MITF：基于图像映射文本特征的跨模态图文检索方法

娄馨月[1,2]，李铀[1,2]，齐睿[1,2]，陈钰枫[1,2]，徐金安[1,2*]
[1]北京交通大学，计算机科学与技术学院，北京，100044
[2]交通数据分析与挖掘北京市重点实验室，北京，100044
{20241254,21241143,20281284,chenyf,jaxu}@bjtu.edu.cn

摘要

减小图文信息间的语义鸿沟，促进跨模态信息的对齐与融合一直是解决跨模态图文检索问题的关键。但现有的双流模型因为训练时图像编码器与文本编码器是分开的，导致图文特征的对齐与融合较难。因此，本文提出图像映射文本特征（MITF）网络将不同模态（图像和文本）的信息映射到单一模态（文本），进一步增强跨模态语义的融合和对齐，提高图文检索的性能。具体地，在冻结预训练的中文视觉语言模型Chinese-CLIP参数的情况下，训练一个MITF网络将图像映射为伪语言标记，在此基础上引入提示词自动学习机制提升模型对于伪语言标记的理解能力。同时，在检索时构建Faiss索引提高检索速度。在三个开源数据集的实验结果表明所提方法相比原始Chinese-CLIP模型检索时的Mean Recall指标平均提高了3.7%，检索速度提高了约4倍。同时，图文特征可视化结果进一步表明所提方法提高了图像特征与文本特征的对齐程度。

关键词： 语义对齐；跨模态；图文检索；图像映射文本特征网络

MITF:Cross-modal Image-text Retrieval Method with Mapping Images to Text Features

Xinyue Lou[1,2], You Li[1,2], Rui Qi[1,2], Yufeng Chen[1,2], Jinan Xu[1,2*]
[1]School of Computer Science and Technology, Beijing Jiaotong University, Beijing, 100044
[2]Beijing Key Lab of Traffic Data Analysis and Mining, Beijing, 100044
{20241254,21241143,20281284,chenyf,jaxu}@bjtu.edu.cn

Abstract

To address the cross-modal image-text retrieval challenge, it's essential to close the semantic gap between vision and language. However, the prevailing dual-stream model which separates image encoder and text encoder during training complicates the problem. In this paper, we propose a mapping network to map information from different modalities (image and text) to a single modality (text), called MITF network, with the purpose of improving the fusion and alignment of cross-modal semantics. Given a frozen pre-trained model, Chinese-CLIP, we train the MITF network to convert the visual embedding into the corresponding pseudo language tokens. Additionally, we also introduce a mechanism that automatically learns prompt to enhance the model's understanding of pseudo language tokens. To speed up the retrieval process, the Faiss index is employed. Experimental findings on three datasets show that our method achieves an absolute 3.7% MR gain and quadruples retrieval speed compared to the

*徐金安（通讯作者）：jaxu@bjtu.edu.cn

第二十三届中国计算语言学大会论文集，第1页-第14页，太原，中国，2024年7月25日至28日。
卷1：主会论文

vanilla Chinese-CLIP model. Additionally, the image-text feature visualization results further show that our method improves the alignment between image features and text features.

Keywords: Semantic alignment , Cross-modal , Image-text retrieval , Mapping images to text features network

1 引言

近年来，随着图像、文本等多模态数据的高速增长，对于访问和利用各种跨模态数据的需求也相应增长，形成了跨模态研究的三个主要领域：跨模态生成、跨模态视觉问答和跨模态检索。跨模态检索是指给定一种模态样本，返回另一模态中相似的样本，如图像-文本检索等，在公安、医疗和电商领域等均有着广泛的应用前景(Kaur et al., 2021)。但由于各模态语义间存在的差异，如何减小差异实现跨模态语义的有效对齐逐渐成为了研究热点。

受预训练语言模型成功的启发(Devlin et al., 2018; Liu et al., 2019; Brown et al., 2020)，大规模预训练视觉语言模型研究(Li et al., 2019a; Radford et al., 2021; Wang et al., 2022a; Liu et al., 2023a)在各种跨模态任务上取得了显著进展，包括图像-文本检索。这些现有方法通常可以根据模型架构大致分为两类：单流模型和双流模型。单流模型通常会添加额外的Transformer层(Vaswani et al., 2017)来模拟图像和文本表示之间的深度交互，从而提高检索性能，如图1(a)所示。但当该模型应用于整个图像集合时会导致检索速度不稳定，因为每当给出新的文本查询时，每个图像样本都需要进行跨模态计算。相比之下，双流模型以完全解耦的方式编码视觉和文本输入，如图1(b)所示。图像表示允许预先计算并独立于文本查询重新使用，检索速度较快。但双流模型也因为图像编码器和文本编码器在训练时参数是独立的，导致两个编码器对图文信息编码后得到的特征不是完全对齐的，多模态语义信息的融合和交互较难。虽然存在一些权衡方法，如ALBEF模型(Li et al., 2021)通过跨模态注意力机制，将图像表征和文本表征先对齐再进行融合，但训练时计算成本较高，且检索速度较慢。

为了解决以上挑战，本文围绕跨模态检索中的语义对齐问题，提出图像映射文本特征（Mapping Images to Text Features，MITF）网络，如图1(c)所示，将图像映射为文本特征后，在文本编码器端模拟跨模态信息间的深度交互，进一步增强跨模态语义的有效对齐和融合，从而提高检索性能。同时，图像表示允许预先计算，检索速度较为稳定。具体来说，首先在冻结预训练的中文视觉语言模型Chinese-CLIP（以下简称CN-CLIP）模型(Yang et al., 2022)参数的情况下，训练一个MITF网络将图像映射为伪语言标记即用文本特征表示图像信息的关键内容。而为了使伪语言标记更好地为文本编码器所利用，引入提示词（prompt）自动学习机制进行优化。然后在检索时构建Faiss索引，有效提升检索速度。该方法的有效性在Flickr30k-CN(Lan et al., 2017)数据集、COCO-CN(Li et al., 2019b)数据集和MUGE(Lin et al., 2021)数据集上均得到了验证。本文的主要贡献总结如下：

