文章编号:1001-9014(2026)01-0103-13

DOI:10.11972/j.issn.1001-9014.2026.01.012

基于知识蒸馏的轻量化遥感多模态大语言模型

张馨月[1,2], 冯世阳[1,2], 王 斌[1,2*]

(1. 复旦大学 电磁波信息科学教育部重点实验室,上海 200433;
2. 复旦大学 信息学院 图像与智能实验室,上海 200433)

摘要:遥感多模态大语言模型融合了丰富的视觉语言模态信息,在遥感图像分析和解译等领域中展现出巨大潜力。然而,现有的知识蒸馏方法多聚焦于单模态大语言模型的压缩,忽视了各模态间的特征对齐,因而阻碍了大语言模型在跨模态任务中的性能表现。针对上述问题,提出一种基于知识蒸馏的遥感多模态大语言模型轻量化方法,通过在特征层对齐各模态的输出,实现了多模态信息的有效对齐;通过引入反向Kullback-Leibler散度作为损失函数,并结合教师混合采样和单步分解的优化策略,进一步提升了学生模型的泛化性与稳定性。实验结果表明,本文方法在遥感图像的场景分类、视觉问答、视觉定位与图像描述四种下游任务上实现了更高的准确性与效率,同时显著减少了模型参数量和对计算资源的需求,为多模态大语言模型在遥感领域的高效应用提供了新的解决方案。

关 键 词:遥感图像; 多模态大语言模型; 知识蒸馏; 反向Kullback-Leibler散度; 特征对齐

中图分类号:TP751 **文献标识码**: A

Lightweight remote sensing multimodal large language model based on knowledge distillation

ZHANG Xin-Yue[1,2], FENG Shi-Yang[1,2], WANG Bin[1,2*]

(1. Key Laboratory for Information Science of Electromagnetic Waves (MoE), Fudan University, Shanghai 200433, China;
2. Image and Intelligence Laboratory, School of Information Science and Technology, Fudan University, Shanghai 200433, China)

Abstract: Remote sensing multimodal large language models (MLLMs), which integrate rich visual-linguistic modal information, have shown great potential in areas such as remote sensing image analysis and interpretation. However, existing knowledge distillation methods primarily focus on the compression of unimodal large language models, neglecting the alignment of features across modalities, thus hindering the performance of large language models in cross-modal tasks. To address this issue, a lightweighting method for remote sensing MLLMs based on knowledge distillation is proposed. This method achieves effective alignment of multimodal information by aligning the outputs across modalities at the feature level. By introducing the reverse Kullback-Leibler divergence as the loss function and combining optimization strategies such as teacher mixed sampling and single-step decomposition, the generalization and stability of the student model are further enhanced. Experimental results demonstrate that the proposed method achieves higher accuracy and efficiency in four downstream tasks of remote sensing image scene classification, visual question answering, visual localization, and image description, significantly reducing the number of model parameters and the demand for computational resources, thereby providing a new solution for the efficient application of MLLMs in the field of remote sensing.

Key words: remote sensing images, multimodal large language models, knowledge distillation, reverse Kullback-Leibler divergence, feature alignment

收稿日期:2024-11-04,修回日期: 2025-02-12 **Received date**:2024-11-04,revised date: 2025-02-12

基金项目:国家自然科学基金(62371140),国家重点研发计划(2022YFB3903404)

Foundation items:Supported by the National Natural Science Foundation of China (62371140), the National Key Researchand Development Program of China (2022YFB3903404)

作者简介(Biography):张馨月(2000-),女,上海人,硕士研究生,主要研究领域为轻量化遥感多模态大语言模型. E-mail:22210720060@m.fudan.edu.cn

*通讯作者(Corresponding author): E-mail: wangbin@fudan.edu.cn

引言

大语言模型(Large Language Model, LLM)是一种基于深度学习的自然语言处理技术,通过学习大规模的文本数据并利用大量参数来理解和生成人类语言,在文本生成、语言翻译、对话系统等多种任务中有广泛应用。近年来,各种具有出色性能的LLM相继被提出,特别是基于Transformer架构的生成式预训练模型GPT[1]和LLaMA[2]系列、双向编码器表示模型BERT[3]等发展迅速且影响深远。然而,传统的LLM只能处理文本数据,无法有效整合其它类型的信息(如图像、视频等),这些局限于单一模态的模型架构在面对复杂的多模态任务时存在明显不足。

为克服这一限制,多模态大语言模型(Multimodal Large Language Model, MLLM)应运而生,并在多个应用领域展现出巨大潜力。不同于单一模态语言模型,MLLM通过融合视觉与语言模态信息,从文本、图像等多种模态数据中提取特征,以实现对多模态任务的全面理解与精准响应,此特性使得MLLM在跨模态问答、图像分类、目标检测等下游任务上取得了显著成果,为遥感、医学成像、智能驾驶等领域的智能化应用提供了坚实的技术基础[4]。

遥感MLLM(如GeoChat[5]、RemoteCLIP[6]、RS-GPT[7]和GRAFT[8]等)通过结合遥感图像的视觉特征与语言信息,能有效应对遥感数据中的多样化信息和复杂场景,高效处理遥感图像中的目标检测、场景分类、变化检测等关键任务,从而实现了对大规模高分辨率遥感图像数据的智能化分析与理解[9]。然而,能够处理多模态数据的模型通常规模庞大,包含了大量参数,并对计算资源的需求极高,导致其推理速度较慢,在存储与部署方面也面临显著瓶颈。在遥感应用中,由于遥感数据通常包含了大量高分辨率图像以及其类内类间的特征差异信息,其处理难度和资源消耗更为显著,上述问题表现得更为突出[10]。为了解决上述瓶颈,知识蒸馏作为一种模型压缩方法,能够有效降低模型规模,为计算资源受限条件下的模型轻量化部署提供支持。该方法通过引入轻量化的学生模型,使其从高性能的教师模型中学习知识,从而在保证性能的前提下,显著减少模型规模并提升推理速度[11]。

当前的大模型轻量化研究主要集中在处理单模态信息的LLM的压缩上[12]。尽管知识蒸馏在压缩LLM方面取得了一定的进展,但在MLLM中应用知识蒸馏以实现模型轻量化的技术尚未得到深入探索。由于多模态任务需要融合多个模态的信息,语言模态与视觉模态之间的特征一致性对于模型的整体性能具有重要影响[13]。然而,当前使用知识蒸馏进行模型轻量化的方法通常将LLM的蒸馏过程与其它模态模块相互独立,这使得学生模型难以达到教师模型在跨模态任务中的表现水平。这种独立的蒸馏策略导致模态信息融合不足,限制了MLLM在多模态任务中的应用和性能表现。

为解决上述问题,本文提出一种基于知识蒸馏的遥感多模态大语言模型轻量化方法,旨在实现多模态任务中各模态特征的对齐,在满足模型轻量化部署需求的同时,能高精度地处理遥感领域中的高分辨率图像分析与解译任务。具体而言,所提出方法在特征层面对齐来自语言与其它模态的输出,确保学生模型在学习过程中能有效捕获并融合多模态信息。我们以遥感MLLM的GeoChat[5]为研究基础,以充分挖掘遥感数据的视觉与语言信息,并通过知识蒸馏的方法将其传递给一个更小的学生模型,使学生模型能更好地理解和处理遥感多模态任务。更进一步,为了提高蒸馏过程的效率和效果,我们引入了反向Kullback-Leibler(KL)散度[14]作为蒸馏损失函数,并结合教师混合采样和单步分解的优化策略[15],以增强学生模型的泛化能力和稳定性。在公开的AID、UCMerced、RSVQA-HRBEN和RSVQA-LRBEN数据集[16][17][18]以及四种特定的遥感任务(场景分类、视觉问答、视觉定位和图像描述)上的实验结果表明,所提出方法取得了优异的定量结果,并显著减小了模型的参数量和运算量。

本文的主要贡献总结如下:

