Exploring Urban Semantics: A Multimodal Model for POI Semantic Annotation with Street View Images and Place Names

Dabin Zhang[1] , **Meng Chen**[1*] , **Weiming Huang**[2] , **Yongshun Gong**[1] , **Kai Zhao**[3]
[1]School of Software, Shandong University
[2]School of Computer Science and Engineering, Nanyang Technological University
[3] Robinson College of Business, Georgia State University
zdb@mail.sdu.edu.cn, mchen@sdu.edu.cn, weiming.huang@ntu.edu.sg, yongshun2512@hotmail.com,
kzhao4@gsu.edu

Abstract

Semantic annotation for points of interest (POIs) is the process of annotating a POI with a category label, which facilitates many services related to POIs, such as POI search and recommendation. Most of the existing solutions extract features related to POIs from abundant user-generated content data (e.g., check-ins and user comments). However, such data are often difficult to obtain, especially for newly created POIs. In this paper, we aim to explore semantic annotation for POIs with limited information such as POI (place) names and geographic locations. Additionally, we have found that the street view images provide extensive visual clues about POI attributes and could be an essential supplement to limited information of POIs that enables semantic annotation. To this end, we propose a novel multimodal model for POI semantic annotation, namely M3PA, which achieves enhanced semantic annotation through fusing a POI's textual and visual representations. Specifically, M3PA extracts visual features from street view images using a pre-trained image encoder and integrates these features to generate the visual representation of a targeted POI based on a geographic attention mechanism. Furthermore, M3PA utilizes the contextual information of neighboring POIs to extract textual features and captures their spatial relationships through geographical encoding to generate the textual representation of a targeted POI. Finally, the visual and textual representations of a POI are fused for semantic annotation. Extensive experiments with POI data from Amap validate the effectiveness of M3PA for POI semantic annotation, compared with several competitive baselines.

1 Introduction

A point of interest (POI) is a specific place that individuals may find helpful or interesting, such as a park, restaurant, or university. In recent decades, services related to POIs have become increasingly popular on the web. Despite the abundance of POIs being generated, the quality of POI data remains questionable. The prevalent problem of missing properties for POIs is particularly challenging, with key information such as categories often missing or incorrect due to uncertainties in the human annotation processes for POIs. POI semantic annotation, which annotates POIs with the most likely category from all categories, not only helps users better understand the characteristics of POIs but also aids in discovering relatedness or similarities between different POIs. Consequently, POI semantic annotation can facilitate many downstream applications, such as POI search, recommendation and the profiling of urban areas [Sun *et al.*, 2021; Huang *et al.*, 2023; Bing *et al.*, 2023; Li *et al.*, 2024; Xu *et al.*, 2023b; Chen *et al.*, 2021; Chen *et al.*, 2020].

Many existing methods for POI semantic annotation generally extract sequential, textual, and visual features from user-generated content data (e.g., check-in logs [Li *et al.*, 2020; Xu *et al.*, 2022], POI descriptions [Lagos *et al.*, 2020; Villegas and Aletras, 2021], and users' comments [Yang *et al.*, 2023]) relevant to a given POI, in order to estimate how probable each category is for annotating the POI. However, these methods heavily depend on abundant user-generated content on POIs, which may be difficult to obtain in reality, particularly for newly created POIs. This raises the need for semantic annotation for POIs with limited information such as POI (place) names and geographical data. In this vein, Zhang *et al.* [2023] address this issue by developing a GCN-based spatial encoder to model spatial correlations among POIs and an attention-based text encoder for POI names. However, their approach struggles to capture the spatially varying context, such as the names of nearby POIs, which are crucial for understanding POI semantics.

With the prosperity of various emerging geospatial data sources, new angles have surfaced to address this problem in ways that more closely align with human perception. Undoubtedly the text information that is intrinsically carried by POIs (in particularly POI names) are indicative to POI semantics and thus categories. From another perspective, when a POI name is inadequate to determine the category of a POI, it is useful to simply look around visually, to help the inference with urban contextual information. In this vein, the emerging urban visual data (e.g., street view images) provide extensive visual clues about POI attributes in the built-up areas. With

[*] Corresponding author.

such data, we can then mimic human perception using "machine eyes", i.e., analyzing the street view images to enhance the determination of POI categories. Consequently, we believe that the street view images could be an essential supplement of POI names to help semantic annotation for POIs.

In this context, we aim to extract semantic features of POIs based on street view images around the POIs and the names of their spatial neighbors, which presents two challenges:

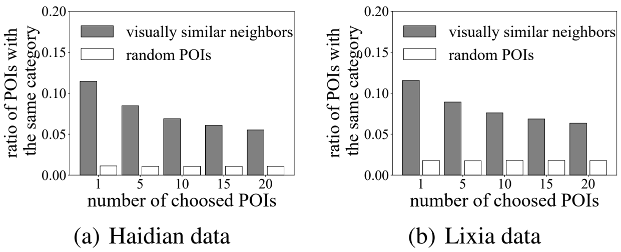
1) **Visual representation:** Presently, current methods using street view images primarily concentrate on streets [Wojna *et al.*, 2017; Li *et al.*, 2022] or a particular area [Liu *et al.*, 2023; Lee *et al.*, 2021]. Many of these methods emphasize the correlation between street view scenes and surrounding environmental factors or the linguistic depiction of street view scenes. Nevertheless, our goal is to extract more detailed semantic information associated with POIs, which poses a significant challenge.

2) **Textual representation:** There are currently many pretrained language models (PLMs) available for successfully extracting textual features. However, it is difficult to effectively capture the spatial semantic information of POIs using these existing models or their extensions. This is because POIs exist in geographic space with their spatial relationships (such as distance and direction), and their distribution does not follow a context linearization representation that is easy for PLMs to understand. Additionally, the established language models pre-trained on general domain corpora need to be adjusted or processed for data on POI names.

