



基于对比学习的社交媒体地理位置预测方法

徐永昌¹, 黄士多², 艾浩军¹

(1. 空天信息安全与可信计算教育部重点实验室, 武汉大学国家网络安全学院, 湖北 武汉 430072;
2. 武汉市互联网舆情研究中心, 湖北 武汉 430014)

摘要: 以往基于社交媒体文本的定位方法主要集中在将文本语义空间映射到地理位置空间, 忽略了文本之间的语义相关性和地理位置之间的距离相关性。提出了一种新的无监督多层次对比学习框架, 并设计了 3 个对比学习模块: 语义学习模块、位置学习模块和跨层次学习模块。首先利用 Transformer 编码器获取文本的语义表示, 以无监督的对比学习方式, 聚拢位置相近文本之间的语义表示和地理表示, 随后进行有监督训练, 输出地理位置分类或回归结果。在 4 个数据集上与 5 个基线模型的对比实验结果表明, 该框架有效地提升了社交媒体地理定位的准确性。

关键词: 社交媒体; 地理定位; 对比学习; 信息挖掘; Transformer

中图分类号: TP391.1

文献标志码: A

doi: 10.11959/j.issn.1000-0801.2023154

A social media geolocation method based on comparative learning

XU Yongchang¹, HUANG Shiduo², AI Haojun¹

1. Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China

2. Internet Public Opinion Research Center of Wuhan, Wuhan 430014, China

Abstract: Previous work on social media text-based geolocation focused on mapping language semantic space to geospatial space, which ignores the semantic correlation between social media texts and the distance correlation between geographical locations. To take advantage of these correlations, mCLF, a new unsupervised multiple-level contrastive learning framework was proposed, three contrastive learning modules were designed: a semantic learning module, a location learning module, and a cross-learning module. Transformer encoder was used to obtain semantic representation of posts, utilizing unsupervised contrastive learning method to decrease the distance of semantic representations and location representations of posts with near locations, and then fine-tuned the model with supervised method for geographic location regression or classification outputs. Compared with five baseline methods, extensive experiments based on four datasets demonstrate the effectiveness of the proposed framework.

Key words: social media, geolocation, contrastive learning, information mining, Transformer

收稿日期: 2023-03-16; 修回日期: 2023-08-07

通信作者: 徐永昌, 940726757@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

移动互联网加速了社交媒体的快速发展,用户生成内容(user generated content, UGC)快速增长, Twitter、Facebook、微博等社交媒体上信息快速传播。研究者可以从中挖掘出与地理位置相关的隐性信息,用于实时应急感知系统^[1]、在线营销^[2]、公共卫生监测^[3]、人口迁移分析等。近年来已经开发了许多社交媒体地理定位方法,包括特征选择、自然语言处理(natural language processing, NLP)和图神经网络(graph neural network, GNN)模型。一般来说,现有的方法分为两类。第一类通过特征选择方法或相应的 NLP 模型对推文中的自然语言文本进行分析;第二类通常采用社会网络知识,利用 GNN 获取位置网络和社会网络的地理嵌入。然后使用地理嵌入预测用户的位置,或者结合帖子文本训练地理定位模型^[4]。但由于社交网络的稀疏性和获取社交网络数据困难,第二类方法的实用性受到了限制,重点研究 NLP 方法。

目前的 NLP 地理定位模型有两个缺点。第一,忽略了文本之间的语义关联,现有方法将社交媒体文本地理定位看作一个从帖子语义空间到地理位置空间的独立映射问题。一方面,帖子内容过于灵活,良好的语义表示能够有效地表达帖子内容,从而预测得到更准确的位置;另一方面,由于社交媒体的快速发展,没有位置标签的 UGC 数据相对更容易获得,因此,挖掘无标签社会文本中的语义关联可以有效地帮助提高定位的准确性。第二,忽视了地点之间的地理空间关联,之前的工作忽略了相似帖子之间的地理坐标相似性,例如,报道同一起交通事故的帖子的用户很可能在同一地区。

为了解决上述问题,提出了一个用于社交媒体地理定位的框架——多层次对比学习框架(multiple contrastive learning framework, mCLF),

该框架可用于分类和回归任务,主要贡献如下。

(1) 提出了一种新颖有效的用于社交媒体定位的对比学习框架 mCLF, 其中, 引入无监督对比学习获得语言语义的相似性, 有监督学习获得社交媒体文本之间的隐性地理位置关系。

(2) 在对比学习框架中引入了语义层、地理位置层和语义-地理的跨层次 3 个不同的模块, 这些模块的协作使得文本语义的地理定位变得更加准确。

(3) 在对比框架下, 提出了一种用于回归任务地理位置层次对比学习模块的相似度表达式 GeoSim, 支持细粒度位置预测, 并在 4 个社交媒体地理定位数据集上进行了大量的实验, 结果表明, 该方法优于主流模型。

