面向多模态预训练的子图匹配式对比学习方法研究

陈公冠[1),2)] 刘 慧[1),2)] 李恒泰[1),2)] 郭 强[1),2)] 张彩明[2),3)]

[1)](山东财经大学计算机与人工智能学院 济南 250014)

[2)](山东省数字经济轻量智算与可视化重点实验室 济南 250014)

[3)](山东大学软件学院 济南 250101)

摘 要 通过图像文本对的联合学习,多模态预训练大模型在各种视觉任务中展现出巨大的潜力,比如在高质量数据集匮乏的医学领域。然而,现有的模态匹配式预训练方法通常使用全局匹配的方式,易受到低质量信息的干扰。尽管少量研究开始关注局部匹配,但这些方法仅仅通过简单的池化操作来缩小匹配范围,忽略了跨模态重要对象之间的内在关系以及跨样本对之间同语义表征的获取。鉴于此,本文在多模态大模型的预训练过程中,提出了一种基于图神经网络的消息传递机制,对多模态数据特征进行节点化和子图化,从而将跨模态的匹配方式由全局匹配转变为子图匹配,减少低质量信息的干扰。同时,利用交叉注意力在单一模态内进行子图级别的差异化处理,使其在跨模态学习中建立更细致的关联和语义理解。此外,提出高维空间的样本对聚类方法,以减少多模态大模型对相同语义的无关联错误表达。在涵盖图像分类、病灶区域目标检测和语义分割任务的七个医学图像数据集上进行了大量实验,验证了本文所提出模型的可行性和优越性能。同时在表情识别任务中进行实验,验证了本文模型的泛化性能。

关键词 多模态预训练大模型;局部匹配;子图匹配;无关联错误;聚类

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2025.00893

Research on Subgraph Matching-Based Contrastive Learning for Multi-Modal Pretraining

CHEN Gong-Guan[1),2)] LIU Hui[1),2)] LI Heng-Tai[1),2)] GUO Qiang[1),2)] ZHANG Cai-Ming[2),3)]

[1)](School of Computer and Artificial Intelligence, Shandong University of Finance and Economics, Jinan 250014)

[2)](Shandong Province Key Laboratory of Lightweight Intelligent Computing and Visualization in Digital Economy, Jinan 250014)

[3)](School of Software, Shandong University, Jinan 250101)

Abstract Multi-modal pretrained large-scale models have gained significant attention due to their ability to jointly learn from image-text pairs, unlocking substantial potential for a wide range of visual tasks. In particular, these models have shown great promise in the medical field, where acquiring comprehensive datasets is often challenging due to privacy concerns, data variability, and the time-intensive nature of medical image annotation. By leveraging the rich information embedded in multi-modal data, these models can help bridge the gap in training data and provide more robust solutions to complex medical image analysis tasks, such as diagnosis,

收稿日期:2024-06-28;在线发布日期:2025-02-11。本课题得到国家自然科学基金(62072274,U22A2033)、中央引导地方科技发展项目(YDZX2022009)、山东省泰山学者特聘专家计划(tstp20221137)、济南市人才发展专项资金(202333037)资助。**陈公冠**,博士研究生,主要研究领域为图像处理、机器学习。E-mail:cgg970411@163.com。**刘 慧**(通信作者),博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究方向为医学图像处理、数据挖掘与机器学习。E-mail:liuh_lh@sdufe.edu.cn。**李恒泰**,硕士研究生,主要研究方向为图像处理和机器学习。**郭 强**,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究方向为计算机视觉、数据挖掘等。**张彩明**,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究方向为数据挖掘、计算机图形学、信息可视化和医学图像处理。

prognosis, and treatment planning. Despite the progress in this area, existing pretraining methods based on modality matching often rely on a global matching approach, where features from the different modalities (e. g., images and text) are matched based on their global similarities. However, this approach has limitations. One of the main drawbacks is that global matching is prone to interference from low-quality information, which can significantly degrade the model's performance. In medical datasets, where data quality can be highly variable, the impact of such interference is particularly pronounced. Furthermore, while some recent studies have shifted towards local matching strategies to address these challenges, they have largely employed simple pooling operations to narrow down the scope of matching. Unfortunately, these methods tend to overlook the intricate relationships between important cross-modal objects and fail to fully capture the semantic representations that exist across different sample pairs. In this paper, we propose a novel approach to address these limitations by introducing a message-passing mechanism based on graph neural networks (GNNs) into the pretraining process of multi-modal large-scale models. The primary aim of this mechanism is to transform the traditional global matching approach into a more refined subgraph matching strategy, which is less susceptible to interference from low-quality data. This enables the model to better align and represent the semantic relationships between image and text pairs, resulting in a more accurate and reliable cross-modal understanding. An important aspect of our approach is the incorporation of cross-modal attention mechanisms. This mechanism performs subgraph-level differentiation within each individual modality, allowing the model to capture detailed and modality-specific associations. This enhances the model's ability to learn complex cross-modal dependencies and achieve superior performance in multi-modal learning tasks. Additionally, we introduce a high-dimensional sample pair clustering method as part of our pretraining framework. This method aims to reduce unrelated errors in expressing the same semantics across different sample pairs, further improving the robustness and accuracy of the model. By clustering similar sample pairs together, the model can more effectively align and refine its understanding of cross-modal relationships, resulting in more coherent and semantically accurate representations. To evaluate the effectiveness of our proposed method, we conducted extensive experiments on seven medical image datasets covering a variety of tasks, including image classification, object detection, and semantic segmentation. The results demonstrated that our model outperforms existing methods in terms of accuracy and robustness, highlighting its ability to better handle low-quality information and improve the reliability of multi-modal learning. Furthermore, we also tested our model on a facial expression recognition task, which is a non-medical application, to verify its generalization ability. The experimental results confirmed that our model can effectively transfer across different domains and maintain high performance even in tasks outside the medical field.

Keywords multi-modal pretrained large-scale models; local matching; subgraph matching; unrelated errors; clustering

1 引 言

近年来，自注意力机制的提出解决了卷积难以实现并行化的问题，极大地推动了大规模预训练模型的发展。BERT、GPT等大语言模型的提出在自然语言处理任务中展现出卓越性能。随着研究的进一步深入，大语言模型逐步向多模态大模型转变。最近的研究表明，构建领域内的预训练多模态大模型能够综合不同类型的数据，提供更全面、准确的特

征表示[1-2]，并在各种单一模态的下游任务中展现出巨大的潜力，因此受到广泛关注。

多模态大模型的预训练过程会受到多个因素的影响，其中包括模态间的对齐和数据集规模等。多模态数据通常涉及不同模态之间的相关性，例如图像中的对象与对应的文本描述之间的关联。准确的模态对齐有助于模型学习到更准确、一致的特征表示。然而，模态之间的对齐并非总是直观简单的，特别是在数据集中存在噪声以及模态之间具有复杂关系的情况下。因此，如何设计更好的对齐策略仍然是一个具有挑战性的任务。其次，多模态数据集的获取和标注相对复杂且成本高昂。例如，在医学领域中，由于医学数据的获取困难性和隐私保护的要求，该领域的多模态数据集规模相对有限。因此，如何在有限的数据集上使用更有效的预训练方法来提高多模态模型的能力，成为一个亟需解决的难题。针对上述问题，在预训练过程中提高多模态模型的性能和适用性成为越来越多研究者的努力方向[3]。

在多模态预训练大模型的研究中，CLIP[4]利用对比学习的方式将两种模态的模型整合为可以处理多模态数据的模型，成为最受欢迎的框架结构。随后的各种工作也大多基于 CLIP 的风格。但在医学领域，在更早的时间就推出了与之类似的 ConVIRT[5]模型。随后诸如 MGCA[6]、MRM[7]和 MLIP[8]等相关模型的出现推动了这一研究的发展。这些工作均采用 ConVIRT 或 CLIP 的风格，在对比学习中使用全局匹配的方式实现多模态间的信息对齐。然而，当重要信息只占图像或文本的一小部分时（例如医学图像），全局匹配的方式会受到过多低质量信息（无用的图像背景、文本中的连接词等）的干扰，导致模型难以捕获细微的医学信息。

为了实现细粒度的对齐，最近的研究开始使用局部匹配的方式，即只针对不同模态的重点区域进行匹配。MedCLIP[9]将图像文本对解耦后进行对比学习，同时引入外部医学知识从而减少假阴性。GLoRIA[10]通过对比配对报告中的图像子区域和单词来学习全局和局部表示。但这些方法只是简单地将重点区域池化连接在一起，忽略了这些区域内重点对象之间的内在关联。简单来讲，对重点区域进行模态间的匹配时，这些方法只是将全局匹配的范围进行了缩小，而没有考虑重要对象之间的相互对应，这在一定程度上会影响模型理解不同模态中重点语义以及重点对象之间的差异，从而形成错误的认识。

图神经网络（Graph Neural Network，GNN）由于其依靠节点之间的关系进行消息传递的机制，可以充分利用多模态数据之间的关联性，从而实现不同模态的信息传递和整合，得到更全面和一致的多模态表示，受到一些研究者的青睐。HetMed[11]利用多层网络结构，将患者的多种非图像特征结合起来，以系统化的方式捕捉患者之间的复杂关系。MMC[12]将表示学习和多模态聚类看作是一个整体而不是两个独立的问题，通过利用伪标签设计了对比损失，探索了跨模态的一致性。但现有的研究还停留在使用图神经网络进行特定下游任务的模块设计，尚缺少有效的途径让图神经网络结合多模态模型在预训练中更好地进行模态间的匹配。

为了解决上述问题，本文构建了图1所示的多模态大模型。首先，为了降低全局匹配带来的低质量信息污染，在预训练阶段，将单模态特征节点化，并通过相似度计算得到各个节点之间的连接关系。同时基于图卷积网络设计节点和边权重的更新条件，实现重要对象和重要关系的建模，并以此为基础为每个关键特征点构造以自身为中心的子图，从而实现跨模态的局部匹配。其次，为了解决大模型在预训练过程中难以捕获重要对象内部关联的问题，本文引入交叉注意力机制对单一模态内部的各个重要对象进行差异化处理，从而促使模型在局部匹配过程中能够正确表达关键特征的语义信息。最后，为了避免模型对不同样本对之间相同语义的表达产生分歧[13]，进而出现假阴性的情况。本文基于图卷积网络设计了内部与外部学习核函数，分别实现同一样本对的高维度融合以及不同样本对之间的高维空间聚类分析，使得模型更加贴近人类对于现实世界的理解方式。

