\mathrm{E}_{(I,T_{\text{mask}})\sim M}(\mathrm{H}(y^{\text{msk}},p^{\text{msk}}(I,T_{\text{mask}}))) \tag{24}$$

其中T_{mask}表示一个被掩蔽的文本标记,$p^{\text{msk}}(I,T_{\text{mask}})$表示模型对被掩蔽标记的预测概率,$y^{\text{msk}}$表示词汇概率分布.

MulGA的预训练总损失可以表示为

$$L_{\text{total}}=\alpha L_{\text{itc}}+\beta L_{\text{itm}}+\gamma L_{\text{mlm}} \tag{25}$$

3　实验分析

本节详细介绍了MulGA在视觉语言检索和分类两种下游任务中的实验测试结果,并与主流的预训练方法进行了全面比较,展示了本文模型在各个任务上的良好表现.

3.1　实验细节

MulGA使用预训练的ViT-B/16作为视觉编码器,将BERT[24]的前六层作为文本编码器,后六层作为多模态融合编码器,视觉标记保留比例λ设置为70%. 在4张NVIDIA Tesla V100 32 GB GPU上预训练了30个周期. 为了公平比较,除了由于计算资源受限,降低了批次大小之外,其余实验设置基本与ALBEF保持一致.

3.2　预训练数据集

本文采用MSCOCO[25](Microsoft Common Objects in Context)、VG[26](Visual Genome)、SBU[27](SBU Captions)和CC3M[28](Conceptual Caption)4个可公开访问的数据集作为基础数据集. 由于四个数据集中的部分图像数据在下载前已缺失,最终4 M大小的基础数据集中,图像总数约为400万张,图像-文本对的数量约为510万套.

此外,本文以随机采样的方式从CC12 M数据集中筛选数据,构造了一个包含6 M图像文本对的数据集,与采用大规模数据集预训练的模型进行对比实验.

3.3　视觉语言检索任务评估

视觉语言检索任务包括图像到文本检索(Image Retrieval,IR)和文本到图像检索(Text Retrieval,TR). 对于此项任务,本文使用召回率R@k(Recall@k)、平均精确率mAP(Average Precision)和F_1分数对模型进行全面评估.

3.3.1　图文检索微调实验

MulGA在MSCOCO和Flickr30k[29]上图文检索微调实验的R@k的评估结果如表1所示. 其中第二列代表预训练数据集中的图像数目,"—"表示该模型结果未报告.

在使用4 M公共预训练数据集的条件下,与ALBEF模型相比,MulGA在MSCOCO和Flickr30 k上的平均召回率分别提升了2.9%和1.2%. 并在Flickr30 k上的文本检索召回率达到了最优水平. 当使用6 M预训练数据集时,MulGA展现了与BLIP(Bootstrapping Language-Image Pre-training)相当的性能,在Flickr30 k上的TR@5和TR@10达到100,相比于在更大数据集上预训练的ALBEF,在MSCOCO数据集的TR@1指标上实现了1.9%的提升.

微调检索实验的mAP与F_1分数评估结果如表2所示,其中mAP为模型在1、5、10不同阈值下的精确率均值. 在4 M数据条件下,MulGA在MSCOCO数据集上的mAP优于TCL(Triple Contrastive Learning),略低于使用14 M数据预训练的ALBEF和BLIP,在Flickr30 k上mAP和F_1分数仅次于BLIP;当预训练数据扩充到6 M时,mAP增长了1.14%,模型性能超过ALBEF,与BLIP模型相近.

为了验证实验结果的可靠性,本文基于Wilcoxon符号秩检验法,以召回率作为各个模型的样本数据,分析MulGA在图文检索微调任务中相较于其他模型的提升是否显著,各个配对样本的P值以热力图的形式展现,如图6所示.

分析结果表明,本文方法与UNITER(UNiversal Image-TExt Representation)、ALBEF、TCL等模型相比,

表1 图文检索微调实验-召回率 单位:%

模型	数据集	MSCOCO						Flickr30 k					
		Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval		
	图像	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
UNITER[6]	4 M	65.7	88.6	93.8	59.2	79.9	88.0	87.3	98	99.2	75.6	94.1	96.8
ImageBERT[30]	6 M	66.4	89.8	94.4	50.5	78.7	87.1	87.0	97.6	99.2	73.1	92.6	96.0
OSCAR[7]	4 M	70.0	91.1	95.5	54.0	80.8	88.5	—	—	—	—	—	—
ALBEF[14]	4 M	73.1	91.4	96.0	56.8	81.5	89.2	94.3	99.4	99.8	82.8	96.7	98.4
METER[31]	4 M	76.2	93.2	96.8	57.1	82.7	90.1	94.3	99.6	99.9	82.2	96.3	98.4
TCL[32]	4 M	75.6	92.8	96.7	59.0	83.2	89.9	94.9	99.5	99.8	84.0	96.7	98.5
ViLTA[33]	4 M	73.3	91.8	95.9	59.5	83.1	89.7	94.5	99.8	99.8	85.2	97.2	98.8
MaskVLM[34]	4 M	76.3	93.8	96.8	60.1	83.6	90.4	95.3	99.8	100	84.9	97.4	98.6
VL-Match[35]	4 M	76.6	93.8	97.1	60.2	83.6	90.1	96.4	99.8	100	86.0	97.5	99.0
ALIGN[9]	1.8 B	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	100	84.9	97.4	98.6
ALBEF[14]	14 M	77.6	94.3	97.2	60.7	84.3	90.5	95.9	99.8	100	85.6	97.5	98.9
BLIP[36]	14 M	80.6	95.2	97.6	63.1	85.3	91.1	96.6	99.8	100	87.2	97.5	98.8
本文方法	4 M	77.1	93.6	97.1	60.0	83.9	90.5	96.4	99.9	100	85.6	97.5	98.9
本文方法	6 M	79.1	95.2	97.8	60.9	84.3	91.3	97.2	100	100	86.5	97.0	98.4