1) 提出一种适用于遥感MLLM的知识蒸馏方法,解决了各模态特征之间的对齐问题,从而提升了模型对多模态信息的理解能力;

2) 在蒸馏过程中,通过引入反向KL散度,并采用教师混合采样和单步分解的优化策略,提高了学生模型的泛化性与稳定性;

3) 在遥感图像的场景分类、视觉问答、视觉定位和图像描述四种多模态任务上进行了性能验证实验,定量结果表明,所提出方法不仅显著减少了模型参数量和对计算资源的需求,还实现了超越教师模型的高准确率。

1 相关工作

1.1 遥感多模态大语言模型

MLLM是以LLM为基础,通过融入其它非文本

的模态信息,实现多模态任务处理的模型。如图 1 所示,典型的 MLLM 架构主要包含三个模块:1)LLM 骨干,对输入的各模态特征进行处理,并输出下游任务所需的 Token;2)视觉编码器,编码输入图像的特征;3)视觉-语言适配器,映射视觉特征到语言空间的模块。现有的通用 MLLM,如 LLaVA[19] 和 MiniGPT[20],已经在自然语言领域数据上展现了强大的多模态任务处理能力。然而,通用 MLLM 在特定领域(如遥感、医学等)中的表现仍存在局限性,因此,针对特定领域需求的研究逐步兴起,从预训练的 LLM 开始,结合特定领域的数据(如遥感图像、医学影像等)对现有模型进行微调[21]。

遥感 MLLM 通过利用大量未标注的遥感图像训练深度学习模型,旨在提取遥感图像中的通用特征表示,提升遥感图像分析任务的性能、效率和通用性[22]。GeoChat[5] 是专为遥感领域高分辨率图像分析和解译任务设计的遥感 MLLM,克服了传统 MLLM 在遥感领域面临的挑战,如图像分辨率多样性、遥感领域数据稀缺等问题[23]。它基于 MLLM 的 LLaVA-v1.5[24] 架构,增加了任务提示以指明所需任务类型,并允许在输入和输出中包含空间位置,支持视觉提示输入和视觉定位输出,并通过低秩自适应策略(LoRA)[15] 进行高效微调,使得它在保留 LLaVA[25] 对话和指令跟随能力的同时,扩展了其遥感任务领域知识。

GeoChat 的核心组件包括可自由调整权重的 LLM Vicuna-v1.5[26],冻结的对比语言-图像预训练模型(Contrastive Language-Image Pre-training,CLIP)[27] 作为视觉编码器,以及多层感知机作为视觉-语言适配器以对齐视觉特征与文本特征。此结构设计使得 GeoChat 具备多任务对话能力,能够基于图像或特定区域回答问题、识别和描述目标,并通过坐标实现图像中的视觉定位,是能同时解决所有任务并具备对话能力的遥感领域通用模型。

1.2　知识蒸馏

知识蒸馏是一种模型压缩技术,旨在通过将大型且复杂的教师模型中的知识传递给较小的学生模型,以提高学生模型的性能。其核心思路是,利用教师模型的输出作为学生模型的训练目标,最小化教师模型和学生模型的输出概率分布间的差异以对齐两者的输出,从而使学生模型能够在保持较小规模的同时,获得更强的泛化能力[28],其基本架构如图 2 所示。

对于教师模型输出概率分布 $p_T(x)$ 以及学生模型输出概率分布 $p_S(x)$,为使两者的输出概率分布更具灵活性和代表性,通常会引入一个温度参数 T[29],

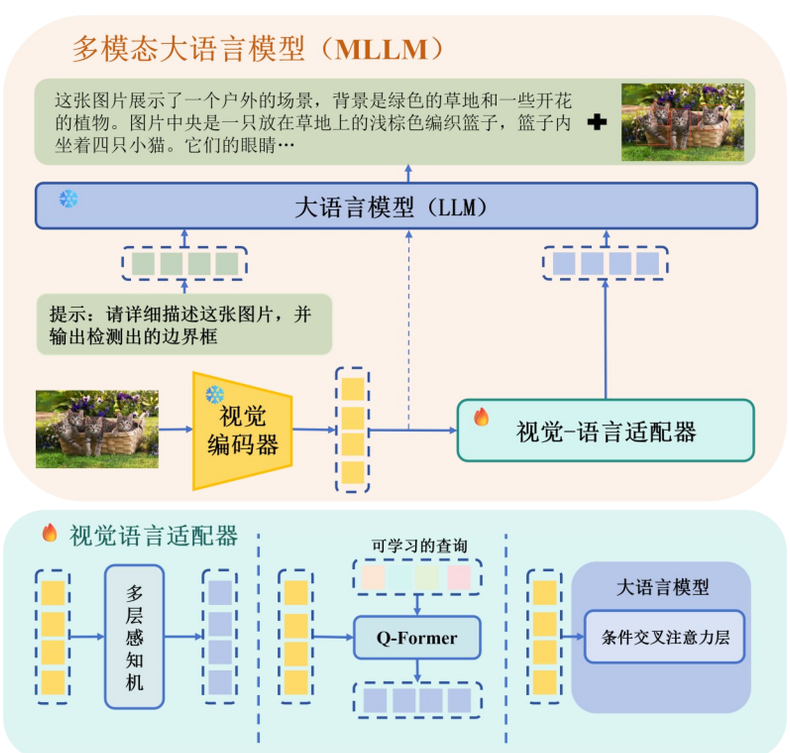


图 1　通用 MLLM 的结构图

Fig. 1　The structure diagram of general MLLM

此时，教师和学生模型的输出经过softmax操作后分别为：

$$p_{\mathrm{T}}^{i}=\frac{\exp\left(Z_{\mathrm{T}_i}/T\right)}{\sum_{k=1}^{C}\exp\left(Z_{\mathrm{T}_k}/T\right)},\ p_{\mathrm{S}}^{i}=\frac{\exp\left(Z_{\mathrm{S}_i}/T\right)}{\sum_{k=1}^{C}\exp\left(Z_{\mathrm{S}_k}/T\right)},\quad(1)$$

其中，z_{T_i}和z_{S_i}分别是教师和学生模型输出层的logits值，C为类别总数。对应地，知识蒸馏的损失L_{KD}可表示为：

$$L_{\mathrm{KD}}=T^2KL\left(p_{\mathrm{T}}\|p_{\mathrm{S}}\right)=T^2\sum_{i=1}^{C}p_{\mathrm{T}}^{i}\log\frac{p_{\mathrm{T}}^{i}}{p_{\mathrm{S}}^{i}},\quad(2)$$

其中，$KL(\bullet)$为KL散度。最后，知识蒸馏损失L_{KD}结合学生模型预测和真实标签之间的交叉熵损失L_{CE}构成了最终损失函数：

$$L=\alpha L_{\mathrm{KD}}+(1-\alpha)L_{\mathrm{CE}},\quad(3)$$

其中，α是一个权重参数，用于平衡知识蒸馏损失与真实标签预测损失之间的重要性。这样，在训练学生模型时，既可学习到教师模型中的知识，又能结合真实标签进行预测训练。

2 模型构建

本文旨在设计并实现一种基于知识蒸馏的遥感多模态大语言模型轻量化方法，其整体结构如图3所示，主要由三部分构成：教师模型、学生模型和知识蒸馏模块。下面，首先介绍作为教师模型和学生模型的遥感MLLM；然后，对遥感MLLM知识蒸馏的设计与具体实现进行详细描述和分析，并给出蒸馏过程中的损失函数和优化策略。

2.1 教师和学生模型架构

在构建适用于遥感图像领域的MLLM框架时，我们精心设计了教师模型与学生模型，它们均遵循GeoChat[5]的模型架构，包含了3个主要组件：LLM、视觉编码器与视觉语言适配器。

对于教师模型，我们选取Vicuna-13B-v1.5[26]作为LLM，并结合视觉主干编码器CLIP-ViT(L-14)[30]和跨模态适配器多层感知机进行训练，最终得到GeoChat-13B。教师模型凭借其高达130亿的参数量，拥有强大的知识存储与表征能力，在处理遥感图像中的复杂视觉信息及多样化任务需求时，它能够充分发挥Vicuna-13B-v1.5强大的语言理解与生成能力。

