International Cartographic Association
Association Cartographique Internationale

LLM-Vision in enhancing the understanding of public spaces

Tom Komar [a,*], Philip James [a]

[a] *School of Engineering, Newcastle University, United Kingdom*

* tom.komar@newcastle.ac.uk

Keywords: large language model, computer vision, urban geography, automation, emergency management

Abstract

Urban geography traditionally focuses on material structure, populations, and services within a city. However, diversity of architectural styles, street layouts, presence of utilities and variety in uses across time has been a difficult aspect to monitor. It was possible to perform scene classification, objected detection and automated observations of traffic, footfall, and crowd dynamics using custom machine learning models requiring large amount of data or manual review. Additionally, performing these operations using available models without pre-training yielded unacceptably inaccurate results as models were not exposed to samples from an unknown context. This has rendered the task of deploying such solutions impractical for all but those engaged in the most advanced research and development efforts. Recent advancements and public availability of Large Language Model with Vision (OpenAI, 2024) have made the task of zero-shot scene understanding attainable, and we will verify it in this work.

Our research examines the role of LLMV as an assistive tool in large scale urban morphology assessment using static imagery, and road safety and public order management using timeseries imagery. Autonomous operations of LLMV enhance existing procedures without replacing or modifying them, instead providing an additional layer of analysis that can create large coverage thematic maps, and generate alerts and automatically log events detected in analysed images.

This is achieved by prompting the LLMV to generate a compliant JSON, facilitating seamless transition into analysis of structured data and integration with other systems. Despite using switches supposedly forcing JSON response, occasional syntax errors in generated JSON required application of another LLM service (Groq, Inc. 2024) to correct these outputs. It has proven to be an effective method of enforcing the format. Testing of a comprehensive set of true/false variables on a small set of test images, we refined the spectrum of reliably reported observations. Use of open-ended questions for explanation of true/false choices made by the model helped identify cause of issues with reliability. Although sometimes use of apostrophes in the open-text message or truncation of the generated text to the desired token limit caused complications in parsing the response.

Method

In our initial experiments we used a live system awaiting messages from a database which captures CCTV image URLs as they become available (i.e., no more than 1 minute after they are captured). The process is illustrated in Fig.1. Upon manual inspection, it was discovered that the model misses or hallucinates observations when presented with an image only once. Repeated submission of images to the captioning model produced different results, causing us to take an alternative approach, and focus on analysing consistency of responses.

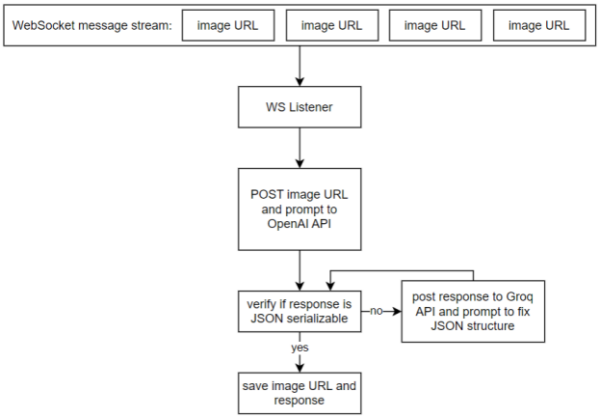

Figure 1. CCTV-LLMV system architecture v1

The final solution presented in Fig.2 is based on 5 submissions of each image to achieve a sample size sufficient for assessing consistency and reducing the influence of outliers on the final results. The method is applied in two scenarios: first one uses images from Google Street View covering selection of streets in city centre of Newcastle Upon Tyne, second uses publicly available CCTV images captured every few minutes. The street view images are taken at 50 meters interval, with 4 images representing each location (looking ahead, back, right, and left), a total of 526 images were used in this scenario. For the CCTV scenario we selected uniformly distributed across time sample of 1000 images from 15 locations.

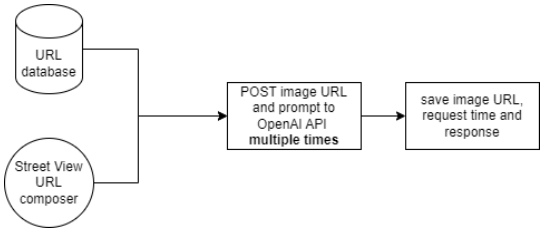

Figure 2. CCTV and Street View URL's submitted multiple times to OpenAI API

LLM-V model evaluated factors such as quality of built environment (architectural styles, land use, presence of utilities) behaviour of people (their presence, specific activities), vehicles (presence, arrangement), location conditions (cleanliness, weather) and usage of visible interactive elements (seatings). The model also reports if the analysed image is complete and error-free.

Results

The inconsistent application of tags often appeared when context was suggestive (e.g. visibility of traffic cones made the model indicate danger or hazard, detecting crowded bus stops only because multiple buses were in view), used tags lacked clear-cut definition for use in visual assessment (e.g. mixed-use area or historic town centre) or misinterpretation (natural features detected when image becomes blurry, group of cyclists wearing high-viz vests taken for construction workers).