1 相关工作

以往关于社交媒体用户地理定位的研究大多采用分析帖子中使用的自然语言的方法。用户在社交媒体上的帖子反映了他们的地理空间语态, 因为不同地区的词汇先验知识不同, 帖子通常包括当地的实体或事件^[5]。此外, 生活在同一个地区的人们可能有一些共同的语言模式或习惯。例如, Han 等^[6]采用特征选择方法发现地理位置指示词; Wing 等^[5]利用语言使用的地理差异, 对不同地区进行了单词分布估计; Rahimi 等^[7]提出从用户帖子中提取词袋特征。然而, 这些方法只是将文本映射到地理位置, 而不考虑句子之间的相关性。除了用户在社交媒体上发布的帖子, 用户与用户之间的社交互动也可以用于社交媒体地理定位, 因为同一地区的相邻用户彼此联系的机会更大。例如, Jurgens^[8]利用了基于图的半监督标签传播, 依赖用户关系中的位置冗余; Wang 等^[9]将从定位网络 and 社交网络中获得的地理嵌入作为社交媒体地理定位的特征。一些方法同时考虑了帖子文本和社会网络知识, 如利用社会网络中的文本频率和异质特征训练^[10-11]。Rahimi 等^[12]利用



文本和网络上下文进行社交媒体地理定位。Scalia 等^[13]提出了利用社交媒体文本和社交网络对紧急事件识别和有关人员的地理定位（context-aware geolocation of emergency-related posts, CIME）方法，然而，公众获取第三方服务商保存的社交网络信息并不容易，空间上相距较远的用户之间也有密切的社交媒体联系。社交网络的稀疏性会导致性能不理想和不稳定^[14]。

句子表示学习，即将句子编码为固定维度的向量，捕捉输入之间的语义关系，是自然语言处理中的一个基本问题，可以有效地帮助执行下游任务。以往的句子表示学习工作可以分为有监督方法^[15-16]和无监督方法^[17-18]。有监督方法采用带标签的自然语言推理（natural language inference, NLI）和并行数据作为监督。Alexis 等^[15]首先提出在 NLI 数据集上对 Siamese 模型进行微调。Reimers 等^[16]和 Cer 等^[19]将该策略扩展到其他编码器和预训练模型。一些研究人员更加关注正则化句子嵌入以缓解表征退化问题^[20-22]，结果表明，该模型较之前的模型有明显的改进。无监督方法能够利用自然语言文本语料库中嵌入的内在语义信息调整训练目标以适应

语义文本相似性（semantic textual similarity, STS）任务。最近的研究多利用对比学习目标将语义差别较大的句子推开，语义相似的句子拉近，从而正则化嵌入空间（embedding space）^[23-25]。一些研究者通过数据增强利用对比目标捕捉句子之间的语义相似性^[26-27]。SimCSE^[17]仅利用了 Transformer 中的丢弃掩盖（Dropout Mask），使相同的句子输入得到不同的隐含层表示，这些隐含层表示作为彼此的增强数据，在有监督和无监督方法下在 STS 任务上表现得更加出色。

在上述工作的激励下，提出了一种无监督方法下的 mCLF，在语义层面提高句子理解和推理能力，在地理位置层面提高句子间地理位置差异的学习能力。

2 mCLF

mCLF 总体架构如图 1 所示。在基于文本的地理定位模型中，首先将文本输入映射到单词嵌入空间，利用 Transformer 编码器生成上下文化表示，使模型能够学习文本的语义知识。随后采用一个线性层将语义表示转化为地理特征表示。通