学生模型则以Vicuna-7B-v1.5[26]作为LLM，搭配视觉主干编码器CLIP-ViT(L-14)和跨模态适配器多层感知机进行训练得到GeoChat-7B。该学生模型仅具有70亿参数量，在确保一定性能的前提下，有效降低了对计算资源的需求。在知识蒸馏过程中，其主体Vicuna-7B-v1.5能够充分利用教师模型传递的知识，在不过度消耗计算资源的情况下，有效地从教师模型学习到知识和表征能力，实现模型的轻量化与高效部署，以满足实际应用中的多样化需求。

教师模型与学生模型采用GeoChat模块化架构设计，以实现灵活处理高分辨率遥感图像信息的能力，并与自然语言生成模型相结合，从而可支持多任务对话需求。与此同时，相同的架构使学生模型能高效继承教师模型的知识，提升学生模型对遥感图像的描述和场景理解能力，优化区域级别的交互和视觉定位表现，进一步增强学生模型在遥感任务中的表现力与适应性。

2.2 损失函数与优化策略

在GeoChat模型的训练过程中，我们采用了知识蒸馏策略以实现教师模型与学生模型之间的有效知识传递。模型蒸馏框架的核心在于通过教师模型提供的软标签来指导学生模型的学习，使学生模型能够在较少参数量的情况下达到或接近教师

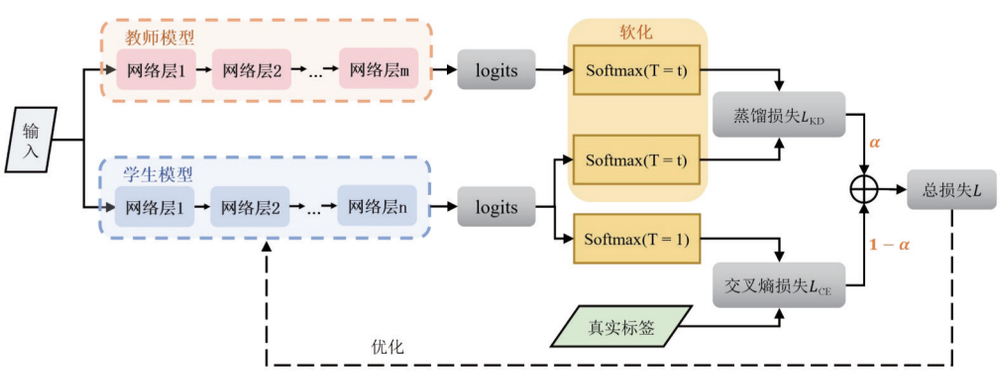


图2 知识蒸馏架构图

Fig. 2 The framework of knowledge distillation

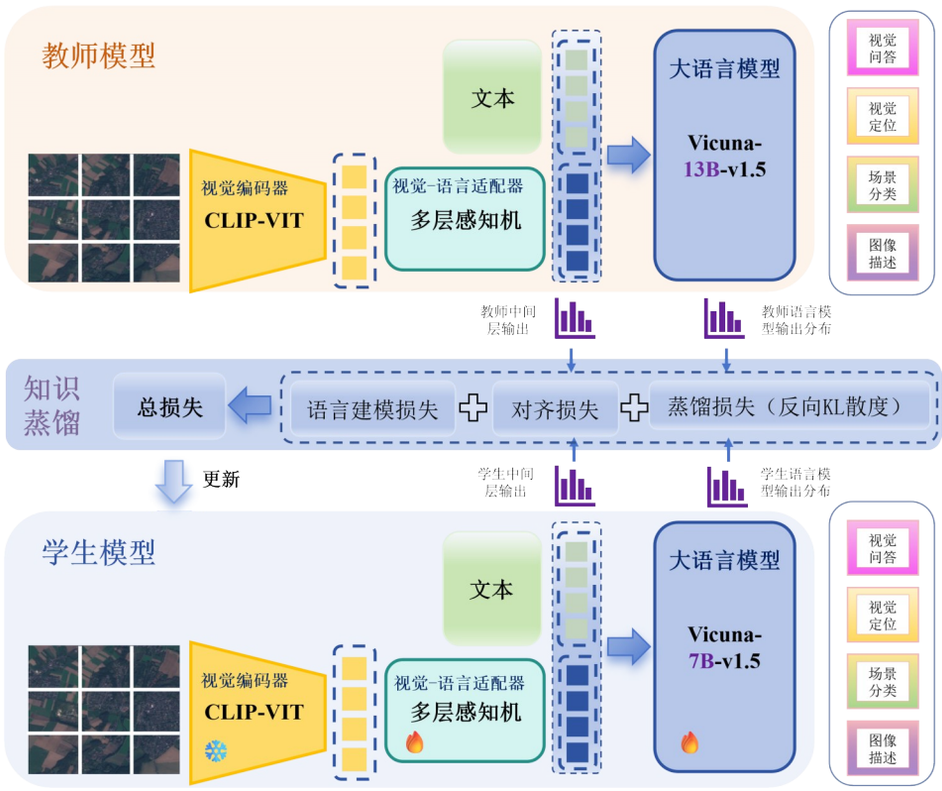


图3　本文方法的整体框架

Fig. 3　The overall framework of our method

模型的性能水平。

在此过程中,视觉特征、语言特征以及LLM输出的对齐关系是我们需要考虑的内容。对于文本输入,无论是来自训练数据中的提示还是真实响应,都经过相应的文本预处理后,与视觉特征一同进入LLM中。

2.2.1　损失函数

在蒸馏过程中,教师模型GeoChat-13B和学生模型GeoChat-7B的架构相似性确保了知识传递的可行性。具体而言,教师模型的输出分布作为学生模型学习的目标,学生模型通过调整自身参数,以在相同输入下生成与教师模型相似的输出,从而实现对教师模型知识的模仿。这种模仿不仅局限于最终的预测结果,还延伸至模型内部的各层特征。

对于从遥感数据分布p_x采样得到的提示x以及对应产生的响应y,教师模型的输出分布为$p(y|x)$,学生模型的输出分布由参数θ表示为$q_\theta(y|x)$。在多模态信息的特征对齐过程中,多层感知机将视觉模态表示映射到语义空间,学生模型通过学习教师模型融合视觉特征与文本特征的信息,从而有效掌握处理多模态信息的能力。

假定经过多层感知机融合后的教师模型表示为$z_T(x)$,学生模型表示为$z_S(x;\theta)$,可采用均方误差(Mean Squared Error, MSE)衡量两个特征向量的相似性,对齐跨模态特征表示,通过最小化该对齐损失L_{align},可使学生模型逐步逼近教师模型的多模态特征融合效果,从而有效完成知识蒸馏任务:

$$L_{\text{align}} = -\frac{1}{2}\left\| z_T(x) - z_S(x;\theta) \right\|^2 \quad . \tag{4}$$

该对齐损失关于学生模型参数θ的梯度δ_{align}为:

$$\delta_{\text{align}} = \nabla_\theta L_{\text{align}} = -\left(z_S(x;\theta) - z_T(x)\right) \cdot \frac{\partial z_S(x;\theta)}{\partial \theta}. \tag{5}$$

对于LLM的输出对齐,我们采用了基于反向KL散度的方法。与传统的前向KL散度$KL[p\|q_\theta]$相比,反向KL散度$KL[q_\theta\|p]$更适用于生成式语言模型的知识蒸馏[31]。传统的前向KL散度在处理文本生成任务时,容易导致学生模型对教师模型输出分布中的低概率区域产生过高估计,从而产生低质量的文本。而反向KL散度能引导学生模型关注教师模型输出分布中的主要模式,避免过度关注低概率区域,从而提高生成内容的准确性和可靠性。具体来说,不同于$KL[p\|q_\theta]$,最小化$KL[q_\theta\|p]$会使得学生模型分布q_θ寻找教师模型分布p的主要模式,并为p的

