Interpretable Sewer Defect Detection with Large Multimodal Models †

Taormina, Riccardo; van der Werf, Job Augustijn

DOI
10.3390/engproc2024069158

Publication date
2024

Document Version
Final published version

Published in
Engineering Proceedings

Citation (APA)
Taormina, R., & van der Werf, J. A. (2024). Interpretable Sewer Defect Detection with Large Multimodal Models †. *Engineering Proceedings*, *69*(1), Article 158. https://doi.org/10.3390/engproc2024069158

engineering proceedings

MDPI

Proceeding Paper

Interpretable Sewer Defect Detection with Large Multimodal Models [†]

Riccardo Taormina * and **Job Augustijn van der Werf**

Department of Water Management, Delft University of Technology, Stevinweg 1, 2826 CN Delft, The Netherlands; j.a.vanderwerf@tudelft.nl

* Correspondence: r.taormina@tudelft.nl

[†] Presented at the 3rd International Joint Conference on Water Distribution Systems Analysis & Computing and Control for the Water Industry (WDSA/CCWI 2024), Ferrara, Italy, 1–4 July 2024.

Abstract: Large Multimodal Models are emerging general AI models capable of processing and analyzing diverse data streams, including text, imagery, and sequential data. This paper explores the possibility of exploiting multimodality to develop more interpretable AI-based predictive tools for the water sector, with a first application for sewer defect detection from CCTV imagery. To this aim, we test the zero-shot generalization performance of three generalist large language-vision models for binary sewer defect detection on a subset of the SewerML dataset. We compared the LMMs against a state-of-the-art unimodal Deep Learning approach which has been trained and validated on >1 million SewerML images. Unsurprisingly, the chosen benchmark showcases the best performances, with an overall F1 Score of 0.80. Nonetheless, OpenAI GPT4-V demonstrates relatively good performances with an overall F1 Score of 0.61, displaying equal or better results than the benchmark for some defect classes. Furthermore, GPT4-V often provides text descriptions aligned with the provided prediction, accurately describing the rationale behind a certain decision. Similarly, GPT4-V displays interesting emerging behaviors for trustworthiness, such as refusing to classify images that are too blurred or unclear. Despite the significantly lower performance from the open-source models CogVLM and LLaVA, some preliminary successes suggest good potential for enhancement through fine-tuning, agentic workflows, or retrieval-augmented generation.

Keywords: artificial intelligence; asset management; generative AI; sewer defect classification

check for updates

Citation: Taormina, R.; van der Werf, J.A. Interpretable Sewer Defect Detection with Large Multimodal Models. *Eng. Proc.* **2024**, *69*, 158. https://doi.org/10.3390/engproc2024069158

Academic Editors: Stefano Alvisi, Marco Franchini, Valentina Marsili and Filippo Mazzoni

Published: 20 September 2024

1. Introduction

Urban drainage systems are large integrated infrastructures often comprising several hundreds of kilometers of underground piped networks. To ensure the continued functioning of these systems, a good understanding of the functional state of the assets is necessary. Traditionally, this is performed through manual CCTV inspection, though this method is labor-intensive and heavily relies on human interpretation [1]. In recent years, Deep Learning (DL) models have become increasingly popular to automate the analysis of CCTV images [2]. However, these models yield outputs that are not easily explainable, limiting their practical utility. Post-hoc methods for explainability do not yet yield significant improvements. Advances in Generative AI, specifically Large Multimodal Models (LMMs), have unlocked the potential for interpreting complex semantic and visual data, an unexplored capability in urban water management. This research investigates whether LLMs can help overcome the challenges of semantic interpretation for automatic sewer defect identification [3], mitigating the lack of interpretability of traditional DL models by delivering accurate predictions coupled with human-intelligible explanations.

2. Methods and Materials

Vision-language models are LLMs that seamlessly integrate visual information with linguistic context. In their simplest form, these models typically utilize distinct pre-trained foundational models for vision and language. The integration is then achieved by aligning the two modalities through a trainable projection matrix, which transforms visual representations into language embeddings, enabling subsequent fine-tuning of the language component to accommodate the combined vision-language data [4]. In this work, we compared the performance of the proprietary OpenAI GPT4-V and two open-source alternatives, LLaVA-v1.6-13B [4] and CogVLM-17B [5]. The LMMs were prompted to function as virtual sewer technicians and asked to analyze CCTV images and summarize their observations before making a classification. The prompt required the assessment of various aspects of the pipes, including material and condition. Basic information on the five types of defects was also provided. The LLMs were tested against the binary Deep Learning classifier in [6].