- 提出图像映射文本特征网络，并引入提示词自动学习机制，利用预训练的视觉语言模型将图像转换为伪语言标记，以减小图文信息间的语义鸿沟。
- 在检索时构建Faiss索引，提高检索效率，在检索精度不变的情况下，检索速度提高了约四倍。
- 经过实验验证，本文所提出的图像映射文本特征网络以较小的参数量实现了与全量微调方法相当的结果，有效降低了模型训练的计算资源和成本。

2 相关工作

2.1 跨模态图文检索

近年来，随着注意力机制的广泛应用，预训练视觉语言模型已被广泛研究并应用于图文检索领域(Tan and Bansal, 2019; Wang et al., 2022a; Liu et al., 2023a)。模型结构大致可分

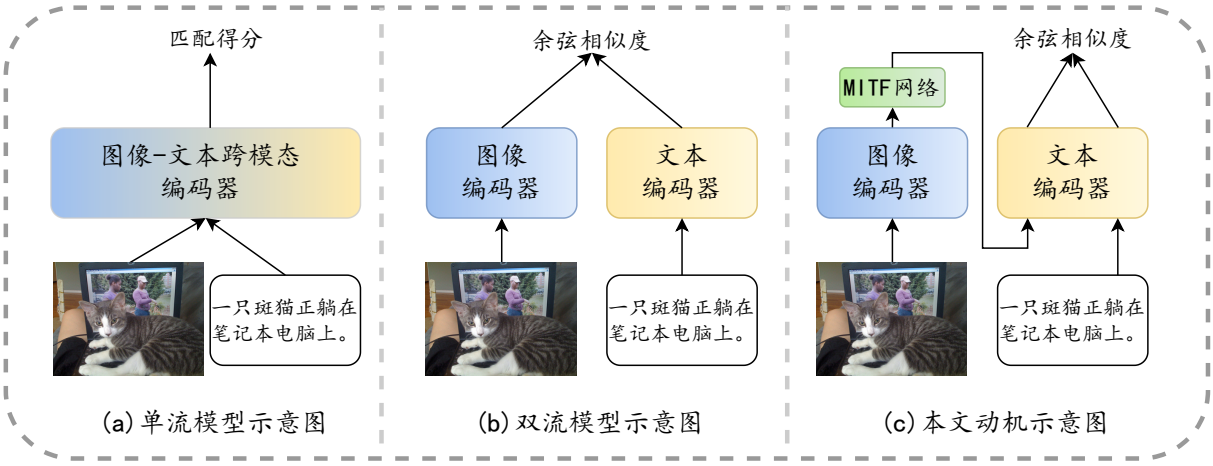


图 1: 图文检索模型与本文动机示意图

为两类：单流模型和双流模型。单流模型通过深度交互模块对图像和文本进行联合编码，并输出融合特征。Lu et al. (2019)提出的早期算法采用对象检测器提取图像特征，但通常忽略了重要的背景信息。随后，Kim et al. (2021)提出ViLT模型将图像编码器和文本编码器统一为Transformer模型(Vaswani et al., 2017)，以充分利用所有信息。然而，这些模型都依赖于跨模态Transfomer编码器来同时融合视觉和文本信号，需要大量的计算成本并降低了推理速度。双流模型主要侧重于学习如何将从独立编码器中获得的视觉特征与文本特征相统一。由于只有一个轻量级交互模块（通常是MLP网络或点乘法）用于匹配图像和文本特征，因此双流模型可以在数十亿实例上进行对比学习，如CLIP模型(Radford et al., 2021)、ALIGN模型(Jia et al., 2021)和BLIP2模型(Li et al., 2023)等。在双流模型中，不同模态的交互只发生在最终的轻量级模块，导致与单流模型相比性能稍差。然而，这种双流模型的后期交互方案允许图像和文本表示的预先计算，从而使它适合于实时搜索。

通常，双流模型由于缺乏有效的跨模态特征融合，表现出较差的检索性能。而单流模型通过深度交互模块在编码过程中整合来自多种模态的信息，通常会导致检索性能较好，但牺牲了灵活性导致检索速度极低。此后虽然一些权衡方法被提出以解决以上问题，如LiT方法(Zhai et al., 2022)训练时冻结图像编码器以整合较为干净的图像数据源和图文对数据源的优点，而ALBEF模型(Li et al., 2021)则通过跨模态注意力机制，先将图像表征和文本表征对齐再融合，但训练时的计算成本较高。受ALBEF模型的启发，本文在CN-CLIP模型(Yang et al., 2022)的基础上，提出图像映射文本特征网络，试图将多模态信息映射到文本模态以减小模态间的差异，进一步增强跨模态语义的对齐和融合。

2.2 用语言标记表示图像方法

曾经有一些研究试图在视觉语言模型的预训练过程中将图像区域映射为语言标记(Wang et al., 2022b; Chen et al., 2019; Li et al., 2020)。典型的框架包括：(1) 使用预训练的物体检测器检测图像中的物体；(2) 将图像区域和相应的文本描述输入文本编码器；(3) 优化多个多模态目标以获得视觉语言模型。但这些方法都需要在预训练阶段使用高性能的物体检测器。还有其他方法如Cohen et al. (2022)使用循环对比损失来学习如何将一组图像转换成一个概念词标记，以解决个性化图像检索和语义分割等问题。Gal et al. (2022)提出用单词标记来表示描绘同一对象的几幅图像，并将其用于文本到图像的生成任务。Saito et al. (2023)提出pic2word方法，使用未标记图像数据集进行无监督对比学习，以便灵活地组合图像和文本查询。受这些方法的启发，本文提出将图像映射为伪语言标记的方法，以实现跨模态语义更有效的对齐，提升图文检索的效率。

3 方法

在本节中，将介绍本文的MITF网络和提示词自动学习机制及检索方法。训练和推理过程的概述如图2所示。所采用的基座模型为预训练的视觉语言模型CN-CLIP(Yang et al.,