空白区域分配低概率。因此，这里我们采用反向KL散度作为蒸馏损失$L(\theta)$，其可表示为

$$L(\theta)=KL\left[q_\theta\|p\right]=E_{x\sim p_x,y\sim p'}\log\frac{q_\theta(y|x)}{p(y|x)}\quad,\quad(6)$$

其中，p'为真实数据分布。

为了保证学生模型在标准自然语言处理基准上的性能，我们还需考虑语言建模损失L_{PT}[32]，采用交叉熵损失梯度来更新语言模型参数：

$$\nabla L_{\mathrm{PT}}=-\frac{1}{|y|}\sum_{k=1}^{|y|}\nabla\log q_\theta\left(y_k|x_{<k}\right)\quad.\quad(7)$$

2.2.2 优化策略

为了提高蒸馏训练的效率和稳定性，我们在优化算法中采用了策略梯度方法[33]。然而，由于策略梯度仍会受到高方差和奖励黑客攻击的影响[34]，因此，在计算蒸馏损失对学生模型参数的梯度时，引入教师混合采样和单步分解策略[15]。

具体而言，教师混合采样策略在每个时间步通过一定概率将教师模型和学生模型的预测进行混合，从而生成新的采样分布。这种方法平衡了教师模型与学生模型的预测输出，避免了学生模型对教师模型的过度依赖，有效提升了模型的鲁棒性。另一方面，单步分解策略则通过将梯度计算分解为当前步的生成质量$(\nabla L)_{\mathrm{Single}}$和长期生成趋势$(\nabla L)_{\mathrm{Long}}$两部分（见公式(8)），以减少训练过程中的方差并加速模型的收敛。这一分解使得学生模型能够更高效地从教师模型中学习，逐步提升自身性能，从而达成优化模型的目标。

$$(\nabla L)_{\mathrm{Single}}=-\underset{x\sim p_x,y\sim p'}{E}\left[\sum_{t=1}^{|y|}w_t\nabla\underset{y_t\sim q_\theta(t)}{E}\left[r_t\right]\right],(\nabla L)_{\mathrm{Long}}=$$

$$-\underset{x\sim p_x,y\sim p'}{E}\left[\sum_{t=1}^{|y|}w_tR_{t+1}\nabla\log q_\theta\left(y_t,x\right)\right]\quad,\quad(8)$$

其中，w_t为重要性权重；r_t为单步生成质量，其累加R_{t+1}用于衡量所有步生成质量，它们分别可根据下式计算求得：

$$w_t=\prod_{k=1}^{t}\frac{q_\theta\left(y_k|y_{<k},x\right)}{p\left(y_k|y_{<k},x\right)},R_{t+1}=\sum_{k=t+1}^{|y|}\log\frac{p\left(y_k|y_{<k},x\right)}{q_\theta\left(y_k|y_{<k},x\right)}\quad.\quad(9)$$

结合上述策略，我们得到最终的优化梯度目标：

$$\nabla L(\theta)=-\underset{x\sim p_x,y\sim p'}{E}\left[\sum_{t=1}^{|y|}w_t\left[(\nabla L)_{\mathrm{Single}}+(\nabla L)_{\mathrm{Long}}^{\mathrm{Norm}}+\delta_{\mathrm{align}}\right]\right]+\nabla L_{\mathrm{PT}}\quad,\quad(10)$$

其中，$(\nabla L)_{\mathrm{Long}}^{\mathrm{Norm}}$为进行归一化后的$(\nabla L)_{\mathrm{Long}}$，以提高模型在处理不同长度序列时的稳定性和准确性。

通过上述蒸馏框架和优化目标设定，教师模型GeoChat-13B生成高质量的软标签，这些标签包含了丰富的跨模态信息，涵盖了视觉特征与语言输出之间的细粒度关系。学生模型GeoChat-7B在学习过程中，通过最小化与教师模型输出之间的差异，逐步优化自身的参数，以学习到与教师模型相似的特征表达和推理能力，从而在遥感图像的多模态处理任务中表现出更好的性能，同时还具备了轻量化和高效推理的特点，满足了遥感图像分析在实际应用中资源受限条件下的需求。

2.2.3 算法总结

表1给出了本文算法的训练过程伪代码。首先，设置学生模型和教师模型的架构及相关参数，确保知识传递的有效性。在蒸馏训练中，循环执行一系列步骤，包括数据采样和梯度计算，通过多重梯度综合更新学生模型参数，逐步优化学生模型的多模态理解与生成能力，最终构建出具备高效性能的蒸馏遥感MLLM模型。

表1 所提出算法训练过程的伪代码

Table 1 The pseudocode for the proposed algorithm in training process

算法训练过程的伪代码：遥感MLLM的知识蒸馏
输入：遥感多模态指令微调数据集D，包含提示和真实响应对；
预训练语料库D_{PT}，包括视觉和文本； 教师模型，具有输出分布p； 学生模型，预训练于D_{PT}，具有输出分布q_{θ_0}； 学习率η；批处理大小M。 在遥感多模态指令数据集D上对学生模型进行微调，以教师模型输出和真实响应同时作为指导， 从q_{θ_0}开始，选择验证损失最小的θ。 重复执行以下步骤，直至收敛并返回q_θ： 步骤1：从数据集D中采样一个小批次的提示，得到响应集合$S=\left\{\left(x^m,y^m\right)\right\}_{m=1}^{M}$； 步骤2：从$D_{PT}$中采样一个小批次$D'_{PT}=\left\{d^m\right\}_{m=1}^{M}$； 步骤3：在集合$S$上，根据式(5)计算跨模态的对齐损失梯度$\delta_{\mathrm{align}}$； 步骤4：在$D'_{PT}$上，根据式(7)计算语言建模损失梯度$\nabla L_{\mathrm{PT}}$； 步骤5：在集合$S$上，根据式(8)计算单步梯度$(\nabla L)_{\mathrm{Single}}$和长序列梯度$(\nabla L)_{\mathrm{Long}}$，以及相应 的规范化长序列梯度$(\nabla L)_{\mathrm{Long}}^{\mathrm{Norm}}$； 步骤6：更新模型参数： $\theta-\eta\left[(\nabla L)_{\mathrm{Single}}+(\nabla L)_{\mathrm{Long}}^{\mathrm{Norm}}+\delta_{\mathrm{align}}+\nabla L_{\mathrm{PT}}\right]\rightarrow\theta$
输出：一个具有输出分布q_θ的学生模型

3　实验结果与分析

3.1　遥感多模态指令数据集

为了应对遥感领域缺乏多模态指令跟随数据集的问题,我们通过遥感多模态指令数据集[5],结合遥感图像与多样化的文本指令,以支持模型在遥感任务中的训练与轻量化。

该数据集[5]包含318,000对图像-指令配对,涵盖多种任务。为了增强数据集的多样性,从多个来源整合了不同类型的遥感图像,涉及多种场景、天气条件和地理特征,从而提升GeoChat模型在不同遥感任务中的适应性和泛化能力。

表2　遥感多模态指令跟踪数据集的指令类型和格式

Table 2　Instruction types and format of remote sensing multimodal instruction dataset

数据集	大小	响应格式提示
NWPU-RESISC-45[17]	31.5k	
RSVQA-LRBEN[18]	56k	用一个单词或短语回答问题
Floodnet[35]	4k	
Detailed Description	30k	详细描述图片
Multi-Round Conversation	65k	-
Complex Questions	10k	-
Grounding Description	45k	[grounding]详细描述图片
Region Captioning	40k	[identify]$b=\{b_{x\_left}, b_{y\_top}, b_{x\_right}, b_{y\_bottom}\mid\theta\}$
Referring Expression	25k	[refer]< p > *Object* < /p >

