基于多模态融合的社交媒体文本地理位置预测方法

黄士多[1]，徐永昌[2]，艾浩军[2]
（1. 武汉市互联网舆情研究中心，湖北 武汉 430014;
2. 武汉大学国家网络安全学院，湖北 武汉 430072）

摘　要：挖掘社交媒体文本的地理位置信息能发现其空间关系，提出了基于多模态融合的社交媒体文本地理位置预测方法，利用文本获取的相关图片作为增强数据，构建图文数据集，以提高地理位置预测的准确性。多模态融合模型分别利用图片通道和文本通道提取两者的地理位置信息。同时，引入图文匹配模块对图文对进行降噪，解决图文不匹配问题。在 Geotext 数据集上进行的地理位置预测实验结果显示，与基线模型相比，中值误差距离降低了 18.8%，平均误差距离降低了 4.5%。

关键词：社交媒体；地理定位；多模态融合；信息挖掘
中图分类号：TP391.1
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2023183

A social media geolocation prediction method based on multimodal fusion

HUANG Shiduo[1], XU Yongchang[2], AI Haojun[2]
1. Wuhan Internet Public Opinion Research Center, Wuhan 430014, China
2. School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China

Abstract: Geographical information extracted from social media text reveals underlying spatial correlations. A geographical location prediction method for social media text based on multimodal fusion was proposed. By utilizing images associated with the text as augmented data, an integrated image-text dataset was constructed to enhance the accuracy of geographical location prediction. The multimodal fusion model employs separate channels for images and text to independently extract their respective geographical location information. Additionally, a text-image matching module was introduced to denoise the image-text pairs, effectively solving the issue of text-image misalignment. Experimental results on the Geotext dataset indicate that compared to the baseline model, the proposed method reduces the median error distance by 18.8% and the average error distance by 4.5%.

Key words: social media, geolocation, multimodal fusion, information mining

收稿日期：2023−07−03；**修回日期：**2023−09−13
通信作者：黄士多，249094702@qq.com
基金项目：国家自然科学基金资助项目（No.61971316）
Foundation Item: The National Natural Science Foundation of China (No.61971316)

0 引言

社交媒体的迅速发展引起了用户生成内容（user generated content，UGC）快速增长，同时加速了信息的传播。国内外学者提出挖掘 UGC 中隐含的地理位置信息的各种方法，广泛应用于实时应急感知系统[1-2]、在线营销[3]、人口迁移分析等方面。尽管文本提供了丰富的地理位置信息并且易于获取，近年来自然语言处理（natural language processing，NLP）领域的研究也日渐深入[4-5]，但现有基于文本的社交媒体地理位置预测方法未能有效解决文本内容所含地理位置信息不足的问题。

本文通过分析帖子文本内容获取相关图片，进一步构建“图片－文本”对照集。这种方法补充了文本中不足的地理位置信息。尽管这些从社交媒体文本获取的图片为基于文本的地理位置预测提供了一种数据增强途径，但该方法也带来了挑战，尤其是在现实生活中社交媒体用户发布的“图片－文本”对并不总是一一对应，有时获取的图片与文本的相关性可能较低。

为应对这些问题和挑战，本文提出了一种基于多模态融合的社交媒体文本地理位置预测框架。该框架利用社交媒体文本作为关键字，通过搜索引擎检索获取相关的增强数据，生成“图片－文本”对照集作为多模态融合模型的输入。该模型包括图片和文本两个独立的通道，分别用于提取图像和文本的地理位置信息。早期融合策略被用于最终预测层以输出具体的地理位置。考虑存在的图文不匹配问题，本文还引入了一个图文匹配模块，用于计算图文对的匹配度，并过滤与文本匹配度较低的图片。

在 Geotext 数据集上的大量实验结果显示，该方法相对于近年来其他方法具有更高的预测精度。本文的主要贡献如下，提出了一种基于多模态融合的社交媒体文本地理位置预测方法，通过将与文本相关的图片作为增强数据的方式，引入了图片中的地理位置信息，并采用多模态融合模型获得了更高的定位精度；设计了图文对的降噪方法，通过引入图文匹配模块计算图文相似度和图文匹配结果，对不匹配图文对进行降噪处理，并将图文相似度引入特征融合阶段调整图片地理信息对预测结果的影响。

1 相关工作

用户在社交媒体上的帖子反映了他们的地理空间语态，不同地区具有不同的词汇先验，帖子通常包括当地的实体或事件[6]，且不同地区的用户在方言、表达方式等内容上存在差异，生活在同一地区的人可能具有一些共同的语言模式或习惯，研究者通过这些特征预测社交媒体文本的地理位置或用户的地理位置。Cheng 等[7]首次提出并评估了仅基于文本内容的社交媒体地理位置预测概率框架，实验结果表明，50%的用户的真实位置与预测位置之间的误差距离小于 160 km，证明了仅基于帖子文本进行地理位置预测的可行性。Han 等[8]采用特征选择的方法获取位置指示词，并结合词袋模型进行地理位置预测。Cha 等[9]提出了基于稀疏编码和字典学习地理位置预测方法，将地理位置预测任务建模为分类问题。Lau 等[10]提出了根据用户的元数据和文本学习其位置指示词，并利用文本发布的时序模式进行地理位置与预测的端到端模型。Miyazaki 等[11]提出在知识库的基础上，利用与其他实体的关系预测用户的地理位置的方法。Fornaciari 等[12]提出了引入注意力机制的多任务卷积神经网络。Milusheva 等[13]提出了用于识别事故发生位置的社交媒体地理位置预测方法及数据集。Ambrosio-Aguilar 等[14]提出了针对西班牙语的社交媒体文本地理位置预测方法。Scherrer 等[15]提出了基于 Transformer 的双向编码器表示（bidirectional encoder representations from transformers，BERT）架构的 geoBERT，并研究了 BERT 架构的不同预训练模型对地理位置预测

