计算机科学
COMPUTER SCIENCE

基于双编码器的多模态融合方法

黄晓飞, 郭卫斌

引用本文

黄晓飞, 郭卫斌. 基于双编码器的多模态融合方法[J]. 计算机科学, 2024, 51(9): 207-213.

HUANG Xiaofei, GUO Weibin. Multi-modal Fusion Method Based on Dual Encoders[J]. Computer Science, 2024, 51(9): 207-213.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230700212

基于双编码器的多模态融合方法

黄晓飞　郭卫斌
华东理工大学信息科学与工程学院　上海 200237
(y30211028@mail.ecust.edu.cn)

摘　要　双编码器模型比融合编码器模型具有更快的推理速度,且能在推理过程中对图像和文本进行预计算。然而,双编码器模型中使用的浅交互模块不足以处理复杂的视觉语言理解任务。针对上述问题,提出了一种新的多模态融合方法。首先,提出一种前交互式桥塔结构(PBTS),在单模态编码器的顶层和跨模态编码器的每层之间建立连接,使得不同语义层次的视觉和文本表示之间能够进行全面、自下而上的交互,从而实现更有效的跨模态对齐和融合。同时,为了更好地学习图像和文本的深度交互,提出了一种两阶段跨模态注意力双蒸馏方法(TCMDD),使用融合编码器模型作为教师模型,在预训练阶段和调优阶段同时对单模态编码器及融合模块的跨模态注意力矩阵进行知识蒸馏。使用400万张图片进行预训练并在3个公开数据集上进行调优来验证该方法的有效性。实验结果表明,所提多模态融合方法在多个视觉语言理解任务中获得了更优的性能。

关键词:多模态融合;双编码器;跨模态注意力蒸馏;桥塔结构
中图分类号　TP391.4

Multi-modal Fusion Method Based on Dual Encoders

HUANG Xiaofei and GUO Weibin
School of Information science and Engineering,East China University of Science and Technology,Shanghai 200237,China

Abstract　The dual encoder model has faster inference speed than the fusion encoder model,and can pre-calculate images and text during the inference process. However,the shallow interaction module used in the dual encoder model is not sufficient to handle complex visual language comprehension tasks. In response to the above issues,this paper proposes a new multi-modal fusion method. Firstly,a pre-interactive bridge tower structure(PBTS) is proposed to establish connections between the top layer of a single mode encoder and each layer of a cross-mode encoder. This enables comprehensive bottom-up interaction between visual and textual representations at different semantic levels,enabling more effective cross-modal alignment and fusion. At the same time,in order to better learn the deep interaction between images and text,a two-stage cross-modal attention double distillation method(TCMDD) is proposed,which uses the fusion encoder model as the teacher model and distills knowledge of the cross-modal attention matrix of the single modal encoder and fusion module simultaneously in the pre-training and tuning stages. Using 4 million images for pre-training and tuning on three public datasets to validate the effectiveness of this method. Experimental results show that the proposed multi-modal fusion method achieves better performance in multiple visual language comprehension tasks.

Keywords　Multi-modal fusion,Dual encoder,Cross-modal attention distillation,Bridge tower structure

1　引言

多模态融合将多个单模态表征整合为一个多模态信息表征,它是多模态信息处理的核心问题。根据多模态融合的阶段,多模态融合方法可简单分为早期融合和晚期融合。

早期融合一般采用融合编码器架构,使用有效但效率较低的Transformer编码器来捕获具有跨模态注意力的图像和文本交互。早期的融合编码器模型依赖于现成的目标检测器来提取图像区域特征,这进一步降低了融合编码器的效率。后来的ViLT[1]模型丢弃检测器,并使用ViT(Vision Transformer)[2]直接对图像补丁进行编码。ViLT在视觉语言理解和检索任务上实现了具有竞争力的性能,同时提高了效率。然而,由于仍然需要同时对图像和文本进行编码,而这需要大量的计算资源,因此推理速度降低,限制了其在大量图像或文本候选任务中的应用。

晚期融合一般采用双编码器架构,包括CLIP[3]和ALIGN[4]等,其分别对图像和文本进行编码。跨模态交互通过浅融合模块建模,通常是多层感知器网络或点积,这与融

到稿日期:2023-07-28　返修日期:2023-11-27
基金项目:几何信息融合的分类学习研究(62076094)
This work was supported by the Research on Classification Learning of Geometric Information Fusion(62076094).
通信作者:郭卫斌(gweibin@ecust.edu.cn)

合编码器模型中的 Transformer 编码器相比非常轻量。此外，分开编码实现了图像和文本的离线计算和缓存，可以很好地扩展到大规模数据。这些变化使其在理解和检索任务中的推理速度更快，使模型在现实场景中变得实用。双编码器模型在图像文本检索任务中取得了良好的性能。然而，其在视觉语言理解任务上的表现远远落后于融合编码器模型，如视觉问答(VQA)[5]，视觉蕴含(SNLI-VE)[6]等。这些任务需要复杂的跨模态推理。

基于此，本文提出了一种新的多模态融合方法来解决双塔模型中由缺乏跨模态交互导致的视觉语言理解任务性能不足的问题，使双塔模型能在拥有更快推理速度的同时，学习到更丰富及更深层次的模态间交互知识。