To tackle the aforementioned challenges, we propose a multimodal model for POI semantic annotation (M3PA) with two key designs. 1) We create a pre-trained street view image encoder to obtain the feature vector of an image aligned to its surrounding POIs' distribution, and use geographic attention to combine visual feature vectors of nearby POIs around the targeted POI, generating the visual representation. 2) We introduce GeoBERT, a language model constructed based on BERT, which combines the name of a targeted POI with its neighbors' names and adds geographical encoding to capture the spatial relationship between POIs, generating the textual representation. Ultimately, we integrate the visual and textual representations for POI semantic annotation.

The contributions of this paper are as follows:

- We introduce street view images to help understand the characteristics of POIs. Such image data is an important supplement to limited POI information (geographic locations and textual names) that enables semantic annotation. To this end, we present a multimodal model (M3PA) which incorporates the textual and visual representations for POI semantic annotation.
- We propose a geographic language model (GeoBERT) which encodes textual names of nearby POIs and the spatial relations between POIs, producing a spatially context-aware representation for each POI.
- We validate the proposed M3PA using two POI datasets obtained from AMap and evaluate its performance through the task of POI semantic annotation. M3PA shows significant performance improvements compared to various baseline methods, as verified by the paired t-test. Data and source codes are available at https://github.com/zdbasdxz/M3PA.



(a) Haidian data

(b) Lixia data

Figure 1: Relationship between categories of POIs and those of their visually similar neighbors in the Haidian and Lixia datasets.

2 Preliminaries

2.1 Problem Statement

Definition 1 (Point of Interest)**.** *The information of a POI* p_i *consist of a POI ID* id_i*, 2D geographic coordinates* g_i*, and a POI name represented as a set of words* $\mathcal{W}_i$*, denoted as* $p_i = \{id_i, g_i, \mathcal{W}_i\}$*. The coordinate* $g_i = (lng_i, lat_i)$ *is comprised of longitude and latitude information.*

Definition 2 (Street View Image)**.** *A street view image* s_i *is captured alongside the road network in urban areas to depict the surrounding and environmental details from a visual perspective, and it is associated with geographic coordinates.*

Street view images are primarily captured by vehicles in urban areas and offer a perspective from a human point of view. Usually, these images are collected from four different directions at a given location, in order to obtain comprehensive coverage.

Definition 3 (POI Category)**.** *A POI category, denoted as* c_i *(e.g., University or Dessert House), signifies the specific topics of the activities afforded at the POI* p_i*.*

Definition 4 (POI Semantic Annotation)**.** *Using the geographical coordinates (locations) and names of POIs, as well as urban street view images, we aim to predict category labels for unlabeled POIs.*

2.2 Data and Motivation

We use the public POI dataset obtained from AMap[1] in Beijing's Haidian District and Jinan's Lixia District [Zhang *et al.*, 2023]. Each POI includes an ID, latitude and longitude, a name, and a category. Furthermore, we gather street view images (sized at 1024×512 pixels) from Baidu[2], a Chinese Internet service company. We obtain a total of 185,388 images from Haidian District and 91,672 images from Lixia District. To aid in understanding POI semantics from images, we filter out POIs without street view images within a 100-meter radius. After preprocessing, the Haidian dataset contains 89,055 POIs across 248 categories, and the Lixia dataset contains 44,514 POIs across 140 categories.

[1] https://lbs.amap.com/api/webservice/guide/api/search/

[2] https://lbsyun.baidu.com/

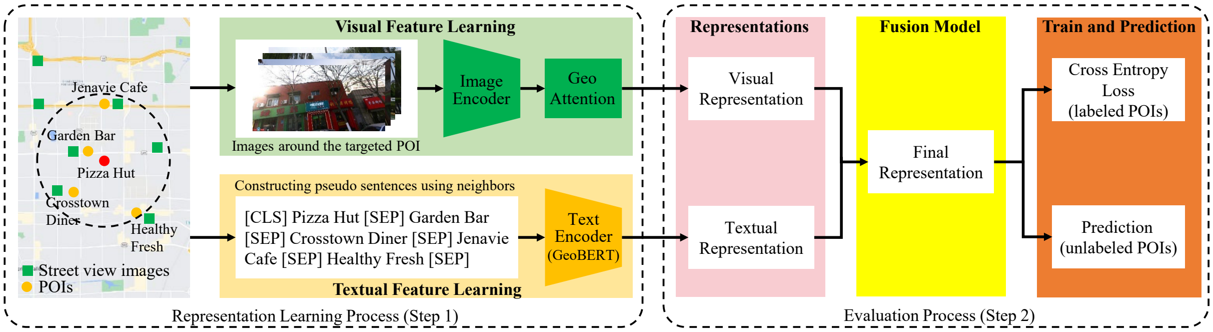

Jenavie Cafe
Garden Bar
Pizza Hut
Crosstown Diner
Healthy Fresh
🟩 Street view images
● POIs
Visual Feature Learning
Images around the targeted POI
Image Encoder
Geo Attention
Constructing pseudo sentences using neighbors
[CLS] Pizza Hut [SEP] Garden Bar [SEP] Crosstown Diner [SEP] Jenavie Cafe [SEP] Healthy Fresh [SEP]
Text Encoder (GeoBERT)
Textual Feature Learning
Representation Learning Process (Step 1)
Representations
Visual Representation
Textual Representation
Fusion Model
Final Representation
Train and Prediction
Cross Entropy Loss (labeled POIs)
Prediction (unlabeled POIs)
Evaluation Process (Step 2)

Figure 2: The framework of M3PA.

