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

Dabin Zhang\textsuperscript{1}, Meng Chen\textsuperscript{*1}, Weiming Huang\textsuperscript{2}, Yongshun Gong\textsuperscript{1}, Kai Zhao\textsuperscript{3}

\textsuperscript{1}School of Software, Shandong University
\textsuperscript{2}School of Computer Science and Engineering, Nanyang Technological University
\textsuperscript{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
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 pre-trained 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.

## 2 Preliminaries

### 2.1 Problem Statement

**Definition 1 (Point of Interest).** The information of a POI \( p_i \) consist of a POI ID \( i_d \), 2D geographic coordinates \( g_i \), and a POI name represented as a set of words \( W_i \), denoted as \( p_i = (i_d, g_i, W_i) \). The coordinate \( g_i = (\text{lng}_i, \text{lat}_i) \) 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.

---

2. https://lbsyun.baidu.com/
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 $V_i$,

$$V_i = \text{ResNet}(s_i).$$

(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 $F_i \in \mathbb{R}^{N_c}$. Here, $N_c$ represents the total number of POI categories, and each dimension in $F_i$ corresponds to the number of POIs with a specific category in the zone.

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

$$L_{\text{image}} = \sum_{i=1}^{\mid S \mid} \text{KL}(F_i, \tilde{V}_i),$$

(2)

$$\tilde{V}_i = W_2 \cdot \text{ReLU}(W_1 \cdot V_i + b_1) + b_2,$$

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

Figure 2: The framework of M3PA.
\( \mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1^* \) and \( \mathbf{b}_2^* \) 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 \( g_{ij} \) of POI \( p_i \) and its \( k \) neighboring POIs \( \{p_{i1}, p_{i2}, \ldots, p_{ik}\} \) as input. The geographic distance between \( g_{i} \) and \( g_{i,j} \) is computed using the Gaussian kernel function,

\[
d(g_{i,j}) = \exp(-\text{geodist}(g_{i}, g_{i,j})\sigma^2/2), \quad (3)
\]

where \( \text{geodist}(\cdot) \) denotes the geodesic distance and \( \sigma \) represents the hyper-parameter governing the decay of similarity with respect to distance.

The distances between \( p_i \) and its \( k \) neighboring POIs 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 \( \mathbf{D}_i \),

\[
w(p_{i}, p_{i,j}) = \exp(\mathbf{H}_k) / \left( \sum_{j=1}^{k} \exp(\mathbf{H}_{ij}) \right), \quad (4)
\]

\[
\mathbf{H}_k = \mathbf{W}_3 \cdot \mathbf{D}_i + \mathbf{b}_3^*,
\]

where \( \mathbf{W}_3 \) and \( \mathbf{b}_3^* \) are the parameters of a fully-connected layer. Lastly, the visual representation of \( p_i \) is calculated as

\[
\mathbf{E}_i^{\text{visual}} = \mathbf{V}_i^p + \mathbf{W}_4 \cdot \sum_{j=1}^{k} w(p_{i}, p_{i,j}) [\mathbf{V}_j^p || \mathbf{c}_j^*], \quad (5)
\]

where \( \mathbf{V}_i^p \) and \( \mathbf{V}_j^p \) are the visual features of POIs generated with the pre-trained image encoder, \( \mathbf{c}_j^* \) 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 \( \mathbf{E}_i^{\text{visual}} \) 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

\[
[\text{CLS}] \text{Pizza Hut} [\text{SEP}] \text{Garden Bar} [\text{SEP}] \text{Crosstown Diner} [\text{SEP}] \text{Jenavie Cafe} [\text{SEP}] \text{Healthy Fresh} [\text{SEP}]
\]

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

$$\mathbf{GE}_i = \frac{1}{\sqrt{M}} \{\cos g_i \cdot W^T_g \parallel \sin g_i \cdot W^T_g\},$$

(6)

where $\parallel$ denotes concatenation, $g_i$ represents the latitude and longitude information of $p_i$, and $W_g \in \mathbb{R}_g^{d \times 2}$ denotes the linear projection matrix. It is important to note that $W_g$ is initialized with a Gaussian distribution following independent and identically distributed properties, denoted as $W_g \sim N(0, \gamma^{-2})$, where $\gamma$ 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., $\mathbf{GE}_i \cdot \mathbf{GE}_j \approx \exp(-\parallel g_i - g_j \parallel^2 \times \gamma^{-2})$. Consequently, the similarity between the geographical embeddings of two POIs can 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 the masked POI (with 15% probability) in a pseudo sentence and predicts the spatial co-occurrence patterns of POIs, we predict the final POI name), in which it completes the full POI names from the masked POI (with 15% probability) in a pseudo sentence and predicts the spatial co-occurrence patterns of POIs, we predict the final POI name, in which it completes the full POI names from the masked POI (with 15% probability) in a pseudo sentence and predicts the spatial co-occurrence patterns of POIs, we predict the final POI name, in which it completes the full POI names from the masked POI (with 15% probability) in a pseudo sentence and predicts the spatial co-occurrence patterns of POIs, we predict the final POI name, in which it completes the full POI names from the masked POI (with 15% probability) in a pseudo sentence and predicts the spatial co-occurrence patterns of POIs, we predict the final POI name, in which it completes the full POI names from the masked POI (with 15% probability) in a pseudo sentence and predicts the spatial co-occurrence patterns of POIs, we predict the final POI name, in which it completes the full POI names from the masked POI (with 15% probability) in a pseudo sentence and predicts the spatial co-occurrence patterns of POIs, we predict the final POI name.