具体而言，本文的主要贡献如下：

（1）提出了子图匹配策略，将全局匹配替换为重要对象之间的匹配，降低低质量信息的干扰，为多维度信息的学习提供更细粒度的监督。

（2）提出了重要对象之间的内部关联策略，实现了图像文本间重要对象的软匹配以及同一模态内部重要对象的差异化处理。

（3）提出了多模态样本对聚类策略，实现了不同样本对之间对于相同语义的统一表达，增强了模型对于真实世界的理解能力。

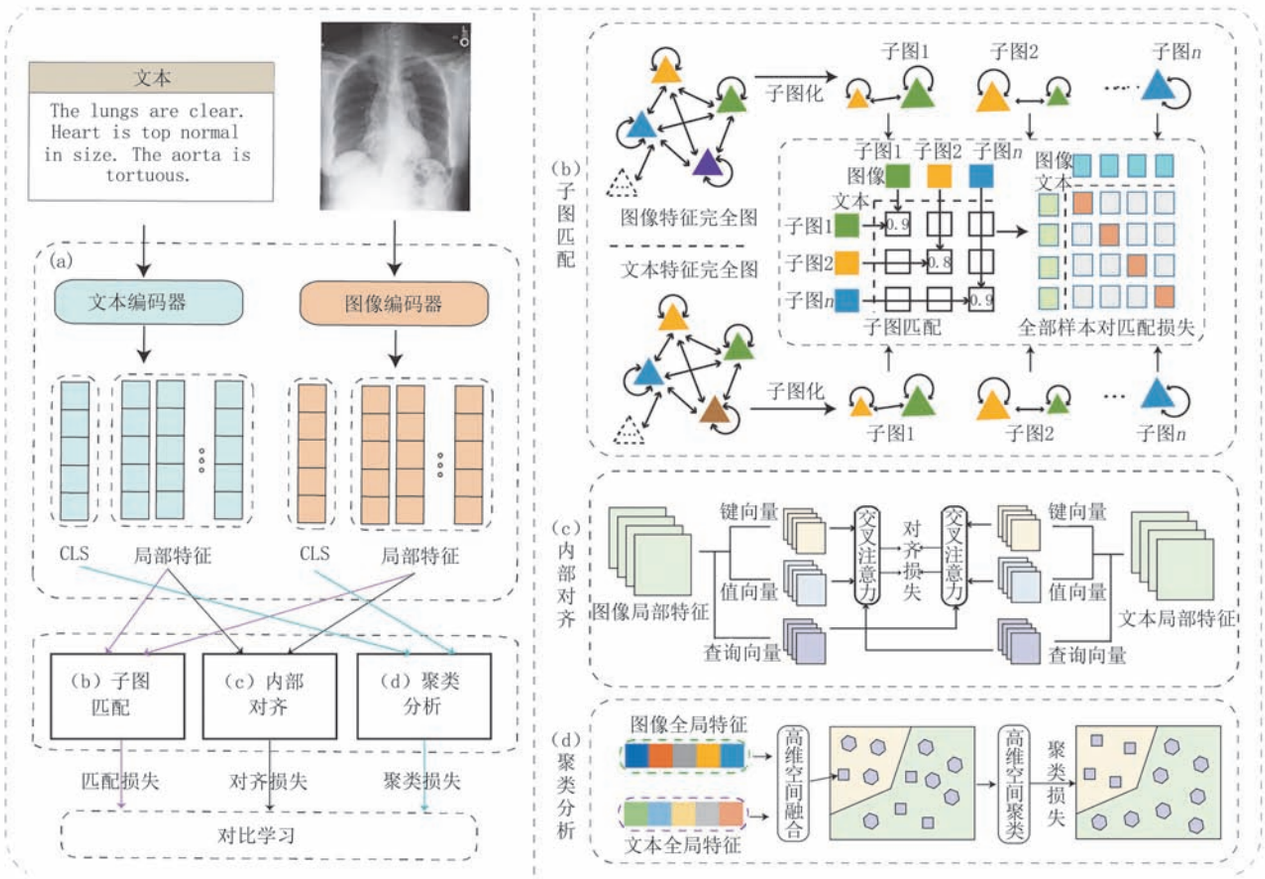


图1 模型的整体架构(基于对不同模态数据的(a)编码和特征提取,利用(b)和(c)分别实现不同模态的子图化匹配以及同一样本对内部的对齐,并利用(d)完成全局特征 CLS 的高维空间聚类,最终(b)中的匹配损失、(c)中的对齐损失以及(d)中的聚类损失共同参与对比学习。)

2 相关工作

2.1 多模态预训练大模型

随着深度学习的快速发展,多模态预训练大模型在计算机视觉领域成为热门研究方向。这些模型能够同时处理多个模态,例如文本、图像和语音等,通过学习不同模态之间的关联性,不仅实现了多模态任务的联合推理[14-16],同时增强了模型在单一模态任务的表征学习能力。CLIP 使用大规模的图像文本对进行预训练,并通过最大化匹配图像和文本的相似性来训练模型。MGCA 通过利用医学图像和放射学报告之间自然展现的语义对应关系,在三个不同层次上进行泛化的医学视觉表示学习。陈波等[17]和陈晓飞等[18]将所有包含相似语义的图像和文本报告视为正匹配。SENSE[19]通过 K-means 对患者数据进行聚类,获取 K 个聚类中心,将样本与其对应的聚类中心视为正对,与其他聚类中心视为负对。X-REM[20]利用预训练的多模态编码器提取图像和文本的嵌入特征,并通过匹配任务中的正类 logit 值作为跨模态对齐的相似性评分。MedM2G[21]通过 InfoNCE 对比损失,以文本特征为引导,将其与其他多种模态对齐到同一空间,为降低资源消耗,采取了两两对齐的策略。但上述方法使用全局匹配的方式完成对比学习,当重要的语义信息只占图像或文本的一小部分时,全局匹配的性能就会受到限制。

近期局部匹配的相关研究相继出现。Med-ST[22]设计了一种模态加权局部对齐方法,针对文本标记与图像空间区域之间的关系,通过 softmax 对文本标记中的每一项与图像块进行余弦相似度评分,从而实现局部匹配。MAVL[23]使用大型语言模型和先验知识,将文本分解为一组共享的视觉特征,并在双头 Transformer 框架下根据这些共享特征直接在图像中进行检索。GLoRIA 将重要的图像子区域与每个成像报告词自动对齐,然后将所有的局部匹配求和以表示整个细粒度对齐。Muller 等[24]提

出将 ResNet[25]最后一层卷积层中每个像素的表示作为图像子区域的局部图像表示。陈志宏等[26]通过交叉注意模块直观地识别出单词对应的最相关图像区域。虽然上述研究推进了局部匹配的进展，但他们均未考虑到重要对象在重点区域中的位置并忽略了重要对象之间的内部关联。本文利用图卷积网络和交叉注意力机制分别实现重点对象之间的跨模态匹配以及重点对象之间的语义差异化处理。

2.2　图神经网络和大模型的融合

图神经网络和大模型的融合已成为近年来机器学习和图数据分析领域的重要研究方向，旨在将自然语言、图像等多种模态信息重构为图结构数据进行处理，以提高模型的表达能力和泛化性能。相关研究主要分为两个方向：(1) 将图卷积网络与大模型进行联合训练，即将图卷积网络嵌入到大模型之中，成为大模型的一部分。CPT-HG[27]提出了两个预训练任务。关系级预训练任务对构成异构图异构基础的关系语义进行编码，而子图级预训练任务对高阶语义上下文进行编码。SGL-PT[28]结合了强大的通用预训练任务，利用了生成和对比方法的优势。PT-HGNN[29]同时考虑节点级和模式级的预训练任务。节点级预训练任务利用节点关系促使 GNN 捕获异构语义属性，而模式级预训练任务利用网络模式促使 GNN 捕获异构结构属性。但这些方法往往受到 GNN 难以深层次建模的影响，导致模型的泛化能力较弱。(2) 将大模型的嵌入作为 GNN 的输入，借助于 GNN 的特性，帮助大模型实现更精确的特征传播和聚合。SimTeG[30]在 TAG(文本属性图)的文本数据集上执行 LLM 的参数高效微调。GraphSAGE[31]提出了一种使用归纳学习为大型图中的节点生成嵌入的方法。Graph-BERT[32]使用基于亲密度和基于跳数的相对位置嵌入来编码子图中的节点位置。Graphormer[33]引入了一种空间编码方法来表示结构节点关系。然而，以上方法往往只针对特定的任务有效。本文借助于 GNN 强大的信息传递机制，将关键关系和关键对象融入到边和节点中。

3　方　法

本节详细介绍了如何将 GCN 与多模态模型相结合，以在重要子图的建模、匹配和聚类中发挥作用，从而得到更加稳健和高泛化性的预训练模型。同时在该节介绍了本文模型的整体架构。如图1所示，模型包括四个重要组成部分：多模态表征提取、局部匹配、跨模态细粒度对齐以及多模态样本对聚类。接下来，将对这四部分进行详细说明。

3.1　多模态表征提取

多模态表征提取是模型中的首要步骤，遵循 CLIP 的设计原则，不同模态分别具有对应的特征提取器，从而获取模态间相互独立的特征向量。在这个过程中，每个模态经过其特征提取器后，可以得到两种不同层次的特征表示。首先，输出结果的第一列是 CLS 令牌，为包含了输入图像或文本整体信息的全局特征矩阵。其次，输出结果的其他列为局部特征矩阵，包含了输入图像或文本的细节信息。图1中的(a)展示了这个过程。

具体来讲，针对图像和文本两种模态的预训练过程，模型的输入为 S 个多模态样本对 $Inp=\{(x_{v,1},x_{t,1}),(x_{v,2},x_{t,2}),\cdots,(x_{v,S},x_{t,S})\}$，其中 $x_{v,i}$ 表示第 i 个图像输入，$x_{t,i}$ 表示其对应的文本输入。随后利用图像编码器 f_v(例如 Vit[34])和文本编码器 f_t(例如 Bert[35])对输入进行特征提取从而获得特征 $F=\{(v_1,t_1),(v_2,t_2),\cdots,(v_S,t_S)\}$，其中 $v_s=f_v(x_{v,s})$，$t_s=f_t(x_{t,s})$。

