Urban2Vec: Incorporating Street View Imagery and POIs for Multi-Modal Urban Neighborhood Embedding

Zhecheng Wang,[1*] **Haoyuan Li,**[2*] **Ram Rajagopal**[1]

[1]Stanford University, [2] Renmin University of China

{zhecheng, ramr}@stanford.edu, 2016201695@ruc.edu.cn

Abstract

Understanding intrinsic patterns and predicting spatiotemporal characteristics of cities require a comprehensive representation of urban neighborhoods. Existing works relied on either inter- or intra-region connectivities to generate neighborhood representations but failed to fully utilize the informative yet heterogeneous data within neighborhoods. In this work, we propose Urban2Vec, an unsupervised multi-modal framework which incorporates both street view imagery and point-of-interest (POI) data to learn neighborhood embeddings. Specifically, we use a convolutional neural network to extract visual features from street view images while preserving geospatial similarity. Furthermore, we model each POI as a bag-of-words containing its category, rating, and review information. Analog to document embedding in natural language processing, we establish the semantic similarity between neighborhood ("document") and the words from its surrounding POIs in the vector space. By jointly encoding visual, textual, and geospatial information into the neighborhood representation, Urban2Vec can achieve performances better than baseline models and comparable to fully-supervised methods in downstream prediction tasks. Extensive experiments on three U.S. metropolitan areas also demonstrate the model interpretability, generalization capability, and its value in neighborhood similarity analysis.

Introduction

More than 50% of the world population live in urban areas and this share is projected to be 68% by 2050 (Ritchie and Roser 2019). Generally, an urban area consists of a large number of spatially-distributed neighborhoods. Each neighborhood is a complex mixture of multiple components including its physical environment, local business, and the people living there. Representing and understanding the characteristics and dynamics of urban neighborhoods are essential for various downstream tasks, such as urban planning, business model development, and social welfare improvement. Survey is a common but costly method to uncover neighborhood characteristics. For example, U.S. Census Bureau spends $250 million per year on the American Community Survey (ACS) which collects demographic statistics of different neighbourhoods (Gebru et al. 2017).

Data gathered through such time-consuming surveys can lag behind real-world urban change particularly in regions where the resources for data collection are limited. To provide decision makers with timely yet comprehensive information, we need a time-saving, cost-efficient, and general approach to characterize urban areas.

Since the intrinsic characteristics of urban neighborhoods are well embedded in their physical appearance, business activities, and even geo-tagged text information, recent data-driven methods have leveraged these widely-available data to predict or interpret the demographic and socioeconomic patterns. For example, point-of-interest (POI) categories have been used to infer region functions (Yuan, Zheng, and Xie 2012); text data from Twitter has been leveraged to predict income, occupation, and unemployment rates (Antenucci et al. 2014; Aletras and Chamberlain 2018); imagery data, including street view images and satellite images, have been utilized to predict perceived safety, house price, and demographic makeup (Naik et al. 2014; Law, Paige, and Russell 2018; Gebru et al. 2017). However, these methods are all based on task-specific supervised learning. Representations learned for one task are not necessarily useful in other tasks. Furthermore, directly applying supervised learning on high-dimensional raw data, such as images, requires massive labeled data for training, which is impractical in the cases when labels are sparse.

Learning multipurpose representations of urban neighborhoods without requiring massive amounts of labeled data has led to research that represents neighborhoods as vectors while preserving the associations with urban attributes. Analog to word embeddings in NLP (Mikolov et al. 2013; Pennington, Socher, and Manning 2014) and graph embeddings in social network analysis (Perozzi, Al-Rfou, and Skiena 2014; Grover and Leskovec 2016), the key idea in neighborhood embeddings is to define a "similarity" metrics between objects, capture it from data, and encode it into vector representations. For example, Zhang et al. (2017) and Yao et al. (2018) extract "similarity" between regions from urban mobility flow, and optimize the neighborhood embedding to represent the neighborhood function via either skip-gram or matrix factorization techniques. By contrast, Wang et al. (2018) obtained the region embedding with the "similarity" measurement derived from the POI categories and inter-POI connectivity inside each neighborhood. Be-

---

*Equal contribution

sides intra-region POI information, Fu et al. (2019) further incorporated the inter-region proximity based on both distance and POI-derived functionality to improve the embedding. However, all these methods either relied on connectivities between neighborhoods regardless of components inside neighborhoods or included only a single type of component such as POI. In reality, neighborhoods are complex systems involving the local natural and built environment, population, businesses and their interconnections. Representations can only capture these relationships if different *modalities* of granular data are combined. To the best of our knowledge, no existing work has fully utilized the rich and multi-modal information inside neighborhoods and incorporated them together to generate the neighborhood representations.