We carried out an experiment using the SewerML open-access dataset [7]. We created a dataset by semi-randomly sampling 200 images from the SewerML validation dataset. The images represented normal conditions (No Defects, ND) and different types of defects (D): cracks, breaks, and collapses (RB), Surface Damage (OB), Production Error (PF), Roots (RO), and Deformations (DE). The subset features 100 images with defects (20 per defect class) and 100 without. We chose the dataset to represent frequently occurring and impactful defects, whilst maintaining an analyzable spread of defect types and considering the cost associated with running the LMMs. We assessed model performances with a thorough set of binary classification metrics, including Accuracy, Recall, Precision, and F1 Score. We also analyzed the paired descriptions yielded by the LMMs with their predictions. It is important to highlight that the used benchmark was trained and validated on >1 M SewerML images, including those used here for testing. On the other hand, the LLMs were assessed in a zero-shot generalization fashion, that is, they were given no prior exposure to the specific domain of sewer images and no specialized training on SewerML.

3. Results and Discussion

Table 1 shows that the benchmark outperforms all tested LMMs. Nevertheless, GPT4-V showcases good detection capabilities, by reaching an overall F1 Score of 0.61. Furthermore, it achieves equal or better performances than the benchmark for defect classes RB and RO, as illustrated in Table 2. On the other hand, all LMMs struggled particularly with DE and OB, which are typically more nuanced defects to classify. When provided with a basic prompt containing limited information on the task and no information on the defects to identify, the performance of GPT4-V decreases, although they are still superior to those of the open-source alternatives.

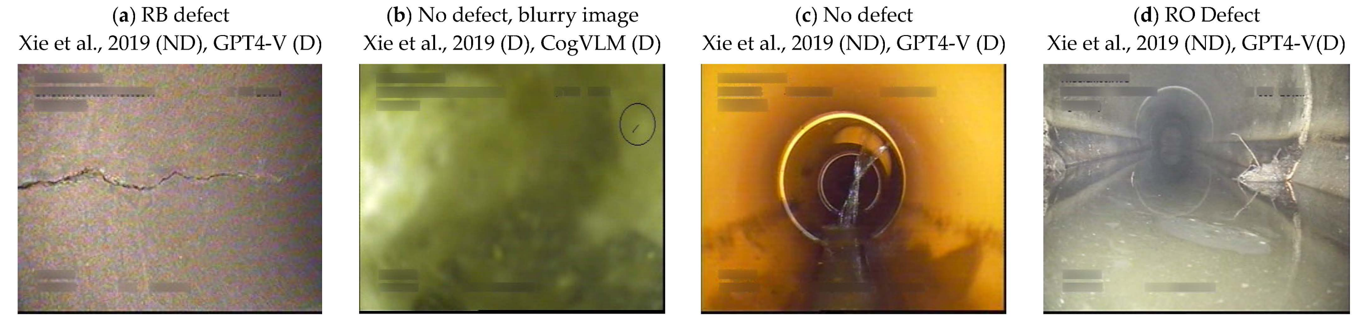
These findings suggest that accurate prompting improves zero-shot generalization, but the capabilities of the underlying LMM play a more decisive role. The gap is also evident when comparing the text provided by the LLMs to justify predictions. GPT4-V provides more accurate descriptions, generally aligned with the predictions. One of the key examples of this can be seen in Figure 1a, which shows a clear crack, which was identified by all the LMMs but missed by the benchmark model. On the other hand, CogVLM is more keen on refusing to predict images that are hard to predict or of poor quality, leading to a higher number of non-predicted images (Table 1). However, this tendency sometimes contradicts related classifications, as shown in Figure 1b, where it labels an image as too blurry, yet still proceeds to identify a defect.

Table 1. Performance metrics of tested models for the identification of observed defects.

Model	Accuracy	Recall	Precision	F1 Score	Not Predicted
Xie et al. [6]	0.81	0.75	0.85	0.80	0
GPT4-V	0.65	0.54	0.70	0.61	3
GPT4-V (basic prompt)	0.61	0.49	0.65	0.56	2
CogVLM	0.60	0.40	0.69	0.50	16
LLaVA	0.52	0.26	0.53	0.35	0

Table 2. F1 Score performance per defect type. The best performance per defect type is highlighted in bold.

Defect Type	Xie et al. [6]	GPT4-V	GPT4-V (Basic Prompt)	CogVLM	LLaVA
DE	**0.75**	0.42	0.33	0.29	0.33
OB	**0.92**	0.40	0.57	0.20	0.18
PF	**0.75**	0.62	0.57	0.57	0.18
RB	**0.95**	**0.95**	0.86	0.86	0.67
RO	0.89	**0.91**	0.82	0.69	0.57
NoDefect	**0.93**	0.87	0.85	0.90	0.87
Overall	**0.80**	0.61	0.56	0.50	0.35



(**a**) RB defect
Xie et al., 2019 (ND), GPT4-V (D)

(**b**) No defect, blurry image
Xie et al., 2019 (D), CogVLM (D)

(**c**) No defect
Xie et al., 2019 (ND), GPT4-V (D)

(**d**) RO Defect
Xie et al., 2019 (ND), GPT4-V(D)

GPT4-V description: The image shows the interior surface of a sewer pipe with a visible longitudinal crack running across the frame and […] surface appears to be concrete. There is no visible presence of water, roots, or blockage around the crack area. The crack itself is a clear sign of structural integrity compromise.