表2 图文检索微调实验-精确率、F_1分数 单位:%

模型	数据集	MSCOCO				Flickr30 k			
	图像	T-mAP	T-F_1	I-mAP	I-F_1	T-mAP	T-F_1	I-mAP	I-F_1
TCL[32]	4 M	34.60	41.36	28.20	34.35	41.59	48.73	37.73	44.71
ALBEF[14]	14 M	35.39	42.23	28.86	35.08	41.59	49.11	38.31	45.34
BLIP[36]	14 M	36.47	43.34	29.76	36.03	42.39	49.56	38.90	45.93
本文方法	4 M	35.18	41.99	28.61	34.82	42.13	49.29	38.33	45.36
本文方法	6 M	35.97	42.87	28.96	35.22	42.43	49.57	38.58	45.58

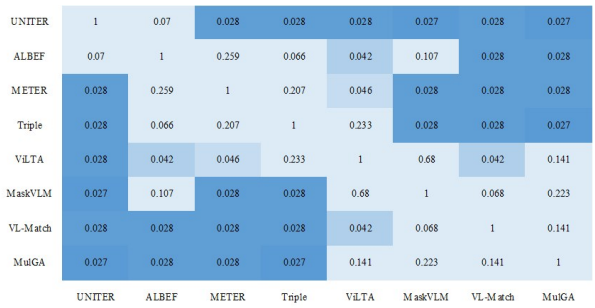


图6 微调检索实验统计显著性验证

图文检索具有显著性提升,在ViLTA(Vision-Language pretraining through Textual Augmentation)和MaskVLM(Masked Vision and Language Modeling)等模型上,提升并不明显.分析结果基本与R@k、mAP和F_1分数三个评估指标的评估结果一致,证明本文实验结果具有一定可靠性.

3.3.2 图文检索零样本实验

为了评估MulGA的零样本迁移能力,本文将在MSCOCO数据集上微调的模型直接迁移到Flickr30 k数据集上进行了图文检索零样本实验.表3展示了MulGA在Flickr30 k上零样本检索的评估结果.

表3 图文检索零样本实验-召回率 单位:%

模型	数据集	Flickr30 k					
	图像	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10
ImageBERT[30]	6 M	70.7	90.2	94.0	54.3	79.6	87.5
UNITER[6]	4 M	83.6	95.7	97.7	68.7	89.2	93.9
MaskVLM[34]	4 M	87.0	97.9	99.3	75.0	92.5	95.8
CLIP[8]	400 M	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN[9]	1.2 B	88.6	98.7	99.7	75.7	93.8	96.8
ALBEF[14]	4 M	90.5	98.8	99.7	76.8	93.7	96.7
METER[31]	4 M	90.9	98.3	99.5	79.6	95.0	97.3
TCL[32]	4 M	93.0	99.1	99.6	79.6	95.1	97.4
VL-Match[35]	4 M	93.3	99.3	99.8	82.0	95.1	97.4
ALBEF[14]	14 M	94.1	99.5	99.7	82.8	96.3	98.1
BLIP[36]	14 M	94.8	99.7	100	84.9	96.8	98.6
本文方法	4 M	91.1	99.1	99.8	80.2	95.3	97.3
本文方法	6 M	93.2	99.6	99.9	82.8	96.0	98.1

相较于ImageBERT、CLIP和ALIGN,MulGA用更少的训练数据获得了更好的性能.与同等训练数据的模型相比,零样本检索性能也优于ALBEF、METER(Multi-modal End-to-end TransformER)等模型.但对比不同数据体量下的两种最优算法VL-Match和BLIP,MulGA在TR@1和IR@1上的表现仍有较大差距,推测可能是固定的视觉压缩比例导致在进行零样本迁移时无法准确压缩图像冗余信息,一定程度的主体结构信息被破坏,最终影响了模型性能.

零样本检索实验的mAP与F_1分数评估结果在表4中显示.MulGA零样本迁移表现与TCL模型相当.将预训练数据集扩充到6 M大小时,模型的零样本检索性能会有显著提升.这一结果表明,通过扩大预训练数据集规模,可以有效增强模型在处理零样本迁移任务时的泛化能力.