表2具体展示了遥感多模态指令跟踪数据集的指令类型和格式。为了消除任务之间的歧义,每个任务分配了唯一的任务标识符,分别为$t\in\{\text{grounding, identify, refer}\}$,对应于视觉定位、图像描述和引用理解。对于视觉问答和场景分类,模型以单个单词或短语输出答案。对于与视觉无关的命令,则无需任务标识符。同时,在视觉定位任务中,模型需要准确识别参考对象的空间位置。为此,我们将边界框的区域位置表示为文本格式:

$$b=\{b_{x\_left}, b_{y\_top}, b_{x\_right}, b_{y\_bottom}\mid\theta\}\quad,\quad(11)$$

其中,(b_{x_left}, b_{y_top})表示边界框的左上角坐标,$(b_{x_right}, b_{y_bottom})$表示边界框右下角的坐标,$\theta$则为边界框的旋转角度。

3.2　实验设置

本文的教师模型和学生模型均在Pytorch框架上构建,使用配备6个Nvidia-A100 GPU(每个GPU内存80GB)的设备进行训练。

教师模型利用预训练的CLIP-ViT作为视觉编码器和Vicuna-v1.5初始化模型权重,应用低秩自适应策略(LoRA)对Vicuna-v1.5特定参数微调,通过调整权重矩阵(指定秩为64)减少训练参数量以提高训练速度与效率。模型训练全程维持504×504图像分辨率,此分辨率利于捕捉遥感图像细节,处理大尺度图像。每次训练步骤结合多模态指令模板优化多种视觉-语言任务训练效果,使用自适应矩估计优化器(Adaptive moment estimation with Weight decay, AdamW)[36]和余弦学习率调度器,全局批量大小设为144,训练分为两阶段:先使用全部数据集训练1个周期(约2 400步),再基于遥感多模态指令跟踪数据集中的定位数据训练1 600步,以优化模型在特定任务上的性能。

学生模型的蒸馏训练过程建立在教师模型的训练基础之上。在蒸馏训练过程中,教师模型的参数被完全冻结,不参与梯度计算和参数更新,仅用于指导学生模型的训练。

3.3　不同任务下的模型性能对比与分析

在场景分类、视觉问答、视觉定位与图像描述四种下游任务中,为验证所提出方法的性能,我们主要选取了三种模型进行性能对比与分析:初始模型GeoChat-7B、教师模型GeoChat-13B和经所提出算法知识蒸馏优化后的学生模型GeoChat-7B模型。

1)初始模型GeoChat-7B,是以Vicuna-7B-v1.5作为微调起点,结合视觉编码器CLIP-ViT和跨模态适配器多层感知机,基于遥感多模态指令数据集[5]进行训练而得到的具有70亿参数量的遥感MLLM,我们以该模型作为基准对比模型[5]。

2)对于教师模型GeoChat-13B,与学生模型具有相同的基础结构,以Vicuna-13B-v1.5为微调起点,以确保其具有高质量的多模态理解与生成能力,同样基于遥感多模态指令数据集[5]进行训练而得到的具有130亿参数量的遥感MLLM。

3)经所提出算法知识蒸馏优化后的学生模型GeoChat-7B①,是在教师模型GeoChat-13B的指导下进行蒸馏训练,通过继承教师模型的特征表示和生成能力,得到一个更为高效的具有70亿参数量的学生模型,用于多模态遥感任务。为了便于使用,经所提出算法蒸馏后得到的学生模型GeoChat-7B可

①https://github.com/I3ab/GeoChat-KD

从本文所提供的网址进行下载和获取。

3.3.1 场景分类

1)实验数据集

实验数据采用了AID①和UCMerced②两个经典的遥感图像数据集合[16][17]，它们采集于不同的场景区域，包含了丰富的场景类别。这两个数据集中的图像空间信息丰富、地物分布复杂，在遥感图像分类领域均具有重要的学术价值与应用潜力。

AID数据集是一个用于遥感图像分类的重要基准数据集，包含10,000张高分辨率航空图像，覆盖30个类别，如河流、城市区域和森林等。该数据集主要来源于Google Earth，旨在评估机器学习和深度学习模型在土地利用和自然景观识别任务中的性能。每张图像的尺寸为600×600像素，其高分辨率特性使得细节信息丰富，适用于深度学习模型的训练与评估。

相较于AID数据集，UCMerced数据集在多个方面具有显著差异。它包含2,100张图像，涵盖21个土地利用类别，主要聚焦于城市与乡村场景。该数据集的图像有相对较低的分辨率，每张图像的尺寸为256×256像素，其细节的提供弱于AID，因而更具挑战性。在数据来源方面，UCMerced主要来自美国地质调查局的航空影像，专注于特定土地利用类型的分类研究。

2)实验结果与分析

我们为模型提供所有类别，并要求仅用一个词或短语来分类图像，例如："将图像分类为：学校、公园等"。在实验中，采用零样本场景分类整体精度(Overall Accuracy, OA)和零样本场景分类平均精度(Average Accuracy, AA)作为模型评价指标，同时利用混淆矩阵提供可视化分类结果。其中，OA通过计算模型正确分类样本数与总样本数的比例评估模型在没有见过目标类别样本的情况下，对新类别场景正确分类的能力，评估模型的泛化性能；AA是对各类别准确率进行平均得到的指标，可更全面地评估模型的性能优劣；混淆矩阵则以矩阵形式展示模型的预测结果和真实标签之间的对比，能够直观地反映模型的分类能力。

表3和表4展示了不同模型在AID数据集和UCMerced数据集上的定量结果，最佳结果用粗体标记。蒸馏后的学生模型GeoChat-7B在AID数据集和UCMerced数据集上的OA分别达到了67.30%和91.24%，尤其是在UCMerced数据集上，显著超越了主流MLLM(MiniGPTv2[20]、LLaVA-1.5[24]和Qwen-VL[37])以及初始模型GeoChat-7B的精度，甚至超越了教师模型GeoChat-13B的精度，展现了更强的泛化能力。同时，AA也进一步反映了蒸馏后模型的精度提升。我们考虑，这主要是因为在知识蒸馏过程中，采用了反向KL散度使得学生模型关注教师模型输出分布中的主要模式，在特征提取和模态信息对齐上得到增强，更好地模仿教师模型的行为，从而更全面地理解遥感多模态大模型的模态信息。

表3 不同模型在AID数据集和UCMerced数据集上的零样本场景分类整体精度比较

Table 3 Zero-shot scene classification OA comparison of different models on AID dataset and UC-Merced dataset

模型	AID	UCMerced
MiniGPTv2[20]	12.90	4.76
LLaVA-1.5[24]	51.00	68.00
Qwen-VL[37]	52.60	62.90
GeoChat-7B	67.20	84.43
GeoChat-13B	**69.53**	90.43
Ours (GeoChat-7B)	67.30	**91.24**

表4 不同模型在AID数据集和UCMerced数据集上的零样本场景分类平均精度比较

Table 4 Zero-shot scene classification AA comparison of different models on AID dataset and UC-Merced dataset

模型	AID	UCMerced
GeoChat-7B	55.88	84.48
GeoChat-13B	**56.59**	90.43
Ours (GeoChat-7B)	55.96	**91.24**

除了上述结果，混淆矩阵进一步验证了我们的模型在场景分类任务中的性能。图4展示了蒸馏后的模型在AID和UCMerced数据集上的零样本场景分类混淆矩阵。从图4(a)可以看出，我们的模型在AID数据集的多数场景类别上均取得了比较优异的分类性能，部分类别由于模型训练数据集的类别分布偏斜，导致模型没有学习到足够的特征，无法进行精准分类，但整体准确率仍然达到了可接受的分

①https://captain-whu.github.io/AID/

②http://weegee.vision.ucmerced.edu/datasets/landuse.html

类结果(67.30%)。从图4(b)中可以看出,尽管UC-Merced数据集分辨率相对较低,更具挑战性,导致模型在类别“agricultural”和“dense residential”上,由于与“runway”以及“medium residential”类别在结构特征上的相似性,取得了低于平均结果的精度,我们模型仍然能够在其他所有类别上取得优异的分类结果,总体准确率达到91.24%。这是由于在模态信息对齐过程中采用的反向KL散度加强了学生模型对教师模型主要模态特征分布的学习,从而超越了教师模型GeoChat-13B的90.43%分类精度。