We posit that street view images surrounding POIs can effectively convey semantic information on POIs. In this context, we have undertaken an analysis of the data to ascertain whether POIs with similar street view images in their vicinity are more likely to share the same category. To obtain the visual feature for each POI, we initially establish a buffer zone (e.g., 100m) for the POI and identify the two closest street view images within the buffer. Subsequently, we utilize ResNet [He *et al.*, 2016] to encode these images and obtain the average visual feature vectors. Furthermore, we employ the visual vector of each targeting POI p_i to calculate the N_k nearest neighbors based on the cosine distance and determine the proportion of neighbors that share the same category as p_i. Similarly, we randomly select N_k POIs from the POI set and compute the ratio of the selected N_k POIs with the same category as p_i. Finally, we calculate the average ratios of all POIs and report the results in Figure 1. We observe that the number of visually similar POIs with the same category as the targeted POI is greater than that of random POIs in the Haidian and Lixia datasets. This observation indicates that the street view images around a POI indeed contain semantic information related to the POI category label, which serves as a crucial motivation for modeling the images to enhance the semantic annotation performance.

3 Multimodal Semantic Annotation Model

3.1 Model Overview

We present an overview of our proposed framework in Figure 2, which is comprised of two primary steps: the representation learning process and the evaluation process. Specifically, we utilize urban images to identify surrounding images for a targeted POI and employ a pre-trained image encoder to produce the visual representation. Additionally, we leverage the POI information to identify geographically neighboring POIs and utilize the newly designed text encoder (GeoBert) to encode the text names and geographical information of these neighboring POIs to create the textual representation. Subsequently, we develop a fusion model to integrate the feature representations from two different perspectives. During the training stage, we optimize the model using cross-entropy loss with the labeled POIs, and during the testing stage, we predict the category labels for the unlabeled POIs.

3.2 Extracting Visual Representations of POIs

We initially develop a pre-trained image encoder to extract visual features from street view images and then combine the visual features of spatially nearby POIs based on an attention mechanism to create the visual representations.

Pre-trained Image Encoder

We have observed that street view images around a POI can convey semantic information related to its category label. As such, we extract street view images within a specified buffer zone (e.g., 100m) of a targeted POI and learn the visual features. Although we could utilize an existing well-trained ResNet to extract visual features from street view images, this type of image encoder is designed to capture general features rather than the semantic functionality of the specific zone where the image is captured. POIs, which denote locations associated with human activity in populated areas, can provide insight into the characteristics of a given zone. Consequently, we propose utilizing the POI information surrounding a street view image to inform the learning process of the image encoder, thereby capturing the inherent semantic functionality of the street view image.

For an image s_i, we first use ResNet to extract its visual feature $\mathbf{V}_i$,

$$\mathbf{V}_i = \text{ResNet}(s_i). \tag{1}$$

Next, we create the POI semantic feature for s_i. Specifically, we build a buffer zone for s_i based on its geographical coordinates and calculate the category distribution of POIs within the buffer zone as the semantic feature, denoted as $\mathbf{F}_i \in \mathbb{R}^{N_c}$. Here, N_c represents the total number of POI categories, and each dimension in $\mathbf{F}_i$ corresponds to the number of POIs with a specific category in the zone.

We proceed to employ a two-layer MLP to transform $\mathbf{V}_i$ into a new feature $\widetilde{\mathbf{V}}_i$ with identical dimensions as $\mathbf{F}_i$. During the learning phase, we utilize the semantic feature $\mathbf{F}_i$ as labels to fine-tune the ResNet-based image encoder and define the loss function as

$$\begin{gathered}\mathcal{L}^{image} = \sum_{i=1}^{|\mathcal{S}|} \text{KL}(\mathbf{F}_i, \widetilde{\mathbf{V}}_i), \\ \widetilde{\mathbf{V}}_i = \mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \cdot \mathbf{V}_i + \vec{b_1}) + \vec{b_2},\end{gathered} \tag{2}$$

where $\text{KL}(\cdot)$ denotes the Kullback Leibler (KL) divergence, $|\mathcal{S}|$ represents the total number of street view images, and

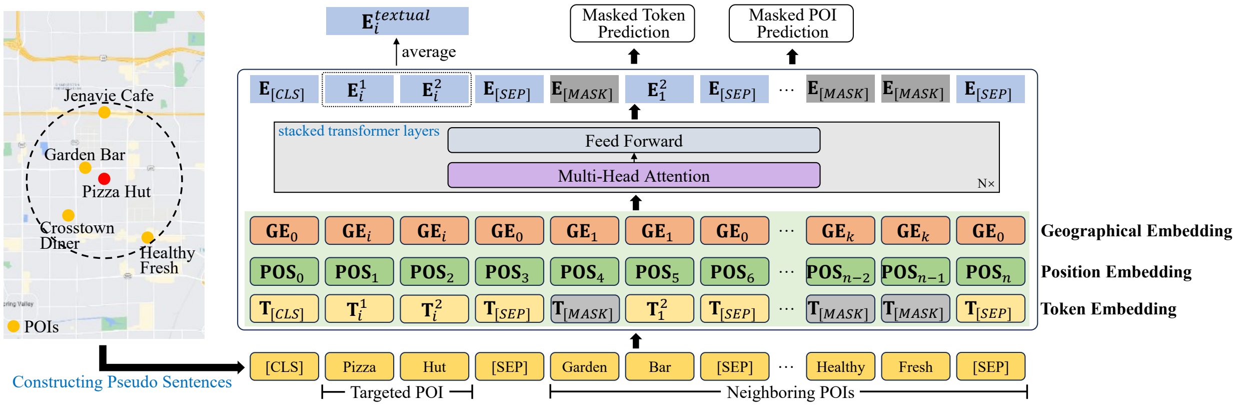

Jenavie Cafe
Garden Bar
Pizza Hut
Crosstown Diner
Healthy Fresh
POIs
Constructing Pseudo Sentences
$E_i^{textual}$
average
Masked Token Prediction
Masked POI Prediction
$E_{[CLS]}$ E_i^1 E_i^2 $E_{[SEP]}$ $E_{[MASK]}$ E_1^2 $E_{[SEP]}$ … $E_{[MASK]}$ $E_{[MASK]}$ $E_{[SEP]}$
stacked transformer layers
Feed Forward
Multi-Head Attention
N×
GE_0 GE_i GE_i GE_0 GE_1 GE_1 GE_0 … GE_k GE_k GE_0
Geographical Embedding
POS_0 POS_1 POS_2 POS_3 POS_4 POS_5 POS_6 … POS_{n-2} POS_{n-1} POS_n
Position Embedding
$T_{[CLS]}$ T_i^1 T_i^2 $T_{[SEP]}$ $T_{[MASK]}$ T_1^2 $T_{[SEP]}$ … $T_{[MASK]}$ $T_{[MASK]}$ $T_{[SEP]}$
Token Embedding
[CLS] Pizza Hut [SEP] Garden Bar [SEP] … Healthy Fresh [SEP]
Targeted POI
Neighboring POIs

Figure 3: The architecture of GeoBERT.