本文的主要贡献如下：

1)使用前交互式桥塔结构 PBTS，在单模态编码器的顶层和跨模态编码器的每层之间建立连接，使不同语义层次的视觉和文本表示之间能够进行全面、自下而上的交互，从而实现更有效的跨模态对齐和融合。

2)提出两阶段跨模态注意力双蒸馏方法 TCMDD，使用融合编码器模型 ViLT 作为教师模型，在预训练阶段和调优阶段同时对单模态编码器及融合模块(跨模态编码器)的跨模态注意力矩阵进行知识蒸馏。通过这种多模态融合方法得到的预训练模型在多个下游任务上比之前的方法取得了更优的性能。

2 相关工作

2.1 跨模态注意力蒸馏

知识蒸馏(KD)是将在强教师模型中学习到的知识迁移到学生模型中，使学生模型具有竞争力。Hinton 等[7]首次利用教师模型中的软标签分布来训练学生模型。之后，通过模仿教师的中间表征，如隐藏状态[8]和注意力分布[9]，学生模型可以进一步改进。最近，Li 等的 VIRT[10]进行了纯文本的句子与句子间交互的注意力蒸馏，首次提出了将融合模型中的交互信息通过知识蒸馏迁移给双塔模型的方法；随后 Wang 等的 Distilled Dual-Encoder 模型[11]提出了迁移多模态融合模型中的文本与图像跨模态交互信息给学生模型(双塔)的跨模态注意力蒸馏方法(CMAD)，使得双塔模型在视觉理解领域的性能得到极大提升。Lu 等提出的 ERNIE-Search 模型[12]使用自我动态蒸馏和级联蒸馏，有效提升了跨架构蒸馏效果。

2.2 桥接层

按照 ViLT 提出的分类方法，最近的视觉语言模型可以统一为 TWO-TOWER(双塔)架构，如图 1(a)－1(d)所示。它们将视觉和文本编码器的最后一层表示提供给顶部的跨模态编码器，并且可以通过文本、视觉和跨模态编码器的深度来区分。

CLIP 和 ALIGN 是代表性模型，他们直接在跨模态编码器中对具有同等表现力的视觉和文本编码器的最后一层表示进行浅融合(如点积)，如图 1(a)所示。其余模型在基于多层 transformer 的跨模态编码器中进行深度融合，但选择具有不同表达水平的单模态编码器。

Chen 等[13]和 Cho 等[14]的工作属于图 1(b)的范畴，因为他们采用不同类型的深度视觉模型(如 Faster R-CNN[15]，ResNet[16]和 ViT 等)作为视觉编码器，获取区域、网格或补丁特征，并将其与词嵌入连接起来，馈送到顶部跨模态编码器中。

还有一类模型(如 ViLT，OFA[17]和 SimVLM[18]等)如图 1(c)所示。这类模型利用轻量级视觉和轻量级文本编码器，并在一个基于 transformer 的跨模态编码器中处理这两种模态。相比之下，图 1(d)类别的模型使用富有表现力的深度视觉和深度文本编码器，并将最后一层表示馈送到顶层的多层跨模态编码器中。

无论使用的是视觉、文本还是跨模态编码器，大多数当前模型都忽略了单模态编码器不同层中的各种级别的语义信息，仅利用最后一层的单模态表示进行跨模态对齐和融合。因此，如图 1(e)所示，BRIDGE-TOWER[19]提出使用多个桥接层将单模态编码器的顶层与跨模态编码器的每一层连接起来，这样不但不会影响模态内的交互，而且可以使不同语义层次的视觉和文本表示在自下而上的方向上进行彻底、温和的交互，同时还可以在跨模态编码器的每一层实现有效的跨模态对齐和融合。

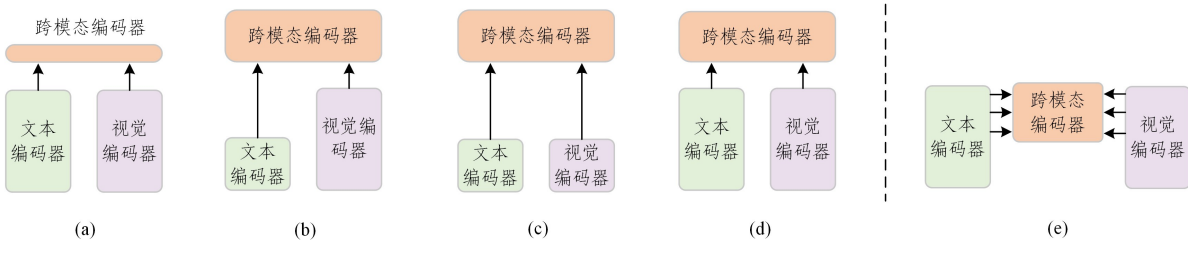

图 1 视觉语言模型分类图

Fig. 1 Classification diagram of visual language model

3 基于双编码器的多模态融合模型

将提出的多模态融合方法应用于双塔架构，可以在学习教师模型中跨模态交互知识的同时，获取单模态编码器内部丰富的不同级别语义信息。本章主要介绍模型结构以及对单模态编码器和桥塔结构的两阶段双蒸馏方法。