表4　图文检索零样本实验-精确率、F_1分数　　单位:%

模型	数据集	Flickr30 k			
	图像	T-mAP	T-F_1	I-mAP	I-F_1
TCL[32]	4 M	40.93	48.04	36.13	42.83
ALBEF[14]	14 M	41.32	48.47	37.30	44.26
BLIP[36]	14 M	41.58	48.74	38.04	45.03
本文方法	4 M	40.30	47.43	36.33	43.17
本文方法	6 M	41.04	48.19	37.27	44.21

为了验证零样本实验结果的可靠性,本文对实验数据进行了显著性分析,统计显著性验证结果如图7所示.

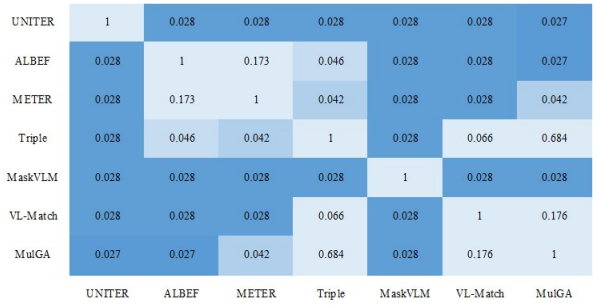


图7　零样本检索实验统计显著性验证

从图中结果可以得出结论,MulGA在零样本检索实验中与UNITER、ALBEF、MaskVLM相比,具有显著的性能提升,但与METER、VL-Match(enhancing Vision-Language with token-level Matching)模型相比提升不具备显著性,总体表现与TCL模型相近.

3.4　视觉语言分类任务评估

为了进一步验证MulGA的性能,本文在视觉问答、视觉推理和视觉蕴含三个视觉语言分类任务上进行了实验.

3.4.1　视觉问答实验

视觉问答(Viual Question Answering,VQA)任务要求模型根据给定的图像和问题,生成或从答案集合中选择出正确的答案.本文在VQAv2[37]数据集上微调和评估MulGA的视觉问答能力.评估结果在表5中展示.

表5　视觉问答和视觉推理实验结果　　单位:%

模型	数据集图像	VQAv2		NLVR2	
		test-dev	test-std	dev	test
UNITER[6]	4 M	72.7	72.9	77.2	77.9
GLIPv2[38]	20 M	73.1	73.3	—	—
OSCAR[7]	4 M	73.2	73.4	78.1	78.4
ALBEF[14]	4 M	74.5	74.7	80.2	80.5
GRIT-VLP[39]	4 M	75.1	75.3	80.7	81.6
VL-Match[35]	4 M	75.1	75.2	82.0	82.2
MaskVLM[34]	4 M	75.5	75.4	81.6	82.0
ALBEF[14]	14 M	75.8	76.0	82.6	83.1
METER[31]	4 M	76.4	76.4	82.4	82.2
BLIP[36]	14 M	77.5	77.6	82.7	82.3
SimVLM[40]	1.8 B	77.9	78.1	81.7	81.8
OFA[41]	54 M	78.0	78.1	—	—
本文方法	4 M	76.9	76.6	82.3	82.7
本文方法	6 M	77.6	77.9	82.7	82.9

基于4 M数据集,MulGA在Test-Dev和Test-Std上的准确率优于所有同等训练体量的对比算法,甚至超越了使用更大数据集进行预训练的GLIPv2(Grounded Language-Image Pre-training v2),其中与基础的ALBEF模型相比,准确率分别提升了3.22%和2.54%.在6 M预训练数据的条件下,MulGA在视觉问答下游任务中的平均准确率达到了77.75%.

3.4.2　视觉推理实验

视觉推理(Natural Language for Visual Reasoning,NLVR)任务侧重于预测文本标题是否与一对图像的内容相符.本文在NLVR2[42]数据集上进行微调实验,并使用准确率进行评估.

如表5所示,MulGA在视觉推理任务上的表现超越了所有对比算法.使用4 M数据预训练的MulGA,在NLVR2的test测试集上准确率达到了最佳水平.与大模型相比,使用6 M数据预训练的MulGA表现出了与ALBEF和BLIP相当的性能,在dev和test测试集上分别比SimVLM(Simple Visual Language Model)提升了1.22%和1.36%.

3.4.3　视觉蕴含实验

视觉蕴含(Visual Entailment,VE)任务是一个判断一幅图像和一段文本之间的关系是蕴含的、中性的还是矛盾的三分类问题.本文在SNLI-VE[43]数据集上微调模型,并评估分类的准确率,结果如表6所示.