**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.

**4.4 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 $\mathbf{E}^{final}$ of the targeted POI $p_i$. Subsequently, we input $\mathbf{E}^{final}_i$ into a two-layer MLP to generate the predicted probability $\hat{\mathbf{y}}_i$, which denotes the probability that $p_i$ is associated with each category label. This is calculated as follows,

$$\hat{\mathbf{y}}_i = \mathrm{softmax}(\mathbf{W}_6 \cdot \text{ReLU}(\mathbf{W}_5 \cdot \mathbf{E}^{final}_i + \mathbf{b}_5) + \mathbf{b}_6),$$

(9)

where $\mathbf{W}_5$, $\mathbf{W}_6$, $\mathbf{b}_5$, and $\mathbf{b}_6$ represent the parameters.

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

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

(10)

where $|P|$ and $|C|$ represent the number of unique POIs and categories in the dataset, $y_{ij}$ represents whether POI $p_i$ is labeled with the $j$th category, and $\hat{\mathbf{y}}_{ij}$ represents the predicted probability that $p_i$ is labeled with the $j$th 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 BERT$_{Base}$, and GeoBERT has an output dimension of 768. We employ the same optimization parameters for

3https://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 $\hat{y}_i$, 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 = $\frac{\sum_{i=1}^{|P_{test}|} (1/\text{rank}_i)}{|P_{test}|}$, where $|P_{test}|$ is the number of POIs in the test set $P_{test}$ 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.

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

<table>
<thead>
<tr>
<th>Data</th>
<th>View</th>
<th>Method</th>
<th>Accuracy</th>
<th>Macro-F1</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haidian</td>
<td>Textual</td>
<td>WTF</td>
<td>60.35</td>
<td>47.88</td>
<td>72.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BERT</td>
<td>64.60</td>
<td>56.88</td>
<td>75.81</td>
</tr>
<tr>
<td></td>
<td>Spatial</td>
<td>GSF</td>
<td>13.59</td>
<td>4.49</td>
<td>23.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GPS2Vec</td>
<td>5.19</td>
<td>0.25</td>
<td>12.26</td>
</tr>
<tr>
<td></td>
<td>Visual</td>
<td>ResNet</td>
<td>6.21</td>
<td>0.15</td>
<td>13.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GeoCLR</td>
<td>6.12</td>
<td>0.65</td>
<td>13.11</td>
</tr>
<tr>
<td></td>
<td>Integrated</td>
<td>EHC</td>
<td>61.18</td>
<td>49.38</td>
<td>73.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WTF+GPS2Vec</td>
<td>60.92</td>
<td>48.77</td>
<td>73.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MM-Gated-XAtt</td>
<td>68.05*</td>
<td>58.20</td>
<td>77.96*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>STPA</td>
<td>67.26</td>
<td>61.75*</td>
<td>77.83</td>
</tr>
<tr>
<td></td>
<td>M3PA</td>
<td>69.43</td>
<td>63.62</td>
<td>79.54</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Improvements</td>
<td>2.03</td>
<td>3.03</td>
<td>2.03</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data</th>
<th>View</th>
<th>Method</th>
<th>Accuracy</th>
<th>Macro-F1</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lixia</td>
<td>Textual</td>
<td>WTF</td>
<td>61.49</td>
<td>52.70</td>
<td>73.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BERT</td>
<td>63.53</td>
<td>55.12</td>
<td>75.41</td>
</tr>
<tr>
<td></td>
<td>Spatial</td>
<td>GSF</td>
<td>13.31</td>
<td>3.34</td>
<td>24.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GPS2Vec</td>
<td>6.97</td>
<td>0.34</td>
<td>15.87</td>
</tr>
<tr>
<td></td>
<td>Visual</td>
<td>ResNet</td>
<td>10.27</td>
<td>1.27</td>
<td>20.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GeoCLR</td>
<td>11.06</td>
<td>1.76</td>
<td>21.41</td>
</tr>
<tr>
<td></td>
<td>Integrated</td>
<td>EHC</td>
<td>61.50</td>
<td>53.21</td>
<td>73.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WTF+GPS2Vec</td>
<td>61.62</td>
<td>53.21</td>
<td>74.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MM-Gated-XAtt</td>
<td>65.65*</td>
<td>54.27</td>
<td>76.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>STPA</td>
<td>65.46</td>
<td>59.72*</td>
<td>76.90*</td>
</tr>
<tr>
<td></td>
<td>M3PA</td>
<td>68.51</td>
<td>63.00</td>
<td>78.91</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Improvements</td>
<td>4.36</td>
<td>5.49</td>
<td>2.61</td>
<td></td>
</tr>
</tbody>
</table>

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.
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.

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


[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


