Harnessing multimodal large language models for traffic knowledge graph generation and decision-making

Downloaded from: https://research.chalmers.se, 2024-12-20 07:22 UTC

Citation for the original published paper (version of record):
Kuang, S., Liu, Y., Wang, X. et al (2024). Harnessing multimodal large language models for traffic knowledge graph generation and
decision-making. Communications in Transportation Research, 4.
http://dx.doi.org/10.1016/j.commtr.2024.100146

N.B. When citing this work, cite the original published paper.

research.chalmers.se offers the possibility of retrieving research publications produced at Chalmers University of Technology. It covers all kind of research output: articles, dissertations, conference papers, reports etc. since 2004. research.chalmers.se is administrated and maintained by Chalmers Library

(article starts on next page)

Communications in Transportation Research 4 (2024) 100146

Contents lists available at ScienceDirect

Communications in Transportation Research

journal homepage: www.journals.elsevier.com/communications-in-transportation-research

Editorial

Harnessing multimodal large language models for traffic knowledge graph generation and decision-making

1. Introduction

Autonomous driving advancements have increased the importance of understanding traffic scenes for intelligent transportation systems. Substantial progress has been made in traditional tasks such as road segmentation and traffic sign recognition (Wandelt et al., 2024). However, these methods are predominantly low-level methods that focus on individual scene elements without fully addressing higher-level comprehension and decision support.

Recent advancements in large models have demonstrated significant potential in complex tasks, including scene understanding and decision-making. This paper introduces a novel task: generating visual traffic knowledge graphs via large models to inform driving decisions. This task transcends traditional traffic scene understanding by incorporating complex relationship parsing, resulting in an end-to-end solution that transforms traffic images into actionable knowledge graphs.

Inspired by the chain-of-think (COT) mechanism (Wei et al., 2022), we design step-by-step reasoning instructions tailored for traffic scenarios. This method leverages large models' reasoning capabilities to dynamically generate knowledge graphs and drive suggestions without relying on large-scale labeled datasets. Our approach not only enhances traffic scene understanding but also expands the application of large models in intelligent transportation.

2. Related works

This section reviews the literature on traffic scene understanding and the development of multimodal large language models, setting the stage for the proposed task.

2.1. Understanding of traffic scenes

The development of intelligent transportation systems (ITSs) has emphasized the need for accurate traffic scene analysis to ensure the safety and efficiency of autonomous driving. Traditional approaches often rely on high-definition maps for detailed positional data (Li et al., 2022), but creating these maps is costly and resource intensive. To overcome these limitations, researchers have increasingly utilized more accessible data sources, such as onboard camera images, for traffic scene analysis. Lane detection, road segmentation, and traffic sign detection remain key research areas, although these methods often require large annotated datasets, raising concerns about their adaptability and generalizability to diverse traffic scenarios (Gu et al., 2019).

2.2. Multimodal large language models

The generalization power of large language models (LLMs) (Devlin et al., 2018) has driven the development of vision foundation models (VFMs), leading to the rise of multimodal large language models (MLLMs). These models combine LLMs and VFMs, enabling advanced visual and textual understanding through end-to-end training frameworks. MLLMs, such as Generative Pre-trained Transformer (GPT) and Segment Anything Model (SAM), demonstrate the adaptability of large-scale models across diverse tasks, with a significant shift toward developing context-aware systems capable of complex traffic scenario analysis with minimal reliance on domain-specific training data.

3. Proposed task

This section outlines the innovative task proposed in this paper, which focuses on leveraging large models for enhanced traffic scene understanding and decision-making, as depicted in Fig. 1.

3.1. Task definition

This paper proposes a novel task: "generating visual traffic knowledge graphs and making decisions on large models". The objective is to utilize large models' generative and reasoning capabilities to analyze a single traffic scene image combined with natural language instructions. The model then generates a comprehensive traffic knowledge graph and offers actionable driving decisions. This task involves two main steps: analyzing the input image to extract traffic information and organize it into a knowledge graph, followed by using this graph to provide driving decisions, such as selecting optimal routes or avoiding hazards.

The task aims to automate the process from image analysis to decision-making, enhancing the intelligence of autonomous driving systems. It extends beyond traditional scene understanding (e.g., road segmentation, lane detection) to include higher-level relationship parsing and decision support, offering a more holistic perception and response to traffic scenes.