图像编码器 f_v 采用 Vit 及其变体作为特征提取器，其输出的全局特征 CLS 令牌记为 $\boldsymbol{I}_*\in\mathbb{R}^D$，局部特征记为 $\boldsymbol{P}_i=\{\boldsymbol{P}_i^1,\boldsymbol{P}_i^2,\cdots,\boldsymbol{P}_i^M\}\in\mathbb{R}^{M\times D}$，其中 M 表示 patch 块的数量。此时 $v_s=[\boldsymbol{I}_*^s,\boldsymbol{P}_i^s]$。文本编码器 f_t 采用 Bert 及其变体模型作为特征提取器，同理将其输出拆分为全局特征 $\boldsymbol{T}_*\in\mathbb{R}^D$ 和局部特征 $\boldsymbol{P}_t=\{\boldsymbol{P}_i^1,\boldsymbol{P}_i^2,\cdots,\boldsymbol{P}_i^N\}\in\mathbb{R}^{N\times D}$，其中 N 为输入文本对应的 patch 块数量。此时 $t_s=[\boldsymbol{T}_*^s,\boldsymbol{P}_i^s]$。

尽管单一模态的编码器(例如 Vit 和 Bert)在各自模态中具有强大的特征提取能力，但独立运行的方式无法充分利用不同模态之间的相关性和互补性。本文的目的就是实现不同模态间的联邦学习，从而提升模型对单一模态任务的处理能力。

3.2　局部匹配

3.2.1　重要对象建模

对于医学多模态数据的对齐，现有方法通常使用全局匹配的方式进行预训练，但其容易受到低质量信息的干扰。为了解决这个问题，本文采用图神经网络来建模重要对象和重要关系。首先将单一模态的特征图转换为完全图并利用图神经网络的信息传递和聚合特性，增强重要对象的表示。其次通过更新边权重和去除低质量边和节点，得到多个包含重要对象的子图。最后在不同模态之间构建跨模态

图，并利用图神经网络来捕捉复杂的跨模态关联和依赖关系。具体来讲，如图2所示，将图像编码器f_v的输出$\boldsymbol{P}_i=\{\boldsymbol{P}_i^1,\boldsymbol{P}_i^2,\cdots,\boldsymbol{P}_i^M\}\in\mathbb{R}^{M\times D}$构建为完全图$G=<\boldsymbol{P}_i,\boldsymbol{A}>$，其中$\boldsymbol{A}\in\mathbb{R}^{M\times M}$表示该完全图的邻接矩阵，计算过程为$A_{mn}=\boldsymbol{p}_i^m(\boldsymbol{p}_i^n)^T$。利用图卷积网络进行浅层次的重要节点增强得到$\boldsymbol{P}_i'=GCN(\boldsymbol{P}_i,\boldsymbol{A})$，其中$GCN$表示图卷积操作。

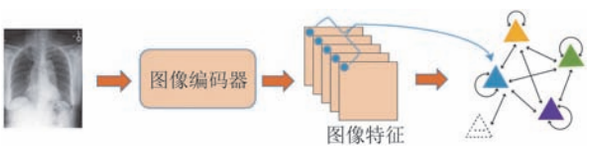


图2 图像特征构图过程

随后进行边权重$\boldsymbol{A}$的更新，为此本文制定了三个条件作为边权重更新的规则：

(1)该边连接的两个节点抗干扰能力强，RKD[36]证明了节点的信息熵对于噪声的敏感程度决定了该节点包含信息的重要程度，为此本文通过加入随机噪声检测节点的重要性：

$$\rho_n(\boldsymbol{p}_i^m)=|\mathrm{E}(G(\boldsymbol{p}_i^{m\prime},\boldsymbol{A}))-\mathrm{E}(G(\boldsymbol{p}_i^m,\boldsymbol{A}))| \quad (1)$$

其中，$\boldsymbol{p}_i^{m\prime}$为$\boldsymbol{p}_i^m$加入随机噪声后的节点表示，$E$表示信息熵的计算，$G(\boldsymbol{p}_i^m,\boldsymbol{A})=\boldsymbol{A}(\boldsymbol{p}_i^m)^T$表示当前节点在全图中发挥的作用，$\rho_n$表示该节点重要性计算结果，数值越小该节点的抗干扰能力越强。

(2)该边本身的抗干扰能力强，通过观察随机噪声对该边连接的两个节点之间交叉熵的影响程度来决定该边的重要程度：

$$\begin{aligned}\rho_a(\boldsymbol{p}_i^m,\boldsymbol{p}_i^n)=&|H(\boldsymbol{p}_i^m,\boldsymbol{p}_i^n)-H(\boldsymbol{p}_i^m,\boldsymbol{p}_i^{n\prime})|\\&+|H(\boldsymbol{p}_i^m,\boldsymbol{p}_i^n)-H(\boldsymbol{p}_i^{m\prime},\boldsymbol{p}_i^n)|, \quad (2)\\H(\boldsymbol{p}_i^m,\boldsymbol{p}_i^n)=&CE(G(\boldsymbol{p}_i^m,\boldsymbol{A}),G(\boldsymbol{p}_i^n,\boldsymbol{A}))\end{aligned}$$

其中，CE表示交叉熵计算，ρ_a表示该边重要性计算结果，数值越小该边的抗干扰能力越强。

(3)该边连接的两个节点距离近，通过计算两个节点间的欧式距离来判断两个节点之间的距离，具体计算过程如下：

$$d(\boldsymbol{p}_i^m,\boldsymbol{p}_i^n)=\sqrt{\Sigma(p_{ic}^m-p_{ic}^n)^2}\ (c=1,2,\cdots,D) \quad (3)$$

其中p_{ic}^m表示第m个节点向量的第c个数，d表示两个节点间的距离，D表示每个节点的长度，结果越小表明两个节点距离越近。

条件1用来区分高低质量节点，条件2用来区分高低质量关系，条件3用来区分该边连接的两个节点是否从属同一个重要对象。邻接矩阵$\boldsymbol{A}$的更新过程如下：$\boldsymbol{A}'=(\boldsymbol{A}+(1-(\rho_n+\rho_a+d)/3)\boldsymbol{A})/2$，利用更新后的邻接矩阵$\boldsymbol{A}'$与$\boldsymbol{P}_i'$进行图卷积操作实现对重要对象的深层次建模$\boldsymbol{P}_i''=GCN(\boldsymbol{P}_i',\boldsymbol{A}')$。$\boldsymbol{P}_i''$内的每个节点在重点关注与自身关联紧密的节点外，也在一定程度上保留了与其他节点的微弱联系，增加了编码器在局部特征提取时的视野。文本采用同样的处理方式得到$\boldsymbol{P}_t''$。

3.2.2 重要关系建模及子图匹配

对于$\boldsymbol{A}'$设置一定的阈值去除低质量边和低质量节点。阈值的设置采用百分比的形式，消融实验中对该参数进行了六种不同情况的对比。此时剩余节点均能形成一个以自身为中心的子图，该子图的中心节点为某一重要对象的其中一个节点，其余节点为该重要对象的剩余节点。此时可以将单一模态的特征图转换为多个包含重要对象的子图，完成重要关系的建模。同时本文构建跨模态图实现跨模态的信息交互。

具体来讲，构造跨模态图的条件包含两条：(1)节点为子图经过图卷积后的中心节点。(2)边的连接只发生在不同模态之间，即每条边连接的两个节点须来自不同的模态。此时会发生两种情况：(1)边两端的节点指向同一重要对象，表达意义相同。(2)边两端的节点指向不同的重要对象，表达意义不同。要完成模态间重要对象的匹配，需要保留第一种情况，去除第二种情况。同时同一对象的不同模态表达可以看作由同一对象产生的不同分布，这满足KL散度的计算条件。因此，本文基于KL散度设计匹配公式将第二种情况去除：

$$\begin{aligned}&T_{ik}^*=Trans(T_{ik}),\\&KL(I_i\rightarrow T_i)=\ln\sum_{e\in IM(I_i)}\coprod_{\max}^{k\in IM(T_i)}\exp\left(\frac{I_{ie}T_{ik}^*}{\sqrt{d}}\right)\\&\qquad-\frac{1}{M}\sum_{e\in IM(I_i)}\coprod_{\max}^{k\in IM(T_i)}\frac{I_{ie}T_{ik}^*}{\sqrt{d}}\end{aligned} \quad (4)$$

其中，$e\in IM(I_i)$表示e为跨模态图中属于图像的节点下标，I_i和T_i分别表示跨模态图中的图像和文本对应的节点，$T_{ik}^*=TransP(T_{ik})$表示T_{ik}^*为T_{ik}的转置，M表示由图像产生的节点数量，d为节点向量的长度。同理，文本向图像的匹配记为$KL(I_i\leftarrow T_i)$。

多模态大模型的预训练需要大量数据的支撑，尽管医学领域的数据集相较于一般图像文本数据集要少几个数量级，但仍难以达到有监督的要求，对比学习是预训练大模型最常用的方式。本文设计了对比损失函数实现局部匹配的训练，计算过程如下：

$$l_i^{I2T}=-\log\frac{\exp(KL(I_i\rightarrow T_i)/\tau_1)}{\sum_{k=1}^{B}\exp(KL(I_i\rightarrow T_k)/\tau_1)},\quad l_i^{T2I}=-\log\frac{\exp(KL(I_i\leftarrow T_i)/\tau_1)}{\sum_{k=1}^{B}\exp(KL(I_i\leftarrow T_k)/\tau_1)}\tag{5}$$

其中，l_i^{I2T}和l_i^{T2I}分别表示图像向文本的匹配损失和文本向图像的匹配损失，τ_1表示温度系数，B表示batch_size的大小。因此对于局部匹配总的损失函数为

$$L_1=\frac{1}{2S}\sum_{i=1}^{S}(l_i^{I2T}+l_i^{T2I})\tag{6}$$

其中，S表示整体样本的数量。

3.3　跨模态细粒度对齐

在子图匹配过程中，针对单一模态构建子图时，模型仍主要依赖于该模态自身的信息，导致在匹配过程中容易缺失跨模态的交互信息，进而无法有效地辅助校正各自模态。同时，子图匹配通常聚焦于较大区域的整体相似性或结构对齐，这在识别全局模式或大尺度关系方面表现突出。然而，子图内部包含着丰富的细节信息，这些细节在对齐过程中至关重要。因此，增加patch级别的交互能够帮助模型进一步解析子图内部的微小差异，使得模型不仅能够处理较大的子图区域，还能在更精细的局部区域内进行特征匹配，从而更准确地捕捉不同子图之间的细节关联。

为此，本文考虑采用一种能够捕获特征图中精细化信息的方法，同时实现跨模态的交互。交叉注意力机制能够通过计算模态间所有patch块的相似度权重，使得模型在进行特征匹配时，可以灵活地调整不同模态之间细粒度特征的连接程度。这种机制不仅提高了模型的鲁棒性，增强了对细节的捕捉能力，也提升了整体对齐的精度。因此，本文引入交叉注意力机制，并设计了相应的对比学习损失函数。