$\mathbf{W}_1$, $\mathbf{W}_2$, $\vec{b_1}$ and $\vec{b_2}$ are model parameters. Given that the semantic feature $\mathbf{F}_i$ represents a distribution of categories, the KL divergence, which quantifies the distinction between probability distributions, stands as an appropriate selection for the loss function. Throughout the training process, we optimize the model using the SGD method. After training, we use the fine-tuned ResNet to encode each street view image and generate the initial visual feature $\mathbf{V}^p$ for each POI by averaging the features of images within the buffer zone.

Geo Attention

The consideration of nearby POIs holds significant importance in semantic annotation for the targeted POI [Zhang *et al.*, 2023]. In line with the spatial interpolation assumption, it is established that closer POIs in the neighborhood hold more weights in determining the category label of the targeted POI compared to those located at a distance. Building on this premise, we introduce a geographic attention mechanism (Geo Attention), which allocates weights to neighboring POIs based on their geographic distance from the targeted POI. Specifically, it takes the geographic coordinates (g_i) of POI p_i and its k neighboring POIs $(p_{i,1}, p_{i,2}, \cdots, p_{i,k})$ as input. The geographic distance between g_i and $g_{i,j}$ is computed using the Gaussian kernel function,

$$d(g_i, g_{i,j}) = exp(-geodist(g_i, g_{i,j})\sigma^2/2), \tag{3}$$

where $geodist(\cdot)$ denotes the geodesic distance and σ represents the hyper-parameter governing the decay of similarity with respect to distance.

The distances between p_i and its k neighbors are computed based on Equation (3), denoted as $\mathbf{D}_i$. Then we calculate the normalized geo-attention weight $w(p_i, p_{i,j})$ for each neighboring POI based on the distances $\mathbf{D}_i$,

$$\begin{gathered} w(p_i, p_{i,j}) = exp(\mathbf{H}_{ij})/(\sum_{j'=1}^{k} exp(\mathbf{H}_{ij'})), \\ \mathbf{H}_i = \mathbf{W}_3 \cdot \mathbf{D}_i + \vec{b_3}, \end{gathered} \tag{4}$$

where $\mathbf{W}_3$ and $\vec{b_3}$ are the parameters of a fully-connected layer. Lastly, the visual representation of p_i is calculated as

$$\mathbf{E}_i^{visual} = \mathbf{V}_i^p + \mathbf{W}_4 \cdot \sum_{j=1}^{k} w(p_i, p_{i,j})[\mathbf{V}_j^p || \vec{c_j}], \tag{5}$$

where $\mathbf{V}_i^p$ and $\mathbf{V}_j^p$ are the visual features of POIs generated with the pre-trained image encoder, $\vec{c_j}$ is the one-hot representation of the category label of neighboring POI p_j, $||$ denotes concatenation, $\mathbf{W}_4$ is the weight matrix. The resulting vector $\mathbf{E}_i^{visual}$ is hence the weighted sum of the visual features concatenated with the category labels of neighboring POIs weighted by the learned geo-attention weights.

3.3 Extracting Textual Representations of POIs

We present GeoBERT, built on a language model, which initially combines a POI's name with its neighbors' names to create a pseudo sentence and then encodes the spatial relations between POIs to generate a spatially context-aware representation for each POI.

Constructing Pseudo Sentences Using POI Names

For a targeted POI, we first identify the k nearest geographical neighbors and transform their names into a BERT-compatible input sequence to create a pseudo sentence. In the example illustrated in Figure 3, the pseudo sentence is constructed as

[CLS] Pizza Hut [SEP] Garden Bar [SEP] Crosstown Diner [SEP] Jenavie Cafe [SEP] Healthy Fresh [SEP]

This sequence begins with the name of the targeted POI, followed by the names of the spatially neighboring POIs. The arrangement of the names is based on their spatial proximity to the targeted POI, in accordance with the first law of geography. Furthermore, we employ the same BERT tokenizer to tokenize the sentence, with [CLS] positioned at the beginning and [SEP] separating the name of each POI.

Encoding Spatial Relations

In order to represent the pseudo sentence, we employ two distinct types of position embeddings in conjunction with token embeddings (cf. Figure 3). The sequence position embedding serves to denote the token order, akin to the original position embedding in BERT. Additionally, we introduce a geographical embedding, designed to capture the spatial relationship between a POI and its neighboring POIs. Specifically, we utilize a multidimensional continuous geographical encoding technique [Li *et al.*, 2021] to map a 2D point to an encoded vector of dimension M based on its coordinate.

We define the geographical embedding $\mathbf{GE}_i$ of POI p_i as

$$\mathbf{GE}_i = \frac{1}{\sqrt{M}}[\cos g_i \cdot \mathbf{W}_g^{\mathrm{T}} || \sin g_i \cdot \mathbf{W}_g^{\mathrm{T}}], \tag{6}$$

where $||$ denotes concatenation, g_i represents the latitude and longitude information of p_i, and $\mathbf{W}_g \in \mathbb{R}^{\frac{M}{2} \times 2}$ denotes the linear projection matrix. It is important to note that $\mathbf{W}_g$ is initialized with a Gaussian distribution following independent and identically distributed properties, denoted as $\mathbf{W}_g \sim N(0, \gamma^{-2})$, where γ controls the spatial kernel bandwidth.