3.1 模型结构

本文模型结构如图 2 所示。该多模态融合模型由一个

视觉编码器、一个文本编码器以及前交互式桥塔结构组成。最终目标是在单模态编码器的顶部和跨模态编码器之间建立一座桥梁，以便其在跨模态编码器的每一层进行全面、详细的交互。

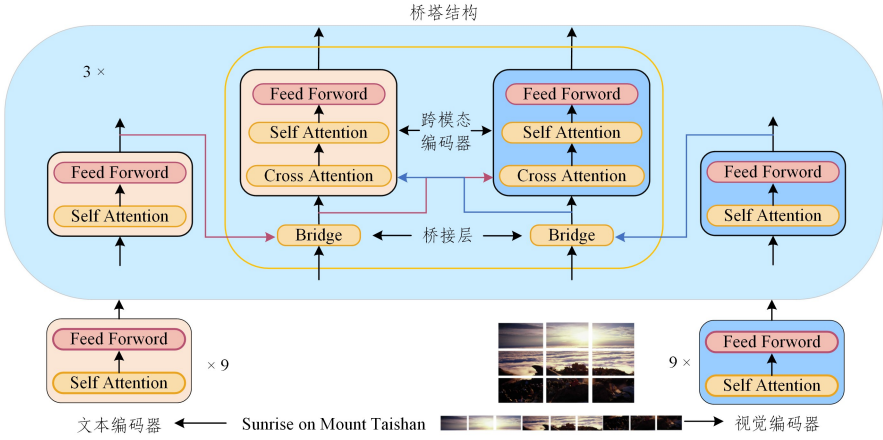

图 2 整体模型结构

Fig. 2 Overall model structure

3.1.1 文本编码器

本文使用 $\text{BERT}_{\text{BASE}}$ 作为文本编码器。每个输入序列 $\boldsymbol{\omega}$ 由字节级字节对编码(BPE)进行标记。[〈s〉]和[〈/s〉]分别作为开始和结束标记添加到序列中。输入文本表示如式(1)所示：

$$T_0=[E_{[\langle S\rangle]};E_{\omega 1};\cdots;E_{\omega M};E_{[\langle /S\rangle]}]+T^{\text{pos}},T_0\in\mathbb{R}^{(M+2)\times D_t} \tag{1}$$

其中，$\boldsymbol{E}$ 为词嵌入矩阵，M 为词元个数，D_t 为文本编码器的维数，$\boldsymbol{T}^{\text{pos}}$ 为位置嵌入矩阵。类似地，将文本编码器的每一层表示为 $Encoder_l^T$，每一层的输出表示为 $T_l=Encoder_l^T(T_{l-1})$，$l=1,\cdots,L_T$，其中 L_T 是文本编码器的层数。

3.1.2 视觉编码器

本文使用 ViT_{BASE} 作为视觉编码器。对于每个输入的 2D 图像 $I\in\mathbb{R}^{H\times W\times C}$，其中(H,W)为输入图像的分辨率，C 为通道数。ViT 将其重塑为一系列平坦的 2D 补丁 $P\in\mathbb{R}^{N\times(P^2C)}$，其中(P,P)为图像补丁分辨率，$N=\dfrac{HW}{P^2}$ 为补丁数。与 BERT 类似，ViT 也将[class]标记前置到补丁序列中，并使用可学习的 1D 位置嵌入 $V^{\text{pos}}\in\mathbb{R}^{(N+1)\times D_v}$，其中 D_v 是视觉编码器的维数。最终可以得到输入的视觉表示，如式(2)所示：

$$V_0=[E_{[\text{class}]};p_1W_p;\cdots;p_NW_p]+V^{\text{pos}},V_0\in\mathbb{R}^{(N+1)\times D_v} \tag{2}$$

其中，$W_p\in\mathbb{R}^{(P^2C)\times D_v}$ 为可训练的线性投影层。ViT 的每一层由多头自注意(MSA)块和前馈网络(FFN)块组成。此处省略了计算细节，并将其简化为 $Encoder^V$。然后，每层的输出表示为 $V_l=Encoder_l^V(V_{l-1})$，$l=1,\cdots,L_V$，其中 L_V 是视觉编码器的层数。

3.1.3 前交互式桥塔结构

为了使桥塔学习到更多更充分的跨模态知识，本文未使用之前的后交互策略(在桥接层融合单模态与多模态信息后，经过多头自注意力层自我学习之后再进行跨模态交互)，而是采用了一种新的前交互策略，在单模态与多模态信息通过桥接层进行交互之后，将学习了文本图像交互信息的单模态信息在自我学习之前直接进行跨模态交互。