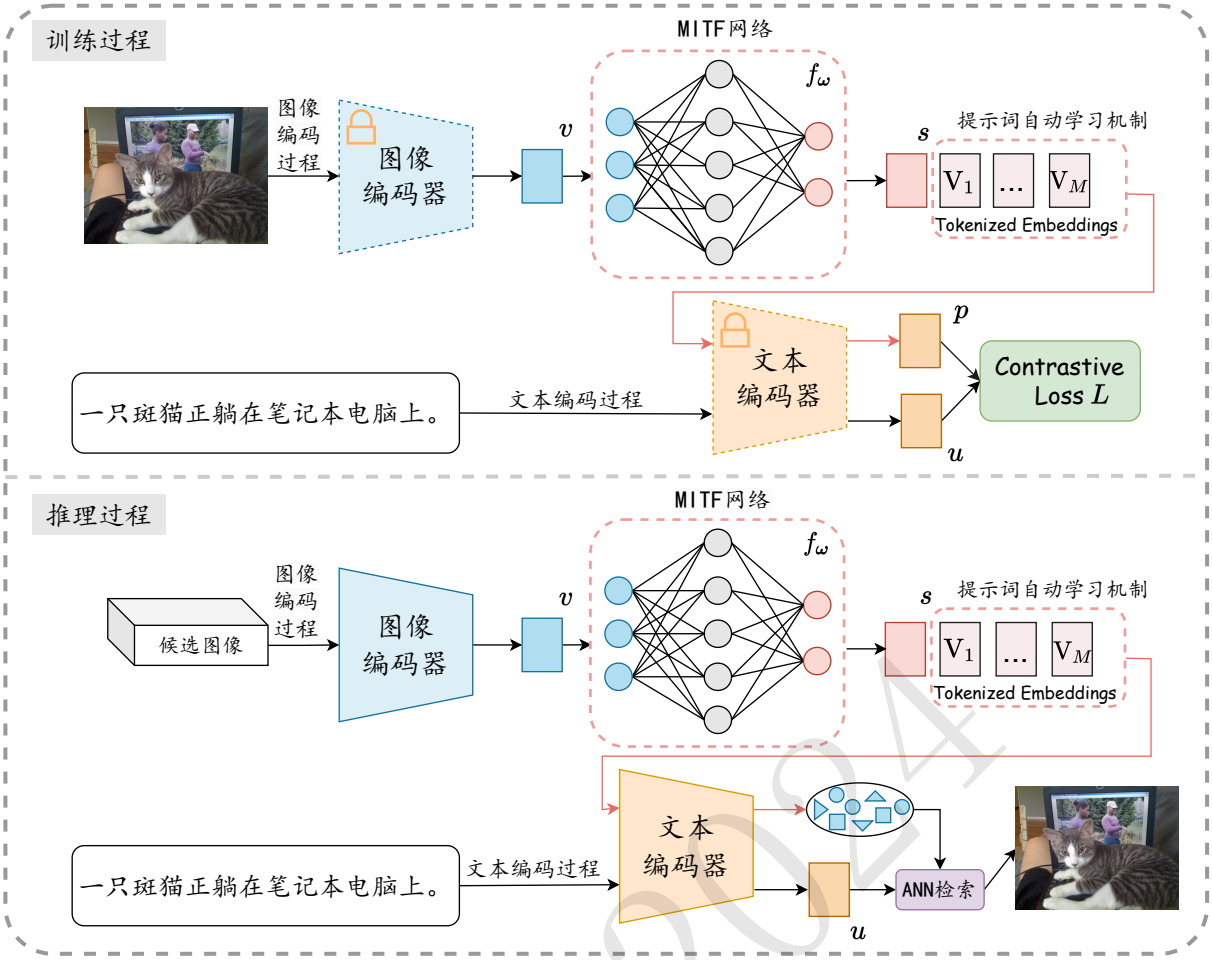


图 2: 训练过程和推理过程概述图

2022)，该模型由文本编码器和图像编码器组成，一共提供了五个不同规模的模型。文本编码器为RoBERTa-wwm系列模型(Liu et al., 2019)，图像编码器为RN50模型(He et al., 2016)或者Vision Transformer系列模型(Dosovitskiy et al., 2020)。在训练时，冻结预训练文本编码器和图像编码器，训练一个映射网络将视觉嵌入转换为相应的伪语言标记。并引入提示词自动学习机制使得到的伪语言标记更好地为文本编码器所利用。在推理时，将图像经过MITF网络得到伪语言标记，将其与待检索文本分别经过文本编码器得到的特征进行比较，通过近似近邻检索得到最相似的图文样本对。我们将在以下小节中详细介绍训练和推理机制。

3.1 对比语言-图像预训练模型（CN-CLIP）

给定一个图像文本数据集$\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{t}_n)\}_{n=1}^{N}$，其中$x \in \mathcal{X}$为图像，$t \in \mathcal{T}$为分词后的语言描述。对于一张图像$x$，以$\theta$为参数的图像编码器$f_{\boldsymbol{\theta}}$对其进行编码得到视觉表征$\tilde{\boldsymbol{v}} \in R^{d\times 1}$：$\tilde{\boldsymbol{v}} = f_{\boldsymbol{\theta}}(\boldsymbol{x})$。对于文本描述$t$，以$\phi$为参数的文本编码器$f_{\boldsymbol{\phi}}$对其进行编码提取语言表征$\tilde{\boldsymbol{u}} \in R^{d\times 1}$：$\tilde{\boldsymbol{u}} = f_{\boldsymbol{\phi}}(\boldsymbol{t})$。CN-CLIP模型(Yang et al., 2022)采用对比学习的方法，使用约2亿对中文图文对进行训练。对于批次$\mathcal{B}$中的第i个图像$\boldsymbol{x}_i$和与其对应的第j个文本描述$\boldsymbol{t}_j$，对其分别进行归一化处理$\boldsymbol{v}_i = \frac{\tilde{\boldsymbol{v}}_i}{\|\tilde{\boldsymbol{v}}_i\|}$和$\boldsymbol{u}_j = \frac{\tilde{\boldsymbol{u}}_j}{\|\tilde{\boldsymbol{u}}_j\|}$。最后，CN-CLIP模型使用对称交叉熵损失进行优化，损失函数如下所示：

$$\min_{\{\boldsymbol{\theta},\boldsymbol{\phi}\}} \mathcal{L}_{\text{con}} = \mathcal{L}_{t2i} + \mathcal{L}_{i2t} \tag{1}$$

