From Geolocated Images to Urban Region Identification and Description: a Large Language Model Approach

Guido Rocchietti
ISTI-CNR, University of Pisa
Pisa, Italy
guido.rocchietti@isti.cnr.it

Chiara Pugliese
ISTI-CNR, University of Pisa
Pisa, Italy
chiara.pugliese@isti.cnr.it

Gabriel Sartori Rangel
Universidade Federal de Santa Catarina
Florianópolis, Brazil
gabriel.sartori.rangel@grad.ufsc.br

Jonata Tyska Carvalho
Universidade Federal de Santa Catarina
Florianópolis, Brazil
jonata.tyska@ufsc.br

Abstract

Urban research faces challenges in understanding and describing city regions, which are essential for urban planning and tourism management. Traditional methods rely on predefined areas and non-human-readable representations. This paper presents a new unsupervised approach that overcomes these limitations using a data-driven method with Instruction-tuned Large Language Models (ILLMs). Our technique dynamically identifies urban regions with similar features and generates human-readable descriptions. We validate this method using Flickr images from Pisa, Italy, and our results show that it effectively captures the semantic features of urban regions and generates comprehensible textual descriptions.

CCS Concepts

• **Information systems** → **Geographic information systems**; • **Computing methodologies** → **Natural language generation**.

Keywords

Urban Regions, Image Captioning, Large Language Models, Summarization

ACM Reference Format:
Guido Rocchietti, Chiara Pugliese, Gabriel Sartori Rangel, and Jonata Tyska Carvalho. 2024. From Geolocated Images to Urban Region Identification and Description: a Large Language Model Approach. In *The 32nd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '24), October 29-November 1, 2024, Atlanta, GA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3678717.3691317

1 Introduction

Understanding and describing urban regions is crucial for various applications in modern society, such as identifying popular areas and understanding their functionalities [5]. However, these tasks raise many challenges, particularly in delineating the boundaries of regions and determining the similarity between different parts of an urban area. For instance, consider two regions within an urban area: the first, located in the northern suburbs, is characterized by parks, greenery, and recreational facilities. Similarly, the second region, in the southern suburbs, features significant green spaces and community gardens. Despite the geographical distance between these two suburban regions, they show similar features, suggesting that it could be useful to consider them as a unique region. Enhancing the region identification process by incorporating the content of the data can lead to more refined results. Furthermore, generating human-readable descriptions of these regions can significantly benefit various tasks. For instance, it can help identify areas that attract significant tourist activity, which may differ from the most popular areas, or occasional events that contribute to the attractiveness of certain regions, such as street markets or concerts. Moreover, this approach detects latent concerns embedded in the data, such as road potholes, water leaks, or other issues requiring attention. Finally, providing human-readable textual descriptions of urban regions ensures that this valuable information is directly accessible to relevant stakeholders and decision-makers.

Numerous studies [8, 13–15] aim to generate dense representations of predefined regions using a combination of different data sources, including points of interest (POIs), satellite images, and social network data. The effectiveness of these representations is typically evaluated through downstream tasks, such as land use distribution and population density estimation. While these approaches effectively capture region features and latent correlations among data, they rely on predefined regions and produce representations that are not directly interpretable by humans. [3] proposes an approach that identifies regions of interest in a city using geolocated images and provides semantic interpretations of these regions based on user tags. However, this method does not consider the content of the images during region identification, and the resulting descriptions of the identified regions remain unreadable and poorly interpretable. A more recent paper [15] introduces the UrbanCLIP framework, an LLM-enhanced framework that integrates textual knowledge into urban imagery profiling. UrbanCLIP generates textual descriptions for each satellite image using an LLM. However, this framework is limited by its focus on individual image-text pairs, which neglects the spatial relationships and dynamic context of urban regions that are crucial for comprehensive urban profiling.

This paper proposes a new, data-driven, unsupervised approach using geolocated images to identify urban regions dynamically.