本文提出的前交互式桥塔结构由桥接层与跨模态编码器共同组成，通过多个桥接层将单模态编码器的顶层与跨模态编码器的每一层连接起来，使得不同语义层次的视觉和文本表示之间能够自下而上地详细交互。桥塔结构可以通过以下方式实现：

$$\widetilde{Z}_l^V=BridgeLayer_l^V(Z_{l-1}^V,V_jW_V+V_{\text{type}}) \tag{3}$$

$$\widetilde{Z}_l^T=BridgeLayer_l^T(Z_{l-1}^T,T_kW_T+T_{\text{type}}) \tag{4}$$

$$Z_l^V,Z_l^T=Encoder_l^Z(\widetilde{Z}_l^V,\widetilde{Z}_l^T) \tag{5}$$

其中，$Encoder_l^Z$ 是第 l 层跨模态编码器，它由视觉和文本两部分组成。每个部分由多头自注意(MSA)块、多头交叉注意(MCA)块和前馈网络(FFN)块组成。$Z_l^{\{V,T\}}$ 是第 l 层跨模态表示的视觉或文本部分，$\widetilde{Z}_l^{\{V,T\}}$ 是每层跨模态编码器的视觉或文本输入。特别指出，由于第 1 层的跨模态编码器缺少前一层的跨模态输入，因此我们直接使用对应层数的单模态表示作为初始输入。$W_V\in\mathbb{R}^{D_V\times D_Z}$ 和 $W_T\in\mathbb{R}^{D_T\times D_Z}$ 为线性投影，V_{type} 和 T_{type} 为视觉和文本类型嵌入。$j\in[1,L_V]$，$k\in[1,L_T]$ 是视觉编码器和文本编码器的单模态表示的索引，L_Z 是跨模态编码器的层数。设置 $L_V=L_T=12$，$L_Z=3$。为了使用顶部 3 层的单模态表示，设置 $j,k=10,11,12$。利用桥接层，顶层单模态表示可以与跨模态编码器的每一层桥接，从而将不同语义级别的单模态表示合并到跨模态交互中。简单地说，桥接层就是加和与层归一化(Add&Norm)，如式(6)所示：

$$BridgeLayer(x,y)=LayerNorm(x+y) \tag{6}$$

3.2 两阶段双蒸馏方法

本节将介绍在预训练阶段和调优阶段对模型的单模态编码器与跨模态编码器同时进行蒸馏的两阶段双蒸馏方法。

3.2.1 蒸馏方法

本文的蒸馏方法分为两个部分：跨模态注意力蒸馏和软标签蒸馏。为了改进双编码器模型以捕获图像和文本更深层次的交互，跨模态注意力蒸馏利用融合编码器模型的跨模态注意力知识来指导双编码器模型的训练，即使用图像到文本和文本到图像的注意力分布来训练双编码器模型。融合编码

器教师模型通过多头注意力机制捕获跨模态交互。整个注意力分布 $A_T^{vl} \in \mathbb{R}^{(N+M)\times(N+M)}$（$N$ 和 M 分别表示图像和文本输入的长度）可以分为两部分。第一部分是单模态注意力分布（$A_T^{v2v} \in \mathbb{R}^{(N\times N)}$，$A_T^{t2t} \in \mathbb{R}^{(M\times M)}$），它为相同模态的词元内的交互建模；第二部分是跨模态注意力分布，包括图像到文本的注意力分布（$A_T^{v2t} \in \mathbb{R}^{(N\times M)}$）和文本到图像的注意力分布（$A_T^{t2v} \in \mathbb{R}^{(M\times N)}$）。跨模态注意力分布捕获视觉和文本特征向量的相互作用。由于双编码器的单独编码只模拟相同模态的词元的相互作用，因此本文引入跨模态注意力蒸馏来鼓励双编码器模型模拟融合编码器模型的图像和文本对齐。双编码器模型 A_S^{v2t}，A_S^{t2t} 的跨模态（图像到文本和文本到图像）注意力分布计算如式(7)和式(8)所示：

$$A_S^{v2t} = \text{softmax}\left(\frac{Q_S^v K_S^{t\,\mathrm{T}}}{\sqrt{dk}}\right) \tag{7}$$

$$A_S^{t2v} = \text{softmax}\left(\frac{Q_S^t K_S^{v\,\mathrm{T}}}{\sqrt{dk}}\right) \tag{8}$$

其中，Q_S^v 和 K_S^v 为自注意力模块的视觉查询和键。Q_S^t，K_S^t 是文本输入的查询和键。本文以同样的方式重新计算教师 A_T^{v2t}，A_T^{t2t} 的跨模态注意力分布，而不是直接分割原本的注意力分布 A_T^{vl}。跨模态注意力蒸馏的损失计算如式(9)所示：

$$\mathscr{L}_{\mathrm{CA}} = D_{\mathrm{KL}}(A_S^{v2t} \parallel A_T^{v2t}) + D_{\mathrm{KL}}(A_S^{t2v} \parallel A_T^{t2v}) \tag{9}$$

其中，D_{KL} 是 Kullback-Leibler 散度。受 Wang 等[20]的启发，本文只迁移了教师模型最后一层的跨模态注意力知识。

软标签蒸馏使用来自教师模型的预测作为软标签来改进学生。软标签损失计算如式(10)所示：

$$\mathscr{L}_{\mathrm{SL}} = D_{\mathrm{KL}}(z_S \parallel z_T) \tag{10}$$

其中，z_S 和 z_T 分别为学生和教师模型的预测对数。

3.2.2 预训练蒸馏

预训练阶段蒸馏的结构如图 3 所示。可以看到，本文一边从教师模型的最后一层中获取多头注意力分布，利用其中的跨模态注意力分布对学生模型进行蒸馏，一边用教师模型分类器输出的软标签对学生模型进行蒸馏。其中 S_H 表示学生模型的注意力头数。在预训练期间，使用预训练好的融合编码器模型 ViLT 作为教师模型，对双编码器学生模型在大规模图像-文本对（400 万张图片，900 万个图片文本对）上进行预训练。一共使用了 3 个预训练任务。