This geographical embedding offers the advantage that the dot product of the embeddings of any two POIs approximates the Gaussian kernel of their original latitude and longitude information, i.e., $\mathbf{GE}_i \cdot \mathbf{GE}_j \approx exp(-\|g_i - g_j\|^2 \times \gamma^{-2})$. Consequently, the similarity between the geographical embeddings of two POIs could reflect their spatial proximity. Furthermore, we employ an embedding $\mathbf{GE}_0$ generated by a special coordinate, which is positioned far away from all the POIs in the dataset, to separate different POIs in the sentence.

Similar to Bert, the token embedding, sequence position embedding, and geographical embedding are summed and then input into stacked Transformer layers, where each layer includes a multi-head self-attention layer and a feedforward layer [Vaswani *et al.*, 2017]. Following this encoder, an output embedding $\mathbf{E}$ is computed for each token.

Pre-training GeoBERT

To train the GeoBERT, we design two tasks to adapt the original BERT backbone to the POI pseudo sentences.

Masked token prediction. The task is designed to sense fine-grained POI semantics in geographic proximity (i.e., partial POI name), in which it completes the full POI names from pseudo sentences with randomly masked tokens. Each token (excluding [CLS] and [SEP]) has a 15% chance of being masked by the special token [MASK], and the masked token is predicted using the remaining tokens and their spatial coordinates. The following example demonstrates the masked input for the masked token prediction task.

[CLS] Pizza [MASK] [SEP] Garden Bar [SEP] [MASK] Diner [SEP] Jenavie Cafe [SEP] [MASK] Fresh [SEP]

Masked POI prediction. Furthermore, taking into account the spatial co-occurrence patterns of POIs, we predict the name of a POI by utilizing the names of its neighboring POIs. In this context, we formulate the task of predicting masked POIs, where we mask all tokens of a random selected POI (with 15% probability) in a pseudo sentence and predict the masked POI name by capturing the spatial relation between the masked POI and its neighboring POIs. The following example illustrates the masked input for the masked POI prediction task.

[CLS] Pizza Hut [SEP] [MASK] [MASK] [SEP] Crosstown Diner [SEP] [MASK] [MASK] [SEP] Healthy Fresh [SEP]

As we only need to predict the masked tokens, we collect a subset based on the output embedding $\mathbf{E}$ (generated by stacked Transformer layers) to form the representations of masked tokens $\mathbf{E}^{mask}$. Subsequently, we proceed to map $\mathbf{E}^{mask}$ into the vocabulary space in order to forecast the probability distributions $\mathbf{p}$ across the entire vocabulary,

$$\mathbf{p} = \text{softmax}(\mathbf{W}^m \mathbf{E}^{mask}), \tag{7}$$

where $\mathbf{W}^m$ is a linear projection matrix. Finally, we define the cross-entropy loss as

$$\mathcal{L}^{mask} = -\frac{1}{N^m} \sum_{i=1}^{N^m} \mathbf{a}_i \log \mathbf{p}_i, \tag{8}$$

where $\mathbf{a}_i$ represents the actual one-hot distribution over the entire vocabulary for the ith masked token and N^m denotes the number of masked tokens in a pseudo sentence.

We optimize the GeoBERT model using stochastic mini-batch gradient descent method. After training, we average the output embeddings $\mathbf{E}$ of tokens belonging to the targeted POI p_i to generate the textual representation $\mathbf{E}_i^{textual}$.

3.4 Training Objective

We now proceed to merge the visual and textual representations to create a more comprehensive final representation for use in the semantic annotation task. We employ three fusion methods (concatenation-based fusion, wide & deep network-based fusion, and attention-based fusion) as delineated in [Zhang *et al.*, 2023] to produce the final representation $\mathbf{E}_i^{final}$ of the targeted POI p_i. Subsequently, we input $\mathbf{E}_i^{final}$ into a two-layer MLP to generate the predicted output $\hat{\mathbf{y}}_i$, which denotes the probability that p_i is associated with each category label. This is calculated as follows,

$$\hat{\mathbf{y}}_i = \text{softmax}(\mathbf{W}_6 \cdot \text{ReLU}(\mathbf{W}_5 \cdot \mathbf{E}_i^{final} + \vec{b_5}) + \vec{b_6}), \tag{9}$$

where $\mathbf{W}_5$, $\mathbf{W}_6$, $\vec{b_5}$, and $\vec{b_6}$ represent the parameters.

During the training stage, we use the classification objective and minimize the cross-entropy loss function,

$$\mathcal{L} = -\frac{1}{|\mathcal{P}|} \sum_{i=1}^{|\mathcal{P}|} \sum_{j=1}^{|\mathcal{C}|} y_{ij} log(\hat{\mathbf{y}}_{ij}), \tag{10}$$

where $|\mathcal{P}|$ and $|\mathcal{C}|$ represent the number of unique POIs and categories in the dataset, y_{ij} represents whether POI p_i is labeled with the jth category, and $\hat{\mathbf{y}}_{ij}$ represents the predicted probability that p_i is labeled with the jth category. During the prediction stage, we simply choose the element with the maximum value in $\hat{\mathbf{y}}_i$ as the predicted category label of p_i.

4 Experiment

4.1 Experimental Settings

Datasets. We utilize the public Haidian and Lixia POI data from AMap as our datasets. Both datasets are randomly divided into two sets, with an 8:2 ratio for training and testing. We perform 5 model runs and present the average results.

Model Parameters. In our experiments, we utilize ResNet18 as the initial image encoder with an output dimension of 512. The weight and tokenizers of GeoBERT are initialized from $\text{BERT}_{\text{Base}}$[3], and GeoBERT has an output dimension of 768. We employ the same optimization parameters for

[3] https://huggingface.co/bert-base-chinese

pre-training tasks on GeoBERT as in [Kenton and Toutanova, 2019], using the AdamW optimizer. To optimize M3PA, we utilize the Adam optimizer and initialize the learning rate at 0.0001 with a linear decay.