综合分析,混淆矩阵结果表明,蒸馏后的GeoChat-7B模型在场景分类任务中不仅在整体准确率上表现优异,同时能够有效应对遥感图像的类间相似性和类内多样性,进一步验证了本文提出的知识蒸馏方法在多模态任务中的有效性。

3.3.2　视觉问答

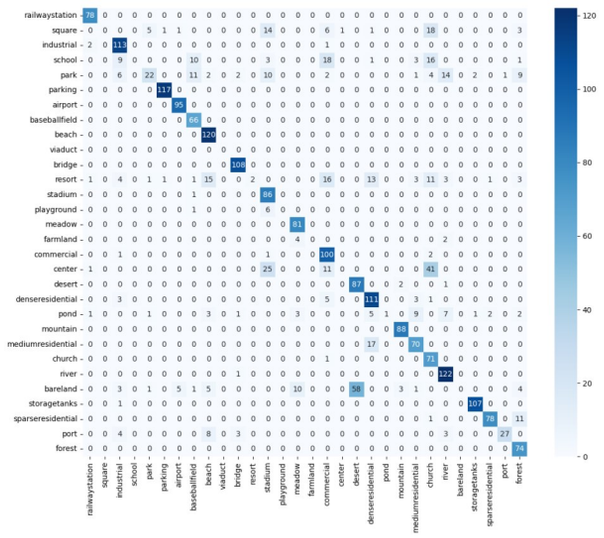
1)实验数据集

我们采用了RSVQA-HRBEN[①]和RSVQA-LRBEN[②]两个用于遥感视觉问答(Remote Sensing Visual Question Answering, RSVQA)任务的基准数据集[18],旨在评估模型对遥感图像的理解和在自然语言指令下生成准确答案的能力。RSVQA-HRBEN数据集包含10,569张高分辨率遥感图像,配有超过1,066,316个多样化的问题-答案对,涵盖存在性、比较性及计数问题,旨在考察模型对高分辨率图像的细粒度理解与推理能力。RSVQA-LRBEN则包含772张低分辨率图像和77232个问题-答案对,问题类型包括存在性、比较性、乡村/城市分类以及计数任务,适用于测试模型在低分辨率条件下的泛化能力。

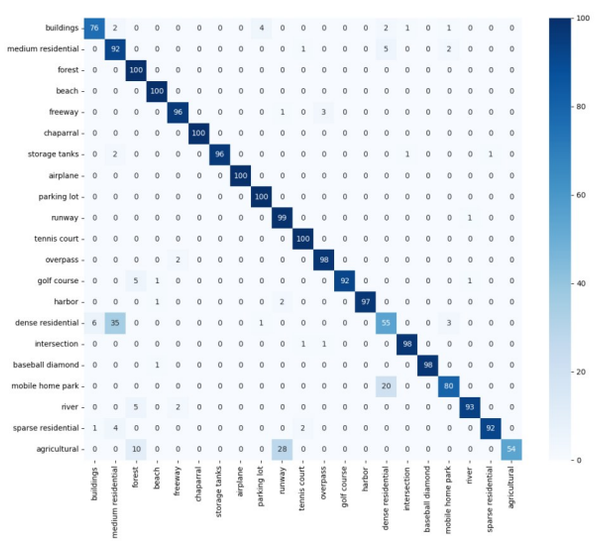
2)实验结果与分析

我们通过计算模型回答正确的问题数量占总问题数量的整体精度OA以及不同任务类别下整体精度的平均值(mean Overall Accuracy, mOA)来衡量模型在视觉问答任务中的准确性。在RSVQA-HRBEN和RSVQA-LRBEN数据集中,分别针对不同模型计算其在不同场景(如存在性、比较性、城乡等)下的准确率,以评估模型对问题的理解和回答能力。

表5展示了不同模型在RSVQA-HRBEN和RSVQA-LRBEN数据集上的定量结果,最佳结果用粗



(a)



(b)

图4　我们模型在AID和UCMerced数据集上的零样本场景分类混淆矩阵: (a) AID数据集;(b) UCMerced数据集

Fig. 4　Zero-shot scene classification confusion matrix of Ours (GeoChat-7B) on AID and UCMerced datasets: (a) AID dataset; (b) UCMerced dataset

体标记。在该视觉问答任务中,通过向模型添加适当的提示,将答案限制为简单的“是/否”和“城市/乡村”形式。总体而言,蒸馏后的学生模型GeoChat-7B在RSVQA-HRBEN和RSVQA-LRBEN上显著超越了初始模型GeoChat-7B以及其他MLLM的精度,甚至超越了教师模型GeoChat-13B的精度,展现了其强大的推理能力,能有效回答与遥感图像相关的

①https://zenodo.org/records/6344367

②https://zenodo.org/records/6344334

问题,验证了蒸馏算法的有效性。

与此同时,所测试的数据集分别聚焦于低分辨率航空图像和高分辨率遥感图像的视觉问答任务,在此任务中,蒸馏后的模型展现出了卓越的适应性,在不同数据集中均能有效学习知识。在测试环节,蒸馏后的模型针对相关问题的回答具有较高准确性。这表明模型可成功地将训练数据中所获取的知识迁移至未曾见过的测试数据中,为蒸馏模型在复杂多样的遥感图像视觉问答应用场景中的有效性提供了有力证据。

表5 不同模型在RSVQA-HRBEN和RSVQA-LRBEN数据集上的定量结果

Table 5 Quantitative results of different models on RSVQA-HRBEN and RSVQA-LRBEN dataset

模型	HRBEN			LRBEN			
	存在	比较	mOA	存在	比较	城市/乡村	mOA
MiniGPTv2[20]	40.79	50.91	46.46	55.16	55.22	39.00	54.96
LLaVA-1.5[24]	**69.83**	67.29	68.40	55.46	68.20	59.00	62.77
Qwen-VL[37]	66.44	60.41	63.06	38.57	67.59	61.00	55.35
GeoChat-7B	57.65	80.84	70.63	90.86	90.25	94.00	90.59
GeoChat-13B	56.05	83.02	71.15	91.20	91.75	**97.00**	91.60
Ours (GeoChat-7B)	60.03	**83.30**	**73.06**	**91.91**	**92.88**	95.00	**92.50**

3.3.3 视觉定位与图像描述

1)实验数据集与评价指标

为评估视觉定位任务,使用遥感图像分割数据集SAMRS①及与遥感多模态指令跟踪数据集相同的构建方法所生成的测试基准[38]。测试基准包括7653个引用问题(refer)、758个定位问题(grounding)和555个定位描述问题。使用accuracy@0.5和accuracy@0.25作为评价指标,具体而言,若预测的边界框与真实边界框的重叠度(Intersection over Union, IoU)分别超过0.5和0.25,则判定为准确。

对于图像描述任务,ROUGE-1衡量生成文本与参考文本中共同的一元组(单个单词)的比例,ROUGE-L关注最长公共子序列的长度;METEOR则综合考虑多种因素,如精确率、召回率和词干提取等,从而更全面地评估模型生成的描述与参考描述在语义和词汇上的匹配程度,以比较不同模型在该任务上的表现。通过上述这些指标来评估图像描述任务中模型生成的区域描述与参考描述之间的相关性和质量。

2)实验结果与分析

视觉定位任务,即在给定特定区域情况下,模型能提供关于该区域的详细信息。从表6可见,蒸馏后的学生模型GeoChat-7B在视觉定位任务上显著优于MiniGPTv2[20]、初始模型GeoChat-7B与教师模型GeoChat-13B。

表6 不同模型在视觉定位任务上的性能比较

Table 6 Performance comparison of different models on visual grounding task

模型	accuracy@0.5	accuracy@0.25
MiniGPTv2[20]	10.8	30.9
GeoChat-7B	11.2	33.9
GeoChat-13B	14.1	35.7
Ours (GeoChat-7B)	**14.4**	**35.9**