精度的影响，并在 VarDial 竞赛取得了最优的效果。随着自然语言处理研究的深入，越来越多的研究者采用 Transformer、BERT 等深度学习模型作为主干。

多模态融合方法缩小模态（如图像、文本、音频等）间的异质性差异，同时保持各模态特定语义的完整性，从融合阶段的角度可以分为 3 类：早期融合[16]、晚期融合[17]以及结合前两者的混合融合方法。Li 等[18]提出图文的多模态编码器－解码器混合模型（bootstrapping language-image pre-training，BLIP），在各种下游任务上实现显著的性能改进，训练过程中使用的图像－文本匹配（image-textmatching，ITM）和图像－文本对比（image-textcontrastive，ITC）两个训练目标有效实现了图文的匹配判断。其中 ITC 的目的是通过促进正向的图像－文本对与负向的图像－文本对有相似的表示，对齐图片和文本的特征空间，通过一个相似度，计算出匹配的“图片－文本”对有更高的相似度值。对于每一个图片 I 和文本 T，相似度计算式如式（1）所示。

$$\begin{aligned}\boldsymbol{p}_m^{\mathrm{i2t}}(I)&=\frac{\exp(\boldsymbol{s}(I,T_m)/\tau)}{\sum_{m=1}^{M}\exp(\boldsymbol{s}(I,T_m)/\tau)}\\ \boldsymbol{p}_m^{\mathrm{t2i}}(T)&=\frac{\exp(\boldsymbol{s}(T,I_m)/\tau)}{\sum_{m=1}^{M}\exp(\boldsymbol{s}(T,I_m)/\tau)}\end{aligned}\tag{1}$$

其中，$\boldsymbol{s}(I,T_m)$ 和 $\boldsymbol{s}(T,I_m)$ 为文献[18]定义的图像和文本的相似性计算方法，M 是类别的总数，τ 是一个可学习参数。以 $\boldsymbol{y}^{\mathrm{i2t}}(I)$ 和 $\boldsymbol{y}^{\mathrm{t2i}}(T)$ 代表真实相似度的独热（One-hot）向量，其中不匹配的标签为 0，匹配的标签为 1。ITC 损失函数如式（2）所示。

$$\begin{aligned}\boldsymbol{L}_{\mathrm{itc}}=&\frac{1}{2}E_{(I,T)\in D}\\ &\left[\boldsymbol{H}(\boldsymbol{y}^{\mathrm{i2t}}(I),\boldsymbol{p}^{\mathrm{i2t}}(I))+\boldsymbol{H}(\boldsymbol{y}^{\mathrm{t2i}}(T),\boldsymbol{p}^{\mathrm{t2i}}(T))\right]\end{aligned}\tag{2}$$

ITM 帮助模型学习图像－文本多模态表示，捕捉图像和语言之间的细粒度对齐，是一种二分类任务，损失函数如式（3）所示。

$$\boldsymbol{L}_{\mathrm{itm}}=E_{(I,T)\in D}\boldsymbol{H}(\boldsymbol{y}^{\mathrm{itm}},\boldsymbol{p}^{\mathrm{itm}}(I,T))\tag{3}$$

其中，$\boldsymbol{y}^{\mathrm{itm}}$ 是一个二维的 One-hot 向量，表示图文匹配的真实标签。

以上述工作为基础，本文试图利用文本检索相关图片，并将其作为文本的增强数据，从而引入图片中的地理位置信息，并利用多模态融合模型预测地理位置。

2　网络框架

本文提出的基于多模态融合的社交媒体文本地理位置预测方法，采用早期融合方式的双通道特征融合方法，包含文本通道和图片通道，将社交媒体地理位置预测任务建模成分类问题。利用以社交媒体文本检索配图组成的“图片－文本”对作为输入，输出对应的地理位置区域。为降低“图片－文本”对中噪声图片的干扰，引入图文匹配判断模块，对相关性低的图片和文本进行降噪处理，防止噪声图片干扰模型，降低模型的预测精度，将处理后的图片和文本分别送入图片通道和文本通道提取特征，不同模态数据的特征在特征融合层进行融合，最后将图文融合特征送入包含一层全连接层的输出层，输出文本对应的地理位置区域或经度、纬度坐标，多模态融合社交媒体文本地理位置预测的网络架构如图 1 所示。