表6 视觉蕴含实验结果 单位:%

模型	数据集	VE	
	图像	Test	Val
UNITER[6]	4 M	78.3	78.6
UniVL[44]	2.84 M	79.7	80.0
ALBEF[14]	4 M	80.3	80.1
MaskVLM[34]	4 M	80.4	80.7
ALBEF[14]	14 M	80.9	80.8
METER[31]	4 M	81.2	80.9
VL-Match[35]	4 M	81.3	80.4
ViLTA[33]	4 M	81.7	81.5
本文方法	4 M	81.2	80.8
本文方法	6 M	81.5	81.6

在4 M数据的条件下,MulGA在test和val上的平均准确率达到了81.0%,相较于具有相似网络架构和训练范式的ALBEF模型,实现了0.99%的性能提升. 使用6 M数据训练的MulGA在val上取得81.6%的最佳表现.

3.4.4 分类任务显著性分析

为了分析MulGA在视觉语言分类任务上是否具有显著性提升,本文将视觉问答、视觉推理和视觉蕴含三个下游任务的测试结果整合作为样本数据,对视觉语言分类任务的数据结果进行了统计显著性验证测试,检验结果如图8所示.

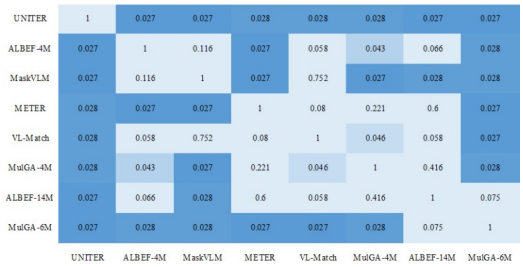


图8 视觉语言分类统计显著性验证

MulGA在4 M训练数据条件下,与其他同体量模型都满足显著性差异条件,当数据扩充到6 M大小时,MulGA相较于4 M的基础模型P值为0.028. 证明适当扩充预训练数据集规模,能够显著提升模型在下游任务上的性能表现.

3.5 细粒度匹配分析

本文通过对融合编码器中交叉注意力层的注意力进行可视化,检验MulGA模型的细粒度匹配能力. 图9展示了MulGA与ALBEF交叉注意力可视化的对比结果,其中左侧为ALBEF,右侧为MulGA.

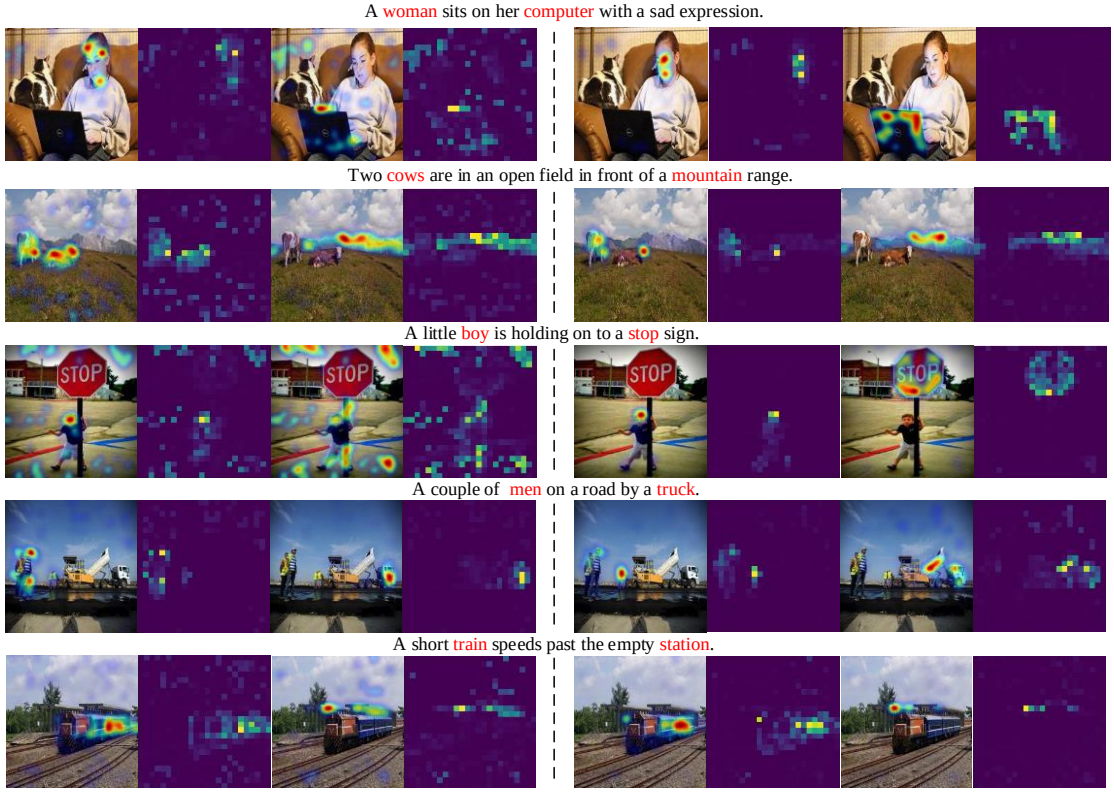


图9 细粒度匹配对比结果

结果表明,ALBEF模型在进行微调后虽然能够正确匹配图像对象和对应的文本字词,但会受到图像背景信息的干扰,注意力权重较为分散. 而MulGA经过视觉序列压缩和细粒度关系推理后,更好地学习了图

