Salient building detection using multimodal deep learning

Eric Lafon, Quentin Potié, Guillaume Touya

International Cartographic Association
Association Cartographique Internationale

Salient building detection using multimodal deep learning

Eric Lafon [a,b], Quentin Potié [b,*], Guillaume Touya [b]

[a] *Master SIGMA, University of Toulouse Jean-Jaurès, Toulouse, France*
[b] *LASTIG, Univ Gustave Eiffel, ENSG, IGN, quentin.potie@ign.fr*

* Corresponding author

Keywords: Human perception, cartographic landmarks, multimodal deep learning

Abstract:

Machine learning has been used for decades in the GIS field, either to classify objects or to segment pixels on an image. A machine learning model is usually trained on either an enriched vector database or on images using CNN. A multimodal architecture tries to join both worlds by giving multiple types of data as an input to a unique model (Baltrušaitis et al., 2019). Multimodal deep learning techniques have been used to predict the price of houses by using photos of the house in combination with its tabular data such as area, number of rooms, etc. (Azizi and Rudnytskyi, 2022).

The goal of this work is to explore the possibilities offered by multimodal deep learning for cartography applications. Our use case is to detect salient buildings in an existing cartographic database. This could be helpful in order to improve the symbolization of the derived map, or to create automatic routing instructions for example. In this work, we define salient buildings as "buildings that are salient in their environment in real life, and which can be used to give some navigation indications".

Figure 1. Studied area: the city of Nantes and its surroundings.

As a first step, we built a dataset of salient buildings. We selected a rectangle of 20x30km around the city of Nantes, France, as our study area to cover urban, semi urban and rural areas. For practical reasons we divided the area in 24 smaller rectangles (Figure 1). We retrieved the building vector data from the BD TOPO, the main geographic database produced by IGN, the French national mapping agency.

The protocol consisted in looking around the city within Google Earth. Out of the 300k polygons in our studied area, we annotated 1129 building polygons as salient (Figure 2).

Because the number of salient buildings is imbalanced compared to the total number of buildings, we keep all the salient buildings and randomly select some non-salient buildings to create our training dataset.

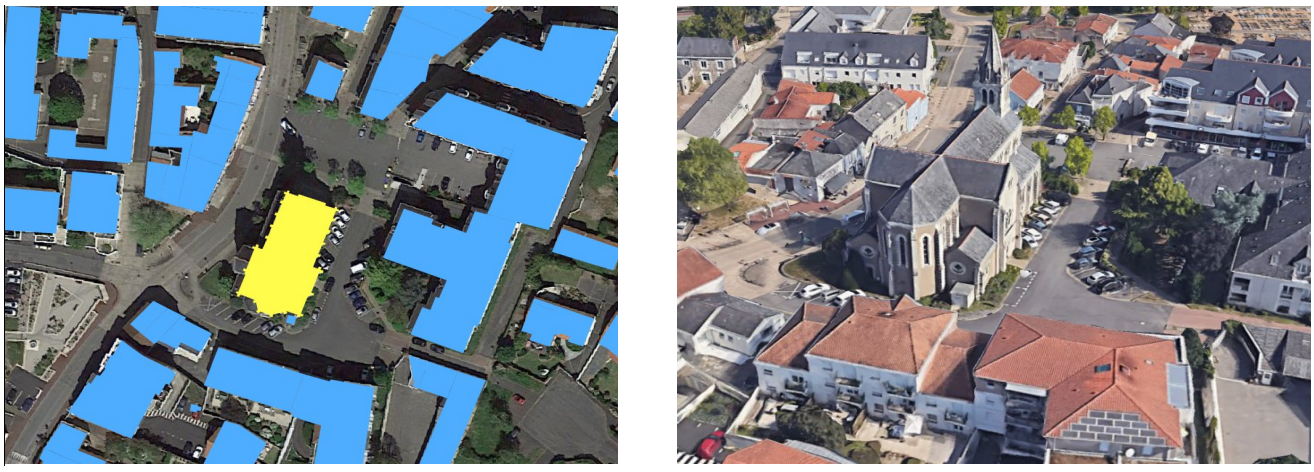

Figure 2. Selected salient building during the database creation (left) and same building on Google Earth (right).