具体来讲，细粒度对齐策略分别以图像和文本为中心进行交叉注意力机制的计算。当以图像为中心时，对于第m个图像的第n个重要patch块，通过线性变换得到查询向量，将文本内的所有重要patch块通过线性变换分别得到键向量和值向量，交叉注意力机制完成第m个图像的第n个重要patch块与文本的所有重要patch块的嵌入，具体计算过程如下：

$$(\boldsymbol{W}_k\boldsymbol{T}_m^l)^*=TransP(\boldsymbol{W}_k\boldsymbol{T}_m^l),\quad Att(I_m^n,\boldsymbol{T})=\sum_{l=0}^{L}\left(\frac{(\boldsymbol{W}_q\boldsymbol{I}_m^n)(\boldsymbol{W}_k\boldsymbol{T}_m^l)^*(\boldsymbol{W}_v\boldsymbol{T}_m^l)}{\sqrt{d}}\right)\tag{7}$$

其中，$(\boldsymbol{W}_kT_m^l)^*$为$\boldsymbol{W}_kT_m^l$的转置，$\boldsymbol{W}_q$、$\boldsymbol{W}_k$和$\boldsymbol{W}_v$表示可学习的参数矩阵，$d$表示patch块向量的长度，$L$表示文本patch块的数量，$Att$表示交叉注意力的计算结果。对文本采用同样的处理过程。

在对比学习中，设计对齐损失函数将不同模态间具有表达意义相同的patch块靠近，将表达意义不同的patch块远离。损失函数的计算过程如下：

$$\boldsymbol{Att}_I^*=TransP(\boldsymbol{Att}(\boldsymbol{I}_s^m,\boldsymbol{T})),\quad \boldsymbol{Att}_T^*=TransP(Att(\boldsymbol{T}_s^m,\boldsymbol{I})),$$
$$L_2=-\frac{1}{SM}\sum_{s=0}^{S}\sum_{m=0}^{M}\frac{\exp(\boldsymbol{I}_s^m\boldsymbol{Att}_I^*)}{\sum_{n=0}^{N}\exp(\boldsymbol{I}_s^n\boldsymbol{Att}_I^*)}-\frac{1}{SN}\sum_{s=0}^{S}\sum_{n=0}^{N}\frac{\exp(\boldsymbol{T}_s^n\boldsymbol{Att}_T^*)}{\sum_{m=0}^{M}\exp(\boldsymbol{T}_s^m\boldsymbol{Att}_T^*)}\tag{8}$$

其中，S表示样本对的数量，M表示图像patch块的数量，N表示文本patch块的数量。

3.4　多模态样本对聚类

前面的章节对局部特征进行了细化分析，但对全局特征$\boldsymbol{I}_*$和$\boldsymbol{T}_*$并未添加约束。全局特征包含了样本对整体信息的表达，且不同样本对之间会存在相似的语义。例如在医学多模态数据中，来自不同患者的图像和报告可能具有类似之处。将同一样本对中的图像和文本归为一类，其他样本对划分为另一类，会造成假阴性的产生。为此，本文为每个样本对检索出与之最为相似的K个样本对，使得该样本对与这K个样本对靠近，与其他样本对远离。这一操作避免了具有类似语义信息的样本对在嵌入空间中被推开。

为了实现上述目的，本文基于图卷积网络的思想设计了两个核函数：(1) 内部学习核函数：内部学习核函数用于单一样本对内不同模态之间的操作，旨在将不同模态的全局特征进行融合。通过这种方式，不同模态的信息可以相互影响和补充，从而提高对样本对的整体理解能力。(2) 外部学习核函数：该核函数作用于样本对之间。目的是计算两个样本对在高维空间的距离，并利用该距离进行聚类。在高维空间中的距离可以更准确地衡量样本之间的相似性，实现更好的数据分离效果。

具体来讲，如图3所示，在内部学习中，对于第m个样本对的全局特征$\boldsymbol{I}_*^m$和$\boldsymbol{T}_*^m$，考虑到两者在融合的结果中占据的权重不同，本文设计了如下非对称

核函数进行融合：

$$\begin{aligned}(\boldsymbol{T}_*^m)^* &= TransP(\boldsymbol{T}_*^m),\\ (\boldsymbol{I}_*^m)^* &= TransP(\boldsymbol{I}_*^m),\\ F(\boldsymbol{I}_*^m,\boldsymbol{T}_*^m) &= TransP((\boldsymbol{T}_*^m)^*\boldsymbol{W}\boldsymbol{I}_*^m(\boldsymbol{I}_*^m)^*),\\ \boldsymbol{C}_*^m &= \exp\left(-\frac{\|F(\boldsymbol{I}_*^m,\boldsymbol{T}_*^m)-\boldsymbol{I}_*^m\|^2}{2\sigma^2}\right)\boldsymbol{I}_*^m\\ &\quad+\exp\left(-\frac{\|F(\boldsymbol{I}_*^m,\boldsymbol{T}_*^m)-\boldsymbol{T}_*^m\|^2}{2\beta^2}\right)\boldsymbol{T}_*^m\end{aligned} \tag{9}$$

其中，$\boldsymbol{W}$为可学习的参数矩阵，σ和β分别表示图像和文本的全局特征向量的标准差。

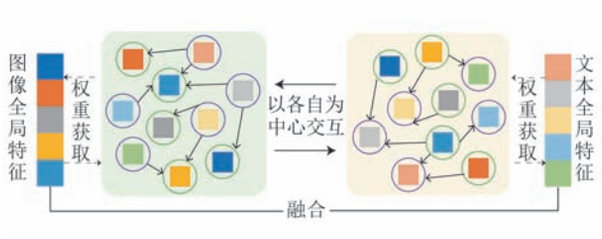


图3 内部学习过程

在外部学习中，对于融合后的两个样本对$\boldsymbol{C}_*^m$和$\boldsymbol{C}_*^n$，本文基于图神经网络的思想设计了如下动态权重核函数将样本对映射到更高的维度空间来进行分类：

$$\begin{aligned}(\boldsymbol{C}_*^m)^* &= TransP(\boldsymbol{C}_*^m),\\ D_{mn} &= \exp\left(-\frac{\left\|\boldsymbol{C}_*^m-\boldsymbol{C}_*^n\right\|^2}{2\alpha^2}\right)(\boldsymbol{C}_*^m W(\boldsymbol{C}_*^n)^*)\end{aligned} \tag{10}$$

其中，α是$TransP(\boldsymbol{C}_*^m)\boldsymbol{C}_*^n$的标准差，$\boldsymbol{W}$是可学习的参数矩阵，$D_{mn}$表示两个样本对间的距离，$\boldsymbol{C}_*^m\boldsymbol{W}(\boldsymbol{C}_*^n)^*$表示基于图卷积网络的动态权重计算。

将样本对$\boldsymbol{C}_*^m$与剩余全部样本对进行上述的距离计算，选取其中的前K个样本对作为$\boldsymbol{C}_*^m$的相似样本对。将$\boldsymbol{C}_*^m$与相似样本对拉近，与其他样本对远离。在对比学习中，对应的损失函数如下：

$$L_3=-\frac{1}{S}\sum_{s=0}^{S}\frac{\sum_{k\in\{K\}}\exp(\boldsymbol{C}_*^s\boldsymbol{TransP}(\boldsymbol{C}_*^k)/\tau_2)}{\sum_{j\in\{S-K\}}\exp(\boldsymbol{C}_*^s\boldsymbol{TransP}(\boldsymbol{C}_*^j)/\tau_2)} \tag{11}$$

其中，S表示样本对的数量，$\{K\}$表示与$\boldsymbol{C}_*^s$距离最近的样本对，$\{S-K\}$表示与$\boldsymbol{C}_*^s$距离不相近的其余样本对。最终，通过联合优化三个损失函数来训练整体模型，鼓励模型学习更加泛化的多模态信息表示。总的损失函数可表示为

$$L=L_1+L_2+L_3 \tag{12}$$

4 实 验

4.1 预训练设置

本文使用MIMIC-CXR 2.0.0[37]数据集的JPG版本对本文模型进行预训练。该数据集是一个大型公开可用的JPG格式胸部X光照片数据集，具有自由文本放射学报告。该数据集包含377 110张JPG格式图像以及与这些图像相关的227 827份自由文本放射学报告的结构化标签。在模型组成方面，文本编码器选用BioClinicalBERT[38]，图像编码器选用Vit[39]的Base版本。由于采用的结构为CLIP形式，对文本编码器和图像编码器的选择并不固定，可以随意搭配各种语言模型和视觉模型作为编码器部分。在预训练过程中，本文使用块A100 40 G显卡进行训练，batch size设置为128，总轮数设置为50，优化器为AdamW，学习率为2e-5，权值衰减为0.05。在超参数设置方面，本文对温度系数进行了多种取值的比较，实验结果如图4所示，通过实验对比最终确定为$\tau_1=0.1$，$\tau_2=0.2$。

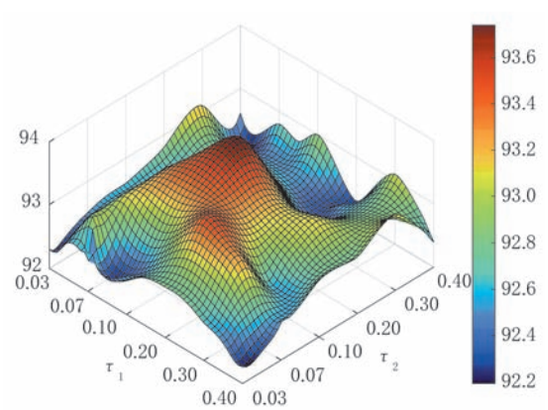


图4 在RSNA数据集上对τ_1和τ_2的不同取值对比

4.2 下游任务设置

本文在以下三个下游任务上进行了实验对比：

(1)医学图像分类：冻结预训练模型参数并在图像编码器上添加一层全连接层完成图像分类任务。包含了三个常用的实验的数据集：CheXpert[40]数据集是一个包含65 240位病人，224 316个胸片图片的大型数据集，数据集中的信息是由标签器Labeler从放射学报告中提取14个观测值中得到，观测值分为正样本、负样本，还有不确定样本。其中验证集由200个胸部X线影像构成，并且标签由3位

专家进行标注。测试集由500个胸部X线影像构成，并且标签由5位专家进行标注。RSNA[41]数据集包含大约29 700张正面胸片。标签分为正常或气胸阳性两个类别。在实验中以70%、15%和15%的比例将数据集分为训练集、验证集和测试集。COVIDx[42]包含三万多张CXR图像数据，这些数据采样自不同国家共16 600个患者，其中含有16 490张COVID-19图像，所有数据被分为三个类别(COVID-19、非COVID-19和正常)。在每个分类数据集上分别使用1%、10%和100%的训练数据来评估我们的模型。其中CheXpert和RSNA采用ROC曲线下面积(AUC)作为评估指标，Covidx采用准确率(ACC)作为评估指标。