Evaluation metrics. POI semantic annotation is a multi-class classification problem. We use the well-known metrics (*Accuracy* and *Macro-F1*) to evaluate model performance. Additionally, using $\hat{\mathbf{y}}_i$, we can create a ranking list of predicted category labels. Therefore, we also utilize the metric *MRR*, which considers the position of real labels in the ranking lists. MRR $= [\sum_{i=1}^{|\mathcal{P}_{test}|}(1/rank_i)]/\,|\mathcal{P}_{test}|$, where $|\mathcal{P}_{test}|$ is the number of POIs in the test set $\mathcal{P}_{test}$ and $rank_i$ is the rank of the real category in the predicted list for POI p_i.

Baselines. In their study on semantic annotation for POIs with geographic locations and POI names, Zhang *et al.* [2023] introduce baselines including textual view methods (WTF), spatial view methods (GSF, GPS2Vec [Yin *et al.*, 2021]), and multi-view methods (EHC [Liu *et al.*, 2020], WTF+GPS2Vec, and SPTA [Zhang *et al.*, 2023]). Detailed information about these methods can be found in [Zhang *et al.*, 2023]. Additionally, we incorporate BERT [Kenton and Toutanova, 2019] as a new textual view method, extracting textual features from POI names using a pre-trained Bert. Furthermore, we compare M3PA with visual view methods.

- **ResNet**: We utilize a pre-trained ResNet18 [He *et al.*, 2016] on ImageNet data to encode street view images and obtain the visual representation of a POI. We then train a two-layer MLP for semantic annotation.
- **GeoCLR**:It generates contrastive samples considering both self-similarity and geographical-similarity across urban images, and trains a contrastive learning model for visual representations [Li *et al.*, 2022]. Subsequently, we acquire the visual representation of a POI and train a two-layer MLP for annotation.
- **MM-Gated-XAtt**: It combines textual features extracted from a tweet and visual features extracted from the twitter image using a cross-attention mechanism for POI annotation, irrespective of the geographic location of POIs [Villegas and Aletras, 2021]. In our setting, we extract textual features from the POI name and visual features for a POI from its nearest street view image.

4.2 Comparison with Baselines

We report the comparative results in Table 1 and observe that:

1) Both the spatial view and visual view methods yield poor performance, as the geographic information or urban images lack sufficient semantic information for POI annotation. GPS2Vec uses the MLP network to encode POI coordinates, leading to the worst performance; GSF utilizes the category information of other POIs in the same grid and outperforms GPS2Vec. ResNet and GeoCLR perform poorly because they only extract urban contextual visual features related to POIs, without considering the semantic features of POI names.

2) The textual view method outperforms the spatial view and visual view methods because the textual features extracted from POI names are more indicative of POI semantics and thus categories. Moreover, BERT performs better than WTF, confirming the effectiveness of using BERT as the backbone to capture textual features from POI names.

Data	View	Method	Accuracy	Macro-F1	MRR
Haidian	Textual	WTF	60.35	47.88	72.56
		BERT	64.60	56.88	75.81
	Spatial	GSF	13.59	4.49	23.36
		GPS2Vec	5.19	0.25	12.26
	Visual	ResNet	6.21	0.15	13.12
		GeoCLR	6.12	0.65	13.11
	Integrated	EHC	61.18	49.38	73.25
		WTF+GPS2Vec	60.92	48.77	73.07
		MM-Gated-XAtt	68.05*	58.20	77.96*
		STPA	67.26	61.75*	77.83
		M3PA	**69.43**	**63.62**	**79.54**
		Improvements	2.03	3.03	2.03
Lixia	Textual	WTF	61.49	52.70	73.90
		BERT	63.53	55.12	75.41
	Spatial	GSF	13.31	5.34	24.61
		GPS2Vec	6.97	0.34	15.87
	Visual	ResNet	10.27	1.27	20.40
		GeoCLR	11.06	1.76	21.41
	Integrated	EHC	61.50	53.21	73.91
		WTF+GPS2Vec	61.62	53.21	74.00
		MM-Gated-XAtt	65.65*	54.27	76.32
		STPA	65.46	59.72*	76.90*
		M3PA	**68.51**	**63.00**	**78.91**
		Improvements	4.36	5.49	2.61

Table 1: Performance comparison of different methods (in percentage), where the performance improvements of M3PA are compared with the best results of these baselines, marked by the asterisk.

3) The integrated view methods demonstrate superior performance compared to single view methods. M3PA exhibits the best performance. For instance, compared with MM-Gated-XAtt, M3PA achieves an average improvement of 4.46% on the Haidian data and 7.95% on the Lixia data across three metrics. Moreover, the results of the superiority paired t-test confirm that M3PA's improvement over baselines is statistically significant, with a p-value less than 0.01.

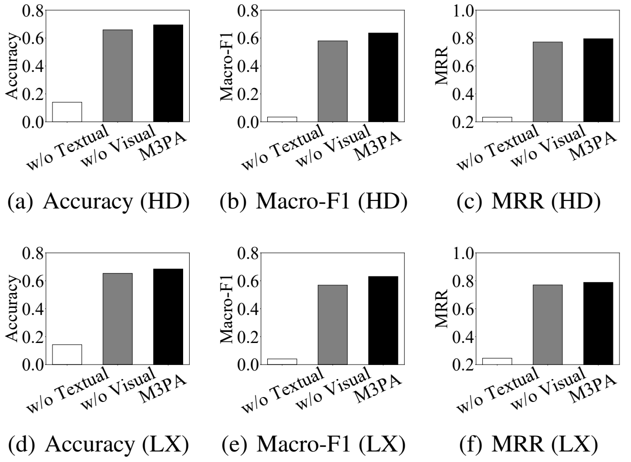
4.3 Ablation Study and Parameter Analysis

Study of Different Variants

We design two variants to explore how each of our modules affects the performance of M3PA: 1) M3PA w/o Textual: it removes the textual feature learning component and only utilizes the visual representation for POI annotation; 2) M3PA w/o Visual: it removes the visual feature learning component and only utilizes the textual representation for POI annotation. As shown in Figure 4, M3PA w/o Textual performs poorly, as urban images offer some but not sufficient information for understanding POI semantics; M3PA outperforms M3PA w/o Visual, indicating that urban images could be a useful supplement to POI names that enables POI annotation.