其中包括2个对比项，温度系数τ用于控制对困难负样本的惩罚力度，如下式所示：

$$\mathcal{L}_{t2i} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\log\frac{\exp\left(\tau \boldsymbol{u}_i^T \boldsymbol{v}_i\right)}{\sum_{j\in\mathcal{B}}\exp\left(\tau \boldsymbol{u}_i^T \boldsymbol{v}_j\right)} \tag{2}$$

$$\mathcal{L}_{i2t} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\log\frac{\exp\left(\tau \boldsymbol{v}_i^T \boldsymbol{u}_i\right)}{\sum_{j\in\mathcal{B}}\exp\left(\tau \boldsymbol{v}_i^T \boldsymbol{u}_j\right)} \tag{3}$$

而在进行图文检索推理时，用余弦相似度来衡量图像文本的相似度大小，给定一张图像x，预测文本描述t与其匹配的概率公式为：

$$p(t=i \mid \boldsymbol{x}) = \frac{\exp\left(\cos\left(\boldsymbol{u_i}, \boldsymbol{v}\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\cos\left(\boldsymbol{u_j}, \boldsymbol{v}\right)/\tau\right)} \tag{4}$$

同理，给定一个文本描述t，预测图像x与其匹配的概率公式为：

$$p(x=i \mid \boldsymbol{t}) = \frac{\exp\left(\cos\left(\boldsymbol{v_i}, \boldsymbol{u}\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\cos\left(\boldsymbol{v_j}, \boldsymbol{u}\right)/\tau\right)} \tag{5}$$

3.2 图像映射文本特征（MITF）网络

该方法的核心是训练一个映射网络来输出伪语言标记，在固定文本编码器和图像编码器的情况下训练该网络，这样获得的伪语言标记能够最大程度地减少图像特征和文本特征之间的对比损失，即强制网络形成一个从图像嵌入到语言嵌入的循环，如图2所示，将图文信息都映射到文本模态以减小模态间特征的差异。对于未归一化的视觉嵌入$\tilde{\boldsymbol{v}}$，参数为$\boldsymbol{\omega}$的映射网络$f_{\boldsymbol{\omega}}$提取伪语言标记嵌入$s = f_{\boldsymbol{\omega}}(\tilde{\boldsymbol{v}})$。$f_{\boldsymbol{\omega}}$为一个深度为3层的多层感知机，其参数大约为17M。如图2右侧所示，在s后添加自动优化的提示句（关于提示句优化的设计在3.3章节中介绍），得到$\hat{s}$。然后将$\hat{s}$输入文本编码器$f_{\boldsymbol{\phi}}$，得到伪语言嵌入$\tilde{\boldsymbol{p}} = f_{\phi}(\hat{s})$，希望$\tilde{\boldsymbol{p}}$能够在表示输入的图像信息的基础上尽可能接近于其对应的原文本的语言嵌入$\tilde{\boldsymbol{u}}$。该映射网络的对比损失为：

$$\min_{M}\mathcal{L} = \mathcal{L}_{t2i}(\boldsymbol{p}, \boldsymbol{u}) + \mathcal{L}_{i2t}(\boldsymbol{p}, \boldsymbol{u}) \tag{6}$$

其中包括2个对比项，如下所示：

$$\mathcal{L}_{t2i}(\boldsymbol{p}, \boldsymbol{u}) = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\log\frac{\exp\left(\tau \boldsymbol{u}_i^T \boldsymbol{p}_i\right)}{\sum_{j\in\mathcal{B}}\exp\left(\tau \boldsymbol{u}_i^T \boldsymbol{p}_j\right)} \tag{7}$$

$$\mathcal{L}_{i2t}(\boldsymbol{p}, \boldsymbol{u}) = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\log\frac{\exp\left(\tau \boldsymbol{p}_i^T \boldsymbol{u}_i\right)}{\sum_{j\in\mathcal{B}}\exp\left(\tau \boldsymbol{p}_i^T \boldsymbol{u}_j\right)} \tag{8}$$

其中$\boldsymbol{p}_i = \frac{\tilde{\boldsymbol{p}}_i}{\|\tilde{\boldsymbol{p}}_i\|}$。在此基础上训练$f_{\boldsymbol{\omega}}$生成$\boldsymbol{s}_i$使$\boldsymbol{p}_i$接近$\boldsymbol{u}_i$。上述公式与式（1、2、3）的不同之处在于：（1）优化是针对参数$\boldsymbol{\omega}$进行的，而其他部分是固定的；（2）使用p代替v，换言之，对比学习的目标是使该图像生成的伪语言标记与其真实文本标记尽可能接近。

3.3 提示词自动学习机制

在3.2节中通过MITF网络得到的伪语言标记是连续的特征，所表达的语义信息较为受限，为了使生成的伪语言标记更好地为文本编码器所利用，并且受提示学习在预训练视觉语言模型上同样取得了令人瞩目的性能的启发(Zhou et al., 2022)，我们引入了提示词自动学习机制。提示工程通过给定具有指示性的离散字符辅助视觉语言模型完成相应的下游任务，如clip模型在进行图像分类时给定一个提示语句：*A photo of a object*（其中object为待分类的物体名称）(Radford et al., 2021)，使其完成图像分类的任务，而提示学习的研究旨在将这一过程自动化，而不是手动设计提示。提示学习的方法为我们如何更好地提升模型对于伪语言标记的理解能力以提高检索性能提供了可供借鉴的思路。它通过从数据中学习端到端连续向量对上下文词

语进行建模，避免手动提示调优。概述如图2所示。具体来说，提供给文本编码器f_ϕ的提示设计为如下形式：

$$\hat{s} = [\mathrm{s}][\mathrm{V}]_1[\mathrm{V}]_2 \dots [\mathrm{V}]_M \tag{9}$$

其中，每个$[\mathrm{V}]_m (m \in \{1, \dots, M\})$是一个维度与词嵌入相同的向量（如CN-CLIP$_{\text{ViT-B}}$模型中的768维），而$\mathcal{M}$是一个超参数，指定了上下文标记的数量，$s$为MITF网络中所学习到的伪语言标记。给定一张图像x，预测文本描述t与其匹配的概率公式为：

$$p(t = i \mid \boldsymbol{x}) = \frac{\exp\left(\cos\left(\boldsymbol{u_i}, \boldsymbol{p}\right)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\cos\left(\boldsymbol{u_j}, \boldsymbol{p}\right)/\tau\right)} \tag{10}$$