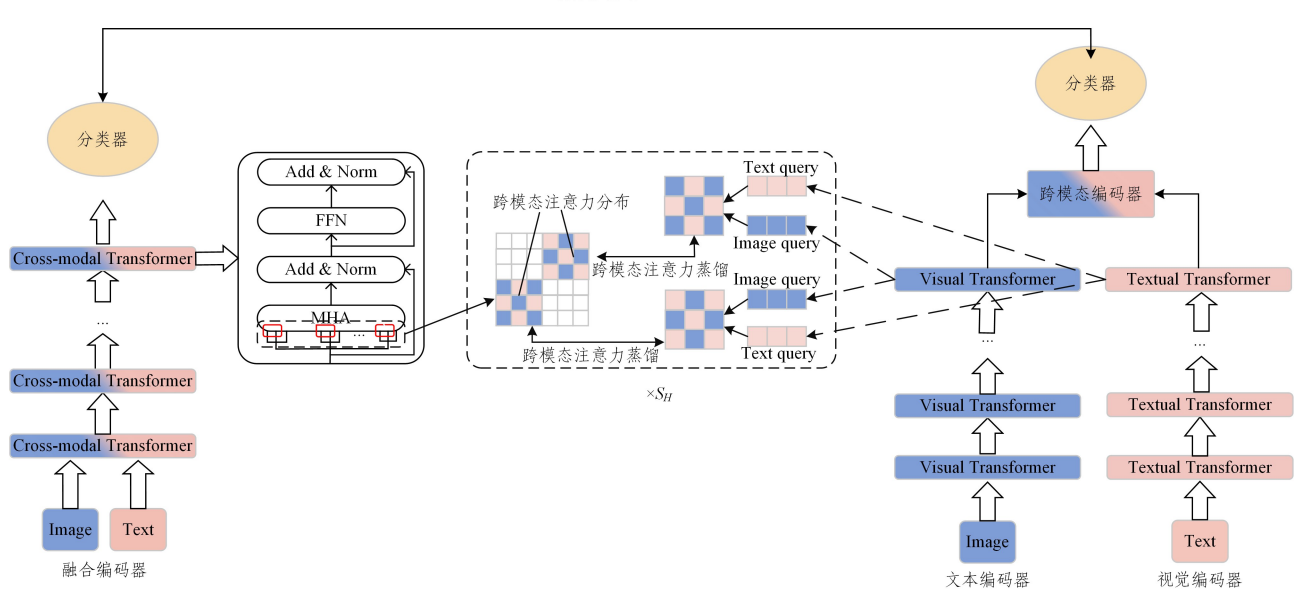

图 3 预训练阶段蒸馏结构图

Fig. 3 Distillation structure diagram during pre-training stage

1）图像文本匹配（Image-Text Matching，ITM）。图像-文本匹配的目标是预测输入的图像和文本是否匹配。和 ViLT 一样，本文以 0.5 的概率替换匹配的图像来构建负样本对，使用 ITM 输入对上的跨模态注意力蒸馏损失和软标签损失来训练双编码器模型。

2）图像-文本对比学习（Image-Text Contrastive Learning，ITC）。本文引入了批量负采样的对比损失来优化视觉和文本表示的共享空间。给定一批给定一批 N 对图像-文本对，可以得到 N 对正样本对和 N^2-N 对负样本对。图像-文本对比学习旨在从所有可能的配对中预测匹配的配对。为了提高训练效率，只考虑在 N 对匹配上计算的跨模态注意力分布。

3）掩码语言建模（Masked Language Modeling，MLM）。掩码语言建模的目标是从所有未被掩码的标记中恢复被掩码标记。本文使用了动态掩码的方式，并且和 BERT 一样，使用了 15%的掩码概率。

以上预训练任务除了蒸馏部分，都使用交叉熵损失函数（Cross Entropy Loss）充当目标函数，如式(11)所示：

$$L_{ce} = -\frac{1}{N}\sum_i \sum_{c=1}^{M} y_{ic} \log(p_{ic}) \tag{11}$$

其中，N 为样本数，M 为类别数，y_{ic} 为样本 i 属于类别 c 的真实概率，p_{ic} 为样本 i 属于类别 c 的预测概率。

3.2.3 调优蒸馏

在调优过程中，本文使用在不同下游任务上调优后的 ViLT 作为教师模型，使用跨模态注意力蒸馏和软标签损失来调优学生模型。

4 数值实验

4.1 预训练和调优数据集

预训练数据集使用在多模态领域常用的 COCO[21]，Conceptual Captions[22]，SBU Captions[23] 和 Visual Genome[24] 4 个公开数据集，共 400 万张图片、900 万个图像-文本对进行模型预训练。数据集组成如表 1 所列。

表 1 预训练数据集统计

Table 1 Pre-training dataset statistics

数据集	图片数	文本数
CC	3.01×10^6	3.01×10^6
VG	108000	5.41×10^6
COCO	113000	567000
SBU	867000	867000