2.1　图文匹配模块

图文匹配模块由图片编码器、文本编码器和融合图片特征的文本编码器 3 个编码器组成，输入为“图片－文本”对，输出图文相似度 sim 和匹配结果 mr。sim 表示图片与文本的相似度，sim 越高，相似度越高。mr=0 说明图文匹配模块认为图片与文本不匹配，mr=1 表示图文匹配，图文匹配模块结构如图 2 所示。

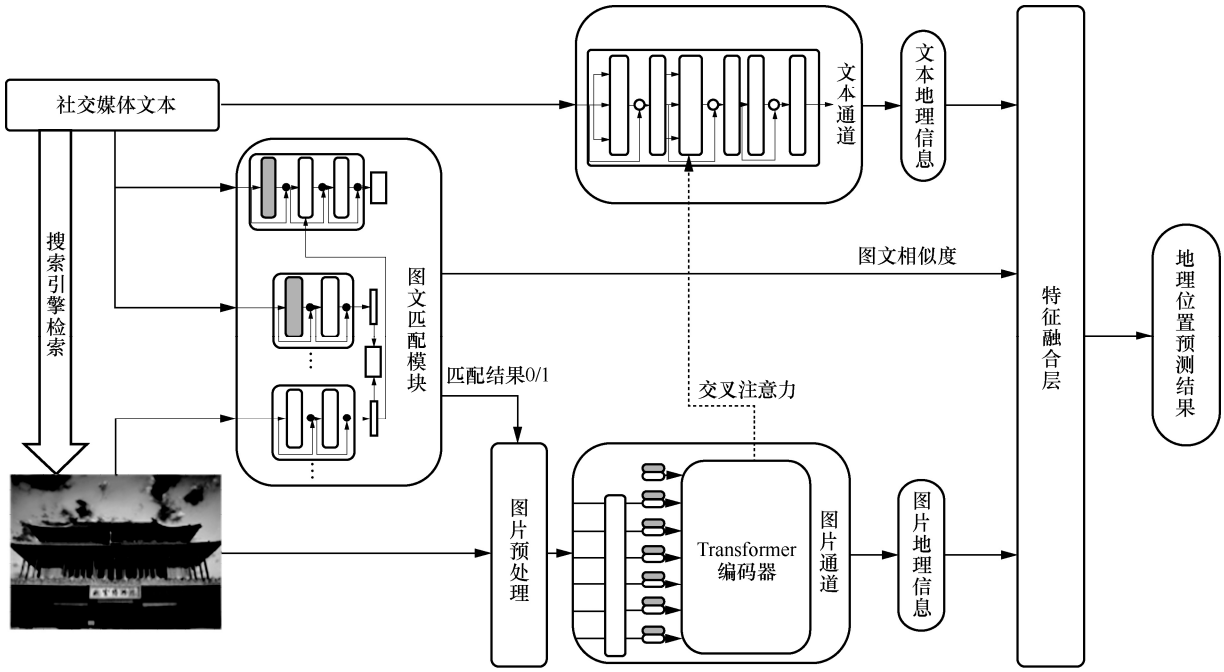


图 1　多模态融合社交媒体文本地理位置预测的网络架构

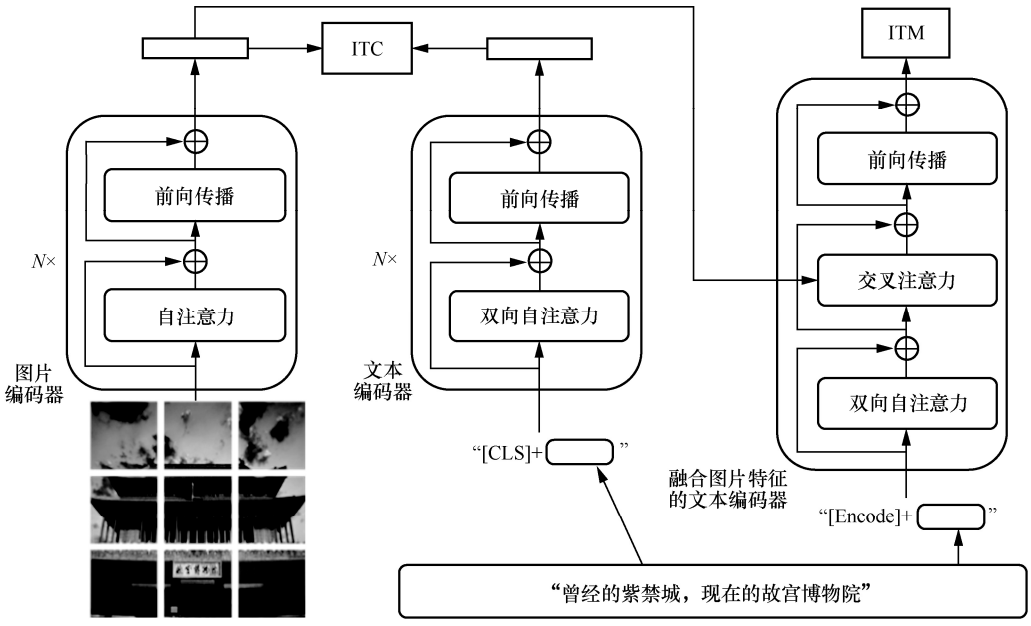


图 2　图文匹配模块结构

图片编码器与视觉 Transformer（visual Transformer，VIT）VIT-B/16 的架构相同，用于提取图像表示。文本编码器与 Transformer 编码器的架构相同，用于提取文本表示。融合图片特征的文本编码器在Transformer编码器的基础上加入了交叉注意力（cross-attention）机制，获取图片编码器的注意力。图文匹配模块的联合了 3 个训练目标，分别是图像－文本匹配 ITM、图像－文本对比 ITC 和掩蔽语言建模（masked language modeling，MLM）训练目标。ITC 和 ITM 已被证明是提高图像和语言理解的有效目标[16]。图文匹配模块的作用体现在图片和文本内容的理解和匹

配能力的提升，为此使用了基于 COCO 数据集的预训练模型，并对其在本文人工标记的图文匹配数据集上进行微调，微调过程中图文匹配模块的训练目标如式（4）所示。

$$\boldsymbol{L}_{\text{match}} = \boldsymbol{L}_{\text{itc}} + \boldsymbol{L}_{\text{itm}} \tag{4}$$