Study of Different Fusion Methods

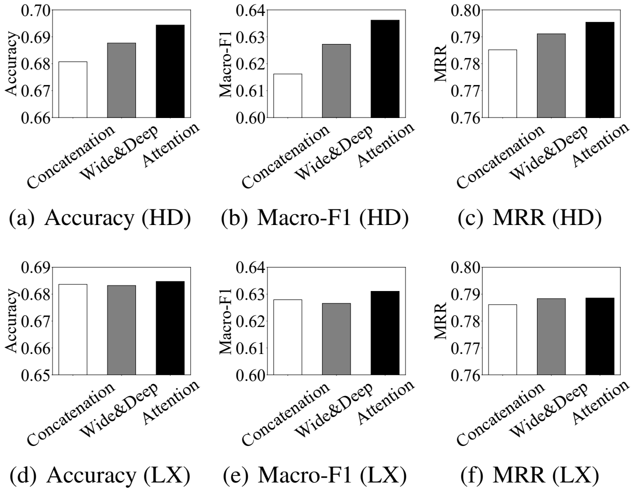
Figure 5 illustrates the results of three fusion methods introduced in Section 3.4. We observe that the attention-based fusion method slightly outperforms the other two, as it can



(a) Accuracy (HD) (b) Macro-F1 (HD) (c) MRR (HD)

(d) Accuracy (LX) (e) Macro-F1 (LX) (f) MRR (LX)

Figure 4: Performance comparison in different variants (HD: Haidian data, LX: Lixia data).



(a) Accuracy (HD) (b) Macro-F1 (HD) (c) MRR (HD)

(d) Accuracy (LX) (e) Macro-F1 (LX) (f) MRR (LX)

Figure 5: Performance comparison of different fusion methods (HD: Haidian data, LX: Lixia data).

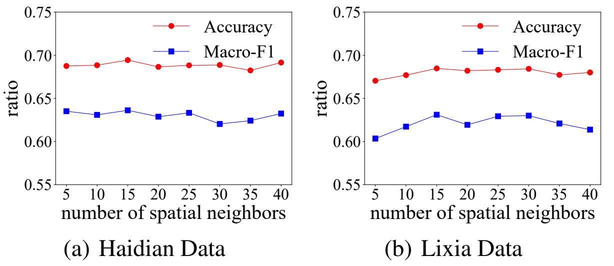
automatically determine the weights of each representation, thus effectively integrating the visual and textual representations to produce the final representation.

Study of Parameter Sensitivity

We analyze how the model's performance is affected by the number of spatial neighbors (k) of a targeted POI. We increment k from 5 to 40 in steps of 5. The results are presented in Figure 6. We note that the performance improves as we raise k from 5 to 15, and then stabilizes as we further increase it.

5 Related Work

Traditional POI semantic annotation methods mainly model users' check-ins to predict the category labels of POIs with manually designed features [Li *et al.*, 2020; Ye *et al.*, 2011]. Further, several studies [Wang *et al.*, 2017; Xu *et al.*, 2022; Xu *et al.*, 2023a] proceed to generate representations of POIs and categories by modeling check-in sequences using embedding techniques. They annotate POIs by calculating the similarity between POI representations and category representations.



(a) Haidian Data (b) Lixia Data

Figure 6: Effect of number of spatial neighbors.

In addition to users' check-ins, some studies utilize external semantic data (e.g., POI descriptions and users' comments including texts and images) to identify POI semantics for enhanced POI annotation. For instance, Meng *et al.* [2017] model the text-image pairs and predict the category labels of venues using a feature-level fusion method. Yang *et al.* [2023] presents a multimodal model to extract the textual and visual features of POIs from user comments for POI tagging. However, these methods do not capture the spatial relationships among POIs and the urban contextual information inherent in street view images.

Meanwhile, several studies have addressed the issue of semantic annotation for POIs using only geographic locations and textual names [Liu *et al.*, 2020; Zhang *et al.*, 2023]. They fail to model the names of spatially adjacent POIs and urban street view images to help better understand POI semantics.

6 Conclusion

In this study, we present a multimodal POI semantic annotation model (M3PA) that utilizes street view images as a crucial supplement to the limited POI information such as POI names and geographic locations. M3PA consists of two key components: 1) it employs a pre-training strategy to guide the image encoder in extracting visual features from street view images and integrates the visual features of nearby POIs to generate the visual representations; 2) it combines the POI names of spatially adjacent POIs and generates the textual representations of POIs via a geographic language model. The two types of representations are fused for POI semantic annotation. We conduct comprehensive experiments using POI data from Amap to demonstrate the effectiveness of M3PA. We find that POI names are more informative than street view images in POI semantic annotation, with the latter serving as a useful supplement.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 42101421, 61906107, and 62202270, the Young Scholars Program of Shandong University. W.H. was supported by the Knut and Alice Wallenberg Foundation.

References

[Bing *et al.*, 2023] Junxiang Bing, Meng Chen, Min Yang, Weiming Huang, Yongshun Gong, and Liqiang Nie. Pretrained semantic embeddings for poi categories based on multiple contexts. *IEEE Transactions on Knowledge and Data Engineering*, 35(09):8893–8904, 2023.

[Chen *et al.*, 2020] Meng Chen, Yan Zhao, Yang Liu, Xiaohui Yu, and Kai Zheng. Modeling spatial trajectories with attribute representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(4):1902–1914, 2020.