Conceptual Captions 数据集和 SBU Captions 数据集中一些图片链接已经失效，训练时使用其仍可用的部分。在调优阶段，使用 3 个视觉语言理解任务来评估本文方法的有效性。第一个任务是视觉推理，使用 NLVR2 数据集[25]，旨在确定文本语句是否描述了一对图像。构建两个图像文本对作为输入，每个图像文本对由一张图像和一段文本描述组成，将两对的最终表示送入分类器层以获得预测。第二个任务是视觉蕴含，使用 SNLI-VE 数据集[26]，旨在预测图像和文本描述之间的关系。与之前的工作一样，本文将视觉蕴含视为三分类任务。第三个任务是视觉问答，使用 VQAv2 数据集[27]，该任务要求模型根据图像回答问题。本文将问答任务转变为具有 3129 个候选答案的分类任务。3 个数据集的组成如表 2 所列。

表 2 视觉语言理解任务数据集组成

Table 2 Composition ofvisual language understanding task datasets

数据集	NLVR2	SNLI-VE	VQAv2
图片数	119000	31000	204000
文本数	100000	565000	1.1×10^6

4.2 实验环境与实现细节

本文使用的 GPU 型号为 Tesla V100，内存大小为 32GB。实验运行环境为 Ubuntu18.04，使用 torch1.8.0-cuda11.1 作为深度学习模型框架。

模型的双编码器部分的 Transformer 架构与 ViLT 相同。视觉和文本编码器的隐藏层大小设置为 768，前馈网络的中间大小设置为 3 072，头部数量设置为 12，文本序列的最大长度设置为 40。使用 AdamW 优化器，初始学习率为 5×10^{-5}，权重衰减为 0.01。在总训练步骤的前 10%中预热学习率，然后在剩余的训练步骤中将学习率线性衰减到零。图像分辨率裁剪为 224×224，补丁大小为 32×32。对于预训练，使用 900 万个图像-文本对训练 25 轮，批大小为 128。在调优期间，对 VQA 任务训练 10 轮，批大小为 64。对于 NLVR2，训练 20 轮，批大小为 128。对于 SNLI-VE，训练 5 轮，批大小为 64。另外，本文使用 RandAugment 来进行图像增强，去除了其中的颜色反转和裁剪。

4.3 实验结果与分析

本文在 3 个下游任务数据集（NLVR2，SNLI-VE 和 VQAv2）上来评估模型的性能，实验结果如表 3 所列。结果表明，相比融合编码器的教师模型 ViLT，本文的双编码器学生模型在 NLVR2，SNLI-VE 和 VQAv2 任务上分别达到了教师模型 98.5%，99.3%和 99.0%的性能。另外，相比其他双编码器基线模型的最优结果，本文模型在 NLVR2 的验证集上的准确率提高了 1.22%，测试集上的准确率提高了 1.08%；在 SNLI-VE 的验证集上的准确率提高了 0.7%，测试集上的准确率提高了 1.18%。在最能体现模型视觉语言理解能力的 VQA 任务上，本文模型相比最好的基线模型准确率提升了 2.05%。本文模型在 3 个视觉语言理解任务上的性能均超过了之前的方法。实验结果证明，本文提出的多模态融合方法让模型学习到了更丰富、更深层次的跨模态交互知识，使得模型获得了更强的视觉语言理解能力。

表 3 本文模型与其他基线模型在 3 个下游任务数据集上的实验结果

Table 3 Experimental results of the proposed model and other baseline models on three downstream task datasets

(%)

模型	NLVR2		SNLI-VE		VQAv2
	dev	test-P	val	test	test-dev
ViLT	69.95	70.14	70.85	71.06	67.45
CLIP	48.51	48.23	60.22	60.34	48.29
ALBEF	67.26	67.34	68.28	68.47	64.18
BRIDGE-TOWER	67.38	67.46	68.87	68.99	64.25
Distilled Dual-Encoder	67.77	68.04	69.17	69.37	64.74
Ours	**68.99**	**69.12**	**70.37**	**70.55**	**66.79**

特别指出，表 3 的数据为模型预训练 25 轮之后调优的实验结果。接下来的对比和消融实验结果为模型预训练 10 轮之后进行调优的结果。

4.3.1 与其他方法的对比实验

本小节将本文提出的两种改进方法（PTBS 和 TCMDD）与之前的方法进行比较来验证其有效性。首先，将 PTBS 模块与之前的方法在 VQA 任务上进行性能对比。为了与之前的方法进行区分，将本文提出的前交互策略记为 PTBS(pre)，将后交互策略记为 PTBS(post)，实验结果如表 4 所列。

表 4 本文方法与其他多模态融合模块在 VQA 任务上的性能

Table 4 Performance of the proposed method compared with other multi-modal fusion modules in VQA tasks

(%)

方法	VQA
Shallow Interaction	52.15
VIRT-Adapted Interaction	55.49
PBTS(post)	62.35
PBTS(pre)	**63.78**

可以看到，不管是前交互策略还是后交互策略，相比之前简单的浅交互方法（Shallow Interaction）和 VIRT 中的适应性交互方法（VIRT-Adapted Interaction），它们所拥有的桥塔结构都能够有效提升模型的跨模态理解能力以及模型在视觉语言理解任务上的性能。另外，相比使用后交互策略，由于前交互策略在每次自我学习之前获得了更多的跨模态交互信息，从而在自我学习之后获得了更丰富的视觉语言理解能力，使得使用前交互策略后模型在 VQA 任务上的准确率提升了 1.43%。