同理，给定一个文本描述t，预测图像x与其匹配的概率公式为：

$$p(x = i \mid \boldsymbol{t}) = \frac{\exp\left(\cos\left(\boldsymbol{p_i}, \boldsymbol{u}\right)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\cos\left(\boldsymbol{p_j}, \boldsymbol{u}\right)/\tau\right)} \tag{11}$$

在交叉熵损失函数的基础上，训练过程中将进行梯度调整以最小化对称多类损失，梯度可以通过文本编码器f_ϕ反向传播，利用参数中编码的丰富知识来优化提示词获得更好的表示。

3.4 使用Faiss索引加速的推理过程

在给定文本检索图像时推理过程概述如图2所示，候选图像经过图像编码器和MITF网络得到伪语言标记，将其与给定文本分别经过文本编码器得到的特征进行对比，通过检索得到最相似的k个样本对，采用余弦相似度函数来衡量样本相似度大小。对于检索过程，本文采用了Faiss框架提供的近似近邻（Approximate Nearest Neighbor，ANN）检索方法来提高检索速度。Faiss*全称是Facebook AI Similarity Search，是由Facebook开发的用于高效相似性搜索的开源库，是目前最为成熟的ANN检索库。对于检索任务，可大致分为以下三步：创建索引列表、聚类训练、查询。我们使用了Faiss的倒排乘积量化方法来加快检索速度，其思路是将原始的高维向量空间划分为多个低维子空间，在每个子空间内进行向量量化编码。最终将所有子空间量化编码依次串联起来作为原向量的编码。在查询时，需要对查询向量进行同样的子空间拆分和量化，然后基于这些量化码在倒排索引中进行近邻搜索。经过处理后，我们得到了经过乘积量化器优化和k-means训练的Faiss索引表I。在给定文本检索图像时的算法伪代码如下所示。对于输入向量u，Faiss 能快速检索出与u最相似的前k个备选向量，表示为$[h_1, h_2, \dots, h_k]$，即候选图像集中与给定文本最匹配的k个样本。同理可得给定图像检索文本的过程。

Algorithm 1: 基于Faiss索引的图文检索算法伪代码

Input：text_features，image_features
Output：top_k_image_ids
for image_feature in image_features **do**
    image_feats=list(image_feature)
index=faiss.IndexFlatIP()
index.add(image_feats)
for text_feature in text_features **do**
    indices = index.search(text_feature, top$_k$)
    Return indices as top_k_image_ids

4 实验

4.1 实验设置

本实验所采用的基座模型为Aliyun开源的CN-CLIP模型†，若无特别指明，模型所采用的基座模型均为CN-CLIP$_{\text{ViT-B}}$模型。MITF网络由深度为3层的多层感知机网络组成，所采用的激活函数为Mish函数(Misra, 2019)（输出层没有使用激活函数，具体的设计细节将在附录A中

*https://github.com/facebookresearch/Faiss
†https://github.com/OFA-Sys/Chinese-CLIP

给出）。使用的优化器为AdamW(Loshchilov and Hutter, 2017)，学习率为$8 * 10^{-4}$，权重衰减为0.1。对比学习的批量大小为576，设置迭代次数为10，采用学习率余弦衰减策略，具体的超参数设置将在附录B中给出。基座模型为CN-CLIP$_{\text{ViT-L}}$和CN-CLIP$_{\text{ViT-H}}$的模型在3块GeForce RTX 3090 GPU上进行训练，其余模型均在6块GeForce RTX 2080 Ti GPU上进行训练。本次实验所选用的数据集为Flickr30k-CN数据集(Lan et al., 2017)、COCO-CN数据集(Li et al., 2019b)和MUGE数据集(Lin et al., 2021)，关于各数据集详细数据将在附录C中给出。首先在对应数据训练集上冻结预训练模型训练MITF网络，选取在验证集上效果最好的模型，然后在对应测试集上测试结果。由于MUGE数据集官方没有提供图到文标注数据和测试集数据，关于MUGE数据集只在验证集上进行文到图检索实验。同时，我们也在相同实验环境下进行了CN-CLIP全量微调实验以进行对比，对MITF网络进行全面分析。

4.2 实验结果

对于图像检索文本和文本检索图像任务，本文使用该任务常用的Recall@1 (R@1)、Recall@5 (R@5)、Recall@10 (R@10) 和Mean Recall (MR)指标来进行评测。

	方法	训练参数量	检索时间	图像检索文本				文本检索图像			
				R@1	R@5	R@10	MR	R@1	R@5	R@10	MR
COCO-CN	CN-CLIP	-	4s	57.0	84.1	93.6	78.2	62.2	86.6	94.9	81.2
	Finetune*	188M	4s	**67.1**	**93.1**	**97.6**	**85.9**	**68.3**	**93.4**	**97.6**	**86.5**
	Ours	**17M**	**1s**	63.7	92.4	97.2	84.4	66.9	91.7	96.8	85.1
Flickr30k-CN	CN-CLIP	-	9s	74.6	93.5	97.1	88.4	62.7	86.9	92.8	80.8
	Finetune*	188M	9s	**87.5**	**97.8**	**99.4**	**94.9**	**71.4**	**92.2**	**96.1**	**86.6**
	Ours	**17M**	**3s**	84.9	96.9	99.1	93.6	70.0	90.9	95.6	85.5
MUGE	CN-CLIP	-	139s	-	-	-	-	52.1	76.7	84.4	71.1
	Finetune*	188M	139s	-	-	-	-	**54.2**	**80.3**	**86.9**	**73.8**
	Ours	**17M**	**41s**	-	-	-	-	52.3	78.2	86.0	72.2
平均	CN-CLIP	-	51s	65.8	88.8	95.4	83.3	59.0	83.4	90.7	77.7
	Finetune*	188M	51s	**77.3**	**95.5**	**98.5**	**90.4**	**64.6**	**88.6**	**93.5**	**82.3**
	Ours	**17M**	**15s**	74.3	94.7	98.2	89.0	63.1	86.9	92.8	80.9