To bridge this gap, we propose **Urban2Vec**, an *unsupervised* and *multi-modal* method for learning compact yet comprehensive representations of urban neighborhoods by incorporating heterogeneous data associated with them. We model each neighborhood as a "container" consisting of inter-correlated components including its physical environment, business activities, and population. While population demographics are expensive to obtain and update, features of physical environment and business activities can be well captured in the widely-available and frequently-updated street views and POI data, respectively. Based on this idea, we develop a multi-stage approach to integrate information from street view images and POIs to generate the neighborhood representations. For street view images, visual features are extracted using convolutional neural network (CNN) while preserving geospatial correlation. For POIs, instead of using POI statistics such as number of POIs in different categories, we model each POI as a bag-of-words by collecting the textual information on its price, rating, and reviews. A major advantage of such modeling is that the semantic correlations lying behind words can be captured and projected into the vector space. Specifically, our contribution is three-fold:

- We develop a multi-modal and multi-stage framework to generate neighborhood representations integrating both image and textual data inside neighborhoods. Unlike previous multi-modal embedding techniques that aim at establishing inter-correlations between different modalities of objects, our major goal is to establish the correlation between the "container" (neighborhood) and its inside objects. In this paper, the inside objects are street view images and POI data, but the framework can be extended to other data such as satellite images and geo-tagged posts.

- We conduct experiments on three metropolitan areas: Bay Area, New York, and Chicago. For demographic and socioeconomic attribute prediction task, Urban2Vec achieves performances better than baseline methods and comparable to fully-supervised methods yet with better generalization capability.

- We map the neighborhoods of New York and Chicago into the same vector space, and illustrate how the embeddings obtained through Urban2Vec can be used to draw similarities between neighborhoods across different cities.

Problem Statement

Definition 1 *(Urban Neighborhood).* *A city or a metropolitan area can be represented by a set of urban neighborhoods* $\mathcal{R} = \{r_1, r_2, ..., r_N\}$. *Each urban neighborhood* r_i *contains a set of street view images* $\mathcal{S}_i = \{s_{i1}, s_{i2}, ..., s_{iM_i}\}$ *taken inside* r_i, *and a set of POIs* $\mathcal{P}_i = \{p_{i1}, p_{i2}, ..., p_{iO_i}\}$ *located in* r_i.

Remark 1. Street view images capture the physical environment in a neighborhood, while POIs, including various types of restaurants, stores, schools, etc., represent the business activities inside the neighborhood. Here we assume the GPS coordinates of both street views and POIs are known.

Definition 2 *(POI Textualization).* *A POI* p *can be textualized as a bag of words* $\{t_1, t_2, ..., t_q\}$, *with the words extracted from* p*'s categories, rating, price and customer reviews. Then, the POI set of neighborhood* r_i *can also be represented by a collection of POI words* $\mathcal{T}_i = \{t_{i1}, t_{i2}, ..., t_{iH_i}\}$, *which is obtained by merging the bags-of-words of all POIs in* r_i.

Remark 2. Rather than generate POI embeddings as an intermediate step, we merge bags-of-words of POIs to obtain a larger bag-of-words for each neighborhood in order to simplify the training process. Furthermore, such bag-of-words modeling can be easily generalized to any other geo-tagged textual data, such as Twitter and Facebook posts. The details of constructing $\mathcal{T}_i$ for each neighborhood will be introduced in Methodology. Here we formulate the problem to investigate as follows:

Definition 3 *(Urban Neighborhood Embedding).* *Given a set of neighborhoods* $\mathcal{R}$, *together with street view collection* $\mathcal{S}_i$ *and POI word collection* $\mathcal{T}_i$ *inside each neighborhood* r_i, *the aim is to learn a vector representation* $z_i \in \mathbb{R}^d$ *for each* r_i, *where* d *is the uniform dimension for all* r_i *in* $\mathcal{R}$.

Methodology

In this section, we introduce the multi-stage framework, Urban2Vec, for urban neighborhood embedding. We first present a framework overview, followed by the detailed steps for incorporating street view images and POI data.

Framework Overview

Each neighborhood r_i is modeled as a "container" consisting of a street view collection $\{s_{ij}\}_{1 \le i \le N, 1 \le j \le M_i}$ and a POI bag-of-words $\{t_{ij}\}_{1 \le i \le N, 1 \le j \le H_i}$. Besides the embedding $z_i \in \mathbb{R}^d$ for each r_i, we also generate embedding $x_{ij} \in \mathbb{R}^d$ for each street view s_{ij}, and embedding $y_{ij} \in \mathbb{R}^d$ for each POI word t_{ij}. To incorporate both semantic and geospatial information extracted from street views and POIs into neighborhood representations, we seek to establish relationships between $\{z_i\}_{1 \le i \le N}$ and $\{x_{ij}\}_{1 \le i \le N, 1 \le j \le M_i}$ as well as $\{y_{ij}\}_{1 \le i \le N, 1 \le j \le H_i}$ in the vector space. Specifically, for each urban neighborhood r_i, we define its street view "context" as the collection of street views located in that neighborhood, which is $\mathcal{S}_i$, and its POI word "context" as the collection of words from POIs in that neighborhood, which is $\mathcal{T}_i$. Our basic assumption is that a neighborhood should be
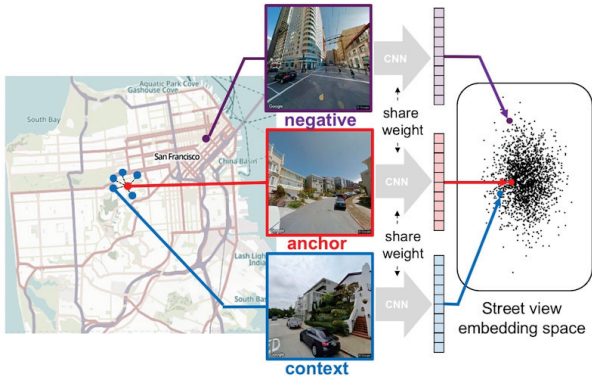
Figure 1: Street view embedding with negative sampling. The anchor image is mapped into closer position to context image than negative image in the vector space.