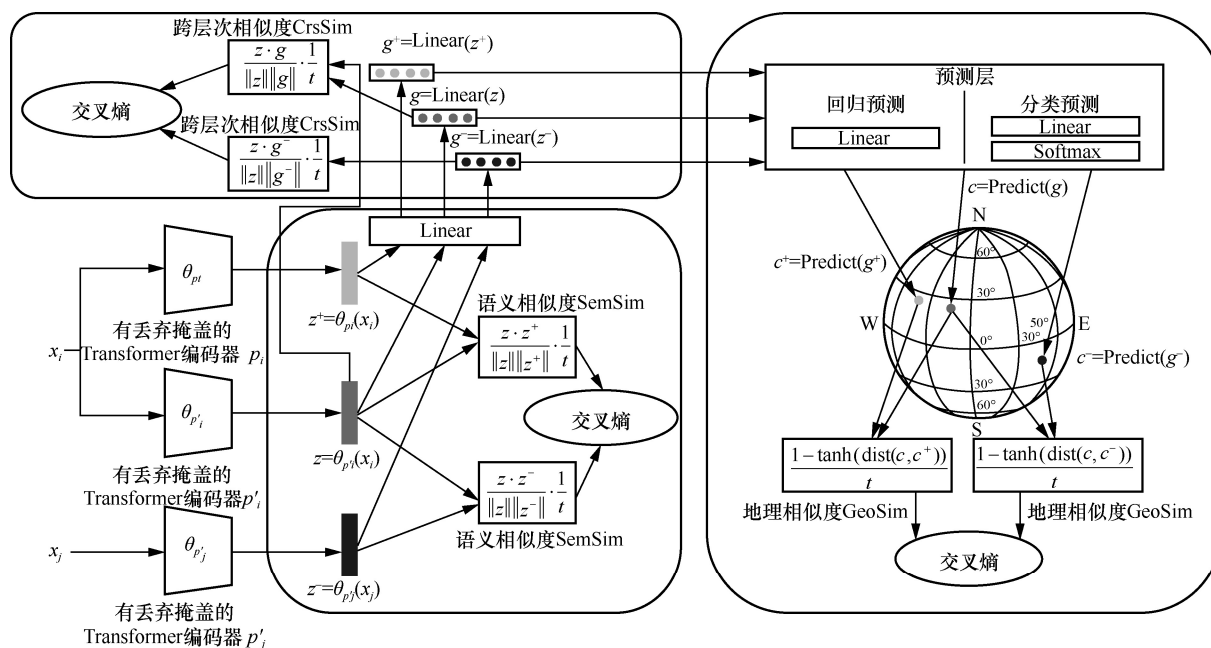


图1 mCLF 总体架构

过一个预测层, 将地理特征表示投影到特定位置的坐标(纬度和经度)。在该模型的基础上, 进一步设计了 3 个对比学习模块, 用于学习文本之间的语义距离、地理位置距离和跨层次距离, 这有助于提高模型对文本之间这些距离的理解和推断能力, 更准确地预测用户的地理位置。

2.1 对比学习

对比学习通过将语义相关的句子拉近、将语义无关的句子推开的方式学习有效表示的方法^[28]。前提是有一个成对的句子数据集 $D = (x_i, x'_i)$, 其中, x'_i 包含 x_i^+ 和 x_i^- , 成对样本中的 x_i 与 x_i^+ 在语义上是高度相关的, 与 x_i^- 的相关度较低。参考 SimCSE^[17] 中的无监督对比框架和 (x_i, x_i^+) 对的构造思想, 将文本两次输入 Transformer 编码器 θ , 不同的 Dropout Mask p 和 p' 可以得到两个相似的嵌入, 这是由于 Transformer 标准训练中放置在全连接层上的 Dropout Mask 和注意力概率 (attention probabilities, prob=0.1)。设置 $z = \theta_{p_i}(x_i)$, 其中, p_i 是随机 Dropout Mask, z 为 x_i 的语义表示。相对于 z , 正面实例 $z^+ = \theta_{p_i}(x_i)$ 具有相似的语义信息, 而负面实例 $z^- = \theta_{p'_i}(x_i)$ 具有不同的语义信息。批 (batch) 处理大小为 N , 且 $j \neq i |_{i,j \in N}$, 其中, 正对为 $z^+ = \theta_{p_i}(x_i)$, 负对为 $z^- = \theta_{p'_i}(x_i)$ 。提出了 3 个对比学习模块, 具体如下。

(1) 语义层次的对比学习 (semantic contrastive learning, SemCL)

为了增强模型学习句子间语义距离的特定能力, Transformer 编码器先将一个句子 x 编码为语义表示 z , 其中包含了对应句子的语义信息。语义相似度表达式 (1) 计算 z 和 z' 之间的余弦相似度表示 x 和 x' 之间的语义距离。

$$\text{SemSim}(z, z') = \frac{z \cdot z'}{\|z\| \|z'\|} \cdot \frac{1}{t} = \frac{\sum_{i=1}^n z_i \times z'_i}{\sqrt{\sum_{i=1}^n (z_i)^2} \times \sqrt{\sum_{i=1}^n (z'_i)^2}} \cdot \frac{1}{t} \quad (1)$$

其中, t 为比例因子, 以调整语义层次对比学习的重要性。语义距离较近的句子编码表示具有较高的相似度, 反之则相似度较低。

(2) 跨层次对比学习 (cross-level contrastive learning, CrsCL)

线性层将语义表示 z 映射到地理特征空间, 得到地理特征表示 g 。为了增强同一句子的语义和地理特征表示之间的相关性, 添加了一个跨层次相似度引导模型保持同一句子的语义和地理一致性, 跨层次相似度如式 (2) 所示。

$$\text{CrsSim}(z, g) = \frac{\sum_{i=1}^n z_i \times g_i}{\sqrt{\sum_{i=1}^n (z_i)^2} \times \sqrt{\sum_{i=1}^n (g_i)^2}} \cdot \frac{1}{t} \quad (2)$$

以句子 x_1 和 x_2 为例, 跨层次距离的设计是为了使模型学习 x_1 的语义表示 z_1 和地理特征表示 g_1 之间有更高的相关性, 以及 x_1 的语义表示 z_1 和 x_2 的地理特征表示 g_2 之间有更大的差异。

(3) 地理位置层次的对比学习 (geolocation level contrastive learning, GeoCL)

利用一个预测层分别预测两个句子的坐标 $c(\text{lat}, \text{lon})$ 和 $c'(\text{lat}', \text{lon}')$ 。为了提高模型学习句子间地理距离的能力, 设计式 (3) 计算句子的地理相似度。

$$\text{GeoSim}(c, c') = \frac{1 - \tanh(\text{dist}(c, c'))}{t} \quad (3)$$

其中, $\text{dist}(c, c')$ 用于计算 c 和 c' 之间的欧氏距离; $\tanh()$ 是一个激活函数, 将欧氏距离的值映射至 0~1。地理位置较近的用户发布的帖子具有较高的地理相似性, 而地理位置较远的用户发布的帖子具有较低的地理相似性。