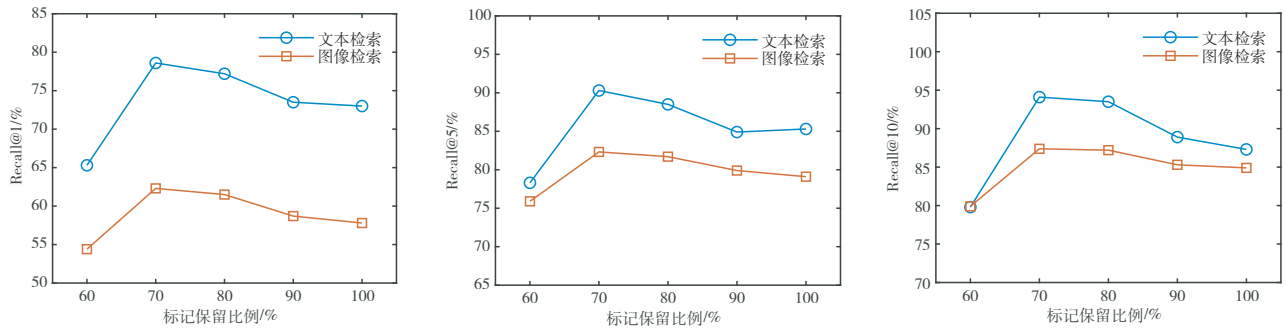
像和文本细粒度对齐关系,对实义单词所指代的具体图像实例聚焦更为准确,在准确识别图像对象、属性、动作和数量等细粒度特征方面具有优越性.

3.6　消融实验

3.6.1　标记保留比例分析

在视觉序列压缩过程中,每层的视觉标记保留比例λ是在训练前预设的超参数. 为了研究不同标记保留比例对于视觉序列压缩的影响,本文使用不同的视觉标记保留λ比例预训练5个周期,并在Flicker30 k上比较零样本图文检索的性能变化,如图10所示.

随着视觉标记保留比例λ的增加,模型在图文检索任务上的总体效果表现为先上升后下降的趋势,当λ减小到60%时,召回率大幅下降,推测是由于视觉标记保留比例过低,图像的大部分信息被压缩,破坏了图像主体信息的完整性,从而影响检索性能. 将λ增加到90%时,视觉序列压缩对模型性能提升不大. 而当λ为70%时,图文检索各指标均优于其他测试值,因此,本文设定每层视觉标记保留比例为70%.



(a) 图文检索R1召回率　(b) 图文检索R5召回率　(c) 图文检索R10召回率

图10　不同保留比例的图文检索性能

3.6.2　图像文本联合引导分析

为了分析视觉序列压缩方法中,图像文本联合引导对模型有效性的影响,本文在视觉问答、视觉推理、视觉蕴含三个下游任务上,分别测试了仅使用视觉信息引导和仅文本信息引导的视觉序列压缩效果. 实验结果在表7中展示.

表7　不同模态信息指导下的实验结果　　单位:%

模态信息	VQAv2	NLVR2	VE
文本信息	74.2	80.4	80.6
视觉信息	73.3	78.1	80.0
视觉信息+文本信息	75.9	81.7	80.4

与使用融合了两种模态的多模态信息引导视觉序列压缩相比,无论是视觉信息还是文本信息,仅依赖单一模态信息会产生较差的结果. 因此,使用图像信息和文本信息来联合指导视觉编码器压缩视觉序列中的冗余信息,在视觉语言多模态学习中更具有效性.

3.6.3　损失函数权重分析

为了研究ITC、ITM和MLM三个损失对模型的贡献程度,本文以1:1:0.8、1:0.8:1、0.8:1:1和1:1:1四种不同的损失函数权重,分别进行5个周期的预训练,并在Flicker30k数据集上做零样本图文检索实验,分析三种损失对模型的影响,实验结果如表8所示.

对比第一行、第三行和第四行数据,当降低ITC损失和MLM损失比重时,对模型零样本检索的性能的影响较大,图像检索和文本检索的召回率都有明显下降;对比第二行与第四行数据,当降低ITM损失比重时,TR@1和TR@5分别降低了1.78%和2.1%,IR@5降低了0.7%. 分析结果证明,三个损失都会对模型作出积极贡献.

3.6.4　视觉序列压缩与细粒度关系推理消融

本文比较了不添加任何组件、单FRIM组件、单VSC组件和VSC+FRIM组合4种不同情况下模型的微调检索性能. 同时,在单张GPU上计算一个训练批次数据前向传播的浮点运算次数(Floating Point Operations,FLOPs),实验结果在表9中展示. 需要注意的是,为了保证先对齐再融合的统一训练范式,单VSC组件测试并没有直接删除FRIM,而是使用原始的ITC进行替换.