highly correlated with the street views and POI words inside its "context", while less correlated with those outside its "context". To this end, we aim to minimize the distances of neighborhood with both its street view "context" and POI word "context" in the vector space:

$$L_{nb\text{-}sv} = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \text{dist}(z_i, x_{ij}) \quad s_{ij} \in \mathcal{S}_i \quad (1)$$

$$L_{nb\text{-}poi} = \sum_{i=1}^{N} \sum_{j=1}^{H_i} \text{dist}(z_i, y_{ij}) \quad t_{ij} \in \mathcal{T}_i \quad (2)$$

where "dist" is a distance function in the vector space. Here we use Euclidean distance. To obtain street view embedding $\{x_{ij}\}$, we apply a CNN on raw images $\{s_{ij}\}$ to extract visual features and project down the dimensionality. By contrast, the embedding of a POI word is retrieved from an embedding matrix $Y \in \mathbb{R}^{|C| \times d}$, where $|C|$ is the size of the POI word corpus. Minimizing $L_{nb\text{-}sv}$ involves much more parameters to train than minimizing $L_{nb\text{-}poi}$, which can make their paces of convergence difficult to be concordant. Therefore simultaneously minimizing them can yield unsatisfactory results (See performance comparison in Appendix A.1.2). Instead, we propose an approach to optimize them separately in two stages. In the first stage, we minimize $L_{nb\text{-}sv}$ leveraging both geospatial relationship between neighborhood and street views, and that between street view and street view. In the second stage, we minimize $L_{nb\text{-}poi}$ to further improve the neighborhood embedding by incorporating textual information of POIs.

Incorporating Street View Imagery

Each street view image can appear in the context of only one neighborhood, thus directly minimizing $L_{nb\text{-}sv}$ will generate multiple *isolated* agglomerations in the vector space, where each agglomeration is formed with a neighborhood embedding surrounded by the embeddings of its contextual street views. This is contrary to the fact that neighborhoods are spatially correlated with each other rather than isolated. To tackle this issue, we further decompose the process by firstly

minimizing the distance between each street view and its contextual street views:

$$L_{sv\text{-}sv} = \sum_{k} \sum_{x_c \in \mathcal{N}_k} \text{dist}(x_k, x_c) \quad (3)$$

and then minimizing $L_{nb\text{-}sv}$ while keeping all x_k fixed. Here $\mathcal{N}_k$ denotes the "context" of street view s_k.

Triplet loss. To optimize $L_{sv\text{-}sv}$, we need to firstly answer the question: What can be the "context" of a street view image? In satellite image tile embedding (Jean et al. 2019), the "context" is defined as the geographic neighbors of a tile, based on the assumption that tiles that are spatially close should have higher semantic similarity, while tiles far apart tend to have less. In our work, we extend this assumption to street view images: Street views with small geographic distance are more likely to share common semantics than those with large distance. Therefore, we also define the "context" of a street view image as its geospatial neighbors. Specifically, to optimize $L_{sv\text{-}sv}$, we train the CNN on a collection of triplets (s_a, s_c, s_n), where s_a is the anchor image, s_c is an image inside s_a's context, and s_n is a negative sample outside s_a's context (See Figure 1). To enforce such relative similarity in the vector space, we minimize the triplet loss to enlarge the margin between anchor-negative distance and anchor-context distance:

$$L_{sv\text{-}sv}^{tri}(x_a, x_c, x_n) = [m + ||x_a - x_c||_2 - ||x_a - x_n||_2]_+ \quad (4)$$

where $[\cdot]_+$ is a rectifier and $||\cdot||_2$ is the Euclidean distance. Margin m is used to prevent infinitely large difference between these two distances. x_a, x_c, and x_n are the embeddings of s_a, s_c, and s_n, respectively, which are outputted by a shared CNN, i.e., $x\cdot = f_\theta(s\cdot)$ where f_θ denotes a CNN parameterized by θ. Note that $L_{sv\text{-}sv}^{tri}$ can be recognized as an implementation of $L_{sv\text{-}sv}$ with negative sampling.