SIGSPATIAL '24, October 29-November 1, 2024, Atlanta, GA, USA

ACM ISBN 979-8-4007-1107-7/24/10
https://doi.org/10.1145/3678717.3691317

Our method addresses the limitations of previous approaches by incorporating image content to refine region boundaries, ensuring that the contents within identified regions are consistent. We employ Instruction-tuned Large Language Models (ILLMs) to generate textual descriptions of these regions. By leveraging the content of geolocated data within these regions, ILLMs generate informative and human-readable textual descriptions, making the semantic features of the regions directly accessible to users. To evaluate the effectiveness of our approach, we use geolocated images captured by users to provide a dynamic representation of the urban environment. Specifically, we employ geolocated images from Flickr in Pisa (Italy). Since our approach is fully unsupervised, we apply a combination of empirical validation techniques and collect feedback from volunteers to assess its performance. These comprehensive evaluation strategies ensure a rigorous assessment of both the region identification process and the quality of the generated descriptions.

2 Method

This section presents our method, which identifies regions with similar characteristics from geolocated image data and generates human-readable descriptions (see Figure 1).

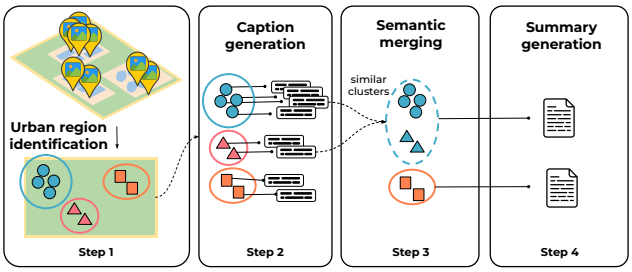


Figure 1: Overview of the method's four key steps.

Step 1: Urban region identification. We identify the urban regions by clustering the set of images $I = \{i_1, \ldots, i_n\}$ within the area of interest A using their geographical coordinates. We employ the HDBSCAN clustering method [1], resulting in the set of clusters $C = \{c_1, \ldots, c_m\}$. We chose HDBSCAN because it is widely used in the literature [4, 12, 16] with geospatial objects due to its ability to identify dense regions in data space, making it effective for spatial data analysis. Moreover, HDBSCAN can detect noise or outliers and discover clusters of arbitrary shapes.

Step 2: Caption generation. We access the semantic contents of each cluster c_i, which, in our case, correspond to the captions generated for each image i within every cluster c. Using a pre-trained model designed to process input images and generate coherent textual captions (further details on the model used are in Section 3), we generate a caption for each image i in every cluster c. To prepare the data for the next step, we concatenate all the captions corresponding to each image in each cluster, resulting in a set of joined captions $D = \{d_1, \ldots, d_m\}$, ensuring that a unique element d_i corresponds to cluster c_i.

Step 3: Semantic merging. We aim to ensure that image contents in each $c \in C$ are similar. Empirical observations suggest that clusters may often contain images with nearly identical descriptions or that refer to the same object. To address this concern, we incorporate the semantic contents (i.e., the set of captions D) when forming the set of the identified regions. Specifically, following the method used in [11], for any two clusters c_i and c_j in C, we consider $c_i \sim c_j$ if their semantic similarity exceeds a threshold value τ. The semantic similarity is calculated on the corresponding set of semantic contents using a similarity measure $sim(d_i, d_j)$. Here, $sim(\cdot, \cdot)$ is a function yielding a value between 0 and 1. This results in a new set of regions $R = \{r_1, \ldots, r_l\}$, where $l \leq m$. Each region r_i is constructed as the union of clusters considered similar to each other and is associated with the concatenation of their respective captions. We call the regions resulting from this step *semantic coherent regions*.

Step 4: Summary generation. This step generates a summary of each $r_i \in R$ by exploiting the set of semantic contents (i.e., the captions). To achieve this, we use an ILLM to generate a set of human-readable summaries denoted as $S = \{s_1, \ldots, s_l\}$, where each summary s_i corresponds to the region $r_i \in R$.

Toy example. Imagine a touristic zone of the city of Pisa, specifically the Piazza dei Miracoli (where the Leaning Tower is). Many people have taken pictures at different locations within the Piazza dei Miracoli and uploaded them to a shared platform.

Urban region identification: We start by using the HDBSCAN method to cluster these images based on their geolocations, resulting in three clusters: one near the church entrance, one around the wall of the Piazza dei Miracoli, and one near the leaning tower.