Repeated requests for captioning images can be translated to confidence scores for the applied tags and changes in model's output can be described as outlier behaviour rather than consistent disfunction. The most changing tag "mixed-use area" was inconsistently applied to 63 (of 526) street view images, presence of "traffic lights" was inconsistently detected in 51 images and tags for "shopping street" were inconsistent in 47 cases. Each of these 3 most inconsistent tags caused issues in no more than 12% of samples. In case of CCTV images the most inconsistent label "danger or hazard detected" was changing in 42 (of 1000) CCTV images, second worse was "natural features present" with 22 inconsistent applications, and "people sitting" with 18. The CCTV use case was more consistent with no more than 4.2% samples receiving inconsistent tags.

https://doi.org/10.5194/ica-abs-8-15-2024

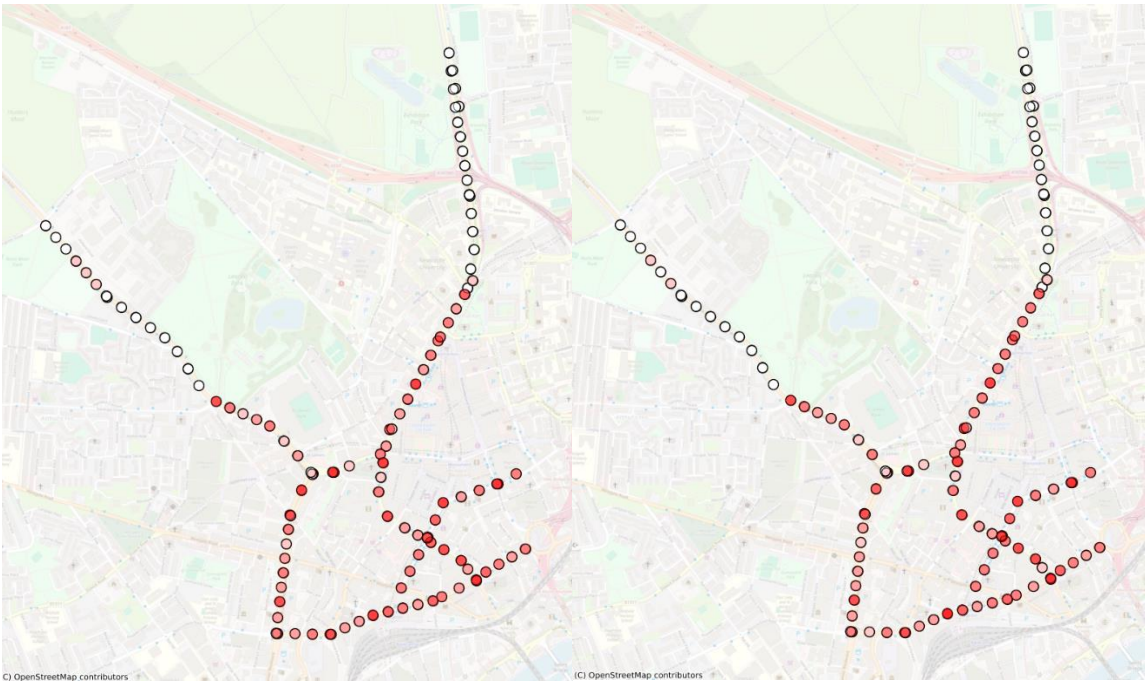

Figure 3. Hardly noticeable changes in distribution of most inconsistently applied label “mixed-use area” (e.g., 3 vs 2 pink points in north-west)

Mapping most variable results of 4-directional Street View images captioning (Fig.3) presents spatial distribution with surprising consistency despite variations in label assignment. The noisy labels appear to concern only a fraction of spatial extents and is partially mitigated by each location being represented by 4 images. As such, the method of automatic assessment of street environment shows promising results.

Figure 4. “Danger or hazard” and “crowded bus stop” tags resulting from suggestive context.

Analysing most variable results of CCTV imagery captioning, appears to be rather over-sensitive than insensitive to presence of requested tags. Mitigating the issue of hallucinated content was not possible using repeated inferences and use of model’s outputs in critical applications is discouraged. Nevertheless, assistive nature of the solution, providing enhancement to performance of rather than replacing the human observer, does allow for some imperfections. It must be highlighted that the model’s outputs should be classed as “suspected” rather than “detected” events.

Conclusions

The ability to automatically analyse urban environments using LLMV presents a valuable tool for planning urban interventions and managing use of public spaces with the use of real-time monitoring and large-scale assessments.

However, variability in tag application observed in this study highlights importance of developing more advanced interpretation and actuation strategies. Model’s tendency to “tag creep” in suggestive context and misinterpret visual cues suggests the solution is not a definitive source of information. Model’s outputs should be treated as “suspected” events, that require verification by human observers.

Abstracts of the International Cartographic Association, 8, 15, 2024.
2024 ICA Workshop on AI, Geovisualization, and Analytical Reasoning – CartoVis24, 7 September 2024, University of Warsaw, Poland.
https://doi.org/10.5194/ica-abs-8-15-2024 | © Author(s) 2024. CC BY 4.0 License.

This research demonstrates potential of LLMV in enhancing the understanding and management of public spaces, particularly through generation of descriptive tags to readily available street view and published in real-time CCTV images. While our findings are promising, we also highlight challenges in regards to consistency and sensitivity in tag application. Future works in this area should focus on developing strategies to mitigate the identified issues.

Acknowledgements

We are grateful for the support this research has received from UK Research and Innovation in partnership with Natural Environment Research Council under “Digital Solutions” programme initiative.

References

Jiang, Albert Q., et al., 2024. Mixtral of experts. arXiv preprint arXiv:2401.04088

Yang, Z., Li, L., et al., 2023. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision)

Witt J., US National Security Agency, Apache Software Foundation, 2006. NiFi Software

OpenAI, 2024. gpt-4o-2024-05-13

Groq, Inc 2024