2.2 损失函数

为了保持语义学习, 同时解决句子间语义距离和地理位置距离被忽略的问题, 设计了一个加权损失函数, 其中包含语义距离、地理距离、跨层次距离和掩码语言模型 (masked language



model, MLM) 损失。MLM 损失帮助 Transformer 编码器学习社交媒体帖子的语义知识。在预训练阶段, 损失函数赋予文本的语义距离和地理位置距离较高的权重。因此, 对比学习训练目标驱动语义和地理上相似句子的语义和地理位置聚集, 而负对样本之间的距离则更远。加权损失函数 L_{total} 由语义距离损失 L_{SemCL} 、跨层次距离损失 L_{CrsCL} 、地理位置距离损失 L_{GeoCL} 和 MLM 损失 L_{MLM} 组成, 损失函数如式 (4) ~ 式 (7) 所示。

$$L_{\text{SemCL}} = -\lg \frac{e^{\text{SemSim}(z_i^{p_i}, z_i^{p_i})}}{\sum_{j=1}^N e^{\text{SemSim}(z_i^{p_i}, z_j^{p_j})}} \quad (4)$$

$$L_{\text{CrsCL}} = -\lg \frac{e^{\text{CrsSim}(z_i^{p_i}, g_i^{p_i})}}{\sum_{j=1}^N e^{\text{CrsSim}(z_i^{p_i}, g_j^{p_j})}} \quad (5)$$

$$L_{\text{GeoCL}} = -\lg \frac{e^{\text{GeoSim}(c_i^{p_i}, c_i^{p_i})}}{\sum_{j=1}^N e^{\text{GeoSim}(c_i^{p_i}, c_j^{p_j})}} \quad (6)$$

$$L_{\text{total}} = L_{\text{MLM}} + L_{\text{CrsCL}} + \alpha \cdot L_{\text{SemCL}} + \beta \cdot L_{\text{GeoCL}} \quad (7)$$

3 实验与分析

本节首先介绍实验的实验设置和实验结果, 并对实验结果进行了详细的分析。

3.1 实验设置

本节详细介绍实验中的数据集、评价指标、基线模型及实验参数设置。

(1) 数据集

采用 4 个数据集评估 mCLF, 分别是 BCMS、DE-AT、CH 和 Geotext, 它们基于两个社交媒体平台: Jodel 和 Twitter, 数据集信息见表 1。Geotext 是 Twitter 用户地理定位常用的数据集^[29], BCMS、CH、DE-AT 数据集由 VarDial 评估活动^[30]提供。参照文献[29]的方法, 对其进行数据过滤和预处理, 每个数据集被分为训练集、验证集和测试集, 文本数据被标记为用户发帖时所在的地理位置坐标, 包含经度和纬度。

表 1 数据集信息

数据集名称	训练集	验证集	测试集	每个帖子 Token 数量
BCMS	353 953	38 013	4 189	13
DE-AT	318 487	2 416	31 515	69
CH	25 261	29 122	2 438	50
Geotext	302 064	37 758	37 758	12

(2) 评价指标与基线模型

为验证模型有效性, 选用如下两个评价指标^[31]评估 mCLF 的性能。

- 中值误差距离 (Median, 单位为 km), 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 (Mean, 单位为 km), 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。

在文本地理位置共享任务中, 这两个指标在大多数情况下显示出相关性。在某些情况下, Median 在小型数据集中比其他数据集具有更大的说服力, 因为离群值对小型数据集的平均误差距离有更大的影响。

为证明本框架的有效性, 选取如下几个较为经典或现阶段定位效果排名靠前的模型作为基线模型。

- geoBERT^[31]: 一个利用 BERT 体系结构学习文本地理特征和地理信息的地理位置共享任务的回归模型, geoBERT-clf 是基于相同体系结构的分类模型。
- MDN^[32]: 在连续向量空间中嵌入二维位置的神经网络模型。
- LR^[33]: 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^[7]: 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^[34]: 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

首先使用无监督方式对模型进行预训练,随后使用有监督方式对模型进行微调。编码器遵循 BERT-base 的设计,包含 12 个隐含层、12 个自注意力头,输出 768 维张量,参数量为 110M ($1\text{M}=1 \times 10^6$)。VarDial 提出的两种场景,受限场景下仅可使用上述提供数据集,不受限场景可以使用其他数据集和预训练好的模型^[31]。

在受限场景中,除了提供的训练数据,任何数据都不能使用,为每个数据集获取标记器,并从头开始预训练 Transformer 编码器。词汇量大的模型比词汇量小的模型产生的距离更小^[31]、收敛速度更快,在受限场景中将词汇量固定设置为 30 000。在不受限的情况下,根据不同的数据集的语言选择公开预训练好的 Transformer 编码器。BCMS 数据集的 Transformer 编码器选择 BERTic1 模型,DE-AT 和 CH 数据集的 Transformer 编码器选择 DBMDZ2 模型,将 Twitter-US^[35]部分训练数据融入 Geotext 数据集作为补充。