Triplet sampling. Rather than use absolute geographic distance to determine the "context", we define the "context" of an anchor image as its K nearest images. This is because street view images can be densely distributed in some regions while sparse in others. Using absolute geographic distance will result in highly variable context size for different anchor images.

Minimizing $L_{nb\text{-}sv}$. After obtaining the street view embedding $\{x_{ij}\}$ by minimizing $L_{sv\text{-}sv}^{tri}$, we then minimize $L_{nb\text{-}sv}$ while keeping all x_{ij} fixed. We do not apply negative sampling in this step, since the relative similarity has already been captured through the geospatial context in street view embedding. Instead, we simply take the average of all x_{ij} located in neighborhood i as the neighborhood embedding z_i, which is exactly the analytical solution of minimizing $L_{nb\text{-}sv}$:

$$z_i = \frac{1}{M_i} \sum_{j=1}^{M_i} x_{ij} \quad s_{ij} \in \mathcal{S}_i \quad (5)$$
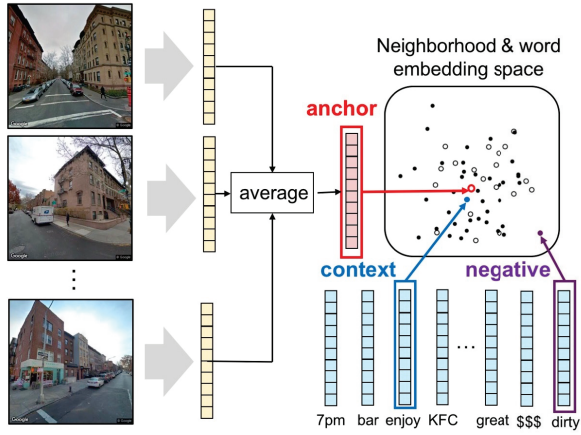
This process is illustrated in Figure 2.

Figure 2: Generating neighborhood embedding by aggregating street view embedding and interacting with POI word embedding.

Incorporating POI Textual Data

In contrast to street view embedding which leverages CNN to project down the dimensionality, the vector representation of each POI word is directly retrieved from the word embedding matrix Y. This is analog to document embedding (Le and Mikolov 2014), where a neighborhood can be regarded as a document containing words. To encode textual semantics regarding local business into neighborhood representations, we define another triplet loss to train both POI word embedding Y and neighborhood embedding $Z = \{z_i\}_{1 \leq i \leq N} \in \mathbb{R}^{N \times d}$. In this stage, each triplet (r_a, t_c, t_n) consists of an anchor neighborhood r_a, a POI word t_c from r_a's context, and a negative sample t_n outside r_a's context. Their embeddings are z_a, y_c, and y_n, respectively. Here we define the POI word "context" of r_a as the collection of words obtained from POIs inside r_a, which is $\mathcal{T}_a$. By minimizing the triplet loss:

$$L_{nb\text{-}poi}^{tri}(z_a, y_c, y_n) = [m' + ||z_a - y_c||_2 - ||z_a - y_n||_2]_+ \quad (6)$$

we make the neighborhood r_a closer to its contextual POI words than to the non-contextual POI words in the vector space. Here m' is another margin value serving the same purpose as in street view embedding. Note that $L_{nb\text{-}poi}^{tri}$ can also be regarded as an implementation of $L_{nb\text{-}poi}$ with negative sampling and Euclidean distance metrics. After training Y and Z on the POI textual dataset, the final neighborhood embedding Z incorporates the semantics from both street views and POIs inside each neighborhood. The whole process is illustrated in Figure 2.

Bag-of-words construction. To construct the POI word "context" $\mathcal{T}_i$ for each neighborhood, we merge the bags-of-words of all POIs located in that neighborhood. For each POI, its bag-of-words includes its categories, rating, price, and words from its customer reviews. In our setting, a category phrase (e.g. "shopping center") is regarded as a single word. Each category word is prefixed with "cat_" to make it distinguishable from the same word appearing in reviews.

Duplicate words are removed from reviews. However, we do not eliminate duplication when merging words from different POIs, as the word frequency is highly indicative of the neighborhood attributes. For example, the neighbourhood with 100 "restaurant" occurrences tends to be more prosperous than that with only 5 "restaurant" occurrences.

Negative sampling. We apply the same negative sampling scheme as in (Mikolov et al. 2013) based on word frequencies in the corpus. Specifically, for word t_k which is outside r_i's context $\mathcal{T}_i$, its possibility of being selected as negative sample is $\frac{F_k^{0.5}}{\sum_{j \in \neg \mathcal{T}_i} F_j^{0.5}}$, where F_k denotes the number of times that word t_k appears in the corpus.