3.2. Task challenges

This task involves several challenges that must be addressed to ensure its success. First, the model must exhibit high scene understanding accuracy by accurately parsing various traffic elements in complex scenes, which requires consideration of factors such as weather, lighting, and traffic conditions. Additionally, the generation of a structured and

https://doi.org/10.1016/j.commtr.2024.100146
Received 28 August 2024; Accepted 17 September 2024
2772-4247/© 2024 The Author(s). Published by Elsevier Ltd on behalf of Tsinghua University Press. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

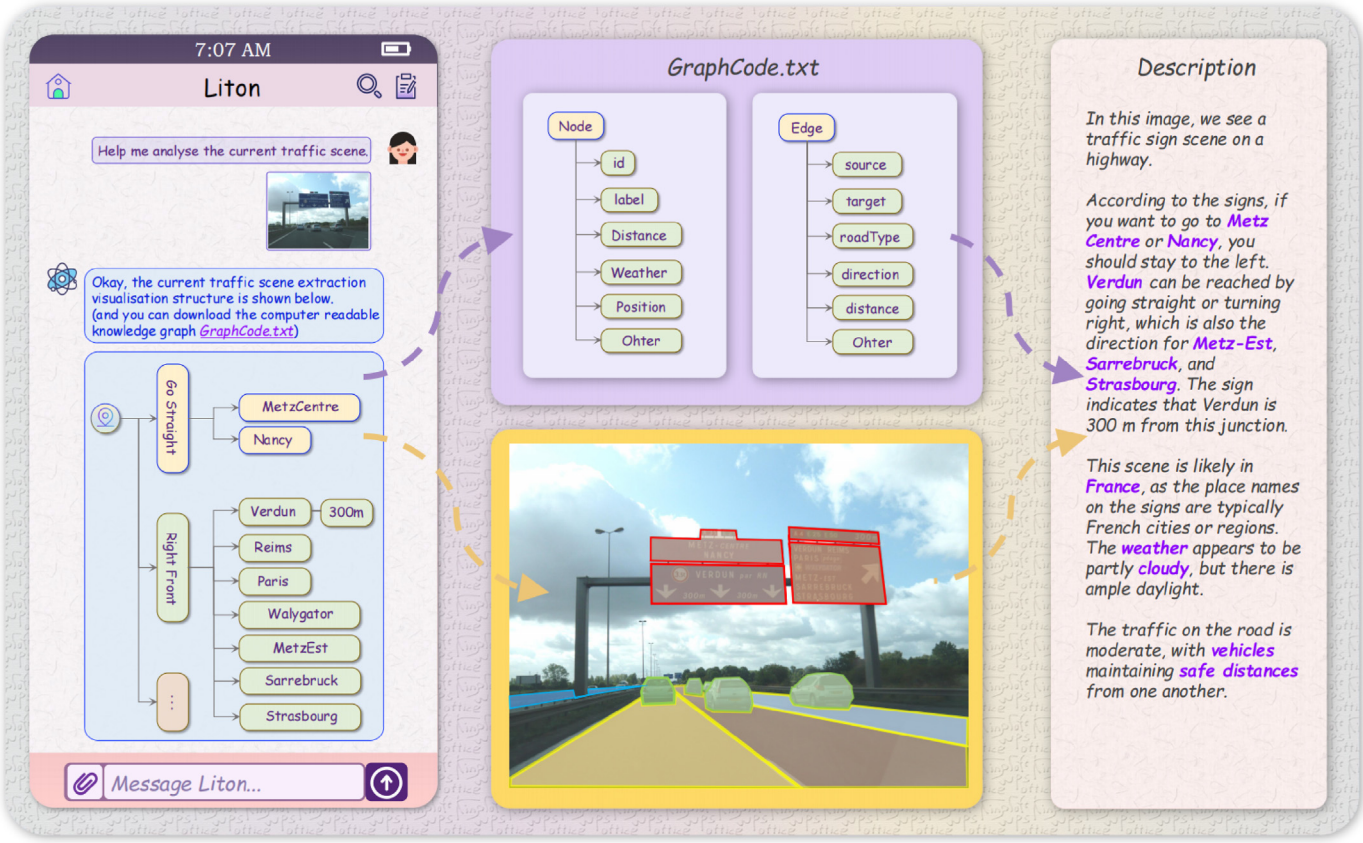

Fig. 1. System architecture for generating visual traffic knowledge graphs and making driving decisions on the basis of the Llava 7b large model. This framework consists of three main stages: input, interaction with the large model, and output. The input stage involves feeding a traffic scene image into the system. In the interaction stage, the large model processes the image via scenario-specific instructions to generate a comprehensive knowledge graph. Finally, the output stage produces the visual traffic knowledge graph and provides actionable driving decisions on the basis of the graph.