回归和分类是社交媒体地理定位任务中两种具体的建模形式。mCLF 同时支持两种建模。在回归建模中,模型以文本为输入,输出对应文本的经度、纬度坐标。在分类建模中,首先采用 K 均值 (K -means) 策略将数据集划分为 75 个聚类作为标签,用于预测地理位置区域。实验是在 NVIDIA RTX-3090 GPU 平台上,利用 PyTorch 工具包实现的。batch 的大小 N 设置为 32。在 3 个相似度表达式中,初始学习率 $\text{lr}=3 \times 10^{-5}$,比例因子 t 参考文献[17]设置为 0.15。

3.2 损失函数权重

定义两个权重 α 和 β 平衡语义层面和地理位置层面对比学习。为了探索 SemCL 和 GeoCL 权重对模型性能的影响,在 BCMS 数据集上进行了实验,为使 3 种对比学习目标的权重组成不同比例,固定 CrsCL 的权重为 1,在 $\{0.5, 1.0, 2.0\}$ 范围内改变 α ,在 $\{0.5, 1.0, 1.5, 2.0, 3.0, 4.0\}$ 范围内改变

β 。不同权重 α 和 β 的影响如图 2 所示。基于文本的社交媒体地理定位的前提是模型对文本有足够的理解和推理能力。随着 α 的增加,模型区分句子语义距离的能力增强。句子的语义表示在映射到地理空间时保持了这种区别性,从而获得了更高的地理定位精度,特别是中位数距离误差。当 $\alpha=2$ 和 $\beta=0.5$ 时,取得了最好的结果,说明语言语义理解和推理能力对社交媒体地理定位的重要性,合理利用语义层面和地理定位层面的对比学习可以有效提高社交媒体地理定位的准确性。因此,在下述实验中设置 $\alpha=2$ 和 $\beta=0.5$ 。

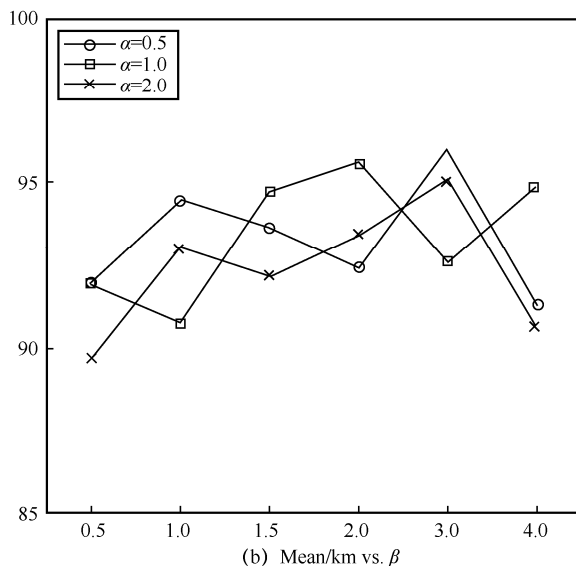
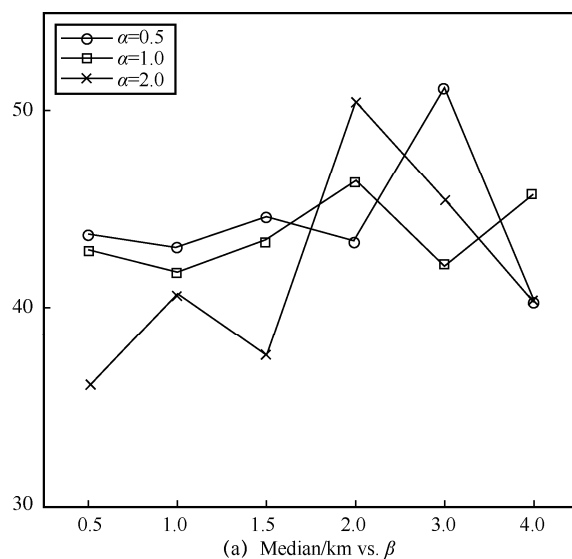


图2 不同权重 α 和 β 的影响



3.3 实验结果

同模型的社交媒体定位精度比较见表 2, 给出了不同方法在 4 个数据集上的实验结果。在回归和分类方法上, mCLF 在所有数据集上都取得了最佳的地理定位效果。在 Geotext 数据集上, MLP、MDN 和 LR 这些基线方法都使用 L_2 范数标准化词袋特征作为给定用户帖子的输入, 未考虑句子的语义信息, 因此无法获得最佳的定位效果。geoBERT 模型使用类似的 Transformer 架构进行语义学习, 但没有考虑样本之间的语义层次、地理位置层次和跨层次的关系。上述结果证明, 该框架有效地帮助模型利用 3 个不同层次关系的知识。除了上述评价指标, 为了更直观地显示本框架的优点, 可视化了 mCLF 和 geoBERT 在 Geotext 测试集上的预测结果和对应样本的真实位置, 如图 3 所示。mCLF 预测的地理位置分布明显比 geoBERT 更接近真实分布, 这表明 mCLF 对于具有复杂分布的数据集有更好的度量, 而不是偏向输出数据集地理分布的中心位置。该框架在社交