Caption generation: Next, we use a pre-trained model to generate captions for each image. For instance, images in the first cluster might have captions like "crowd gathering at the church entrance", "welcome banner at the church", etc. Images in the second cluster might be captioned "the leaning tower of Pisa", "a view of a tower", etc. The third cluster might have captions such as "people lining up at the entrance of the leaning tower", "view of the leaning tower", etc. We then concatenate these captions within each cluster.

Semantic merging: In this step, we look for clusters with similar captions. For instance, if we look at the captions of the second and third clusters, they reflect the fact that most of the photos in these two clusters are images of the leaning tower, so we apply the semantic merging (according to a similarity function).

Summary generation: Finally, we generate a summary for each region using an ILLM. The summary for the first cluster might be "A large crowd gathered at the church entrance, welcomed by a banner". The second cluster's summary might be "Capturing the leaning tower from various perspectives".

3 Experiments

This section outlines the experiments conducted to assess the proposed approach. Initially, we describe the datasets, the experimental setup, and the evaluation metrics employed throughout the experiments. As the proposed method comprises unsupervised steps, we evaluate it with image data from Pisa, Italy. For this city, we have the dataset of region summaries generated by volunteers. The code and datasets used are available on the GitHub repository[1].

[1]Github: https://github.com/giuid/UrbanRegionDescription

3.1 Experimental setup

Dataset. The datasets used for the experiments are obtained from Flickr[2], a popular online photo management and sharing application. The data includes geolocation information, user tags, and timestamp data. The dataset obtained for Pisa comprises 2013 photos. Notably, we do not use the user's tags because they are noisy and more personally related than content-related.

Clustering parameter selection. To cluster the images, we employ the HDBSCAN method. The number of clusters is mainly influenced by two parameters: `min_cluster_size` and `min_samples`. We vary the former from 2 to 50 and the latter from 3 to 28.

Semantic similarity merging. To merge clusters that exhibit a higher similarity than the threshold τ in terms of contents, we observe the overlap of the captions associated with each pair of clusters. The similarity function we use is a metric that considers the unigrams shared between two texts, i.e., ROUGE-1 [9]. We vary the threshold τ within the range [0.7, 0.75, 0.8, 0.85, 0.9].

Models for text generation. For the image captioning phase, we use BLIP-2[3] [7], a state-of-the-art model for image captioning. We input the model with the images to produce the required captions. Concerning the summarization phase with ILLMs, we test three different models to generate summaries of the identified regions, i.e., the instructed versions of LLama-2[4] in both the 7 and 13 billion parameters and SOLAR-10.7B [6], which were, at the time we started the current research, the best-performing models on the HuggingFace's Leaderboard[5]. Following the methodology described in previous studies [2], to optimize the effectiveness of the model, we prompt each of the selected models with five different inputs. The rationale behind this choice is to test the variability of the summaries given different instructions. The prompts are reported in the Github repository.

Evaluation Dataset. To evaluate the descriptions generated automatically, we need a gold standard to compare with. After the semantic merging phase, we create a dataset of summaries written by three volunteers with different backgrounds related to the city of Pisa for each emerging cluster. This dataset is created to establish whether we could achieve a good standard to evaluate a task such as the summarization of urban regions. In fact, as we observe in the literature [10], the more common way to evaluate summarization is to have a target standard to compare with.

Evaluation metrics. To evaluate the quality of the generated summaries, we use the dataset with the handmade summaries of all semantic coherent regions we obtain in Pisa. Regarding metrics, as commonly found in the summarization-related literature, we select ROUGE-1, ROUGE-L [9], and BERTScore [17]. All metrics compare the generated summary with a provided one that serves as the gold standard. Henceforth, *precision*, *recall*, and *F1-score* will be denoted as P, R, and $F1$, respectively.