当 mr=1 时，将 sim 作为权重，在特征融合层对图片特征进行调整，从而根据图文相似度调整图文匹配样本中的图片特征对地理位置预测结果的影响。在该情况下，融合模型充分利用文本特征，有效降低噪声图片对社交媒体地理位置预测任务的影响。mr=0 表明“图片−文本”对数据集中该样本的图片与社交媒体文本不匹配，为噪声图片，在图片预处理阶段对该图像进行降噪处理，即将对应的图片的所有图像像素信息初始化为 0，成为一张纯黑色图片，作为一个特殊图片，降低了噪声图片特征对地理位置预测结果的影响。

2.2　图片通道和文本通道

图片通道和文本通道模型结构如图 3 所示。

（1）图片通道

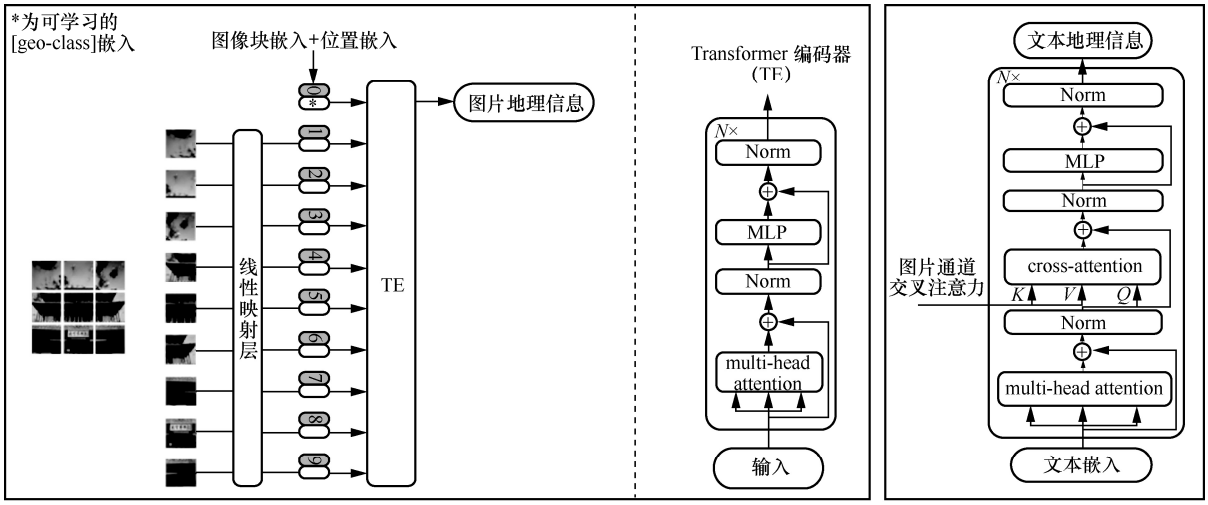
图片通道采用 VIT 架构，用于学习并提取对应图片的图片地理信息特征，并加入了[geo-class]向量用于提取图像的地理信息特征，该模型主要包括网格线性映射层、Transformer 编码器和[geo-class]特征提取层，如图 3（a）所示。其中[geo-class]向量的作用类似于自然语言处理中[CLS]向量，用于整合图片地理信息，可作为图片整体特征用于后续特征融合。Transformer 编码器由 N 个编码块重复堆叠组成，每个编码块包含多头注意力（multi-head attention）层、层标准化（layer normalization）层和多层感知机（multilayer perceptron，MLP）块，MLP 块包含两层线性全连接层、GELU 激活函数和 Dropout 层。

（2）文本通道

文本通道用于提取社交媒体文本中隐含的地理位置信息，采用如图 3（b）所示基于 BERT 框架的模型。首先对文本处理得到文本嵌入（embedded text），经过 N 个堆叠的编码块，同时将图片通道中的注意力作为 cross-attention 层的 K 和 V，每个文本通道编码块相比图片通道编码块多了一个 cross-attention 层。