comprehensive traffic knowledge graph is crucial, as it involves managing heterogeneous elements and understanding complex relationships within the scene. Another critical challenge is decision accuracy and safety, where the model must execute natural language instructions precisely and provide safe driving decisions, especially in emergencies or uncertain environments. Finally, reducing data dependency is essential, as traditional methods rely heavily on large labeled datasets. This task aims to achieve accurate results by leveraging the generalization and reasoning capabilities of large models, thereby minimizing the need for extensive labeled data. Overcoming these challenges will introduce a new approach to intelligent transportation and autonomous driving, enabling systems to handle complex and dynamic traffic scenes effectively.

4. Methodology

This section details the methodology employed to achieve the proposed task, focusing on the system architecture and the mechanisms that ensure accurate and reliable decision-making.

4.1. System architecture

The proposed system architecture consists of three main stages: (1) input, (2) large model interaction, and (3) output.

1) Input stage: In this stage, a traffic scene image is provided as input, representing real-world driving scenarios and capturing elements such as roads, vehicles, and traffic signs.
2) Large model interaction stage: Here, the system engages the large model via scenario-specific instructions, such as generating a traffic knowledge graph or providing driving decisions on the basis of the graph. The model's reasoning and generative capabilities are leveraged to interpret the scene and respond to the instructions.
3) Output stage: The large model generates a visual traffic knowledge graph, organizing the scene's information and reflecting relationships between traffic elements. It also provides a scene description and driving decision suggestions, aiding in safe and efficient navigation.

4.2. Chain of thought mechanism

To improve the reasoning accuracy, we implement a COT mechanism tailored for traffic scenarios. This mechanism guides the large model through intermediate reasoning steps, breaking down complex tasks into manageable parts. For example, the model first identifies and categorizes scene elements before inferring their relationships and generating a complete knowledge graph.

By using the COT mechanism, the system ensures a thorough and methodical reasoning process, leading to reliable knowledge graph generation and decision-making. This approach also reduces the model's cognitive load by focusing on smaller, sequential tasks, resulting in a robust and adaptable system capable of handling diverse traffic

scenarios.

5. Conclusions and prospects

This paper introduces a novel task of generating visual traffic knowledge graphs and making driving decisions on the basis of large models. We design a systematic methodology, encompassing the input stage, large model interaction stage, and output stage, with a particular emphasis on the COT mechanism that inspired our approach. This mechanism guides large models through a step-by-step reasoning process in complex traffic scenarios, enabling the generation of accurate knowledge graphs and the provision of sound driving decision suggestions. The proposed approach highlights the significant potential of large models in the realms of intelligent transportation and autonomous driving, particularly in managing diverse traffic scenarios effectively without the need for large-scale labeled datasets. This research offers new insights and a technical foundation for the continued advancement of intelligent transportation systems.

Future research should focus on optimizing the COT mechanism to enhance the model's reasoning process, making it more efficient and accurate for complex tasks. Testing the approach in real-world scenarios is essential for continuous refinement and improving its practicality and reliability. Furthermore, integrating this method with other intelligent transportation technologies, such as vehicle-to-everything (V2X) communication systems, could lead to a more comprehensive and intelligent traffic solution, addressing the current challenges and expanding its applicability.

CRediT authorship contribution statement

Senyun Kuang: Data curation, Investigation, Methodology. **Yang Liu:** Supervision. **Xin Wang:** Methodology. **Xinhua Wu:** Formal analysis, Funding acquisition. **Yintao Wei:** Investigation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant Nos. 51761135124, 11672148, 52003142, and 51775293).

References

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. BERT: pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805v2.

Gu, S., Zhang, Y., Tang, J., Yang, J., Kong, H., 2019. Road detection through CRF based LiDAR-camera fusion. In: 2019 International Conference on Robotics and Automation (ICRA), pp. 3832–3838.