CogVLM description: The image appears to be a blurred and unclear due to the lack of clarity and lighting conditions. It is difficult to make out specific details or defects in the image. Therefore, I would classify this image as DEFECTIVE.

GPT4-V description: The image shows an interior view of a sewer pipe with a small amount of water at the bottom [...]. There is a notable presence of root infiltration near the center of the pipe, where a cluster of roots is visible hanging from the top and extending downward, partially obstructing the passage. [...]

GPT4-V description: The image shows the interior of a sewer pipe with moderate lighting conditions [...]. There is an object that resembles an insect on the right side of the pipe. There are visible cracks on the top right side of the pipe, where a piece of the pipe wall has broken away, creating an irregular edge and an opening [...].

Figure 1. An overview of various images that were assessed by the LMMs, showing the data and the classifications according to the benchmark model in [6], GPT4-V and CogVLM. (**a**) shows a crack that was identified correctly by most of the LMMs but not by the reference AI model, (**b**) shows a blurry image that was correctly identified as such by CogVLM, but still classified (**c**,**d**) to show examples of CCTV images that resulted in LMM hallucinations.

One of the drawbacks of zero-shot LLM generalization is their propensity to hallucinate, which is very prominent for LLaVA, but it is also observed for GPT4-V and CogVLM. For example, the image in Figure 1c was classified as a (RO) defect by GPT4, which identifies the lateral inflow as roots. Showing similar hallucinations, Figure 1d was reported as containing "large insect on the right side" instead of the visible roots. This hallucinating behavior, although inherent in LLMs and LMMs, is exacerbated here as none of the LMMs have seen this type of imagery during their training/fine-tuning process. No trend in hallucination extent and frequency could be observed in the dataset.

We did not perform extensive prompt optimization, aiming to show the relative baseline power of LMMs for sewer classification. Fine-tuning the models on domain-specific data, retrieval augmented generation, and advanced prompting techniques will likely result in improved performances and reduced hallucinations. We do anticipate that the performance of the LMMs can be significantly improved beyond the performance of the non-specialist models we present here.

4. Outlook

This work showed the performance assessment of three generalist Large Multimodal Models (LMMs) prompted to assess if a CCTV image of a sewer pipe included defects and if so, to classify that defect. Although the LMMs were overall outperformed by a state-of-the-art CNN model used for benchmarking, their relative performance was encouraging for future optimization. In future work, the LMMs should be fine-tuned with a build-for-purpose curated dataset, based on CCTV imaging to improve the performance and minimize hallucination in the output. Alternatively, retrieval-augmented generation can be considered to further optimize the performance potential of LMMs for sewer defect identification. Additionally, the LMM linguistic output can be automatically interpreted and corrected via an additional AI agent.

Author Contributions: Conceptualization, R.T.; methodology, R.T. and J.A.v.d.W.; formal analysis, R.T. and J.A.v.d.W.; software, R.T.; data curation, R.T. and J.A.v.d.W.; writing; R.T. and J.A.v.d.W.; visualization, R.T. and J.A.v.d.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The full SewerML dataset is available from [7]. Code for LLaVA and CogVLM is available from [4,5]. A demo with all results is available at https://huggingface.co/spaces/rtaormina/LMM_sewerML (accessed on 23 March 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dirksen, J.; Clemens, F.H.L.R.; Korving, H.; Cherqui, F.; Le Gauffre, P.; Ertl, T.; Plihal, H.; Müller, K.; Snaterse, C.T.M. The consistency of visual sewer inspection data. *Struct. Infrastruct. Eng.* **2013**, *9*, 214–228. [CrossRef]
2. Haurum, J.B.; Moeslund, T.B. A survey on image-based automation of CCTV and SSET sewer inspections. *Autom. Constr.* **2020**, *111*, 103061. [CrossRef]
3. Tscheikner-Gratl, F.; Caradot, N.; Cherqui, F.; Leitão, J.P.; Ahmadi, M.; Langeveld, J.G.; Le Gat, Y.; Scholten, L.; Roghani, B.; Rodríguez, J.P.; et al. Sewer asset management–state of the art and research needs. *Urban Water J.* **2019**, *16*, 662–675. [CrossRef]
4. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. In Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023), New Orleans, LA, USA, 10–16 December 2023.
5. Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. Cogvlm: Visual expert for pretrained language models. *arXiv* **2023**, arXiv:2311.03079.
6. Xie, Q.; Li, D.; Xu, J.; Yu, Z.; Wang, J. Automatic detection and classification of sewer defects via hierarchical deep learning. *IEEE Trans. Autom. Sci. Eng.* **2019**, *16*, 1836–1847. [CrossRef]
7. Haurum, J.B.; Moeslund, T.B. Sewer-ML: A multi-label sewer defect classification dataset and benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13456–13467.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.