3.2 Results

This section presents the results from the experimental setups detailed in the previous section. We outline the parameters selected in the urban region identification step. We begin by varying the `min_cluster_size` parameter from 2 to 50. Through our knowledge of the city and empirical observations, we note that between 28 and 50, a significant group of photos geolocated in the city center is classified as outliers. Consequently, we set `min_cluster_size` to 28. Then, we explore the range of the other parameter, `min_samples`, from 3 to 28. Upon examination of the city plot, we discern that only two values (3 and 15) enable HDBSCAN to capture the outliers accurately. Hence, we set `min_samples` to 3. These parameter selections result in 27 different clusters (after removing outliers).

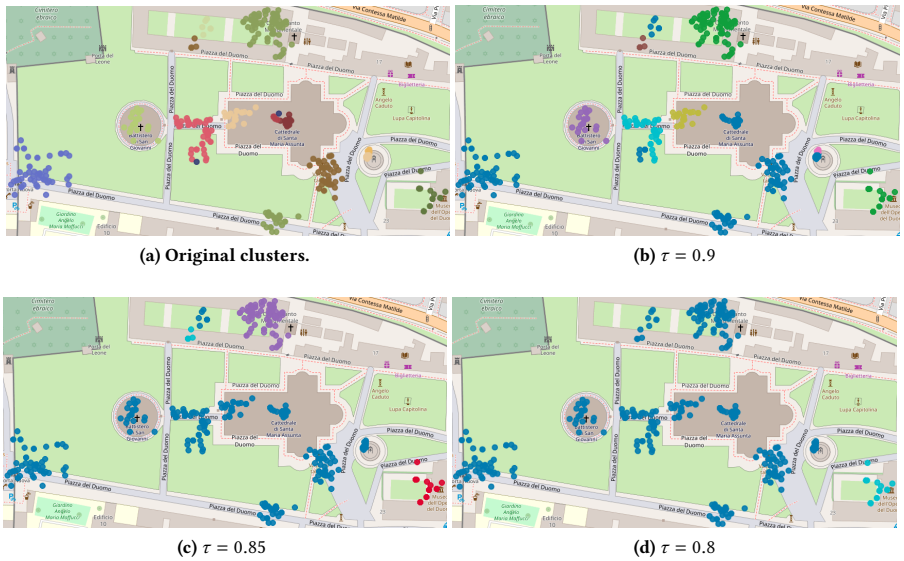
Then, we merge clusters that exhibit significant semantic similarity to each other based on the generated captions. Here, we employ both the $P_{ROUGE-1}$ and $R_{ROUGE-1}$ scores as our similarity function and compute them between each pair of clusters. We vary the τ threshold value using the values described in the previous section. We consider two clusters semantically similar whenever one of $P_{ROUGE-1}$ or $R_{ROUGE-1}$ is higher than the threshold τ. Examining the photos within the clusters obtained in the densest zone of Pisa (i.e., Piazza dei Miracoli shown in Figure 2a), we select a value of τ of 0.9. This choice allows us to merge clusters containing a significant portion of photos where the primary object is the Leaning Tower and maintain separated clusters containing photos taken inside buildings, depicting objects like ceiling paintings, the altar, and the statues inside the cathedral (see in Figure 2b purple and green clusters). The semantic merging step results in the final number of semantic coherent regions for Pisa equal to 20. Then, we generate summaries of the resulting semantic coherent regions. As described in the previous section, we use three different ILLMs with five different prompts. We input the models by concatenating the prompts with the captions associated with the selected region. The performance of these models is evaluated in terms of ROUGE-1, ROUGE-L, and BERTScore, which involve comparing the generated summaries against the reference dataset created from volunteer descriptions. The model SOLAR-10.7B consistently emerges as the best-performing model on average in all three metrics.

Finally, the quality of the generated summaries is enhanced by using image content to refine and describe urban regions, allowing the summaries to capture both the primary features and occasional events depicted in user-taken photos. For example, the summary produced for the semantic coherent region in which there are Keith Haring's mural, restaurants, and shops, not only describes the graffiti and touristic attractions ("[...] *numerous artistic attractions, such as murals, sculptures, and historic churches* [...]") but also includes details about the Christmas street markets and the common practice of people frequenting local restaurants and cafes ("[...] *bustling markets throughout the year, including during the Christmas season. Visitors can enjoy cafes, restaurants* [...]"). Based on the metric values and our knowledge of the city of Pisa, we can affirm that the automatically generated urban region summaries accurately reflect the images contained in the clusters and effectively capture the areas' semantic features. All the generated summaries and metric values are available at the GitHub repository.