Li, Q., Wang, Y., Wang, Y., Zhao, H., 2022. Hdmapnet: an online hd map construction and evaluation framework. In: 2022 International Conference on Robotics and Automation (ICRA), pp. 4628–4634.

Wandelt, S., Sun, X., Zhang, J., 2024. GraphCast for solving the air transportation nexus among safety, efficiency, and resilience. Commun. Transp. Res. 4, 100120.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., et al., 2022. Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural Inf. Process. Syst. 35, 24824–24837.

Senyun Kuang received the B.Sc. degree from the School of Information Science and Technology, Southwest Jiaotong University, China, in 2023. She is now pursuing the Ph.D. degree in the State Key Laboratory of Automotive Safety and Energy, Tsinghua University. Her research interests include computer vision and pattern recognition and intelligent transportation.

Yang Liu received the Ph.D. degree from Southeast University, China, in 2021. Currently, he is a Marie Curie Fellow and a Research Assistant Professor at the Department of Architecture and Civil Engineering, Chalmers University of Technology, and a Visiting Scholar at the School of Vehicle and Mobility, Tsinghua University. His research interests include intelligent transportation systems, AI techniques, and data mining, with applications in complex real-world problems such as autonomous vehicles and urban mobility. He serves as a Young Editor for two top tier academic journals, including *The Innovation* and *IEEE/CAA Journal of Automatica Sinica*. In addition, he is an Associate Editor of *IEEE Transactions on Intelligent Vehicles* and *Journal of Intelligent and Connected Vehicles*. He leads a European Union project and a project funded by the Swedish Innovation Agency. He has received numerous awards for his research, including the IEEE ITSM Best Paper High Commendation Award, Honorable Mention of the COTA Best Dissertation Award, ICME Grand Challenge Second Runner-up Award, and the China Highway Society Outstanding Doctoral Dissertation Award. He is experienced in the practice of AI techniques and has won several world prizes in AI competitions organized by leading international AI conferences or research institutes (e.g., KDD, IJCAI, NeurIPS, CVPR, ICME, and TRB), including the 1st place of KDD Cup, the most well-known algorithm competition in data mining.

Xin Wang received the Ph.D. degree at Beijing Institute of Technology, China, in 2022. Currently, he is a post-doctoral studying intelligent suspension and tires at the School of Vehicle and Mobility, Tsinghua University. His research interests include vehicle terra-mechanics, vehicle dynamics, and AI technologies. He applies intelligent technology to the recognition of road roughness and adhesion ability. His main research topic now is vehicle electromagnetic suspension control technology based on road preview.

Xinhua Wu received the B.S. degree in transportation engineering from the School of Transportation, Nanjing Tech University, Nanjing, China. He is currently pursuing the Ph.D. degree with the Department of Civil and Environmental Engineering, Northeastern University, USA. His research interests include transportation big data analysis and modeling, machine learning, data mining, and intelligent transportation systems.

Yintao Wei received the B.S. and M.S. degrees in computational mechanics and the Ph.D. degree in composite materials from the Harbin Institute of Technology, Harbin, China, in 1992, 1994, and 1997, respectively. He is currently a tenure Professor with the School of Vehicle and Mobility, Tsinghua University, and the Chair of the Smart Chassis Division. He has managed more than 50 research projects, published more than 100 articles, and awarded more than 30 patents. His research interests include intelligent tires for road perception, smart materials and suspensions, and their industrial applications. He is currently an Associate Editor of *Tire Science and Technology* and a reviewer of a number of international journals.

Senyun Kuang
School of Vehicle and Mobility, Tsinghua University, Beijing, 100084, China

Yang Liu
Department of Architecture and Civil Engineering, Chalmers University of Technology, Gothenburg, SE-41296, Sweden

Xin Wang*
School of Vehicle and Mobility, Tsinghua University, Beijing, 100084, China

Xinhua Wu
Department of Civil and Environmental Engineering, Northeastern University, Boston, MA, 02115, USA

Yintao Wei**
School of Vehicle and Mobility, Tsinghua University, Beijing, 100084, China

* Corresponding author.

** Corresponding author.
E-mail address: wangxin1991@mail.tsinghua.edu.cn (X. Wang).
E-mail address: weiyt@tsinghua.edu.cn (Y. Wei).