接下来，将 TCMDD 方法与之前的蒸馏方法进行对比，实验结果如表 5 所列。其中 CMD 为跨模态注意力蒸馏方法，其只在预训练阶段对模型进行蒸馏；TCMD 为两阶段跨模态注意力蒸馏方法，其在预训练和调优阶段对由单模态编码器模拟的跨模态注意力矩阵进行蒸馏。本文提出的 TCMDD 方法为两阶段跨模态注意力双蒸馏方法，其在预训练和调优阶段同时对单模态编码器和跨模态编码器的跨模态注意力矩阵进行蒸馏。可以看到，提出的 TCMDD 方法在 VQA 任务上达到了 63.09%的准确率，比之前提出的 CMD 和 TCMD 方法分别提高了 3.11%和 1.64%。实验结果表明，在预训练和调优阶段将教师模型的模态间交互知识蒸馏给多模态融合模块能够进一步提升模型在视觉语言理解任务上的表现。

表 5 本文方法与其他蒸馏方法在 VQA 任务上的性能

Table 5 Performance of the proposed method and other distillation methods in VQA tasks

(%)

方法	VQA
CMD	59.98
TCMD	61.45
TCMDD	**63.09**

4.3.2 消融实验

为了验证所提出的两个改进对模型最终性能的实际影响，本文做了大量消融实验。首先，对所提出的 TCMDD 方法的蒸馏对象进行了实验来验证同时对桥塔结构和单模态编码器进行跨模态注意力蒸馏的有效性，其中 Baseline 为不对模型进行蒸馏。实验结果如表 6 所列，可以看到，在两阶段蒸馏过程中同时对跨模态编码器和单模态编码器进行蒸馏后模型在 VQA 任务上取得了最高 63.09%的准确率，相比只对跨模态编码器进行蒸馏的结果提高了 0.94%，比只对单模态编码器进行蒸馏的结果提高了 1.64%，证明了本文提出的 TCMDD 方法可以在两阶段的蒸馏过程中学习到更多的模态间交互信息，获得更深层次的跨模态理解能力。另外，只对跨模态编码器进行蒸馏的准确率比只对单模态编码器蒸馏的准确率提高了 0.7%，表明跨模态编码器能更好地学习从教师模型传递来的模态间交互知识。

表 6 TCMD 方法的消融实验结果

Table 6 Ablation experimental results of TCMD method

(%)

方法	VQA
Baseline	57.37
+Single Modal Distillation	61.45
+Bridge Tower Distillation	62.15
+Both	**63.09**

接着对本文提出的桥塔结构与单模态编码器的连接方式进行探索，考虑了跨层映射(Cross-layer mapping)与累加映射(Cumulative mapping)对性能提升的可能性，实验结果如表 7 所列。可以看到，相比跨层映射和累加映射，更简单的顶层直接映射在 VQA 任务上的准确率为 60.04%，比跨层映射和累加映射的准确率分别提升了 1.46%和 1.07%。研究认为，这是由于模型顶部汇聚了大量的高级语义信息，如果使用跨层映射和累加映射，可能会丢失部分有价值的语义信息并且使原本的语义信息受到污染，从而降低模型性能。因此，本文采用顶层直接映射的方式连接单模态编码器与跨模态编码器。

表 7 映射方法的消融实验结果

Table 7 Results of ablation experiments using mapping methods

(%)

方法	VQA
Cross-layer mapping	58.59
Cumulative mapping	58.97
Top mapping	**60.04**

最后，实验探索了 PBTS 结构和 TCMDD 方法共同作用下对模型性能的影响，结果如表 8 所列。可以看到，在基线模型上应用 PBTS 和 TCMDD 方法后，模型在 VQA 任务上的表现分别提升了 3.41%和 2.72%，而在同时使用这两个方法后，模型的性能得到了进一步的提升，比基线模型的性能提升了 4.47%。实验结果表明，PBTS 和 TCMDD 方法都能够有效地提升模型的视觉语言理解能力，而将二者组合应用后模型性能得到了进一步的提升。

表 8 两种方法对模型性能的消融实验结果

Table 8 Results of ablation experiments on model performance using two methods

(%)

方法	VQA
Baseline	60.37
+ PBTS	63.78
+ TCMDD	63.09
+ Both	**64.84**

4.3.3 模型规模及推理速度

本文在 Tesla V100 GPU 上评估了所提出的双编码器模型和融合编码器模型 ViLT 在视觉语言理解任务上的推理速度差异。由于本文模型采用双编码器架构，因此可以缓存图像文本表示来减少冗余计算。最终实验结果如表 9 所列。

表 9 本文模型和 ViLT 模型推理速度的比较

Table 9 Comparison of inference speed between the proposed model and ViLT model

(s)

模型	NLVR2	SNLI-VE	VQA
ViLT	160.5	195.2	1295.6
Ours	57.3	67.2	462.5
	(2.8×)	(2.9×)	(2.8×)

结束语 为了提升双塔模型在视觉语言理解任务上的性能，本文提出了一种多模态融合方法。将融合编码器中丰富的跨模态交互知识在预训练阶段和调优阶段通过跨模态注意力蒸馏迁移到单模态编码器和跨模态编码器中；又由于跨模态编码器独特的桥塔结构，使得模型在预训练过程中不仅学习到了丰富的不同层级的单模态知识，也学习到了多个模态间的交互知识。实验结果表明，所提出的多模态融合方法不仅大大提升了模型的视觉语言理解能力，而且相比融合编码器模型具备更快的推理速度。但当前的研究只使用了单一教师蒸馏，未来将研究探索多教师蒸馏对模型性能提升的可能性。

参 考 文 献

[1] KIM W，SON B，KIM I. Vilt：Vision-and-language supervision [C]// International Conference on Machine Learning. PMLR，2021：5583-5594.