Experiments

By conducting experiments on real-world dataset, we demonstrate: (1) the performances of using the neighborhood embedding derived by Urban2Vec for downstream prediction tasks, (2) the interpretability underlying the vector representation, and (3) the usage of the embedding for neighborhood similarity analysis.

In this work, we use census tracts to define "neighborhoods". Census tracts are small and relatively homogeneous units regarding their demographics, and the census-tract-level data of these attributes is readily available, which is favorable for our model evaluation. Our framework is easy to be extended to other geographic units such as block groups or counties, which is a part of our future work.

Experimental Setups

We collect street view images and POI information in three major metropolitan areas in the US: San Francisco Bay Area, Chicago, and New York. Street view images are retrieved using Google Street View Static API. POI information is obtained with Yelp Fusion API. The statistics of our datasets are shown in Table 1.

	# street views	# POIs	# unique words	# census tracts
Bay Area	44174	69765	14279	1198
Chicago	64739	38445	10013	1317
New York	67271	50697	11386	1371

Table 1: Dataset statistics.

To obtain the street view embedding, we use context size $K = 5$. The CNN used in our work is an Inception-v3 architecture (Szegedy et al. 2016) with a final linear layer projecting 2048-dimensional feature into the d-dimensional embedding space. We use $d = 200$ for result analysis, and the dimension sensitivity analysis is shown in Appendix A.2. We use parameters pre-trained on ImageNet to intialize the CNN model. In POI word embedding, we use the pretrained GloVe (Pennington, Socher, and Manning 2014) to initialize the embeddings of words in reviews, while embeddings of other words like price and rating are randomly initialized.

Baseline Models

We compare Urban2Vec with other unimodal and multimodal representations of urban neighborhoods. Specifically,

(a) Average R^2

(b) Average household income (**AHI**)

(c) Average years of education (**AYE**)
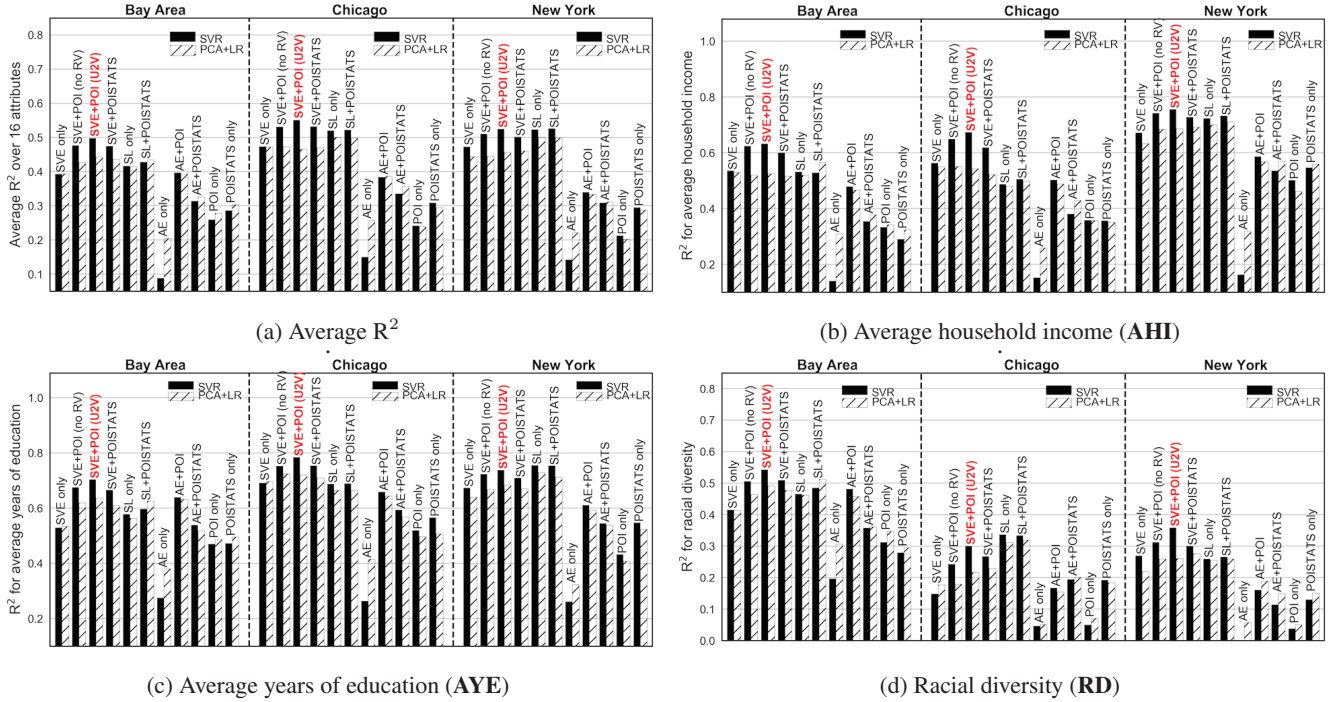
(d) Racial diversity (**RD**)