表 1: 与CN-CLIP全量微调方法的对比（*为我们复现的实验结果），检索时间为在相应数据集的测试集（验证集）检索所有数据所需的时间

与CN-CLIP全量微调方法的对比。对比学习的效果与学习时的批量大小有较大关系，在内存资源一定的情况下，模型参数量越多，对比学习的批量数就越小。因此，为保证实验的公平性，我们在相同实验环境条件下对CN-CLIP$_{\text{ViT-B}}$模型进行了全量微调方法的实验，对比结果如表1所示，在3个数据集上的平均实验结果表明，训练时MITF网络参数量为全量微调参数量的9.4%，但与全量微调实验结果MR值的误差仅为1.4%。而与原始CN-CLIP模型零样本预测的表现相比，本文所提出的模型在3个数据集上均达到了较好的效果，在COCO-CN数据集上相比原始CN-CLIP模型MR指标提升了5.0%，在Flickr30k-CN数据集上提升了5.0%，在MUGE数据集上提升了1.1%。同时，在进行图文检索时检索速度提高了约4倍。证明了本文方法的有效性。较低的参数量意味着占用较少的内存和较短的训练时间，利于加快模型在指定下游任务上的微调和模型的部署，并且降低了模型对于数据量较少的数据产生过拟合现象的风险。

与其他中文视觉语言预训练模型对比。对比结果如表2所示，我们选取了FILIP(Yao et al., 2021)模型与Wukong(Gu et al., 2022)模型在下游数据集全量微调的结果进行对比。由于FILIP模型没有提供源代码，实验结果均来自WuKong论文(Gu et al., 2022)中提供的结果。实验结果表明，本文方法优于目前的其他中文预训练视觉语言模型。

4.3 消融实验

提示词自动学习机制的影响。为探究该模块的影响，我们在三个下游数据集上使用手动设计的提示句代替提示词自动学习，在其他条件都相同的情况下进行实验。手动设计的提示句为：“_____是这张图片的内容”,下划线部分为MITF网络所学习到的伪语言标记。实验结果如表3所示。分析可知提示词自动学习的方法普遍优于手动设计提示词的方法，且该方法在MUGE数据集上的效果更为显著，可能原因为MUGE数据集专注于电商领域，领域内图文信

	方法	图像检索文本				文本检索图像			
		R@1	R@5	R@10	MR	R@1	R@5	R@10	MR
COCO-CN	FILIP	52.7	81.3	88.3	74.0	56.2	86.8	94.3	79.1
	WuKong	**65.8**	90.3	96.6	84.2	**67.0**	91.4	96.7	85.0
	Ours	63.7	**92.4**	**97.2**	**84.4**	66.9	**91.7**	**96.8**	**85.1**
Flickr30k-CN	FILIP	72.1	91.3	95.8	86.4	57.5	84.3	90.6	77.5
	WuKong	83.9	**97.6**	99.0	93.5	67.6	89.6	94.2	83.8
	Ours	**84.9**	96.9	**99.1**	**93.6**	**70.0**	**90.9**	**95.6**	**85.5**
MUGE	FILIP	-	-	-	-	30.6	58.2	70.2	53.0
	WuKong	-	-	-	-	39.2	66.9	77.4	61.2
	Ours	-	-	-	-	**52.3**	**78.2**	**86.0**	**72.2**

表 2: 与其他中文预训练视觉语言模型的对比结果

息相似度较高，自动学习的提示词对于提升模型在该领域对于伪语言标记的理解能力更为有效。

	方法	图像检索文本				文本检索图像			
		R@1	R@5	R@10	MR	R@1	R@5	R@10	MR
COCO-CN	手动设计	59.9	89.6	96.4	82.0	66.5	91.2	**97.0**	84.9
	自动学习	**63.7**	**92.4**	**97.2**	**84.4**	**66.9**	**91.7**	96.8	**85.1**
Flickr30k-CN	手动设计	84.5	**97.0**	**99.1**	93.5	69.9	90.7	95.4	85.3
	自动学习	**84.9**	96.9	**99.1**	**93.6**	**70.0**	**90.9**	**95.6**	**85.5**
MUGE	手动设计	-	-	-	-	51.7	77.8	85.4	71.6
	自动学习	-	-	-	-	**52.3**	**78.2**	**86.0**	**72.2**

表 3: 提示词自动学习机制的影响

Faiss索引的影响。为探究构建Faiss索引进行ANN检索对实验结果的影响，我们在三个下游数据集上使用原KNN检索方式代替ANN检索，在其他条件都相同的情况下进行实验，实验结果如表4所示。结果表明，构建Faiss索引进行ANN检索较原检索速度提高了约4倍，并且检索精度没有下降，对于数据量较大的MUGE数据集，提升效果更为明显。

	ANN检索	检索时间	图像检索文本				文本检索图像			
			R@1	R@5	R@10	MR	R@1	R@5	R@10	MR
COCO-CN	✗	4s	63.7	92.4	97.2	84.4	66.9	91.7	96.8	85.1
	✓	**1s**	63.7	92.4	97.2	84.4	66.9	91.7	96.8	85.1
Flickr30k-CN	✗	9s	84.9	96.9	99.1	93.6	70.0	90.9	95.6	85.5
	✓	**3s**	84.9	96.9	99.1	93.6	70.0	90.9	95.6	85.5
MUGE	✗	139s	-	-	-	-	52.3	78.2	86.0	72.2
	✓	**41s**	-	-	-	-	52.3	78.2	86.0	72.2

表 4: Faiss索引的影响，检索时间为在相应数据集的测试集（验证集）检索所有数据所需时间

图像映射文本特征网络规模的影响。为探究MITF网络规模对于实验结果的影响，选取了模型参数量分别为17M（隐藏层维度为1200）、19M（隐藏层维度为1400）、22M（隐藏层维度为1600）、25M（隐藏层维度为1800）的四个规模网络在Flickr30k-CN数据集上进行实验。实验结果如表5所示，本文方法的实验结果均优于原始CN-CLIP模型，充分体现了该方法的有效性。

基座模型规模的影响。为探究本文所提出的方法是否受模型参数量大小的影响，在Flickr30k-CN数据集上使用CN-CLIP的4个不同规模的基座模型进行实验。实验结果如表6所

方法	图像检索文本				文本检索图像			
	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR
CN-CLIP	74.6	93.5	97.1	88.4	62.7	89.7	92.8	81.7
Ours(17M)	84.9	96.9	99.1	93.6	70.0	90.9	95.6	85.5
Ours(19M)	85.0	97.0	99.1	93.7	70.1	91.0	95.6	85.6
Ours(22M)	**85.3**	**97.2**	99.3	**94.0**	**70.4**	**91.3**	**95.9**	**85.9**
Ours(25M)	85.1	97.1	**99.4**	93.9	70.2	91.2	95.7	85.7