2.3　特征融合层

在特征融合过程中引入了图文匹配模块的输出结果 sim，其数值范围为 0～1，sim 值越大说明图片和文本的匹配度与相关性越高，反之亦然。将 sim 引入特征融合过程，作为图片特征的权重，sim 值越高，图片特征在融合特征的重要性就越



（a）图片通道　　（b）文本通道

图 3　图片通道和文本通道模型结构

高。特征融合层将图片通道提取到的图片地理信息I_{geo}和文本通道提取到的文本地理信息T_{geo}融合为图文融合特征F，首先将图文相似度 sim 乘以I_{geo}，并将该结果与T_{geo}相加得到图文融合特征F，如式（5）所示。输出结果为地理位置区域类别，特征融合层如图 4 所示。

$$F = I_{geo} \cdot \mathrm{sim} + T_{geo} \tag{5}$$

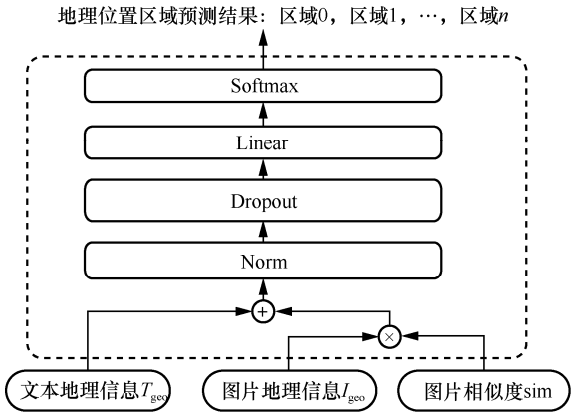


图 4　特征融合层

图文匹配结果 mr=0 表明“图片−文本”对数据集中该样本的图片与文本不匹配，无须调整融合特征F的计算方式。在这种情况下，融合模型充分利用文本地理信息特征预测其地理位置，有效降低了噪声图片对地理位置预测精度的影响。对于地理位置区域预测的任务，采用交叉熵作为训练目标，式（6）为损失函数表达式。

$$L = -\sum_{i=1}^{N} L_i = -\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} \lg \hat{y}_{ij} \tag{6}$$

其中，N是样本数量，M是类别数量，y_{ij}是第i个样本的真实类别j的指示变量（若第i个样本属于类别j，则$y_{ij}=1$，否则$y_{ij}=0$），$\hat{y}_{ij}$是模型预测的第i个样本属于类别j的概率。

3　实验与分析

3.1　实验设置

（1）数据采集和预处理

为了解决部分社交媒体样本仅包含文本，地理位置信息不充分的问题，本文对仅包含文本的社交媒体地理位置数据集 Geotext 进行了拓展，利用文本获取对应图片，自建“图片－文本”对数据集。文本数据的标签为用户发帖时所在的地理位置坐标，坐标中包含经度和纬度。

为此编写了一个利用 Bing 图片搜索引擎的 Python 爬虫程序，将经过预处理后 Geotext 数据集的文本作为关键字，爬取每一条文本对应的 JPG 或 PNG 格式图片，每条文本爬取一幅图片组成“图片－文本”对，图片的大小不低于 10 KB，不高于 300 KB。由于图片通道采用的模型的输入维度是[244, 244, 3]，将该图像大小放缩至长为 244 pixel，宽为 244 pixel 的包含 3 个颜色通道的 RGB 图像。

爬虫程序对所有文本获取了对应的图片后，生成了“图片－文本”对样本，每个样本的标签为原本文本对应的经度和纬度值，最终得到了 Geotext 图文对数据集，除部分文本样本在预处理阶段被删除和部分文本无法通过搜索引擎获取有效的配图外，相较于 Geotext 数据集，该图文数据集的训练集、验证集的数量有所减少，测试集的数量基本不变。数据集详细信息见表 1。

数据集中的用户位置以经度、纬度坐标进行标识，为了将社交媒体地理位置预测任务转为分类任务，采用k均值（k-means）聚类算法将样本数据分为 75 个类别，将地理位置预测任务转化为一个分类任务，每个类别代表一个地理位置区域，并将该类别作为“图片－文本”对样本的标签，

表 1　数据集详细信息（单位：条）

数据集名称	训练集样本量	验证集样本量	测试集样本量	每个帖子 Token 数量
Geotext 数据集	302 064	37 758	37 758	12
Geotext 图文对数据集	293 003	36 029	37 580	13

将每个区域内的经度和纬度的中位数作为该区域的预测位置用于计算后续评价指标。

由于需要对图文匹配模块进行微调，从 Geotext 图文对数据集的训练集中提取出 5%的样本数，共 14 650 条，人工标记这些样本中图文匹配结果，1 为匹配，0 为不匹配，该人工标记数据集用于微调图文匹配模块，提升图文匹配模块对图片和文本中地理信息的匹配性判断。

（2）评价指标与基线模型

为验证模型有效性，选用如下的两个评价指标[15]评估多层次对比学习框架（multiple contrastive learning framework，mCLF）的性能。