[Chen *et al.*, 2021] Meng Chen, Lei Zhu, Ronghui Xu, Yang Liu, Xiaohui Yu, and Yilong Yin. Embedding hierarchical structures for venue category representation. *ACM Transactions on Information Systems*, 40(3):1–29, 2021.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[Huang *et al.*, 2023] Weiming Huang, Daokun Zhang, Gengchen Mai, Xu Guo, and Lizhen Cui. Learning urban region representations with pois and hierarchical graph infomax. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196:134–145, 2023.

[Kenton and Toutanova, 2019] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[Lagos *et al.*, 2020] Nikolaos Lagos, Salah Ait-Mokhtar, and Ioan Calapodescu. Point-of-interest semantic tag completion in a global crowdsourced search-and-discovery database. In *Proceedings of the 24th European Conference on Artificial Intelligence*, pages 2993–3000. 2020.

[Lee *et al.*, 2021] Jihyeon Lee, Dylan Grosz, Burak Uzkent, Sicheng Zeng, Marshall Burke, David Lobell, and Stefano Ermon. Predicting livelihood indicators from community-generated street-level imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 268–276, 2021.

[Li *et al.*, 2020] Yanhui Li, Xiangguo Zhao, Zhen Zhang, Ye Yuan, and Guoren Wang. Annotating semantic tags of locations in location-based social networks. *GeoInformatica*, 24(1):133–152, 2020.

[Li *et al.*, 2021] Yang Li, Si Si, Gang Li, Cho-Jui Hsieh, and Samy Bengio. Learnable fourier features for multi-dimensional spatial positional encoding. *Advances in Neural Information Processing Systems*, 34:15816–15829, 2021.

[Li *et al.*, 2022] Tong Li, Shiduo Xin, Yanxin Xi, Sasu Tarkoma, Pan Hui, and Yong Li. Predicting multi-level socioeconomic indicators from structural urban imagery. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 3282–3291, 2022.

[Li *et al.*, 2024] Zechen Li, Weiming Huang, Kai Zhao, Min Yang, Yongshun Gong, and Meng Chen. Urban region embedding via multi-view contrastive prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8724–8732, 2024.

[Liu *et al.*, 2020] Shaopeng Liu, Jifan Yu, Juanzi Li, and Lei Hou. Geographical information enhanced poi hierarchical classification. In *International Conference on Web Information Systems and Applications*, pages 108–119, 2020.

[Liu *et al.*, 2023] Yu Liu, Xin Zhang, Jingtao Ding, Yanxin Xi, and Yong Li. Knowledge-infused contrastive learning for urban imagery-based socioeconomic prediction. In *Proceedings of the ACM Web Conference 2023*, pages 4150–4160, 2023.

[Meng *et al.*, 2017] Kaidi Meng, Haojie Li, Zhihui Wang, Xin Fan, Fuming Sun, and Zhongxuan Luo. A deep multi-modal fusion approach for semantic place prediction in social media. In *Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes*, pages 31–37, 2017.

[Sun *et al.*, 2021] Huimin Sun, Jiajie Xu, Kai Zheng, Pengpeng Zhao, Pingfu Chao, and Xiaofang Zhou. Mfnp: A meta-optimized model for few-shot next poi recommendation. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, pages 3017–3023, 2021.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[Villegas and Aletras, 2021] Danae Sánchez Villegas and Nikolaos Aletras. Point-of-interest type prediction using text and images. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7785–7797, 2021.

[Wang *et al.*, 2017] Yue Wang, Meng Chen, Xiaohui Yu, and Yang Liu. Lce: A location category embedding model for predicting the category labels of pois. In *Proceedings of the 2017 International Conference on Neural Information Processing*, pages 710–720, 2017.

[Wojna *et al.*, 2017] Zbigniew Wojna, Alexander N Gorban, Dar-Shyang Lee, Kevin Murphy, Qian Yu, Yeqing Li, and Julian Ibarz. Attention-based extraction of structured information from street view imagery. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 844–850, 2017.

[Xu *et al.*, 2022] Haoran Xu, Ronghui Xu, Meng Chen, Yang Liu, and Xiaohui Yu. Cave-sc: Inferring categories for venues using check-ins. *Information Sciences*, 611:159–172, 2022.

[Xu *et al.*, 2023a] Ronghui Xu, Meng Chen, Yongshun Gong, Yang Liu, Xiaohui Yu, and Liqiang Nie. Tme: Tree-guided multi-task embedding learning towards semantic venue annotation. *ACM Transactions on Information Systems*, 41(4):1–24, 2023.

[Xu *et al.*, 2023b] Ronghui Xu, Weiming Huang, Jun Zhao, Meng Chen, and Liqiang Nie. A spatial and adversarial representation learning approach for land use classification with pois. *ACM Transactions on Intelligent Systems and Technology*, 14(6):1–25, 2023.

[Yang *et al.*, 2023] Jingsong Yang, Guanzhou Han, Deqing Yang, Jingping Liu, Yanghua Xiao, Xiang Xu, Baohua Wu, and Shenghua Ni. M3pt: A multi-modal model for poi tagging. *arXiv preprint arXiv:2306.10079*, 2023.

[Ye *et al.*, 2011] Mao Ye, Dong Shou, Wang-Chien Lee, Peifeng Yin, and Krzysztof Janowicz. On the semantic annotation of places in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 520–528, 2011.

[Yin *et al.*, 2021] Yifang Yin, Ying Zhang, Zhenguang Liu, Sheng Wang, Rajiv Ratn Shah, and Roger Zimmermann. Gps2vec: Pre-trained semantic embeddings for worldwide gps coordinates. *IEEE Transactions on Multimedia*, 24:890–903, 2021.

[Zhang *et al.*, 2023] Dabin Zhang, Ronghui Xu, Weiming Huang, Kai Zhao, and Meng Chen. Towards an integrated view of semantic annotation for pois with spatial and textual information. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 2441–2449, 2023.