媒体地理定位任务中无论在误差数值还是可视化方面都有优异的性能, 并且对分类和回归方法都具有很高的适应性。

在不受限场景下, mCLF 在所有数据集上都实现了最佳性能。与受限场景相比, 使用针对数据集的特定语言的预训练好的 Transformer 编码器进一步提升了在 4 个数据集的定位精度。这反映了预训练语言模型提高的文本理解能力也可以用于增强使用回归和分类方法的社交媒体地理定位任务。这主要是因为额外的训练数据有助于扩展样本之间语义关系的知识, 最终表现为 Median 和 Mean 距离误差值的降低。此外, 从不同数据集上的实验结果来看, 该框架在小型和大型数据集上都有效。

(1) 对比学习的零次效果 (zero-shot effect)

由于对比学习属于无监督度量学习, 本文对其在没有微调的情况下是否学习到一些相对位置信息进行验证。首先, 将 BCMS 数据集按照经度、纬度划分为多个网格, 每个网格为 $0.5^\circ \times 0.5^\circ$ 。取

表 2 不同模型的社交媒体定位精度比较 (单位: km)

方法	BCMS		DE-AT		CH		Geotext	
	Median	Mean	Median	Mean	Median	Mean	Median	Mean
回归+受限								
geoBERT	52.06	98.74	161.13	184.97	20.7	29.62	424.90	593.08
本文分类方法	37.35	88.63	160.11	182.53	20.25	28.55	410.01	566.52
回归+不受限								
geoBERT	26.07	83.61	149.83	174.55	19.13	27.01	403.84	588.07
本文分类方法	20.88	81.33	149.28	173.57	18.76	26.74	379.75	561.26
分类+受限								
MLP	—	—	—	—	—	—	389	844
MDN	—	—	—	—	—	—	412	865
LR	—	—	—	—	—	—	397	880
Sparse-Coding	—	—	—	—	—	—	425	581
geoBERT-cla	41.84	107.38	183.17	210.04	22.18	32.77	380.08	578.89
本文分类方法	15.75	89.62	181.09	208.62	18.80	30.59	326.24	557.84
分类+不受限								
geoBERT-cla	15.00	90.96	170.25	197.40	21.61	30.77	356.43	572.77
本文分类方法	12.32	89.62	166.32	193.92	18.79	29.12	316.70	555.23

其中一个网格作为基准 (G.base), 然后根据到 G.base 的距离选择其他 3 个网格 (G.near、G.medium、G.far), 共 4 组数据。每组包含 30 个帖子, G.near、G.medium、G.far 与 G.base 之间的距离逐渐增大 (平均距离分别为 65.16 km、222.05 km 和 347.56 km)。在没有对 BCMS 数据集进行微调的情况下, 使用 geoBERT 和 mCLF 在受限场景下在 4 组样本上进行预测。根据模型的二维输出, 计算 G.near、G.medium、G.far 和 G.base 组之间的欧氏距离。值得注意的是, geoBERT 和 mCLF 输出的欧氏距离值不能直接与参考距离进行比较, 因为它只表示相对位置信息, 并没有在真实坐标系中进行校准。

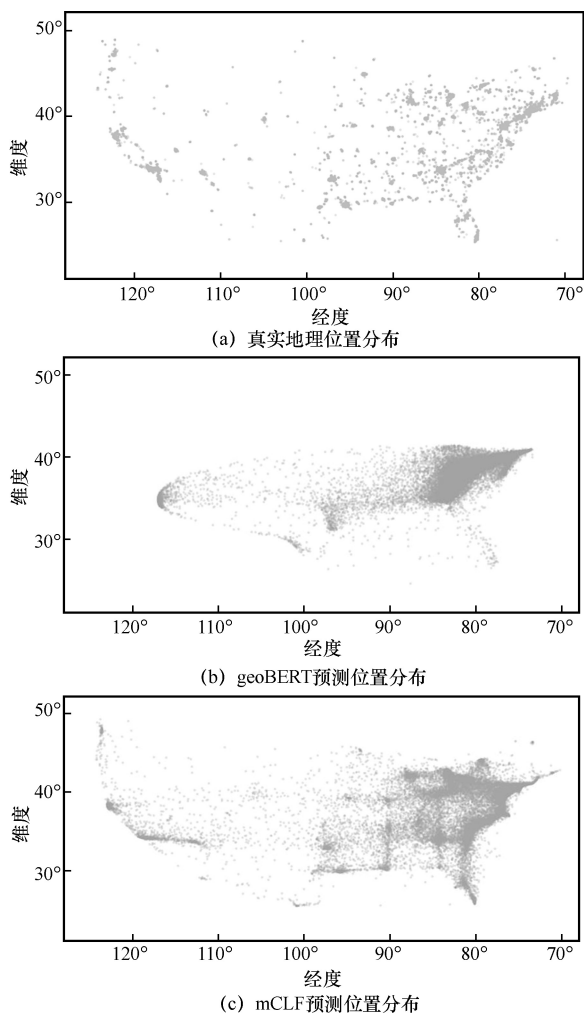


图3 Geotext 真实地理位置分布及 geoBERT 和 mCLF 预测分布结果的可视化

不同距离组之间的零次效果见表 3, 对比了使用 mCLF 预训练的模型和仅使用蒙面语言建模预训练的 geoBERT 模型在的平均欧氏距离。在上述 3 个对比中, mCLF 输出的相对位置与趋势中的真实相对位置是一致的, 这表明该框架确实能够使相似的例子学习到相似的距离度量, 并且, 预训练阶段具有 zero-shot 位置关系的学习能力。然而, 基于 MLM 预训练的 geoBERT 并没有取得良好的效果, 预测得到 G.base 与 G.far 之间的欧氏距离甚至小于其他组合, 这说明它没有在预训练中学习地理位置之间的距离度量。