表 5: 图像映射文本特征网络模型规模的影响

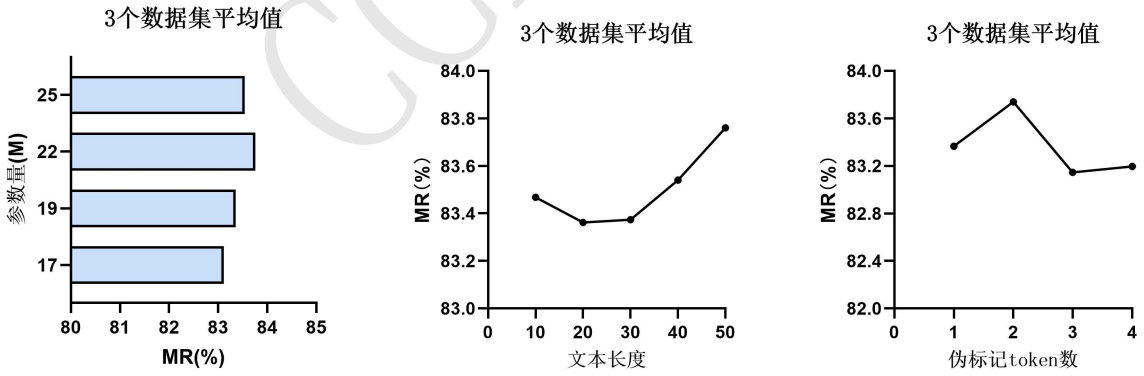
示，分析可知，本文方法在4个不同规模的基座模型的实验结果均优于原始CN-CLIP模型，表明所提出的MITF网络是一种提升跨模态检索性能的通用方法，可作为即插即用模块加入其他网络。

方法	图像检索文本				文本检索图像			
	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR
CN-CLIP$_{\text{RN50}}$	60.0	85.9	92.0	79.3	48.8	76.0	84.6	69.8
Ours$_{\text{RN50}}$	**72.7**	**91.7**	**96.5**	**87.0**	**57.7**	**83.7**	**90.5**	**77.3**
CN-CLIP$_{\text{ViT-B}}$	74.6	93.5	97.1	88.4	62.7	89.7	92.8	81.7
Ours$_{\text{ViT-B}}$	**84.9**	**96.9**	**99.1**	**93.6**	**70.0**	**90.9**	**95.6**	**85.5**
CN-CLIP$_{\text{ViT-L}}$	80.2	96.6	98.2	91.7	68.0	89.7	94.4	84.0
Ours$_{\text{ViT-L}}$	**89.2**	**98.5**	**99.7**	**95.8**	**74.4**	**93.0**	**96.3**	**87.9**
CN-CLIP$_{\text{ViT-H}}$	81.6	97.5	98.8	92.6	71.2	91.4	95.5	86.0
Ours$_{\text{ViT-H}}$	**89.9**	**99.1**	**99.9**	**96.3**	**76.1**	**94.0**	**97.0**	**89.0**

表 6: 基座模型规模的影响

4.4 分析与讨论

为了更好地体现本文方法各参数选取的科学性与合理性，本节对实验结果进行如下分析与讨论。



(a) 图像映射文本特征网络规模　(b) 提示词自动学习文本长度　(c) 伪语言标记token数

图 3: 关于图像映射文本特征网络各参数选取的讨论

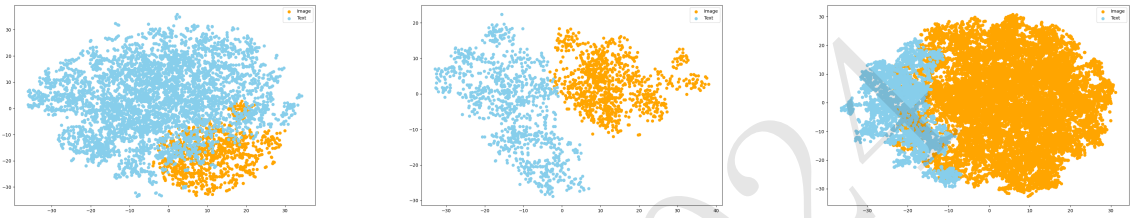
图像映射文本特征网络规模的选取：在MR值相差0.64%（参数量22M vs 17M）的情况下，我们选择参数量更少的17M规模的网络。如图3a所示，通常来说，较多的参数可以保留更多的信息从而取得更好的效果，但也提高了模型过拟合的风险。实验表明，当模型规模增加到25M时，模型性能有所下降，推测模型此时开始出现过拟合现象。

提示词自动学习文本长度的选取：当提示词自动学习文本长度为50时，3个数据集的MR平均值最高，因此我们选择了50作为提示词自动学习文本的长度。如图3b所示，当提示词文本长度从10等间距变到50时，MR值先下降后上升。这表明具有更多上下文内容会导致更好的性能。但当上下文长度为10时效果也较好，可能原因是较短的长度更有利于域泛化。

伪语言标记的token数选取：当MITF网络学习到的伪语言标记的token数为2时，3个数据集的MR平均值最高，因此伪语言标记的token数设置为2。如图3c所示，当伪语言标记的token数从1变到4时，MR值先上升后下降。直观上，更多的token表示更多的信息，而Gal et al. (2022)通过实验表明一个token的嵌入特征就足以捕获足够的语义信息，这与我们的实验结果相一致：使用1个伪token已经能得到较好的检索结果。