[2] DOSOVITSKIY A，BEYER L，KOLESNIKOV A，et al. An image is worth 16x16 words：Transformers for image recognition at scale[J]. arXiv：2010. 11929，2020.

[3] RADFORD A，KIM J W，HALLACY C，et al. Learning transferable visual models from natural language supervision[C]// International Conference on Machine Learning. PMLR，2021：8748-8763.

[4] JIA C，YANG Y，XIA Y，et al. Scaling up visual and vision-language representation learning with noisy text supervision[C]// International Conference on Machine Learning. PMLR，2021：4904-4916.

[5] ANTOL S，AGRAWAL A，LU J，et al. Vqa：Visual question answering[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015：2425-2433.

[6] XIE N，LAI F，DORAN D，et al. Visual entailment：A novel task for fine-grained image understanding[J]. arXiv：1901. 06706，2019.

[7] HINTON G，VINYALS O，DEAN J. Distilling the knowledge in a neural network[J]. arXiv：1503. 02531，2015.

[8] ROMERO A，BALLAS N，KAHOU S E，et al. Fitnets：Hints for thin deep nets[J]. arXiv：1412. 6550，2014.

[9] ZAGORUYKO S，KOMODAKIS N. Paying more attention to attention：Improving the performance of convolutional neural networks via attention transfer[J]. arXiv：1612. 03928，2016.

[10] LI D，YANG Y，TANG H，et al. VIRT：Improving Representation-based Text Matching via Virtual Interaction[C]// Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022：914-925.

[11] WANG Z，WANG W，ZHU H，et al. Distilled dual-encoder model for vision-language understanding[J]. arXiv：2112. 08723，2021.

[12] LU Y，LIU Y，LIU J，et al. Ernie-search：Bridging cross-encoder with dual-encoder via self on-the-fly distillation for dense passage retrieval[J]. arXiv：2205. 09153，2022.

[13] CHEN Y C，LI L，YU L，et al. Uniter：Universal image-text representation learning[C]// European Conference on Computer Vision. Cham：Springer International Publishing，2020：104-120.

[14] CHO J，LEI J，TAN H，et al. Unifying vision-and-language tasks via text generation [C]// International Conference on Machine Learning. PMLR，2021：1931-1942.

[15] GIRSHICK R. Fast r-cnn[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015：1440-1448.

[16] HE K，ZHANG X，REN S，et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016：770-778.

[17] WANG P，YANG A，MEN R，et al. Ofa：Unifying architectures，tasks，and modalities through a simple sequence-to-sequence learning framework [C]// International Conference on Machine Learning. PMLR，2022：23318-23340.

[18] WANG Z，YU J，YU A W，et al. Simvlm：Simple visual language model pretraining with weak supervision [J]. arXiv：2108. 10904，2021.

[19] XU X，WU C，ROSENMAN S，et al. Bridgetower：Building bridges between encoders in vision-language representation learning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2023：10637-10647.

[20] WANG W，WEI F，DONG L，et al. Minilm：Deep self-attention distillation for task-agnostic compression of pre-trained transformers[J]. Advances in Neural Information Processing Systems，2020，33：5776-5788.

[21] LIN T Y，MAIRE M，BELONGIE S，et al. Microsoft coco：Common objects in context[C]// Computer Vision－ECCV 2014：13th European Conference，Zurich，Switzerland，Part V 13. Springer International Publishing，2014：740-755.

[22] 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）. 2018：2556-2565.

[23] ORDONEZ V，KULKARNI G，BERG T. Im2text：Describing images using 1 million captioned photographs[C]// Proceedings of the 24th International Conference on Neural Information Processing Systems. 2011：1143-1151.

[24] KRISHNA R，ZHU Y，GROTH O，et al. Visual genome：Connecting language and vision using crowdsourced dense image annotations[J]. International Journal of Computer Vision，2017，123：32-73.

[25] SUHR A，ZHOU S，ZHANG A，et al. A corpus for reasoning about natural language grounded in photographs [J]. arXiv：1811. 00491，2018.

[26] XIE N，LAI F，DORAN D，et al. Visual entailment：A novel task for fine-grained image understanding[J]. arXiv：1901. 06706，2019.

[27] GOYAL Y，KHOT T，SUMMERS-STAY D，et al. Making the v in vqa matter：Elevating the role of image understanding in visual question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017：6904-6913.

HUANG Xiaofei，born in 1994，postgraduate. His main research interests include knowledge distillation and multi-modality.

GUO Weibin，born in 1968，Ph. D，professor. His main research interests include high performance computing，computer application and software engineering.

（责任编辑：何杨）