表8　损失函数不同权重占比下的实验结果　　单位:%

$\alpha:\beta:\gamma$	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10
1 : 1 : 0.8	73.0	84.7	90.4	60.5	80.2	85.3
1 : 0.8 : 1	77.2	88.4	94.3	62.3	81.7	87.7
0.8 : 1 : 1	73.3	84.9	88.9	58.7	75.9	80.9
1 : 1 : 1	78.6	90.3	94.1	62.3	82.3	87.4

表9　在Flickr30 k上使用不同模块组合进行预训练的图文检索评估

ITC	VSC	FRIM	FLOPs/G	TR@1/%	TR@5/%	IR@1/%	IR@5/%
√			988.6	88.1	95.5	74.8	92.9
		√	1 031.3	91.2	98.5	76.9	94.9
√	√		917.7	89.2	96.4	76.3	94.1
	√	√	931.9	91.6	99.1	79.1	95.7

第一行与第二行相比,将ITC更换为FRIM,虽然增加了少量计算量,但图文检索的召回率都获得了提升;第三行在第一行的基础上添加了VSC,显著降低了模型的计算复杂度,在提升模型效率的同时改善了检索性能;第四行组合了VSC与FRIM,与第三行相比,计算成本略微增加,在文本检索和图像检索的性能分别提升了2.81%和2.6%,实验结果表明,本文提出的VSC和FRIM两个关键组件能够有效提升多模态模型的性能和效率.

3.7 可视化

为了验证本文提出的图像文本联合引导的视觉序列压缩方法的有效性,从MSCOCO数据集中随机采样了部分数据作为样本输入,可视化视觉编码器中3个视觉序列压缩层所选择的相关视觉标记,如图11所示.

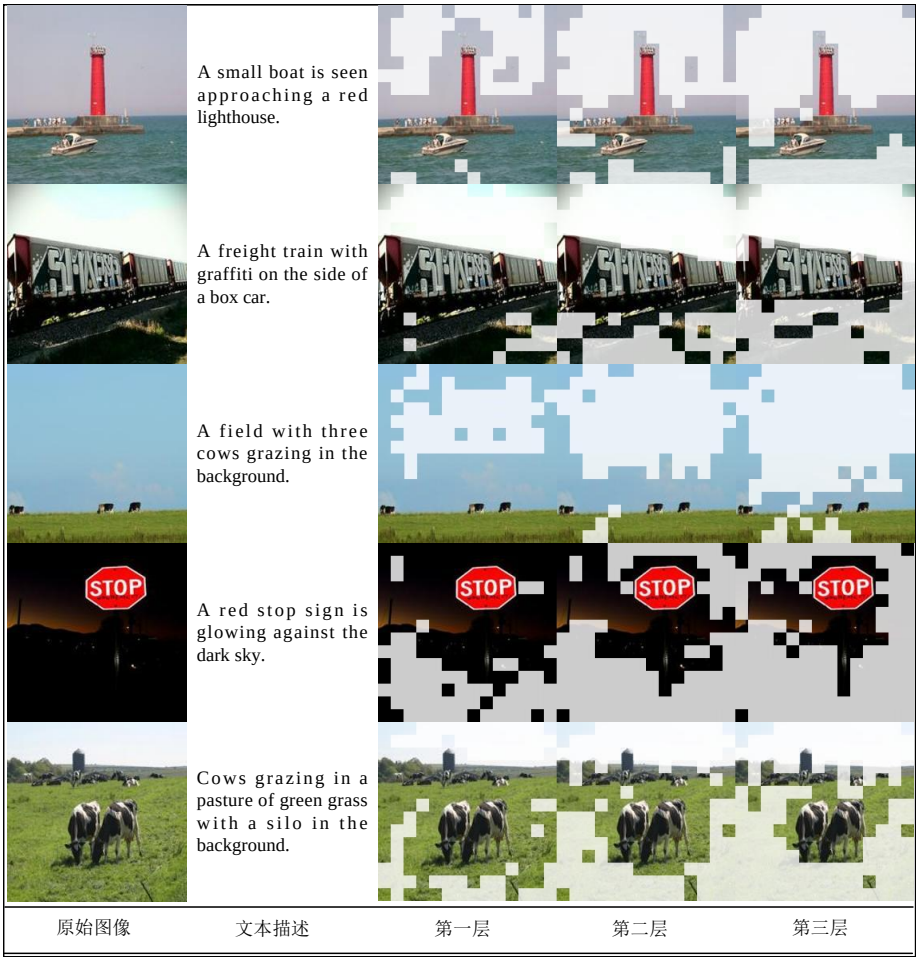


图11 相关视觉标记选择可视化

随着层数增加,视觉相关标记逐渐减少,图像中与文本描述不相关的背景信息被逐渐过滤,模型更加关注文本所描述的主题内容.证明本文的视觉序列压缩方法能够在图像信息和文本信息的联合指导下,有效地减少视觉序列中的冗余信息,保留与文本描述相关的主题内容.

4 结论

本文提出了一种基于跨模态引导和对齐的视觉语言多模态预训练方法,成功解决了现有模型在模态间缺乏显式细粒度对齐机制的问题.通过引入视觉序列压缩模块和细粒度关系推理模块,显著提升了模型在多个视觉语言任务中的性能,尤其在图文检索和视觉问答任务中表现出色.该方法在增强模态间细粒度特征对齐方面展现出强大能力,为多模态预训练模型的发展提供了新的思路.