4 Conclusion

This paper introduces a new unsupervised method for identifying semantic coherent urban regions using user-taken geolocated

[2]https://www.flickr.com/
[3]Model available at https://huggingface.co/Salesforce/blip2-opt-2.7b
[4]https://huggingface.co/meta-llama
[5]https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard



(a) Original clusters. **(b)** $\tau = 0.9$

(c) $\tau = 0.85$ **(d)** $\tau = 0.8$

Figure 2: Clusters obtained varying τ values in Pisa. Each color corresponds to a different cluster.

images from Flickr and generating human-readable summaries of these areas. The approach demonstrates effective semantic merging of clusters with varying contents, even if they are spatially distant but similar in nature. Evaluation reveals that the SOLAR-10.7B ILLM performs best on average despite occasional low metric scores. Future work will focus on refining evaluation methods through crowdsourcing and manual assessment, integrating additional data sources like POIs and public transport, and fine-tuning summarization models to improve description accuracy and consistency.

Acknowledgments

This work was partially supported by the project CAMEO, PRIN 2022 n. 2022ZLL7MW. Funding for this research has been provided by the European Union's Horizon Europe research and innovation program EFRA (Grant Agreement Number 101093026). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or European Commission-EU. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 160–172.

[2] Elnara Galimzhanova, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, and Guido Rocchietti. 2023. Rewriting Conversational Utterances with Instructed Large Language Models. *2023 IEEE International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)* (2023), 56–63.

[3] Yingjie Hu, Song Gao, Krzysztof Janowicz, Bailang Yu, Wenwen Li, and Sathya Prasad. 2015. Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems* 54 (2015), 240–254.

[4] Rami Ibrahim and M Omair Shafiq. 2018. Mining trajectory data and identifying patterns for taxi movement trips. In *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*. IEEE, 130–135.

[5] Shan Jiang, Gaston A Fiore, Yingxiang Yang, Joseph Ferreira Jr, Emilio Frazzoli, and Marta C González. 2013. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing*. 1–9.

[6] Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2023. SOLAR 10.7B: Scaling Large Language Models with Simple yet Effective Depth Up-Scaling. *ArXiv* abs/2312.15166 (2023). https://api.semanticscholar.org/CorpusID:266550918

[7] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 814, 13 pages.

[8] Yi Li, Weiming Huang, Gao Cong, Hao Wang, and Zheng Wang. 2023. Urban region representation learning with OpenStreetMap building footprints. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1363–1373.

[9] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[10] Lingfeng Lu, Yang Liu, Weiqiang Xu, Huakang Li, and Guozi Sun. 2023. From task to evaluation: an automatic text summarization review. *Artificial Intelligence Review* 56, Suppl 2 (2023), 2477–2507.

[11] Chiara Pugliese, Francesco Lettich, Fabio Pinelli, and Chiara Renso. 2023. Summarizing Trajectories Using Semantically Enriched Geographical Context. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*. 1–10.

[12] Lianhui Wang, Pengfei Chen, Linying Chen, and Junmin Mou. 2021. Ship AIS trajectory clustering: An HDBSCAN-based approach. *Journal of Marine Science and Engineering* 9, 6 (2021), 566.

[13] Zhecheng Wang, Haoyuan Li, and Ram Rajagopal. 2020. Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 1013–1020.

[14] Yanxin Xi, Tong Li, Huandong Wang, Yong Li, Sasu Tarkoma, and Pan Hui. 2022. Beyond the first law of geography: Learning representations of satellite imagery by leveraging point-of-interests. In *Proceedings of the ACM Web Conference 2022*. 3308–3316.

[15] Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2023. When urban region profiling meets large language models. *arXiv preprint arXiv:2310.18340* (2023).

[16] Dongzhi Zhang, Kyungmi Lee, and Ickjai Lee. 2018. Hierarchical trajectory clustering for spatio-temporal periodic pattern mining. *Expert Systems with Applications* 92 (2018), 1–11.

[17] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).