表 3 不同距离组之间的零次效果

数据组	参考距离/km	欧氏距离	
		mCLF	geoBERT
G.base-G.near	65.16	1.536 5	4.38
G.base-G.medium	222.05	1.808 5	4.69
G.base-G.far	347.56	1.954 2	3.07

其次, 为了进一步验证上一步的分析, 对文本地理定位数据微调后的模型在 4 组样本上的预测位置进行了评估。在 G.base、G.near、G.medium、G.far 数据集上的地理定位效果见表 4。结果表明, mCLF 对 4 个数据集的预测的 Median 和 Mean 显著低于 geoBERT, 这表明在预训练阶段相对位置信息的度量学习有助于在微调阶段的绝对位置学习。

(2) 实例分析

使用 geoBERT 和 mCLF 对 Geotext 数据集中包含的单个或多个地理位置词的帖子进行小规模实验。地理位置词的实例实验结果见表 5, 由相应的经度、纬度坐标推断获得的参考城市, 以及由真实坐标与 geoBERT 和 mCLF 预测的坐标计算的误差距离。可以看出, 对于同时包含有效位置词和误导性位置词的情况, mCLF 的误差距离比 geoBERT 低。同样的结果也适用于只包含误导性位置词的情况。这说明该对比学习框架具有较好的捕获有效位置信息的能力, 并且对误导位置信息具有抗干扰能力。



表 4 在 G.base、G.near、G.medium、G.far 数据集上的地理定位效果 (单位: km)

方法	G.base		G.near		G.medium		G.far	
	Median	Mean	Median	Mean	Median	Mean	Median	Mean
geoBERT	127.71	156.34	54.2	139.1	53.46	92.58	72.31	91.63
mCLF	107.54	147.63	48.38	126.31	37.53	81.14	71.39	88.8

表 5 地理位置词的实例实验结果

实例	帖子内容	参考城市	误差距离/km	
			geoBERT	mCLF
1	Welcome to New York Antonio Cromartie, my friends from Ohio!	New York	413.03	157.37
2	@USER_2ded78c9 and Miami still beat them in December, beat up and all. If Miami gets these upgrades (NT, LB, S, WR) it'll be an even field.	Jacksonville	538.48	169.06
3	& @USER_84b6a0b7 You should to come to FAMU. We're having a forum on Texas Black Women and their various roles in society.	Tallahassee	999.68	351.09

(3) 消融实验

为了进一步说明不同的对比学习层次 SemCL、GeoCL 和 CrsCL 的影响,在 BCMS 数据集上进行了消融实验。在 BCMS 上的消融实验结果见表 6。与 wo/GeoCL+SemCL 相比,wo/GeoCL 的中位数和平均距离误差分别减小了约 26.6%和 6.9%, wo/SemCL 的中位数和平均距离误差分别降低了约 7.4%和 2.0%,这是因为权重 α 高于权重 β 。与 wo/GeoCL+SemCL 相比, mCLF 的中位数和平均距离误差由 (52.58, 97.42) 减小到 (37.35, 88.63)。跨层次对比学习有助于将表示的语义空间和地理空间对齐,从而帮助提高地理定位精度。综上所述,对比学习框架在提高社交媒体地理定位的准确性方面效果显著。

表 6 在 BCMS 上的消融实验结果 (单位: km)

方法	Median	Mean
mCLF	37.35	88.63
跨层次模块 (wo/CrsCL)	37.75	89.64
地理位置模块 (wo/GeoCL)	38.57	90.73
语义模块 (wo/SemCL)	48.67	95.45
地理位置模块-语义模块 (wo/GeoCL+SemCL)	52.58	97.42

4 结束语

本文提出了一种新的用于社交媒体地理定位

的多层次对比学习框架。为了获取样本之间的语义层次、地理位置层次和跨层次的联系和信息,在社交媒体地理定位模型中引入了 3 个对比学习模块。在 4 个数据集上的实验表明,该框架在受限和不受限场景下的性能都优于基线模型。此外,可视化展示了 3 个对比学习模块的 zero-shot 距离度量学习能力,这为社交媒体地理定位带来了显著的增强。

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] 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.
- [3] PAUL M J, DREDZE M. You are what your Tweet: analyzing twitter for public health[J]. Artificial Intelligence, 2011(38): 265-272.
- [4] DO T H, NGUYEN D M, TSILIGIANNI E, et al. Multiview deep learning for predicting twitter users' location[J]. arXiv preprint, 2017, arXiv: 1712. 08091.
- [5] WING B, BALDRIDGE J. Simple supervised document geolocation with geodesic grids[C]//Meeting of the Association for Computational Linguistics: Human Language Technologies. DBLP, 2012.
- [6] HAN B, COOK P, BALDWIN T. Geolocation prediction in social media data by finding location indicative words[J]. 24th International Conference on Computational Linguistics - Pro-