参考文献

[1] ABDU S A, YOUSEF A H, SALEM A. Multimodal video sentiment analysis using deep learning approaches, a survey[J]. Information Fusion, 2021, 76: 204-226.

[2] ACOSTA J N, FALCONE G J, RAJPURKAR P, et al. Multimodal biomedical AI[J]. Nature Medicine, 2022, 28(9): 1773-1784.

[3] 樊琳, 龚勋, 郑岑洋. 基于文本引导下的多模态医学图像分析算法[J]. 电子学报, 2024, 52(7): 2498-2512.
FAN L, GONG X, ZHENG C Y. A Multi-modal medical image analysis algorithm based on text guidance[J]. Acta Electronica Sinica, 2024, 52(7): 2498-2512. (in Chinese)

[4] HUANG K L, SHI B T, LI X, et al. Multi-modal sensor fusion for auto driving perception: a survey[EB/OL]. (2022-02-06) [2024-03-25]. https://doi. org/10.48550/arXiv. 2202. 02703.

[5] 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). Stroudsburg: Association for Computational Linguistics, 2019: 5100-5111.

[6] CHEN Y C, LI L J, YU L C, et al. UNITER: Universal image-text representation learning[M]//Computer Vision — ECCV 2020. Cham: Springer International Publishing, 2020: 104-120.

[7] LI X J, YIN X, LI C Y, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks[M]//Computer Vision — ECCV 2020. Cham: Springer International Publishing, 2020: 121-137.

[8] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[EB/OL]. (2021-02-26) [2024-03-25]. http://arxiv. org/abs/2103.00020.

[9] JIA C, YANG Y F, XIA Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision[C]//Proceedings of the 38th International Conference on Machine Learning. New York: PMLR, 2021, 139: 4904-4916.

[10] YU J H, WANG Z R, VASUDEVAN V, et al. Coca: Contrastive captioners are image-text foundation models[EB/OL]. (2022-06-14)[2024-03-25]. https://doi.org/10.48550/arXiv.2205.01917.

[11] BAO H B, WANG W H, DONG L, et al. VLMo: Unified vision-language pre-training with mixture-of-modality-experts[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS). New York: Curran Associates, 2022, 35: 32897-32912.

[12] 李志欣, 凌锋, 张灿龙, 等. 融合两级相似度的跨媒体图像文本检索[J]. 电子学报, 2021, 49(2): 268-274.
LI Z X, LING F, ZHANG C L, et al. Cross-media image-text retrieval with two level similarity[J]. Acta Electronica Sinica, 2021, 49(2): 268-274. (in Chinese)

[13] YAO L W, HUANG R H, HOU L, et al. FILIP: fine-grained interactive language-image pre-training[C/OL]// The Tenth International Conference on Learning Representations. (2022-01-29)[2024-03-25]. https://openreview.net/forum?id=cpDhcsEDC2.

[14] LI J N, SELVARAJU R, GOTMARE A, et al. Align before fuse: Vision and language representation learning with momentum distillation[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS). New York: Curran Associates, 2021, 34: 9694-9705.

[15] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. (2020-10-22) [2024-03-25]. https://doi.org/10.48550/arXiv.2010.11929.

[16] LI C L, XU H Y, TIAN J F, et al. mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2022: 7241-7259.

[17] TANG Y H, HAN K, WANG Y H, et al. Patch slimming for efficient vision Transformers[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 12165-12174.

[18] RAO Y M, ZHAO W L, LIU B L, et al. DynamicViT: Efficient vision Transformers with dynamic token sparsification[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS). New York: Curran Associates, 2021, 34: 13937-13949.

[19] LIANG Y W, GE C J, TONG Z, et al. EViT: Expediting vision Transformers via token reorganizations[C/OL]// The Tenth International Conference on Learning Representations. (2022-01-29)[2024-03-25]. https://openreview.net/forum?id=BjyvwnXXVn_.

[20] WEI S Y, YE T Z, ZHANG S, et al. Joint token pruning and squeezing towards more aggressive compression of vision Transformers[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 2092-2101.

[21] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31th International Conference on Neural Information Processing Sys-

tems (NeurIPS). New York: Curran Associates, 2017: 6000-6010.

[22] BAO H B, DONG L, PIAO S H, et al. BEIT: BERT pre-training of image Transformers[C/OL]//The Tenth International Conference on Learning Representations. (2022-01-29) [2024-03-25]. https://openreview. net/forum? id=p-BhZSz59o4.

[23] 汤嘉, 郭燕, 叶名玮, 等. 面向多视角对比学习和语义增强的多模态预训练方法[J]. 计算机科学, 2024, 51(1): 168-174.
TANG J, GUO Y, YE M W, et al. Multimodal pre-training method for multi-view contrastive learning and semantic enhancement[J]. Computer Science, 2024, 51(1): 168-174. (in Chinese)

[24] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional Transformers for language understanding[C]//Proceedings of NAACL-HLT. Stroudsburg: ACL, 2019: 4171-4186.