Then, for each polygon of the training dataset, we extract an image of Plan IGN (a topographic multi-scale map) which shows the surroundings of the buildings, with the building at the center of the image (Figure 3).

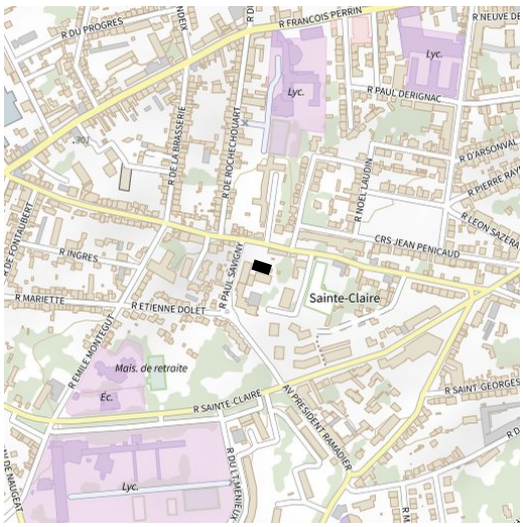

Figure 3. Sample of an input image from Plan IGN.

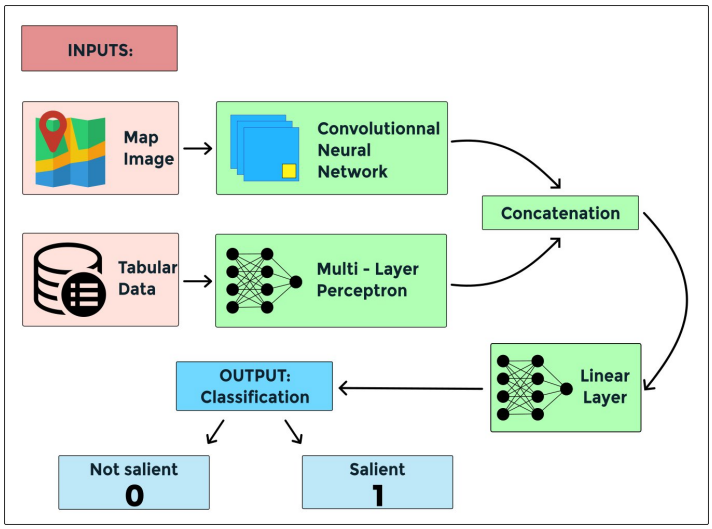

Figure 4. Proposed adaptation of Rosenfelder's multimodal deep learning architecture for our classification problem.

Finally, we intend to adapt the multimodal deep learning architecture proposed by Rosenfelder (2020). Originally designed for predicting housing prices using tabular data (e.g., house location, number of rooms) and corresponding house images, this model has demonstrated superior performance compared to models that rely solely on images, even with small datasets. Our goal is to train this adapted architecture on our dataset to predict whether a building is salient or not (Figure 4). We will compare the results with other machine learning models trained separately on tabular data only and images only to assess the effectiveness of such a multimodal architecture for classification problems. The feature we compute in the tabular data are the following: height, area, perimeter, nature and usage of the building, number of floors, number of housings, geometric granularity, geometry, shape index, mean height of the neighbours, mean area of the neighbours.

Based on the success of multimodal deep learning techniques in predicting housing prices and the potential of leveraging both tabular data and images for building saliency classification, we have strong indications that the results of our work will be highly promising. When it comes to classifying or segmenting cartographic data, we believe that multimodal models can overcome the current limitations of models only based on images or on tabular information.

References:

Azizi I, Rudnytskyi I. Improving Real Estate Rental Estimations with Visual Data. *Big Data and Cognitive Computing*. 2022; 6(3):96. https://doi.org/10.3390/bdcc6030096

Baltrušaitis, T., Ahuja, C., Morency, L.P.. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence. 2*019*;* 41(2): 423-43. https://doi.org/10.1109/TPAMI.2018.2798607.

Rosenfelder, M. Multi-Input Deep Neural Networks with PyTorch-Lightning - Combine Image and Tabular Data. rosenfelder.ai, 2020. https://rosenfelder.ai/multi-input-neural-network-pytorch/