(2)病灶区域目标检测：对于目标检测任务，将YOLO[43]的骨干部分替换为预训练好的本文模型，同时冻结该部分参数，借助于YOLO模型完成目标检测任务。进行实验的数据集包含两个：RSNA肺炎数据集包含29 700个样本，目标是对肺炎区域进行检测，在实验过程中将数据集随机分为训练集(16 010个样本)、验证集(5337个样本)和测试集(5337个样本)。Object CXR[44]包含9000张X光成像，目标为异常区域检测，如肺内体外异物的识别。训练集包含6400个样本，验证集包含1600个样本，测试集包含1000个样本。同样通过1%、10%和100%的训练数据对模型进行评估。

(3)医学影像语义分割：对于语义分割任务，将U-Net[45]模型的主干部分替换为预训练好的本文模型，同时冻结该部分参数，借助于U-Net模型完成语义分割任务。数据集包含两个：SIIM[46]数据集包含12 047张胸片，训练接、验证集和测试集分别占比70%、15%和15%。RSNA与目标检测任务的处理方式相同，将目标检测的真实标签转换为掩码，用于语义分割。同样通过1%、10%和100%的训练数据对模型进行评估。

4.3　实验结果

在医学图像分类任务中，将本文模型与ConVIRT、MedCLIP、MGCA-ResNet、MGCA-ViT、RECLF和MLIP进行了对比。这些方法均采用了CLIP形式的对比学习方法。其中ConVIRT作为CLIP的前身研究，与CLIP的处理方式基本相同。RECLF[47]通过关系增强的对比学习框架对局部匹配之间的关系进行建模。MLIP利用领域特定的医学知识作为引导信号，通过图像文本对比学习将语言信息整合到视觉领域中。为了进行公平对比，上述方法的预训练数据集均设置为与本文相同。实验结果如表1所示。从实验结果可以看出，本文模型在所有数据集上均取得了最好的结果。这表明本文模型能够利用图神经网络和动态权重核函数的设计，从医学图像中提取出更丰富和区分度更高的特征表示。相比其他方法，本文模型在医学图像分类任务中具有更强的性能。

表1　医学图像分类任务对比

方法	CheXpert			RSNA			COVIDx		
	1%	10%	100%	1%	10%	100%	1%	10%	100%
ConVIRT[5]	85.9	86.8	87.3	77.4	80.1	81.3	72.5	82.5	92.0
MedCLIP[9]	87.1	87.4	88.1	88.3	89.3	90.1	-	-	-
MGCA-ResNet[6]	87.6	88.0	88.2	88.6	89.1	89.9	72.0	83.5	90.5
MGCA-ViT[6]	88.8	89.1	89.7	89.1	89.9	90.8	74.8	84.8	92.3
RECLF[47]	88.0	89.3	89.6	89.3	89.6	90.9	-	-	-
MLIP[8]	89.0	89.4	90.0	89.3	90.0	90.8	75.3	86.3	92.5
本文	**89.8**	**90.1**	**90.9**	**91.6**	**92.5**	**93.7**	**75.6**	**87.2**	**92.9**

在病灶区域目标检测任务中，将本文模型与ConVIRT、GLoRIA、MGCA、GLoRIA-M、MedKLIP和MLIP进行了对比。其中MedKLIP[48]基于Transformer的融合模型，在图像块级别上实现实体描述与视觉信号的空间对齐。实验结果如表2所示。根据实验结果，本文模型在病灶区域目标检测任务中表现出了卓越的性能，显示出了明显的优势。特别当训练数据占比为1%时，本文模型相较于其他模型有一个较大幅度的提升。可以看出，本文模型利用重要对象匹配的方式代替全局匹配，能够更好地捕捉医学图像中的异常目标特征，并提供更准确的检测结果。

在医学影像语义分割任务中，将本文模型与ConVIRT、GLoRIA、MGCA、GLoRIA-M、

表2 医学目标检测任务对比

方法	RSNA			Object CXR		
	1%	10%	100%	1%	10%	100%
ConVIRT[5]	8.20	15.6	17.9	-	8.60	15.9
GLoRIA[10]	9.80	14.8	18.8	-	10.6	15.6
MGCA[6]	12.9	16.8	24.9	-	12.1	19.2
GLoRIA-M[10]	11.6	16.1	24.8	-	8.90	16.6
MedKLIP[48]	8.90	16.3	24.5	-	7.10	11.6
MLIP[8]	17.2	19.1	25.8	4.6	17.4	20.2
本文	**18.3**	**19.4**	**25.9**	**5.1**	**17.4**	**20.8**

MedKLIP和MLIP进行了对比。实验结果如表3所示。综合来看，本文模型在六种比较情况下共取得了五个第一。医学图像中的边缘信息对于准确的语义分割至关重要，结果表明本文模型能够准确地捕捉医学图像中的复杂结构和病变区域。这使得分割结果更加准确，能够更好地揭示病变的边界和形态特征，同时具备较强的泛化能力。

表3 医学语义分割任务对比

方法	SIIM			RSNA		
	1%	10%	100%	1%	10%	100%
ConVIRT[5]	25.0	43.2	59.9	55.0	67.4	67.5
GLoRIA[10]	35.8	46.9	63.4	59.3	67.5	67.8
MGCA[6]	49.7	59.3	64.2	63.0	68.3	69.8
GLoRIA-M[10]	37.4	57.1	64.0	60.3	68.7	68.3
MedKLIP[48]	50.2	60.8	63.9	66.2	69.4	71.9
MLIP[8]	51.6	60.8	**68.1**	67.7	68.8	73.5
本文	**52.3**	**61.6**	66.9	**68.5**	**69.1**	**73.8**

此外，本文还对图像编码器进行了实验，将Vit-base替换为参数更少的Deit-base。通过本文提出的训练策略对视觉模型进行训练，并将其与MedCLIP和MGCA两种方法进行比较。实验结果如表4所示。实验结果表明，通过本文提出的训练策略得到的视觉模型在分类任务中取得了更加优异的效果。即使使用参数量更小的Deit-base作为图像编码器，本文的训练方法仍能够充分捕捉图像中的重要信息，从而提升了模型的性能。与MedCLIP和MGCA方法相比，本文的训练方法能够更好地利用输入图像的特征，实现更准确地分类。这一结果进一步验证了本文提出的训练方法的有效性。

4.4 泛化性实验

进一步地，为了验证所提出的子图匹配式方法的泛化性能，我们对两种基于CLIP的表情识别方法EmoCLIP[49]和DFER-CLIP[50]的对齐机制进行了替换实验。具体而言，将其原有对齐方式更改为子图匹配式方法，并与原始模型进行了实验对比。实验使用了动态面部表情任务中常用的两个数据集DFEW[51]和FERV39k[52]。其中DFEW数据集包含从全球1500多部电影中收集的11 697个视频片段。在专家的指导下，十名注释员将这些视频分为七种基本面部表情(快乐、悲伤、中性、愤怒、惊讶、厌恶和恐惧)。FERV39k数据集包含38 935个视频片段，是目前最大的真实场景下的DFER数据集，这些视频片段来自四个场景，包括犯罪、日常生活、演讲和战争等22个细分场景，并由30位注释者标注了基本面部表情。使用的指标为UAR和WAR。其中UAR是未加权平均召回率，主要用于多分类问题。它表示各类别的平均召回率，在计算时不考虑各类别的样本数量。WAR是加权准确率，是一种考虑每个类别样本数量的准确率计算方式。实验结果如表5所示。

表4 不同视觉编码器对比

方法	CheXpert		RSNA	
	10%	100%	10%	100%
MedCLIP-Vit	87.2	87.6	88.7	89.2
MGCA-Vit	89.1	89.7	89.9	90.8
本文-Vit	90.1	90.9	92.5	93.7
MedCLIP-Deit	84.3	85.2	84.6	86.3
MGCA-Deit	83.7	86.6	87.2	88.7
本文-Deit	88.6	90.4	90.6	91.7

表5 子图匹配泛化性实验

方法	DFEW		FERV39k	
	UAR	WAR	UAR	WAR
EmoCLIP	58.04	62.12	31.41	36.18
EmoCLIP+子图匹配	58.65	62.37	31.86	37.04
DFER-CLIP	59.61	71.25	41.27	51.65
DFER+子图匹配	59.72	70.58	41.83	51.66

从实验结果中可以看出，替换为子图匹配式方法后，模型在表情识别任务中的性能获得了一定提升。具体而言，对于EmoCLIP，替换后的模型在所有评估指标上均较原方法有所提升；而对于DFER-CLIP，替换后在DFEW数据集的WAR指标上有一定程度的下降，但其余指标均有提升。子图匹配式方法在这些实验中的表现优异，主要归功于其局部对齐机制，有助于模型更细致地捕捉面部表情的细节特征，同时提升了对跨模态噪声的鲁棒性。通过模态间更精确的区域级对齐，子图匹配式方法在

跨模态信息传递过程中能够捕捉到更多精细特征。这一实验结果表明，子图匹配式方法不仅在医学领域的多模态数据融合任务中具有显著优势，还在表情识别等视觉任务中展现了良好的泛化能力。

4.5　消融实验

为了分析本文模型中的各个模块在预训练时发挥的作用，这一章节对局部匹配、跨模态细粒度信息对齐以及多模态样本对聚类分别进行了消融实验。

(1)局部匹配

为了验证局部匹配的作用，本文分别对该模块进行了定量和定性分析。在定量分析中，本文针对局部匹配模块进行了三种设置：不使用匹配、使用全局匹配和使用局部匹配。针对这三种情况进行了实验，并将结果列在表6中。实验结果显示，如果丢弃局部匹配模块，分类准确率会显著下降。无论是使用全局匹配还是局部匹配，其目的都是为了提高模型对图像和文本之间关联性的理解能力。这一能力的丢失会造成多模态信息的交互困难，导致准确率的下降。对比全局匹配和局部匹配的实验结果，可以观察到局部匹配在分类准确率上表现更好。这表明局部匹配能够更好地捕捉图像和文本之间的细粒度语义关系，从而提高模型在细节信息和局部关联性方面的性能。

表6　局部匹配消融实验

方法	CheXpert			RSNA		
	1%	10%	100%	1%	10%	100%
无匹配	86.4	87.1	87.8	88.2	89.5	90.7
全局匹配	88.7	89.3	89.6	89.9	90.8	92.4
子图匹配	89.8	90.1	90.9	91.6	92.5	93.7