[25] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[C]//Computer Vision — ECCV 2014. Cham: Springer International Publishing, 2014: 740-755.

[26] KRISHNA R, ZHU Y K, 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.

[27] ORDONEZ V, KULKARNI G, BERG T L. Im2Text: Describing images using 1 million captioned photographs [C]//Proceedings of the 25th International Conference on Neural Information Processing Systems. New York: Curran Associates, 2011, 24: 1143-1151.

[28] 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). Stroudsburg: Association for Computational Linguistics, 2018: 2556-2565.

[29] 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.

[30] 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) [2024-03-25]. https://doi. org/10.48550/arXiv.2001.07966.

[31] DOU Z Y, XU Y C, GAN Z, et al. An empirical study of training end-to-end vision-and-language Transformers [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 18166-18176.

[32] YANG J Y, DUAN J L, TRAN S, et al. Vision-language pre-training with triple contrastive learning[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 15671-15680.

[33] WANG W H, YANG Z, XU B, et al. ViLTA: Enhancing vision-language pre-training through textual augmentation [C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2023, 3158-3169.

[34] KWON G, CAI Z W, RAVICHANDRAN A, et al. Masked vision and language modeling for multi-modal representation learning[EB/OL]. (2022-08-03) [2024-3-25]. https://doi.org/10.48550/arXiv.2208.02131.

[35] BI J Y, CHENG D X, YAO P, et al. VL-match: Enhancing vision-language pretraining with token-level and instance-level matching[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2023, 2584-2593.

[36] LI J N, LI D X, XIONG C M, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//Proceedings of the 39th International Conference on Machine Learning. New York: PMLR, 2022, 162: 12888-12900.

[37] AGRAWAL A, LU J S, ANTOL S, et al. VQA: Visual question answering[J]. International Journal of Computer Vision, 2017, 123(1): 4-31.

[38] BYUN J, HWANG T, FU J L, et al. GRIT-VLP: Grouped mini-batch sampling for efficient vision and language pre-training[C]//Computer Vision — ECCV 2022. Cham: Springer Nature Switzerland, 2022: 395-412.

[39] ZHANG H T, ZHANG P C, HU X W, et al. GLIPv-2: Unifying localization and VL understanding[EB/OL]. (2022-06-12) [2024-3-25]. https://doi. org/10.48550/arXiv.2206.05836.

[40] WANG Z R, YU J H, YU A W, et al. SimVLM: Simple visual language model pretraining with weak supervision [EB/OL]. (2021-08-24) [2024-3-25]. https://doi. org/10.48550/arXiv.2108.10904.

[41] WANG P, YANG A, MEN R, et al. OFA: Unifying archi-

tectures, tasks, and modalities through a simple sequence-to-sequence learning framework[C]//Proceedings of the 39th International Conference on Machine Learning. New York: PMLR, 2022, 162: 23318-23340.

[42] SUHR A, ZHOU S, ZHANG A, et al. A corpus for reasoning about natural language grounded in photographs [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 6418-6428.

[43] XIE N, LAI F, DORAN D, et al. Visual entailment: A novel task for fine-grained image understanding[EB/OL]. (2019-01-20) [2024-03-25]. https://doi. org/10.48550/arXiv.1901.06706.

[44] 刘天义, 吴祖煊, 陈静静, 等. 面向视觉语言理解与生成的多模态预训练方法[J]. 软件学报, 2023, 34(5): 2024-2034.
LIU T Y, WU Z X, CHEN J J, et al. Multimodal pre-training method for vision-language understanding and generation[J]. Journal of Software, 2023, 34(5): 2024-2034. (in Chinese)

作者简介

才　华　男, 1977 年 2 月出生于吉林省辉南县. 现为长春理工大学副教授、博士生导师. 获吉林省科技进步奖 1 项. 在国内外发表学术论文 100 余篇. 研究方向为计算机视觉与自然语言处理, 目前主持国家级省部级项目多项.
E-mail: caihua@cust.edu.cn

易亚希　男, 2001 年 2 月出生于湖南省湘潭市. 现为长春理工大学电子信息工程学院硕士研究生. 主要研究方向为计算机视觉和视觉语言多模态.
E-mail: 2022100885@mails.cust.edu.cn

付　强　男, 1984 年 8 月出生于吉林省长春市. 现为长春理工大学空间光电技术研究所副所长. 主要研究方向为光学传输特性测试与多维度成像探测.
E-mail: fuqiang@cust.edu.cn

冉　越　男, 2000 年 4 月出生于河南省南阳市. 现为长春理工大学电子信息工程学院硕士研究生. 主要研究方向为计算机视觉和视觉语言多模态.
E-mail: ry13523068581@163.com

孙俊喜　男, 1971 年 6 月出生于河北省唐山市. 现为东北师范大学信息科学与技术学院教授、博士生导师. 主要研究方向为 AI 视觉与智能感知技术.
E-mail: sunjx100@nenu.edu.cn