Figure 3: Performances in demographic attribute prediction based on both SVR (black bar) and PCA+LR (hatching bar). See abbreviation description in **Experiments-Baseline Models** section.

we investigate four methods of incorporating street view images: (1) street view embedding (**SVE**) proposed in this paper, (2) convolutional autoencoder (**AE**) proposed in (Law and Neira 2019), which uses the encoder to embed street views into vector space, (3) supervised learning (**SL**), and (4) none. Note that for supervised learning, we use the same Inception-v3 architecture with demographic attributes as supervisory signals, which serves as a benchmark for prediction tasks. Moreover, we compare three methods of incorporating POI information: (1) POI word embedding (**POI**) proposed in this paper, (2) statistics of the POI categories using tf.idf scheme (**POISTATS**), and (3) none.

Finally, we obtain 10 different neighborhood representations using their combinations: **SVE only**, **SVE+POI** (Urban2Vec, abbreviated as **U2V**), **SVE+POISTATS**, **SL only**, **SL+POISTATS**, **AE only**, **AE+POI**, **AE+POISTATS**, **POI only**, **POISTATS only**.

We do not compare **SL+POI** since the POI word corpus cannot be split into training and test sets for POI word embedding, while evaluating supervised learning methods requires strict training/test set partition to prevent overfitting. Besides, we also compare the Urban2Vec model excluding words from reviews, abbreviated as **SVE+POI (no RV)**.

Experimental Results

Predicting demographic and socioeconomic attributes.
We use neighborhood embeddings as inputs to predict their demographic attributes at census tract level. The demographic attributes are obtained from ACS and include 16 variables. Both Support Vector Regression (SVR) and lin-

ear regression with PCA preprocessing (PCA+LR) are used as regressors, since they are less subjective to model hyper-parameters than more complex models such as Neural Network regressor. Census tracts in each dataset are split into training (70%), validation (15%), and test sets (15%). For each variable we calculate the average R^2 on test set over 20 random splits. For supervised learning (**SL**), we directly use these 16 attributes for supervision during training.

Figure 3 shows the prediction results of three major variables: average household income (**AHI**), average number of years of education (**AYE**), racial diversity (**RD**), together with the overall R^2 of all these 16 variables on three datasets (See the results for the rest of variables in Appendix A.3). As is shown, when using SVR as regressor, Urban2Vec outperforms all other models in all three areas, except for predicting **AYE** in New York and predicting **RD** in Chicago on which **SL+POISTATS** performs slightly better by 0.02-0.03. The overall R^2 of **SVE+POISTATS** is higher than that of **SL+POISTATS** in Bay Area but lower in New York for both regressors, while in Chicago, SVR favors **SVE+POISTATS** but PCA+LR favors **SL+POISTATS**. Such results indicate that the representation obtained with our unsupervised method can achieve comparable performances with fully-supervised method in downstream prediction tasks. For regions where granular demographic features are not readily available or out-of-date, the unsupervised learning proposed in this paper will be a good approach to obtain the neighborhood representations. Moreover, all **SVE**-based methods have significantly higher R^2 than their **AE**-based counterparts. This may be because **SVE**

Figure 4: **(a)** k-means clustering on the embeddings in three areas (k=4). **(b)** Street views with extreme values in 1st PCA dimension. **(c)** Street views with extreme values in 2nd PCA dimension.

leverage the geospatial correlation to derive representation while autoencoder only learns visual features from images themselves regardless of the relationships between images.

Regarding the POI representations, although using **POISTATS** only can yield generally better prediction performances than using **POI** only, **SVE+POI** has a ~0.02 gain in average R^2 compared with **SVE+POISTATS** when using SVR, suggesting that in the multi-modal setting, POI word embedding is a better way to learn representations from unstructured POI data than using POI statistics as it captures the semantic nuances from the textual data. Comparing **SVE+POI (no RV)** with **SVE+POI** shows that including customer reviews leads to a ~0.02 increase in average R^2. Except for **SL**-based methods, incorporating both street views and POIs yields a remarkably better R^2 than incorporating only a single mode, indicating the strength of multi-modal learning for urban neighborhood representations.

	Urban2Vec	AE+POI	SL+POISTATS	demographic
Avg apartment sale price	**0.487**	0.391	0.417	0.445
Avg office sale price	**0.443**	0.308	0.328	0.337

Table 2: Real estate sale price prediction R^2.

We also evaluate embedding methods on predicting real estate sale price in Bay Area (Table 2). The data is obtained from Cushman & Wakefield and aggregated at census tract level. For **SL+POISTATS**, in order to compare its generalization capability with Urban2Vec, we still use the vector representation obtained from the demographic prediction as inputs, instead of re-training it with the real estate sale price as supervision. Result shows that for both apartment and office sale price predictions, Urban2Vec outperforms other methods including directly using demographic attributes as inputs. Though the performances of supervised method and Urban2Vec are comparable in demographic prediction, when transferring the learned representations to another task, Urban2Vec performs significantly better, indicating its good generalization capability.