本文还对MGCA和Med-ST两种与本文方法结构相似的模型进行了分析。同时为了验证本文提出的子图匹配式局部对齐方案的有效性，将MGCA和Med-ST中原有的对齐方案替换为子图匹配式局部对齐，并与原始方法进行了对比。实验结果如表7所示。结果表明，当对齐方案进行替换后，模型性能均得到了一定程度的提升。这一改动进一步验证了子图匹配式局部对齐方法在特征对齐上的有效性。

在定性方面，本文绘制了三种情况下的注意力图，实验结果如图5所示。可以看出在没有匹配的情况下，图像编码器因为在其他领域的数据集上进行过预训练，使得其仍能够大致辨别重点区域(例如有小部分红色区域和绿色区域集中在肺部)，但对重点对象的敏感度较低，甚至有错误的情况。使用全局匹配会促使模型对重点区域的进一步关注，但受到低质量信息的干扰，使得模型对重点对象的关注产生一定的偏差。局部匹配能够更加关注模态内的重要对象，降低背景等低质量信息的干扰。

表7　子图匹配消融实验

方法	RSNA		COVIDx	
	1%	10%	1%	10%
MGCA	89.1	89.9	74.8	84.8
MGCA＋子图匹配	89.8	90.4	74.9	85.7
Med-ST	90.8	91.4	71.2	87.7
Med-ST＋子图匹配	91.2	91.9	72.2	87.7
本文方法	91.6	92.5	75.6	87.2

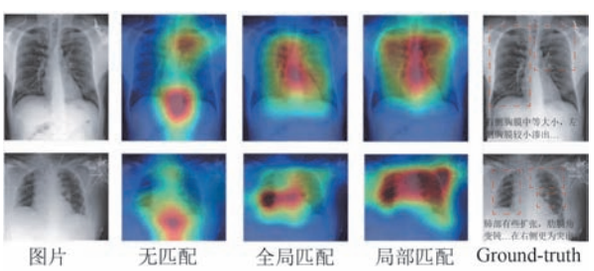


图5　不同匹配方式的注意力图

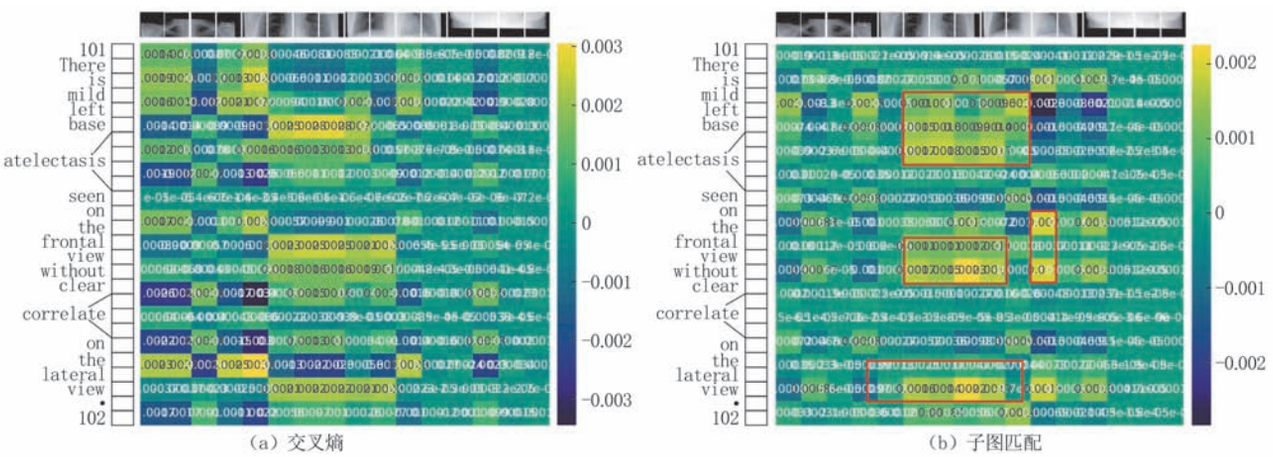
当针对3.2.1节中对重要关系建模时使用的三种边权重更新规则进行消融实验时，本文采用了不同的组合方式来评估它们对语义分割任务的影响。下面是实验的详细描述：首先构建一个基准模型，即没有应用任何边权重更新规则的模型。然后独立地应用三种边权重更新规则来对重要关系进行建模。最后将这三种规则进行两两组合，得到三个组合模型。表8展示了不同模型在语义分割任务中权重更新规则的消融实验结果。实验结果表明，每种边权重更新规则都对语义分割任务产生了积极的影响。相比于基准模型，应用任何一种规则都能够提升性能。这表明每种规则都能够有效地捕捉重要关系，并引导模型更好地理解多模态数据之间的依赖关系。此外，将两种规则进行组合时，性能进一步提升。这表明这些规则在协同作用方面具有互补性，从而提供更全面的建模能力。通过消融实验的结果，我们进一步验证了这些边权重更新规则在建模重要关系方面的有性。它们能够帮助模型更好地理解多模态数据中的关联和依赖关系，从而提升模型的性能。

同时，为了验证信息对齐的精度对下游任务的影响，本文设计了两个预训练模型：一个模型直接使用交叉熵进行局部信息对齐，另一个模型则采用我

表8 权重更新规则的消融实验

方法	SIIM			RSNA		
	1%	10%	100%	1%	10%	100%
无规则	49.3	58.9	64.2	64.8	66.3	70.2
规则1	50.7	60.1	64.9	66.5	67.5	71.3
规则2	51.1	60.1	64.8	66.8	67.6	71.8
规则3	50.4	59.8	65.0	66.9	67.5	71.4
规则1+2	51.7	61.1	65.8	67.8	68.6	73.1
规则1+3	51.5	60.9	65.7	67.9	68.8	72.8
规则2+3	51.2	60.7	65.2	67.4	68.1	72.2

们提出的子图匹配式局部信息对齐。此实验旨在验证两方面内容:第一,本文提出的模型在下游任务中指标提升是否源于更精准的局部信息对齐;第二,信息对齐的偏差对下游任务性能的影响是否显著。为此,本文对比了两个模型的图像编码器和文本编码器的输出特征,并对图像和文本之间的注意力矩阵进行了可视化。同时,为了使结果更加直观易于理解,图中也对文本编码结果中每个单词占用的位置进行了展示。实验结果如图6所示。



(a) 交叉熵 (b) 子图匹配

图6 不同匹配方式的注意力矩阵

结果显示,采用子图匹配式局部信息对齐的模型,注意力矩阵中的文本关键词与图像关键区域的对齐更加准确。例如右图中画出的四个红框,均属于句子中的重要信息与图像相匹配的关键区域。相较于左图,增加子图匹配后,模型对这四个区域的关注程度表现出明显的提升。相反,仅依赖交叉熵损失进行局部信息对齐的模型,表现出较低的对齐精度。同时,在下游任务中,我们使用 COVIDx 进行了定量分析。仅依赖交叉熵损失在 COVIDx 数据集上的准确率为89.5%,而使用子图匹配式局部信息对齐,其准确率为92.9%,这一结果进一步验证了子图匹配式局部信息对齐的有效性。

在子图构建过程中,本文采用了百分比形式的阈值来对子图进行稀疏化,以去除低质量边和低质量关系。该章节的实验探究了不同阈值的选取对模型性能的影响。实验结果如表9所示。实验结果显示,当不进行稀疏化时,模型的性能最差。这是因为子图中存在大量的低质量边和低质量关系,干扰了模型的学习过程。本文共尝试了六种不同的阈值选取情况。结果表明,丢弃30%的边所得到的模型性能最好。这意味着保留70%的边能够在一定程度上去除低质量的信息,并保留了更重要的边和关系,从而提升了模型的性能。通过实验分析,本文验证了使用百分比形式的阈值进行子图稀疏化的有效性。选择适当的阈值可以帮助去除低质量的信息,从而使模型能够更加准确地学习和理解图像中的关键信息。

表9 阈值敏感性分析

阈值	CheXpert			RSNA		
	1%	10%	100%	1%	10%	100%
0%	88.2	88.5	89.1	89.1	90.3	90.9
10%	88.4	89.5	90.2	90.6	92.1	92.9
30%	89.8	90.1	90.9	91.6	92.5	93.7
50%	89.1	89.2	90.3	90.7	91.7	92.3
70%	89.3	89.7	89.8	90.8	91.5	92.5
90%	88.8	89.1	89.5	89.6	90.7	91.6

(2)跨模态细粒度信息对齐

为了验证交叉注意力机制能够实现重要对象之间更加细粒度化的对齐,本文同样在定量和定性方面进行了分析。在定量方面,本文对比了丢弃该模块与保留该模块两种情况下的结果。实验结果如表10所示。可以看出,引入交叉注意力机制能够提升

下游任务的性能。同时为了更加直观地观察到交叉注意力机制能够有效区分同一模态内的各个重要对象，从而能够帮助模型更好地进行模态间的匹配。

表10　跨模态细粒度信息对齐消融实验

方法	CheXpert			RSNA		
	1%	10%	100%	1%	10%	100%
丢弃对齐	89.1	89.3	89.8	90.9	92.0	93.2
保留对齐	89.8	90.1	90.9	91.6	92.5	93.7

在定性方面，本文跟踪了两种情况下模型在预训练过程中的跨模态匹配准确率，实验结果如图7所示。可以看出，丢弃对齐时，模型的训练过程相对不够稳定，在前期出现明显跳跃的区间，后期的匹配准确率提升也较为缓慢。增加对齐操作，模型的训练过程更加稳定，由横坐标可以看出，此时的模型仅仅使用80 k步的训练过程就达到了较为先进的性能。保留对齐能够在各个重要对象之间产生独立的差异化信息，促使模型增强对重要对象的识别和表达能力。