- 中值误差距离（单位为 km，下文记为 Median），数据集中用户的真实位置和预测位置之间误差距离的中位数。
- 平均误差距离（单位为 km，下文记为 Mean），数据集中用户的真实位置与预测位置之间的误差距离的平均数。

为验证基于多模态融合的社交媒体文本地理位置预测方法有效性，选用如下 5 个基线模型。

- geoBERT-cls[15]：一个基于 BERT 结构学习文本隐含的地理位置信息的地理位置预测分类模型。
- MDN[20]：一种在连续向量空间中嵌入二维位置，结合高斯分布的基于神经网络的方法。
- LR[21]：一种基于文本预测的用于彼此无连接的用户的混合分类方法。
- MLP[22]：一种基于神经网络的简单有效的文本用户地理定位模型。
- Sparse-Coding[23]：一种基于稀疏编码和字典学习的数据驱动的地理定位和区域分类方法。

（3）参数设置

由于图文匹配模块需要很强的理解能力，采用在 COCO 数据集上预训练过的 BLIP 的预训练模型“blip-itm-base-coco”，该模型具有较好的图片匹配分辨能力。为提升社交媒体的文本和配图的匹配度，本文从 Geotext 图文对训练集中采集了 5%的样本，对图文匹配度做了人工标记，按照 3:1:1 的比例分为训练集、验证集和测试集。图文匹配模块经过微调后，直接提升了在社交媒体地理位置预测任务中的图文匹配分辨能力，在后续模块的训练中不更新图文匹配模块的网络参数。

Geotext 图文对数据集的样本数量相对较少，图片通道采用在 ImageNet 数据集上预训练模型作为初始化模型，文本通道采用基于 BERT 架构的“bert-base-uncased”模型作为初始化模型，随后利用 Geotext 图文对数据集进行针对社交媒体地理位置预测任务的微调。

在微调过程中，训练参数设置如下，训练轮次为 30，采用 Adam 优化器，整体学习率为 4×10^{-5}。值得注意的是，由于图片通道的 VIT 预训练模型的中间网络层已经包含丰富的图片特征信息，因此本文将最后 3 层的学习率设置为 5×10^{-4}，学习率的增加能够更快地帮助图片通道模型适应针对图片中地理位置信息特征的学习。网络中的 Dropout 层的参数统一设置为 0.5。

3.2　实验结果与分析

为验证该方法在基于文本的社交媒体地理位置预测任务上的有效性、图文匹配模块的有效性，探究图片通道和文本通道采用不同模型以及该方法中不同模块对地理位置预测精度的影响，设计了如下实验：地理位置预测实验、图文匹配模块有效性验证实验、对比实验和消融实验。

（1）地理位置预测实验

本文提出的基于多模态融合的社交媒体文本地理位置预测方法在 Geotext 图文对数据集上进行了地理位置预测实验，并在 Median 和 Mean 两个指标中都取得了较好的成绩，分别是 308.54 km 和 552.89 km，基线模型实验结果见表 2。由于图片通道采用的是基于 VIT 架构的模型，文本通道

表 2　基线模型实验结果

方法	Geotext 数据集		Geotext 图文对数据集	
	Median/km	Mean/km	Median/km	Mean/km
MLP	389	844	—	—
MDN	412	865	—	—
LR	397	880	—	—
Sparse-Coding	425	581	—	—
geoBERT-cls	380.08	578.89	—	—
VIT+BERT	—	—	308.54	552.89

采用的是基于 BERT 架构的模型，下文将该方法记为 VIT+BERT。

由表 2 可以看出，本文方法在地理位置预测的两个指标上都优于基线模型，geoBERT-cls 采用了 BERT 为主干网络，与本文提出的基于多模态融合的社交媒体文本地理位置预测模型 VIT+BERT 有最接近的地理位置预测效果。可以看出，相比 geoBERT-cls 模型，VIT+BERT 模型在 Median 误差上降低了 23.2%，在 Mean 误差上降低了 4.5%，大幅度的性能提升归功于 VIT+BERT 模型利用搜索引擎获取与文本内容相关的配图作为数据增强，丰富了模型可以学习的知识。

（2）图文匹配模块有效性验证

利用文本内容作为关键字，通过图片搜索引擎获取的这些图片可能与对应文本的相关性较低，本文对图文匹配模块的有效性进行了实验，将图文匹配模块去除，记为 VIT+BERT-ITM，并与完整的基于多模态融合的社交媒体文本地理位置预测方法进行了对比，图文匹配模块有效性验证实验结果见表 3。

表 3　图文匹配模块有效性验证实验结果

方法	Median/km	Mean/km
VIT+BERT	308.54	552.89
VIT+BERT-ITM	415.53	607.01

去除图文匹配模块后，Median 误差增加了 106.99 km，约为 VIT+BERT 方法的 34.6%，Mean 误差增加了 54.12 km，约为 VIT+BERT 方法的 9.8%。该实验结果可以从两个角度理解，一是由于图文匹配模块通过判断是否匹配对数据集中的图片进行了降噪，降低了无效图片对模型的影响；二是由于图文匹配模块通过利用图文匹配模块输出的图文相似度，调节特征融合过程中图片特征的权重，实现对样本中图片的细粒度降噪。