Interpreting the learned representations. We apply k-means clustering on the embeddings obtained with Urban2Vec in three areas. Figure 4a shows the results with k=4. In Bay Area, regions with higher street enclosure and denser population, such as San Francisco and Oakland, are clustered into the same group. Similar cluster also occurs in

Chicago and New York (i.e. Central Chicago, most part of Manhattan). In each area, regions with less dense population and buildings are clustered into another group (green). Such clustering results align well with human intuition on the city structures of these three areas.

To uncover the semantic meaning learned in neighborhood representations, we apply PCA on the embedding matrix obtained in Bay Area, and visualize the street views and POI category words with extreme values along the directions of first and second principle components. They are shown in Figure 4 and Table 3, respectively. We find that the 1st component is associated with street enclosure (average building heights divided by average width between buildings of the same street), while the 2nd component is correlated with vegetation. Interestingly, we also observe that the 2nd component has strong positive correlation with average household income and education level when we use PCA+LR to predict demographic attributes in Bay Area, suggesting that denser greenery is indicative of high income/education level in Bay Area. Moreover, POI categories such as "gallery" and "winery" have high values in the 2nd dimension, while "discount store" has low value (Table 3), indicating that they are potentially associated with high income/education level and low income/education level, respectively.

dim1	min	dance club	comedy club	popup restaurant
	max	horse back riding	winery	pool cleaning
dim2	min	water store	discount store	
	max	gallery	bed breakfast	winery

Table 3: POI category words with extreme values in 1st and 2nd PCA dimensions.

Neighborhood similarity analysis. Vector representation of neighborhoods can be used to draw similarity efficiently between neighborhoods subject to geographic constraints. Specifically, we seek to answer questions like this: Which neighborhood in Chicago is most similar to neighborhood X in New York? To illustrate the value of Urban2Vec in addressing such problems, we train a joint embedding using Urban2Vec for all neighborhoods in New York and Chicago. Given a source neighborhood in one city, we can compute the cosine similarity between its embedding vector and all embedding vectors in another city, and then rank the neighborhoods according to the similarity metrics. Figure 5 shows
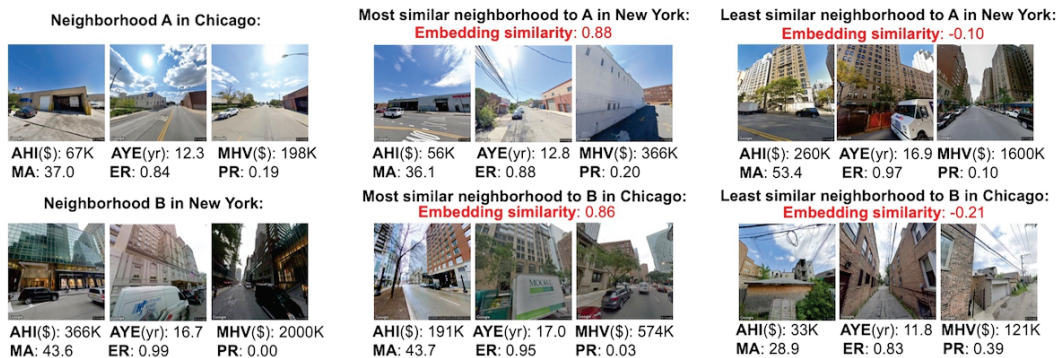
Figure 5: Searching for the most/least similar neighborhoods in another city with Urban2Vec embedding. Abbreviation: **AHI**: average household income. **AYE**: average years of education. **MHV**: median housing unit value. **MA**: median age. **ER**: employment rate. **PR**: ratio of families below poverty level.

two instances of similar neighborhood search (See more instances in Appendix A.4). Neighborhood A in Chicago and its most similar neighborhood in New York obtained through Urban2Vec share the similar physical appearance and demographic attributes: warehouse-like buildings, sparse vegetation, as well as medium-to-low income and employment rate. By contrast, the least similar neighborhood to A in New York features tall buildings, high income/housing unit value/employment rate, and a more aging community. Also, neighborhood B in New York and its "nearest neighbor" in Chicago both feature tall apartments, high income/education level/housing unit values, and near-zero poverty rate, while the least similar neighborhood to B in Chicago is characterized by its dilapidated buildings and opposite demographic features. To summarize, Urban2Vec transforms the complicated urban system comparison into a simple vector comparison problem, which can facilitate commercial site selection, real estate recommendation, and urban planning.

Conclusion

In this work, we propose Urban2Vec, an unsupervised multimodal learning framework to learn neighborhood representations by incorporating both street view and POI data. We demonstrate its high accuracy in downstream prediction tasks, interpretability underlying the embedding space, and its potential usage in neighborhood similarity analysis. Future work includes (1) incorporating other imagery, textual, and spatiotemporal data, such as mobility and social media data, to increase the information richness of neighborhood representations, (2) using user study to evaluate neighborhood similarity ranking performance, and (3) exploring the potential usage of Urban2Vec in analyzing urban change.