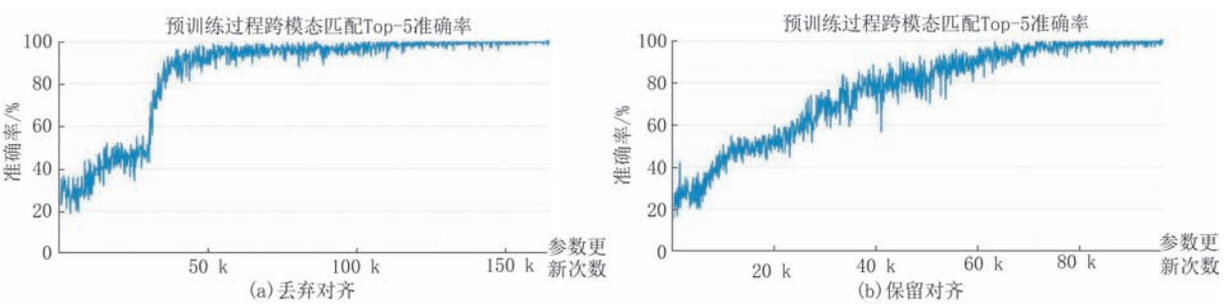


图7　不同对齐方式的匹配准确率

同时，为了更方便地展示交叉注意力机制在patch级别进行匹配的作用，本文在消融实验中增加了交叉注意力机制对下游任务的影响分析。具体而言，本文构建了两种模型进行对比：第一种模型基于交叉熵损失进行细粒度对齐，第二种模型则采用交叉注意力机制进行更加精细的特征对齐。实验数据选用了COVIDx数据集，本文在测试集中选取了600例样本，每个类别各包含200张图像，确保了类别分布的均衡性，从而使不同模型之间的比较更加公平。实验结果如图8所示通过ROC曲线进行展示。从实验结果可以观察到，引入交叉注意力机制后，虽然“无肺炎”类别的线下面积基本保持不变，但剩余两个类别的线下面积均有一定程度的提升。这一结果表明，交叉注意力机制通过在特征对齐过程中对模态间的patch块相似性进行权重调整，使得模型不仅能够关注子图的全局特征，还能精准捕捉到细粒度层面的局部特征，进一步提升了跨模态的对齐精度。

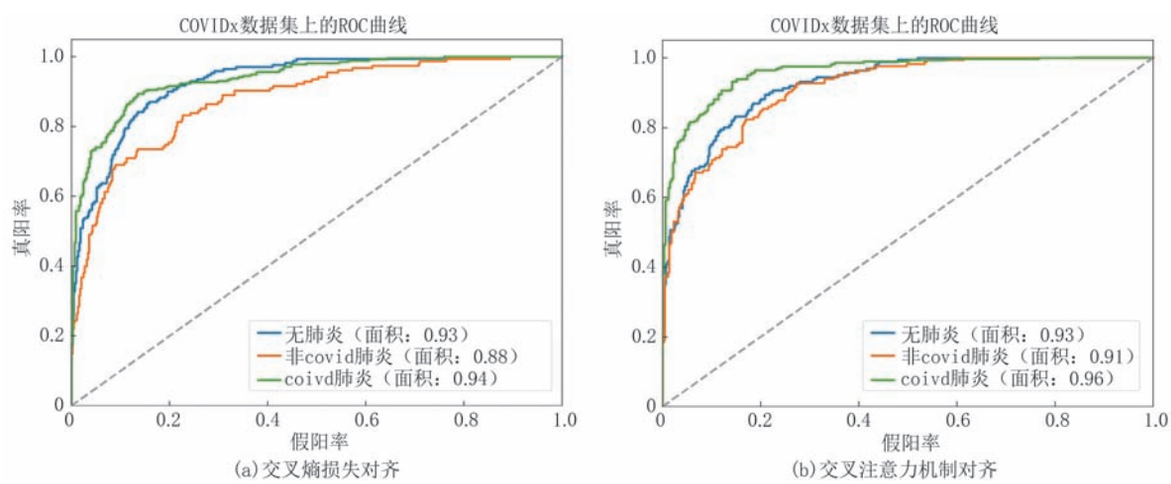


图8　不同对齐方式的ROC曲线.

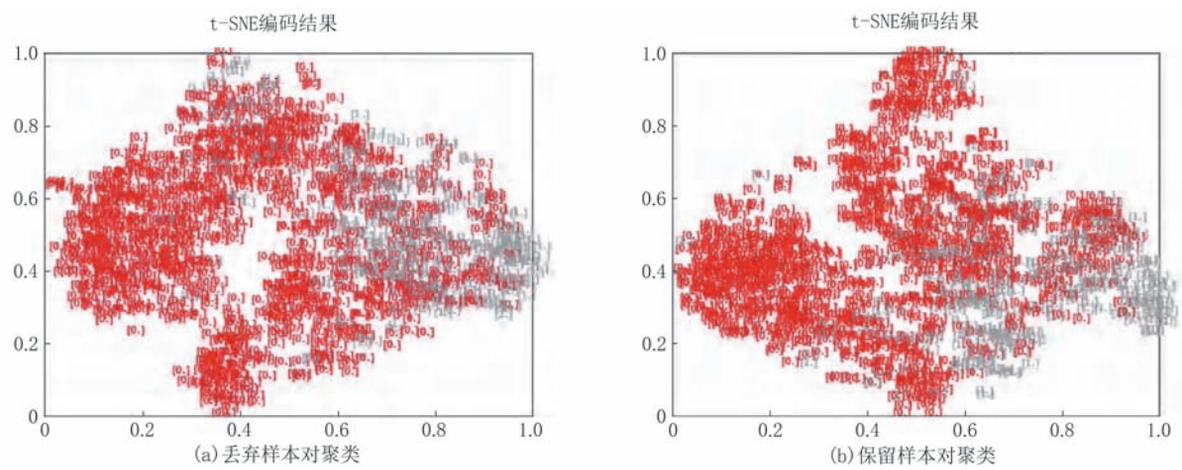
(3)多模态样本对聚类

针对该模块的消融实验包含两部分，一是通过丢弃与保留该模块验证该模块的有效性，实验结果如表11所示。实验结果显示，保留该模块能够显著提升下游任务的效果。这表明多模态样本对聚类模块在模型中起到了有效的作用。

表11 多模态样本对聚类消融实验

方法	CheXpert			RSNA		
	1%	10%	100%	1%	10%	100%
丢弃聚类	88.5	89.2	90.2	89.7	91.4	92.3
保留聚类	89.8	90.1	90.9	91.6	92.5	93.7

此外，为了更加直观地观察多模态样本对聚类任务中视觉编码器的影响，本文利用t-SNE算法对丢弃聚类与保留聚类两种情况下的视觉编码器输出进行了可视化。RSNA数据集作为下游分类任务的可视化分析对象。实验结果如图9所示。在图中，横纵坐标分别表示不同图像经过预训练编码器提取特征后，将特征经过t-SNE算法降维到二维空间中的对应坐标值。红色和灰色分别代表不同类别的图像输出结果。观察左侧部分，可以明显看出，在未应用聚类的情况下，视觉编码器的输出结果呈现出一定的混乱性，不同类别的样本在降维空间中缺乏明显的区分。具体表现为红色部分和灰色部分之间存在显著的交叉。然而，在引入聚类后，右侧图中红色部分和灰色部分之间的区别变得明显。聚类方法有效地将不同类别的样本聚集到特征空间中的相应区域，增强了类别之间的可分性。具体而言，灰色部分在右图中显示出更为集中的特征分布，显著减少了与红色部分的交叠。这一变化不仅提升了样本的可分性，也为下游任务的性能改善奠定了基础。这一实验结果表明，增加聚类具有正向的影响，能够提升视觉编码器的聚类效果。



(a) 丢弃样本对聚类 (b) 保留样本对聚类

图9 视觉编码器输出特征的可视化

二是对可调节参数K，即聚类过程中选择相似样本的数量进行了敏感性分析。在选择相似样本数量K时，batch_size的大小直接决定了模型在每次训练中能够同时处理的样本数量。较大的batch_size使得模型在每次训练更新时能够考虑更多样本，从而在选择相似样本时具备更丰富的参考依据。为了确保代码的移植性并验证K在不同环境下的最佳取值，本文分别在单块显卡上对batch_size等于32和24两种情况下进行了分类任务的实验。实验结果如图10所示。可以看出，两种情况具有相同的走势。同时实验结果显示，当K取batch_size的1/6时，模型的下游任务表现都达到了最优。使用过多数量会导致所有样本同一化，难以对各个样本对进行准确的区分。使用过少的数量导致聚类效果不够显著，难以学习到重要对象本身的语义表达。这表明K和batch_size之间存在一定比例的优化关系。同时可以观察到，即使在较小的batch_size下，模型的性能损失也非常有限。

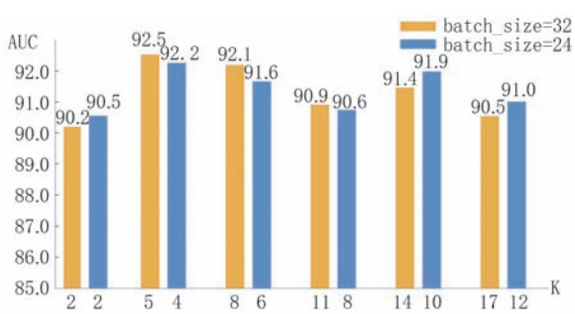


图10 参数K的敏感性分析

5 结 论

在多模态大模型的预训练过程中，常常采用全局匹配的方式进行对比学习。然而，这种全局匹配方法容易受到低质量信息的干扰，特别是当关键信息只集中在图像或文本的一小部分时。为了解决这个问题，本文首先提出了基于图卷积网络的局部匹

配方法使得模型更加精确地关注重要对象。其次提出了样本对聚类方法以提供更符合人类思维的分类策略。同时引入交叉注意力机制实现更加细粒度信息的对齐。促进了医学领域视觉任务的进一步发展。但本文并未将该模型应用于跨模态或者多模态任务，因此，未来的研究会在此基础上设计更加符合图卷积网络的多模态融合方法，并将其应用到各种跨模态和多模态任务中去，探索该大模型更多的应用场景。同时，在接下来的工作中，基于图神经网络设计更好的匹配和聚类方式，将为众多下游任务提供更加可靠的多模态预训练模型。

参 考 文 献

[1] Jeffrey D, Joseph R, Bernardino R, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine, 2018, 24(9):1342-1350

[2] Qiang Wei, Du Yu, Li Xinjin, et al. Auxiliary diagnosis for Parkinson's disease using multimodel feature analysis. Journal of Software, 2024, 35(5): 1-16 (in Chineae)
(强薇，杜宇，李信金 等. 多模态特征分析的帕金森病辅助诊断方法. 软件学报，2024，35(5):1-16)

[3] Liu Hui, Li Shanshan, Gao Shanshan, et al. Research on dual-adversarial MR image fusion network using pre-trained model for feature extraction. Journal of Software, 2023, 34(5): 2134-2151 (in Chineae)
(刘慧，李珊珊，高珊珊 等. 预训练模型特征提取的双对抗磁共振图像融合网络研究. 软件学报，2023，34(5):2134-2151)

[4] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision//Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2021: 8748-8763