为了更直观地表现图文匹配模块的效果，图文匹配模块实例见表 4，第一组图文对匹配结果为 1，即图片与文本匹配，图文相似度为 0.81，根据文本和图片内容可以看出地理位置都是 Hollywood（好莱坞），图文匹配模块的输出结果是正确的。而第二组样本的文本内容是“The night of New York is intoxicating, hoping to live here forever（纽约的夜色令人陶醉，真希望一直生活在这里）”，图片中的画面是旧金山金门大桥在白天的景色，并不一致，图文匹配模块给出的匹配结果为 0，即图文不匹配，图文相似度为 0.22。这说明图文匹配模块能有效辨别不匹配的图文对。

（3）对比实验

以下对图片通道和文本通道采用不同的主流模型在 Geotext 图文对数据集上进行了实验，并从不同角度进行了评估。实验采用的图片通道的模型包括 VIT 和 ResNet18，采用的文本通道的模型包括 BERT 和 BiLSTM。

对比实验结果见表 5，随着模型结构复杂度的

表 4　图文匹配模块实例

图片	文本	匹配结果	图文相似度
	The sky is blue here, waiting for you in Hollywood	1	0.81
	The night of New York is intoxicating, hoping to live here forever	0	0.22

表 5　对比实验结果

方法	模型参数量	推理速度/（条·秒$^{-1}$）	Median/km	Mean/km
RESNET18+BiLSTM	37.96×106	116.53	421.63	672.44
VIT+BiLSTM	112.78×106	114.44	413.49	649.07
RESNET18+BERT	121.11×106	86.04	344.25	567.10
VIT+BERT	195.53×106	69.50	308.54	552.89

增加，地理位置预测准确度也有所提升，在地理位置预测准确度上表现最好的仍是 VIT+BERT 方法，采用 BiLSTM 作为文本通道模型的表现明显弱于采用 BERT 模型，可能有两个原因，一是采用的 BiLSTM 的预训练模型不适用于该任务，二是 BiLSTM 模型的参数较少，导致其语义理解能力弱于 BERT。

显示随着模型参数量的上升，推理速度逐渐降低，可以根据对实时性和地理位置精度的不同需求选择合适的图片通道和文本通道模型。值得注意的是，RESNET18+BiLSTM 模型参数量大幅度减少，而推理速度却仅有 2.09 条/秒的提升，这可能是因为模型对图文数据的预处理过程较长。

（4）消融实验

根据基于多模态融合的社交媒体文本地理位置预测方法的结构，本文在 Geotext 图文对数据集上进行了消融实验，实验结果包含无图文匹配模块、无图片通道模型、无 cross-attention 的结果，消融实验结果见表 6。

表 6　消融实验结果

方法	Median/km	Mean/km
VIT+BERT	307.38	552.91
无 cross-attention	310.37	553.41
无图片通道	381.51	580.17
无图文匹配模块	409.37	609.21

可以得出如下结论。

- 当去除 cross-attention 时，模型性能略有下降，Median 变为 310.37 km，Mean 变为 553.41 km。这表明 cross-attention 帮助模型更好地融合图像和文本信息。
- 当去除图片通道时，模型性能出现明显下降，Median 变为 381.51 km，Mean 变为 580.17 km。这说明利用图片作为社交媒体文本的增强数据，并引入图片通道在整个模型中起到重要作用，提供了更多有关地理位置的关键信息。
- 当去除图文匹配模块时，模型性能下降最为明显，Median 变为 409.73 km，Mean 变为 609.21 km。这表明图文匹配模块在整个

模型中起到了核心作用，有效地降低了噪声图片的影响，更好地融合图像和文本信息，极大地提高了模型的性能。

消融实验结果表明，VIT+BERT 模型中的 cross-attention、图片通道和图文匹配模块都对模型性能有优化效果，这些组件共同帮助模型实现了更好的图文融合和地理位置预测性能。

4 结束语

本文提出了一种基于多模态融合的社交媒体文本地理位置预测方法，利用文本获取相关图片作为增强数据，生成“图片－文本”对数据集，并引入图文匹配模块对生成的“图片－文本”对进行降噪。多模态融合模型结合文本和图片多模态特征获得了更加丰富的语义信息和地理位置信息，在 Geotext 数据集上显著提高了地理位置预测精度。模型的可扩展性便于引入更先进的图片通道模型、文本通道模型特别是大模型和多模态分析将进一步提升社交媒体地理位置预测的性能。

参考文献：

[1] SAKAKI T, OKAZAKI M, MATSUO Y. Earthquake shakes Twitter users: real-time event detection by social sensors[C]// Proceedings of the 19th International Conference on World Wide Web. New York: ACM Press, 2010: 851-860.

[2] 邬柯杰, 吴吉东, 叶梦琪. 社交媒体数据在自然灾害应急管理中的应用研究综述[J]. 地理科学进展, 2020, 39(8): 1412-1422.
WU K J, WU J D, YE M Q. A review on the application of social media data in natural disaster emergency management[J]. Progress in Geography, 2020, 39(8): 1412-1422.