- ceedings of COLING 2012: Technical Papers, 2012: 1045-1062.
- [7] RAHIMI A, COHN T, BALDWIN T. A neural model for user geolocation and lexical dialectology[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume2: Short Papers). Stroudsburg: Association for Computational Linguistics, 2017: 209-216.
 - [8] 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.
 - [9] WANG F J, LU C T, QU Y Z, et al. Collective geographical embedding for geolocating social network users[M]//Advances in Knowledge Discovery and Data Mining. Cham: Springer International Publishing, 2017: 599-611.
 - [10] HUANG B X, CARLEY K. A hierarchical location prediction neural network for twitter user geolocation[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg: Association for Computational Linguistics, 2019: 4731-4741.
 - [11] RAHIMI A, COHN T, BALDWIN T. Semi-supervised user geolocation via graph convolutional networks[J]. 2018: 2009-2019. DOI: 10.18653/v1/P18-1187.
 - [12] RAHIMI A, VU D, COHN T, et al. Exploiting text and network context for geolocation of social media users[J]. 2015: 1362-1367. DOI: 10.3115/v1/N15-1153.
 - [13] SCALIA G, FRANCALANCI C, PERNICI B. CIME: context-aware geolocation of emergency-related posts[J]. *GeoInformatica*, 2022, 26(1): 125-157.
 - [14] ZHENG C, JIANG J Y, ZHOU Y C, et al. Social media user geolocation via hybrid attention[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2020: 1641-1644.
 - [15] ALEXIS C, KIELA D. SentEval: an evaluation toolkit for universal sentence representations[J]. arXiv preprint, 2018, arXiv: 1803.05449.
 - [16] REIMERS N, GUREVYCH I. Sentence-BERT: sentence embeddings using Siamese BERT-networks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg: Association for Computational Linguistics, 2019: 3980-3990.
 - [17] GAO T Y, YAO X C, CHEN D Q. SimCSE: simple contrastive learning of sentence embeddings[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 6894-6910.
 - [18] GIORGI J, NITSKI O, WANG B, et al. DeCLUTR: deep contrastive learning for unsupervised textual representations[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2021: 879-895.
 - [19] CER D, YANG Y, KONG S Y, et al. Universal sentence encoder[J]. 2018: 169-174. DOI: 10.18653/v1/D18-2029. 2018.
 - [20] HUANG J J, TANG D Y, ZHONG W J, et al. Whitening-BERT: an easy unsupervised sentence embedding approach[C]//Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2021. Stroudsburg: Association for Computational Linguistics, 2021: 238-244.
 - [21] LI B H, ZHOU H, HE J X, et al. On the sentence embeddings from pre-trained language models[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics, 2020: 9119-9130.
 - [22] SU J, CAO J, LIU W, et al. Whitening sentence representations for better semantics and faster retrieval[J]. arXiv preprint, 2021, arXiv: 2103.15316.
 - [23] CARLSSON F, GYLLENSTEN A C, GOGOULOU E, et al. Semantic re-tuning with contrastive tension[C]//In 9th International Conference on Learning Representations, Austria: Virtual Event, 2021.
 - [24] KIM T, YOO K M, LEE S G. Self-guided contrastive learning for BERT sentence representations[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2021: 2528-2540.
 - [25] YAN Y M, LI R M, WANG S R, et al. ConSERT: a contrastive framework for self-supervised sentence representation transfer[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2021: 5065-5075.
 - [26] MENG Y, XIONG C, BAJAJ P, et al. COCO-LM: correcting and contrasting text sequences for language model pretraining[J]. arXiv preprint, 2021 arXiv: 2102.08473.
 - [27] WU Z, WANG S, GU J, et al. CLEAR: contrastive learning for sentence representation [J]. arXiv preprint, 2020, arXiv: 2012.15466.
 - [28] HADSELLR, CHOPRA S, LECUN Y. Dimensionality reduction by learning an invariant mapping[C]//Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). Piscataway: IEEE Press, 2006: 1735-1742.
 - [29] EISENSTEIN J, O'CONNOR B, SMITH N A, et al. A latent variable model for geographic lexical variation[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Lan-



- guage Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL. Stroudsburg: Association for Computational Linguistics, 2010.
- [30] 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@EACL 2021, Kiyv, Ukraine, 2021. 1-11.
- [31] SCHERRER Y, LJUBEŠIĆ N. Social media variety geolocation with geobert[C]//Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects. Stroudsburg: Association for Computational Linguistics, 2021.
- [32] RAHIMI A, BALDWIN T, COHN T. Continuous representation of location for geolocation and lexical dialectology using mixture density networks[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2017. 167-176.
- [33] RAHIMI A, VU D, COHN T, et al. Exploiting text and network context for geolocation of social media users[C]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2015: 1362-1367.
- [34] 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.
- [35] ROLLER S, SPERIOSU M, RALLAPALLI S, et al. Supervised text-based geolocation using language models on an adaptive

grid[J]. EMNLP-CoNLL2012- 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Proceedings of the Conference, 2012: 1500-1510.

[作者简介]



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。



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



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与室内定位。