[5] Zhang Yuhao, Hang Jiang, Miura Y, et al. Contrastive learning of medical visual representations from paired images and text//Proceedings of the Machine Learning for Healthcare. North Carolina, USA, 2022: 2-25

[6] Wang Fuying, Zhou Yuyin, Wang Shujun, et al. Multigranularity cross-modal alignment for generalized medical visual representation learning//Proceedings of the Advances in Neural Information Processing Systems 35. New Orleans, USA, 2022: 33536-33549

[7] Zhou Hongyu, Lian Chenyu, Wang Liansheng, et al. Advancing radiograph representation learning with masked record modeling//Proceedings of the International Conference on Learning Representations. Kigali, Rwanda, 2023: 1-16

[8] Li, Zhe, Laurence T, Ren Bocheng, et al. MLIP: Enhancing medical visual representation with divergence encoder and knowledge-guided contrastive learning. arXiv preprint arXiv: 2402.02045, 2024

[9] Wang Zifeng, Wu Zhenbang, Agarwal D, et al. Medclip: Contrastive learning from unpaired medical images and text//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates, 2022: 3876-3887

[10] Huang Shih-Cheng, Shen liyue, Lungren M P, et al. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition//Proceedings of the International Conference on Computer Vision. Montreal, Canada, 2021: 3942-3951

[11] Kim S, Lee N, Lee J, et al. Heterogeneous graph learning for multi-modal medical data analysis//Proceedings of the AAAI Conference on Artificial Intelligence. Washington, USA, 2023: 5141-5150

[12] Wei Xia, Wang Tianxiu, Gao Quanxue, et al. Graph embedding contrastive multi-modal representation learning for clustering. IEEE Transactions on Image Processing, 2023, 32: 1170-1183

[13] Yu Xiao, Liu Hui, Lin Yuxiu, et al. Consensus guided auto-weighted multi-view clustering. Journal of Computer Research and Development, 2022, 59(7): 1496-1508 (in Chinese)
(于晓，刘慧，林毓秀 等. 一致性引导的自适应加权多视图聚类. 计算机研究与发展，2022，59(7): 1496-1508)

[14] Yin Zhenfei, Wang Jiong, Gao Jianjian, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark//Proceedings of the Advances in Neural Information Processing Systems 36. OrleansNew, USA, 2023: 26650-26685

[15] Tu Rongcheng, Mao Xianling, Kong Weijie, et al. CLIP based multi-event representation generation for video-text Retrieval. Journal of Computer Research and Development, 2023, 60(9): 2169-2179 (in Chinese)
(涂荣成，毛先领，孔伟杰 等. 基于CLIP生成多事件表示的视频文本检索方法. 计算机研究与发展，2023，60(9): 2169-2179)

[16] Shao Rui, Wu Tianxing, Liu Ziwei. Detecting and grounding multi-modal media manipulation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 6904-6913

[17] Liu Bo, Lu Donghuan, Dong Wei, et al. Improving medical vision-language contrastive pretraining with semanticsaware triage. IEEE Transactions on Medical Imaging. 2023, 42(12): 3579-3589

[18] Chen Xiaofei, He Yuting, Cheng Xue, et al. Knowledge boosting: Rethinking medical contrastive vision-language pre-training//Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Vancouver, Canada, 2023: 405-415

[19] Liu Bo, Xin Luze, Wang Yan. Towards medical vision-language contrastive pre-training via study oriented semantic exploration//Proceedings of the ACM Multimedia. Melbourne, Australia, 2024

[20] Jeong J, Tian K, Li A, et al. Multimodal image-text matching improves retrieval-based chest x-ray report generation//Proceedings of Machine Learning Research. 2024: 978-990

[21] Zhan Chenlu, Lin Yu, Wang Gaoang, et al. MedM2G: Unifying medical multi-modal generation via cross guided

diffusion with visual invariant//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 11502-11512

[22] Yang Jinxia, Su Bing, Zhao Wayne Xin, et al. Unlocking the power of spatial and temporal information in medical multimodal pre-training. arXiv preprint arXiv: 2405.19654, 2024

[23] Phan V M H, Xie Yutong, Qi Yuankai, et al. Decomposing disease descriptions for enhanced pathology detection: A multi-aspect vision-language pre-training framework//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 11492-11501

[24] Muller P, Kaissis G, Zou C, et al. Joint learning of localized representations from medical images and reports//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 685-701

[25] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[26] Chen Zhihong, Du Yuhao, Hu Jinpeng, et al. Multi-modal masked autoencoders for medical vision-and-language pre-training//Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Singapore, 2022: 679-689

[27] Jiang Xunqiang, Lu Yuanfu, Fang Yuan, et al. Contrastive pre-training of GNNs on heterogeneous graphs//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. Queensland Gold Coast, Australia, 2021: 803-812

[28] Zhu Yun, Guo Jianhao, Tang Siliang. SGL-PT: A strong graph learner with graph prompt tuning. arXiv preprint arXiv: 2302.12449, 2023

[29] Jiang Xunqiang, Jia Tianrui, Fang Yuan, et al. Pre-training on large-scale heterogeneous graph//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. Singapore, 2021: 756-766

[30] Duan Keyu, Liu Qian, Tat-SengChua, et al. Simteg: A frustratingly simple approach improves textual graph learning. arXiv preprint arXiv: 2308.02565, 2023

[31] Hamilton W, Ying Zhitao, Leskovec J. Inductive representation learning on large graphs//Proceedings of the Advances in Neural Information Processing Systems 30. California, USA, 2017: 1-11

[32] Zhang Jiawei, Zhang Haopeng, Xia Congying, et al. Graph-bert: Only attention is needed for learning graph representations. arXiv preprint arXiv: 2001.05140, 2020

[33] Ying Chengxuan, Cai Tianle, Luo Shengjie, et al. Do transformers really perform badly for graph representation?// Proceedings of the Advances in Neural Information Processing Systems 34. Online, 2021: 28877-28888

[34] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale//Proceedings of the Eleventh International Conference on Learning Representations. Vienna, Austria, 2021: 1-22

[35] Devlin J, Chang Mingwei, Lee K, et al. Bert: Pretraining of deep bidirectional transformers for language understanding// Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Florence, Italy, 2019: 4171-4186

[36] Wu Lirong, Lin Haitao, Huang Yuefei, et al. Quantifying the knowledge in gnns for reliable distillation into mlps// Proceedings of the International Conference on Machine Learning. Hawaii, USA, 2023: 37571-37581

[37] Johnson A E, Pollard T J, Berkowitz S J, et al. Mimiccxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data, 2019, 6(1):1-8

[38] Alsentzer E, Murphy J R, Boag W, et al. Publicly available clinical BERT embeddings. arXiv preprint arXiv: 1904.03323, 2019

[39] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention. arXiv preprint arXiv: 2012.12877, 2020

[40] Irvin J, Rajpurkar P, Ko M, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison//Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii, USA, 2019: 590-597

[41] Shih G, Wu C C, Halabi S S, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiology: Artificial Intelligence, 2019, 1(1):e180041

[42] Wang Linda, Lin Zhongqiu, Wong A. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. Scientific Reports, 2020, 10(1): 1-12

[43] Redmon J, Farhadi A. Yolov3: An incremental improvement. arXiv preprint arXiv: 1804.02767, 2018

[44] Healthcare J. Object-cxr-automatic detection of foreign objects on chest x-rays//Proceedings of the Medical Image with Deep Learning. Montreal, Canada, 2020

[45] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation//Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention. Munich, Germany, 2015: 234-241

[46] Zawacki A, Carol W, Shih G, et al. Siim-acr pneumothorax segmentation, 2019

[47] Li Mingjian, Meng mingyuan, Fulham M, et al. Enhancing medical vision-language contrastive learning via inter-matching relation modelling. arXiv preprint arXiv: 2401.10501, 2024

[48] Wang Zifeng, Wu Zhenbang, Agarwal D, et al. MedCLIP: Contrastive learning from unpaired medical images and text// Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Dublin, Ireland, 2022: 3876-3887

[49] Foteinopoulou N M, Ioannis P. Emoclip: A visionlanguage method for zero-shot video facial expression recognition// Proceedings of the IEEE 18th International Conference on Automatic Face and Gesture Recognition. Istanbul, Turkey, 2024:1-10

[50] Zhao Z, Patras I. Prompting visual-language models for dynamic facial expression recognition//Proceedings of the British Machine Vision Conference. Aberdeen, UK, 2023:1-16

[51] Jiang Xingxun, Zong Yuan, Zheng Wenming, et al. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild//Proceedings of the 28th ACM International Conference on Multimedia. New York, USA, 2020:2881-2889

[52] Wang Yan, Sun Yixuan, et al. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022:20922-20931

CHEN Gong-Guan, Ph. D. candidate. His main research interests include image processing and machine learning.

D., professor, Ph. D. supervisor. Her main research interests include medical image processing, data mining and machine learning.

LI Heng-Tai, Master candidate. His main research interests include image processing and machine learning.

GUO Qiang, Ph. D., professor, Ph. D. supervisor. His main research interests include computer vision, data mining, and other related fields.

ZHANG Cai-Ming, Ph. D., professor, Ph. D. supervisor. His main research interests include data mining, computer graphics, information visualization and medical image processing.

Background

Multimodal pretraining models are an approach that leverages information from different modalities to enhance the performance of models in single-modality tasks. Over the past two years, researchers have shown increasing interest in this approach. Current methods mostly adopt the network structure of CLIP (Contrastive Language-Image Pretraining), but their performance is unsatisfactory when applied to the medical domain. This is primarily because medical datasets are more challenging to obtain compared to datasets in other domains, leading to a scarcity of medical data for CLIP training. One method to reduce the reliance on data is to replace CLIP global matching with local matching to achieve more precise learning. However, existing principles of local matching are limited to simple concatenation, and there is a shortage of relevant research in this area.

In this paper, we propose a method that utilizes the powerful message-passing mechanism of graph neural networks to disperse important information from different modalities into multiple subgraphs, achieving subgraph representation for each modality. We achieve more accurate local matching through subgraph matching, thereby reducing the dependence on data quantity. Additionally, we employ a cross-attention mechanism to accomplish intra-modality matching at the patch level, addressing the issue of sample homogeneity caused by relying solely on local matching. Furthermore, we leverage the ideas of graph neural networks and design a high-dimensional space clustering method to ensure proximity between similar sample pairs and distance between dissimilar sample pairs.

This work was supported by the National Natural Science Foundation of China (62072274, U22A2033), the Central Guidance on Local Science and Technology Development Project (YDZX2022009), the Special Funds of Taishan Scholars Project of Shandong Province(tstp20221137), and the Special Funds for Talent Development in Jinan(202333037).