Related Work

Image embedding. Unsupervised representation of images, or image embedding, aims to project high-dimensional image data into compact but informative representations. In terms of specific applications in urban study, Law and Neira (2019) used convolutional autoencoder followed by PCA to generate representations of both street views and street network images. However, directly applying autoencoder on images does not establish any geospatial correlation between street views at different locations. Jean et al. (2019) developed an algorithm to learn vector representations of satellite images by using geospatial information as a weak supervision, which provided a paradigm for learning the representation of spatially distributed image data.

Text and graph embedding. In text embedding, the target is to enforce single words (Mikolov et al. 2013) or word sequences (Le and Mikolov 2014) in the same context to be close in the vector space, In graph embedding, each node is treated as a word with its "context" defined as a set of other nodes along a random walk (Perozzi, Al-Rfou, and Skiena 2014; Grover and Leskovec 2016). Inspired by them, existing neighborhood embedding methods modeled mobility flow as graphs or origin-destination pairs either within a region (Wang et al. 2018; Fu et al. 2019) or among regions (Zhang et al. 2017; Yao et al. 2018). Compared to these works, we fully utilize the heterogeneous data in each neighborhood and incorporate multiple modalities of components to obtain the comprehensive neighborhood representations.

Multi-modal embedding. Mapping different modalities of data into same latent space has been studied before. It has been shown that images and text can be projected into a single space using neural networks (Wang, Li, and Lazebnik 2016; Zhang et al. 2017), so did the combination of point-cloud, text, and robot manipulation trajectories (Sung, Lenz, and Saxena 2017), and the combination of text, location, and time (Zhang et al. 2017). However, instead of aiming at embedding data of different modalities into one space, the question we are interested in is how to obtain the embedding of a "container" (neighborhood) by integrating multiple modalities of data inside the "container".

Acknowledgements

References

Aletras, N., and Chamberlain, B. P. 2018. Predicting twitter user socioeconomic attributes with network and language information. In *Proceedings of the 29th on Hypertext and Social Media*, 20–24. ACM.

Antenucci, D.; Cafarella, M.; Levenstein, M.; Ré, C.; and Shapiro, M. D. 2014. Using social media to measure labor market flows. Technical report, National Bureau of Economic Research.

Fu, Y.; Wang, P.; Du, J.; Wu, L.; and Li, X. 2019. Efficient region embedding with multi-view spatial networks: A perspective of locality-constrained spatial autocorrelations.

Gebru, T.; Krause, J.; Wang, Y.; Chen, D.; Deng, J.; Aiden, E. L.; and Fei-Fei, L. 2017. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences* 114(50):13108–13113.

Grover, A., and Leskovec, J. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 855–864. ACM.

Jean, N.; Wang, S.; Samar, A.; Azzari, G.; Lobell, D.; and Ermon, S. 2019. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3967–3974.

Law, S., and Neira, M. 2019. Learning from discovering: An unsupervised approach to geographical knowledge discovery using street level and street network images. *arXiv preprint arXiv:1906.11907*.

Law, S.; Paige, B.; and Russell, C. 2018. Take a look around: using street view and satellite images to estimate house prices. *arXiv preprint arXiv:1807.07155*.

Le, Q., and Mikolov, T. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, 1188–1196.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Naik, N.; Philipoom, J.; Raskar, R.; and Hidalgo, C. 2014. Streetscore-predicting the perceived safety of one million streetscapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 779–785.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Perozzi, B.; Al-Rfou, R.; and Skiena, S. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 701–710. ACM.

Ritchie, H., and Roser, M. 2019. Urbanization. *Our World in Data*. https://ourworldindata.org/urbanization.

Sung, J.; Lenz, I.; and Saxena, A. 2017. Deep multimodal embedding: Manipulating novel objects with point-clouds, language and trajectories. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2794–2801. IEEE.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Wang, P.; Fu, Y.; Zhang, J.; Li, X.; and Lin, D. 2018. Learning urban community structures: A collective embedding perspective with periodic spatial-temporal mobility graphs. *ACM Transactions on Intelligent Systems and Technology (TIST)* 9(6):63.

Wang, L.; Li, Y.; and Lazebnik, S. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5005–5013.

Yao, Z.; Fu, Y.; Liu, B.; Hu, W.; and Xiong, H. 2018. Representing urban functions through zone embedding with human mobility patterns. In *IJCAI*, 3919–3925.

Yuan, J.; Zheng, Y.; and Xie, X. 2012. Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 186–194. ACM.

Zhang, C.; Zhang, K.; Yuan, Q.; Peng, H.; Zheng, Y.; Hanratty, T.; Wang, S.; and Han, J. 2017. Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning. In *Proceedings of the 26th International Conference on World Wide Web*, 361–370. International World Wide Web Conferences Steering Committee.