表7展示了不同模型在图像描述任务上性能比较的定量结果,其最好结果用粗体标记。可以发现,在区域级别的描述任务中,蒸馏后的学生模型GeoChat-7B在ROUGE和METEOR分数方面均优于MiniGPTv2[20]、初始模型GeoChat-7B与教师模型GeoChat-13B。

上述结果表明,蒸馏后的学生模型GeoChat-7B能够成功进行更精确的视觉定位与描述生成,与目标区域的特征完美匹配,进一步验证了本文所提议的蒸馏算法在遥感图像分析中的有效性,展示了其在多模态任务中的应用潜力。

表7 不同模型在图像描述任务上的区域级描述性能比较

Table 7 Region level captioning performance comparison of different models

模型	ROUGE-1	ROUGE-L	METEOR
MiniGPTv2[20]	32.10	31.20	10.00
GeoChat-7B	86.64	86.59	62.27
GeoChat-13B	87.10	87.05	63.18
Ours (GeoChat-7B)	**87.80**	**87.75**	**63.55**

我们考虑,在训练过程中确保视觉和语言模态特征在同一语义空间中有效对齐,这使得我们模型能够更好地理解图像与文本之间的关系,弥补了特征对齐不足可能导致的学生模型在处理多模态任

①https://github.com/ViTAE-Transformer/SAMRS

务时表现不稳定的问题,从而使蒸馏后模型在视觉定位和图像描述任务中表现更为优异。

另外,我们对视觉定位任务的效果进行可视化展示与分析。图5展示了三种模型在视觉定位任务中几种不同图像上的目标检测边界框输出。其中,图5的第一列来自测试数据集中的原始图像,第二、三和四列分别展示了初始模型、教师模型以及蒸馏后的学生模型所对应的视觉定位边界框输出。

通过对比图5中的各图片,可观察到不同模型在视觉定位任务上边界框输出的表现差异。具体而言,教师模型GeoChat-13B通常能够生成较为精确的边界框,反映出其在特征提取和视觉定位方面的优势。而初始模型GeoChat-7B的边界框输出可能存在一定的偏差,尤其在复杂场景中的目标定位上表现不足,显示出其特征提取能力的局限性。蒸馏后的学生模型GeoChat-7B在教师模型GeoChat-13B和初始模型GeoChat-7B上表现出改进,在Airplane、Baseball field等遥感图像上,可直观清晰地看到蒸馏后的学生模型GeoChat-7B的边界框定位更为精准,能在保持较高准确率的同时,展现出更好的灵活性和适应性。

3.4　模型计算效率分析

为进一步验证本文所提知识蒸馏方法在计算效率上的提升效果,我们对模型的参数量和每秒浮点运算次数(Floating Point Operations Per Second,FLOPS)进行了定量分析。表8列出了本文三种模型在参数量和FLOPS方面的对比情况。结果表明,GeoChat-13B作为教师模型具有约130亿的参数量和较高的运算量,蒸馏后的GeoChat-7B在参数量上大幅减少,仅不到70亿,相较于教师模型减少了约48%,这直接反映了所提轻量化策略在降低模型复杂度方面的优势;在运算量方面,蒸馏后的GeoChat-7B模型的FLOPS也显著下降,减少了约48%,进一步验证了其在更优任务性能下所需计算资源的显著降低。

结合表8的结果可以看出,虽然蒸馏后的GeoChat-7B相较于教师模型在模型规模上大幅缩减,但其在多模态任务中的性能并未因规模减小而受到明显影响,甚至在多数任务上超越了教师模型GeoChat-13B。这表明本文提出的知识蒸馏策略不仅能够显著降低模型参数量和运算量,还能有效提升多模态任务的性能,对资源受限场景中的实际应

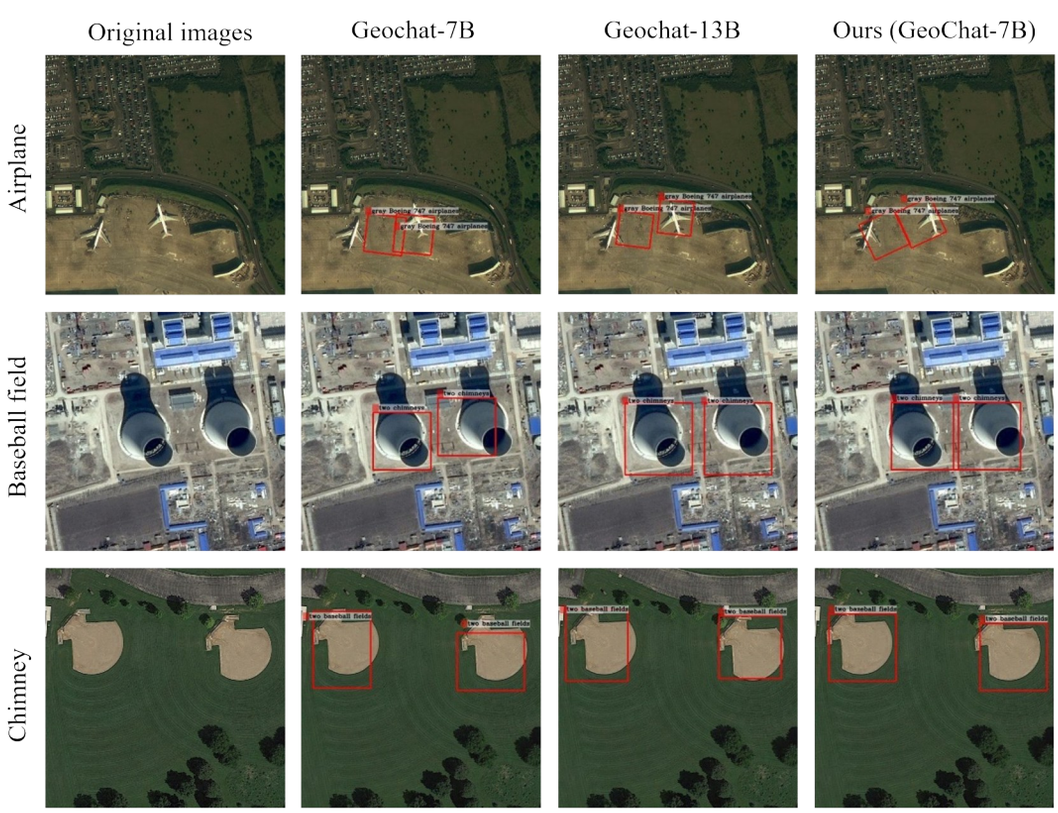


图5　视觉定位任务的可视化结果

Fig. 5　Visual results on visual grounding task

用具有重要的实践价值。

表8 不同模型的参数量和FLOPS

Table 8 Parameters and FLOPS of different models

模型	参数量	FLOPS
GeoChat-7B	6 760.72 M	6 904.53 G
GeoChat-13B	13 048.66 M	13 377.05 G
Ours (GeoChat-7B)	6 760.72 M	6 904.53 G

3.5 小结

由以上实验结果可知,经本文所提出方法蒸馏后的学生模型GeoChat-7B在多个多模态任务(场景分类、视觉问答、视觉定位和图像描述任务)中的性能测试中,其结果均优于现有的初始模型GeoChat-7B,甚至超越了教师模型GeoChat-13B,其主要特点可总结如下:

1) 精度高:所提出方法得到的蒸馏后的学生模型在各项任务中均达成了较高的定量指标,显著超越了初始模型,甚至超越了教师模型。我们认为,这主要得益于所提出方法在特征层面对齐不同模态的输出,深度学习到多模态信息;更进一步,蒸馏引入了反向KL散度作为损失函数,使得学生模型关注教师模型输出分布中的主要模式,从而使得特征提取和模态信息对齐得到增强,并结合教师混合采样和单步分解的优化策略,使模型能够更精准地处理多模态任务,有效提升了模型的准确性,实现了超越教师模型的性能。