[3] KINSELLA S, MURDOCK V, O’HARE N. “I’m eating a sandwich in Glasgow”: modeling locations with tweets[C]// Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents. New York: ACM Press, 2011: 61-68.

[4] 杨腾飞, 解吉波, 闫东川, 等. 基于深度学习的社交媒体情感信息抽取及其在灾情分析中的应用研究[J]. 地理与地理信息科学, 2020, 36(2): 62-68, F0002.
YANG T F, XIE J B, YAN D C, et al. Extracting sentiment information from social media based on deep learning and the research on disaster reduction[J]. Geography and Geo-Information Science, 2020, 36(2): 62-68, F0002.

[5] 徐永昌, 黄士多, 艾浩军. 基于对比学习的社交媒体地理位置预测方法[J]. 电信科学, 2023, 39(8): 58-68.
XU Y C, HUANG S D, AI H J. A social media geolocation method based on comparative learning[J]. Telecommunications Science, 2023, 39(8): 58-68.

[6] WING B, BALDRIDGE J. Simple supervised document geolocation with geodesic grids[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. OR: The Association for Computer Linguistics, 2011: 955-964.

[7] CHENG Z Y, CAVERLEE J, LEE K. You are where you tweet: a content-based approach to geo-locating twitter users[C]// Proceedings of the 19th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2010: 759-768.

[8] HAN B, COOK P, BALDWIN T. Geolocation prediction in social media data by finding location indicative words[C]// COLING. Indian: Indian Institute of Technology Bombay, 2012: 1045-1062.

[9] CHA M, GWON Y, KUNG H. Twitter geolocation and regional classification via sparse coding[J]. Proceedings of the International AAAI Conference on Web and Social Media, 2021, 9(1): 582-585.

[10] LAU J H, CHI L, TRAN K N, et al. End-to-end network for twitter geolocation prediction and hashing[J]. arXiv preprint, 2017, arXiv:1710.04802.

[11] MIYAZAKI T, RAHIMI A, COHN T, et al. Twitter geolocation using knowledge-based methods[C]//Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text. Stroudsburg: Association for Computational Linguistics, 2018: 7-16.

[12] FORNACIARI T, HOVY D. Geolocation with attention-based multitask learning models[C]//Proceedings of the 5th Workshop on Noisy User-generated Text(W-NUT 2019). Stroudsburg: Association for Computational Linguistics, 2019: 217-223.

[13] MILUSHEVA S, MARTY R, BEDOYA G, et al. Applying machine learning and geolocation techniques to social media data (Twitter) to develop a resource for urban planning[J]. PLoS One, 2021, 16(2): e0244317.

[14] AMBROSIO-AGUILAR A D, BÁRCENAS E, MOLERO-CASTILLO G, et al. Geolocation of tweets in Spanish with transformer encoders[C]//Proceedings of 2021 9th International Conference in Software Engineering Research and Innovation(CONISOFT). Piscataway: IEEE Press, 2021: 227-231.

[15] SCHERRER Y, LJUBESIC N. Social media variety geolocation with geobert[C]//Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial). Asso-

ciation for Computational Linguistics, 2021: 135-140.
[16] KARPATHY A, TODERICI G, SHETTY S, et al. Large-scale video classification with convolutional neural networks[C]// Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 1725-1732.
[17] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv preprint, 2015, arXiv: 1503.02531.
[18] LI J, LI D, XIONG C, et al. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//Proceedings of Machine Learning Research: volume 162 ICML. PMLR, 2022: 12888-12900.
[19] CHAKRAVARTHI B R, GAMAN M, IONESCU R T, et al. Findings of the vardial evaluation campaign 2021[C]//Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial). Association for Computational Linguistics, 2021: 1-11.
[20] BACKSTROM L, SUN E, MARLOW C. Find me if you can: improving geographical prediction with social and spatial proximity[C]//Proceedings of the 19th International Conference on World Wide Web. New York: ACM Press, 2010: 61-70.
[21] DAVIS C A Jr, PAPPA G L, DE OLIVEIRA D R R, et al. Inferring the location of twitter messages based on user relationships[J]. Transactions in GIS, 2011, 15(6): 735-751.
[22] JURGENS D. That's what friends are for: inferring location in online social media platforms based on social relationships[J]. Proceedings of the International AAAI Conference on Web and Social Media, 2021, 7(1): 273-282.
[23] ROUT D, BONTCHEVA K, PREOŢIUC-PIETRO D, et al. Where's @wally? : a classification approach to geolocating users based on their social ties[C]//Proceedings of the 24th ACM Conference on Hypertext and Social Media. New York: ACM Press, 2013: 11-20.

[作者简介]

黄士多（1965− ），男，武汉市互联网舆情研究中心副研究员、主任，主要研究方向为网络舆情和社交媒体分析。

徐永昌（1998− ），男，武汉大学国家网络安全学院硕士生，主要研究方向为普适计算。

艾浩军（1972− ），男，博士，武汉大学国家网络安全学院副教授，主要研究方向为普适计算和室内定位。