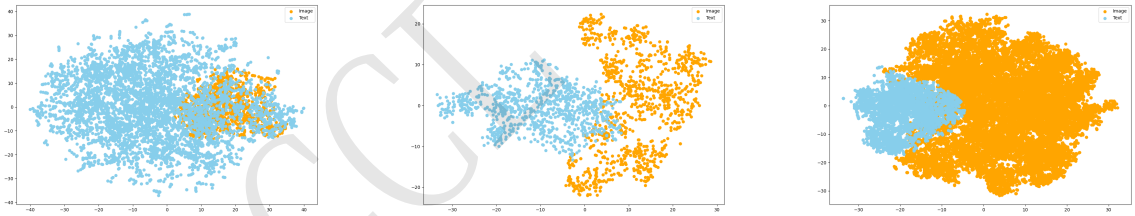
4.5 图文特征可视化结果

使用T-SNE算法(Van der Maaten and Hinton, 2008)将CN-CLIP全量微调方法和本文方法得到的图文特征分别进行降维处理，所得到的可视化结果分别如图4和图5所示。相比CN-CLIP全量微调的结果，本文方法得到的图像特征与文本特征更为接近和集中，表明该方法提高了两种模态的特征的对齐程度，缓解了不同模态间的分布差异的问题。



(a) Flickr30k-CN (b) COCO-CN (c) MUGE

图 4: CN-CLIP全量微调方法得到的图文特征可视化结果



(a) Flickr30k-CN (b) COCO-CN (c) MUGE

图 5: 本文方法得到的图文特征可视化结果

5 结论与展望

在本文中，提出了一种用于多模态特征对齐的图像映射文本特征网络，该网络试图将多模态信息融合到文本模态以减少双流编码器所带来的模态间的差异。实验结果表明，该方法以较少的参数量提升了CN-CLIP模型图文检索任务的准确率和检索速度，并且有效减少了模型微调训练时占用的内存资源。但还存在缺乏有效的中文预训练数据集导致该网络依赖于下游数据集的问题。未来将针对此问题加以改进和探索。

致谢

本研究受国家自然科学基金面上项目（No.62376019, 61976015, 61976016, 61876198, 61370130）资助。作者们还对匿名评审专家给予的宝贵建议表示衷心的感谢。

参考文献

Alexandr Andoni, Piotr Indyk, Thijs Laarhoven, Ilya Razenshteyn, and Ludwig Schmidt. 2015. Practical and optimal lsh for angular distance.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations.(2019). *arXiv preprint arXiv:1909.11740*.

Niv Cohen, Rinon Gal, Eli A Meirom, Gal Chechik, and Yuval Atzmon. 2022. "this is my unicorn, fluffy": Personalizing frozen vision-language representations. In *European conference on computer vision*, pages 558–577. Springer.

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

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

Matthijs Douze, Hervé Jégou, and Florent Perronnin. 2016. Polysemous codes. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 785–801. Springer.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. 2022. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35:26418–26431.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Michael E. Houle and Michael Nett. 2015. Rank-based similarity search: Reducing the dimensional dependence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):136–150.

Piotr Indyk and Rajeev Motwani. 1998. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Armand Joulin, Laurens Van Der Maaten, Allan Jabri, and Nicolas Vasilache. 2016. Learning visual features from large weakly supervised data. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 67–84. Springer.

Parminder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. 2021. Comparative analysis on cross-modal information retrieval: A review. *Computer Science Review*, 39:100336.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.

Weiyu Lan, Xirong Li, and Jianfeng Dong. 2017. Fluency-guided cross-lingual image captioning. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1549–1557.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019a. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. 2019b. Coco-cn for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21(9):2347–2360.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, et al. 2021. M6: A chinese multimodal pretrainer. *arXiv preprint arXiv:2103.00823*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Chong Liu, Yuqi Zhang, Hongsong Wang, Weihua Chen, Fan Wang, Yan Huang, Yi-Dong Shen, and Liang Wang. 2023a. Efficient token-guided image-text retrieval with consistent multimodal contrastive training. *IEEE Transactions on Image Processing*, 32:3622–3633.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Diganta Misra. 2019. Mish: A self regularized non-monotonic activation function. *arXiv preprint arXiv:1908.08681*.

Marius Muja and David G. Lowe. 2014. Scalable nearest neighbor algorithms for high dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Gonzalo Navarro. 2002. Searching in metric spaces by spatial approximation. *The VLDB Journal*, 11:28–46.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. 2023. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022a. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

Teng Wang, Wenhao Jiang, Zhichao Lu, Feng Zheng, Ran Cheng, Chengguo Yin, and Ping Luo. 2022b. Vlmixer: Unpaired vision-language pre-training via cross-modal cutmix. In *International Conference on Machine Learning*, pages 22680–22690. PMLR.

Patrick Wieschollek, Oliver Wang, Alexander Sorkine-Hornung, and Hendrik Lensch. 2016. Efficient large-scale approximate nearest neighbor search on the gpu. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2027–2035.

Chunyu Xie, Jincheng Li, Heng Cai, Fanjing Kong, Xiaoyu Wu, Jianfei Song, Henrique Morimitsu, Lin Yao, Dexin Wang, Dawei Leng, et al. 2022. Zero and r2d2: A large-scale chinese cross-modal benchmark and a vision-language framework. *arXiv preprint arXiv:2205.03860*.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

A 图像映射文本特征网络结构设计

表A总结了基于PyTorch的MITF网络结构描述，一层MLP网络包括一个线性输入层、Dropout层、激活函数层和一个线性输出层，相邻两层之间设计了残差块结构，共3层MLP网络组成整个MITF网络。

Layer	Module
Output	nn.Linear(embed_dim,output_dim)
FC2	nn.Linear(middle_dim,embed_dim)
Mish	nn.Mish()
Dropout	nn.Dropout(0.01)
FC1	nn.Linear(embed_dim,middle_dim)

表 A: 基于PyTorch的图像映射文本特征网络结构描述

B 实验超参数设置

实验时具体的超参数设置如表B所示。对于MUGE数据集，我们报告验证集上的最佳结果。对于Flickr30K-CN数据集和COCO-CN数据集，我们选择在验证集上表现最佳的检查点，报告其测试集上的结果。

	批量大小	峰值学习率	迭代次数	热身步数	权重衰减
COCO-CN	512	1e-3	10	20	0.1
Flickr30k-CN	576	8e-6	10	70	0.1
MUGE	576	8e-6	10	70	0.1

表 B: 实验超参数设置

C 数据集信息

关于各数据集的训练集、验证集、测试集详细数据如表C所示，各数值代表该数据集中匹配图文对数据的数量。

数据集	训练集	验证集	测试集
COCO-CN	20K	1K	1K
Flickr30k-CN	148K	5K	5K
MUGE	250K	30K	-

表 C: 实验数据集表