2) 轻量化:相较于参数量达130亿的教师模型,经蒸馏后的学生模型参数量大幅缩减至70亿。经蒸馏后的学生模型使GeoChat模型更适合在边缘设备或资源有限的环境中进行部署,响应了轻量化部署的实际需求,并提升了其在各种应用场景中的适用性和灵活性。

3) 泛化能力强:我们的方法在不同类型数据集以及不同任务上均达成了较高的定量指标,表现出强大的泛化能力。我们考虑,这主要归因于我们方法在设计时充分考虑了多模态信息的通用性,以及通过知识蒸馏过程使学生模型能学习到教师模型的通用特征表示,从而在不同类型数据集以及不同遥感任务中都能发挥良好性能。

另外,在我们方法中,教师模型和学生模型的LLM部分均基于同一系列,这一选择取得了出色的定量实验结果,但一定程度上也限制了模型性能的进一步提升与知识迁移的多样性。由于同一系列LLM在结构和特性上高度相似,学生模型可能会过度依赖教师模型的特定模式和特征表示,从而难以学习到更加广泛或具有创新性的知识与处理方法。具体而言,如果教师模型在面对某些特殊语言结构或语义理解场景时存在局限,学生模型可能由于结构上的相似性,难以有效突破这一局限,从而影响学生模型在遥感图像多模态任务中的灵活性和适应性。因此,未来研究可考虑引入不同家族或类型的语言模型,探索更多样的模型组合,借助于异构模型引入的优势以克服这一局限,有望进一步提升模型的性能与泛化能力。

4 结论

提出了一种基于知识蒸馏的遥感MLLM轻量化方法,以提升压缩后模型在遥感多模态任务中的性能与效率。在遥感MLLM压缩过程中,所提出方法通过在特征层对齐各模态的输出,实现了多模态信息的有效对齐;通过采用反向KL散度作为损失函数,并结合教师混合采样和单步分解的优化策略,进一步提升了学生模型的泛化性与稳定性。在多个不同数据集以及四种下游多模态任务(场景分类、视觉问答、视觉定位和图像描述)中的实验结果表明,经所提出方法蒸馏后模型在分类精度、准确率和定位精度等指标上表现出色,具有精度高、轻量化和泛化能力强的优点,这些特点对实际应用有重要价值和意义。

在未来工作中,将深入探索异构LLM家族的知识蒸馏,拓宽学生模型学习知识的范围,实现更有效的遥感MLLM轻量化,以进一步适应复杂多变的遥感图像多模态任务需求。

References

[1] Radford A. Improving language understanding by generative pre-training[J]. [Online] Available at https://openai.com/research/language-unsupervised, 2018.

[2] Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models [J]. arXiv: 2302.13971, 2023.

[3] Devlin J. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv: 1810.04805, 2018.

[4] Yin S, Fu C, Zhao S, et al. A survey on multimodal large language models [J]. National Science Review, 2024: nwae403.

[5] Kuckreja K, Danish M S, Naseer M, et al. Geochat: Grounded large vision-language model for remote sensing [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 27831-27840.

[6] Liu F, Chen D, Guan Z, et al. Remoteclip: A vision language foundation model for remote sensing[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-16, Art no. 5622216.
[7] Hu Y, Yuan J, Wen C, et al. Rsgpt: A remote sensing vision language model and benchmark [J]. arXiv: 2307.15266, 2023.
[8] Mall U, Phoo C P, Liu M K, et al. Remote Sensing Vision-Language Foundation Models without Annotations via Ground Remote Alignment [C]. The Twelfth International Conference on Learning Representations, 2024.
[9] Wang H, Liu X, Qiao Z, et al. Multimodal Remote Sensing Data Classification Based on Gaussian Mixture Variational Dynamic Fusion Network[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-14, Art no. 5621214.
[10] Gómez-Chova L, Tuia D, Moser G, et al. Multimodal classification of remote sensing images: A review and future directions[J]. Proceedings of the IEEE, 2015, 103(9): 1560-1584.
[11] Hinton G. Distilling the Knowledge in a Neural Network [J]. arXiv:1503.02531, 2015.
[12] Zhu X, Li J, Liu Y, et al. A survey on model compression for large language models[J]. Transactions of the Association for Computational Linguistics, 2024, 12: 1556-1577.
[13] Kang C, Xiang S, Liao S, et al. Learning consistent feature representation for cross-modal multimedia retrieval [J]. IEEE Transactions on Multimedia, 2015, 17(3): 370-381.
[14] Lee H, Park Y, Seo H, et al. Self-knowledge distillation via dropout[J]. Computer Vision and Image Understanding, 2023, 233: 103720.
[15] Gu Y, Dong L, Wei F, et al. MiniLLM: Knowledge distillation of large language models[C]. The Twelfth International Conference on Learning Representations, 2024.
[16] Xia G S, Hu J, Hu F, et al. AID: A benchmark data set for performance evaluation of aerial scene classification [J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(7): 3965-3981.
[17] Yang Y, Newsam S. Bag-of-visual-words and spatial extensions for land-use classification[C]. Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems, 2010: 270-279.
[18] Lobry S, Marcos D, Murray J, et al. RSVQA: Visual question answering for remote sensing data [J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(12): 8555-8566.
[19] Liu H, Li C, Wu Q, et al. Visual instruction tuning[J]. Advances in neural information processing systems, 2024, 36.
[20] Chen J, Zhu D, Shen X, et al. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning[J]. arXiv:2310.09478, 2023.
[21] Wang X, Chen G, Qian G, et al. Large-scale multi-modal pre-trained models: A comprehensive survey[J]. Machine Intelligence Research, 2023, 20(4): 447-482.
[22] Zia U, Riaz M M, Ghafoor A. Transforming remote sensing images to textual descriptions[J]. International Journal of Applied Earth Observation and Geoinformation, 2022, 108: 102741.
[23] Zhang B, Chen T, Wang B. Curriculum-style local-to-global adaptation for cross-domain remote sensing image segmentation [J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-12.
[24] Liu H, Li C, Li Y, et al. Improved baselines with visual instruction tuning[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 26296-26306.
[25] Liu H, Li C, Wu Q, et al. Visual instruction tuning[J]. Advances in neural information processing systems, 2024, 36.
[26] Chiang W L, Li Z, Lin Z, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality[J]. [Online] Available at https://vicuna. lmsys. org (accessed 14 April 2023), 2023, 2(3): 6.
[27] 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.
[28] Tang R, Lu Y, Lin J. Natural language generation for effective knowledge distillation[C]. Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019), 2019: 202-208.
[29] Cho J H, Hariharan B. On the efficacy of knowledge distillation[C]. Proceedings of the IEEE/CVF international conference on computer vision, 2019: 4794-4802.
[30] Tay Y, Phan M C, Tuan L A, et al. Learning to rank question answer pairs with holographic dual lstm architecture [C]. Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval, 2017: 695-704.
[31] Gou J, Yu B, Maybank S J, et al. Knowledge distillation: A survey [J]. International Journal of Computer Vision, 2021, 129(6): 1789-1819.
[32] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback [J]. Advances in neural information processing systems, 2022, 35: 27730-27744.
[33] Williams R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning [J]. Machine learning, 1992, 8: 229-256.
[34] Skalse J, Howe N, Krasheninnikov D, et al. Defining and characterizing reward gaming[J]. Advances in Neural Information Processing Systems, 2022, 35: 9460-9471.
[35] Rahnemoonfar M, Chowdhury T, Sarkar A, et al. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding[J]. IEEE Access, 2021, 9: 89644-89654.
[36] Loshchilov I. Decoupled weight decay regularization [J]. arXiv:1711.05101, 2017.
[37] Bai J, Bai S, Yang S, et al. Qwen-vl: A frontier large vision-language model with versatile abilities [J]. arXiv: 2308.12966, 2023.
[38] Wang D, Zhang J, Du B, et al. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model [J]. Advances in Neural Information Processing Systems, 2024, 36. https://github.com/mbzuai-oryx/GeoChat