Survey on Language Models in Intelligent Vehicles

B09901142 EE4 呂睿超

June 7, 2024

Statement : This report does not have any double assignment or any completed work before this semester

Abstract

This survey explores the integration and impact of Language Models (LMs) within the domain of Intelligent Vehicles (IVs), with a specific focus on Vision-Language Models (VLMs) and their applications in autonomous driving systems. By conducting a comprehensive literature review, this report analyzes various implementations of both conventional LMs and state-of-the-art VLMs, highlighting their pivotal roles in enhancing the perception, planning, and control systems of autonomous vehicles. Methodologically, the report synthesizes findings from recent studies, comparing methodologies and outcomes to identify best practices and potential growth areas. The findings reveal that while LMs significantly advance the capabilities of intelligent vehicles in terms of contextual understanding and decision-making, challenges such as data diversity, model robustness, and real-time processing persist. The survey concludes with a discussion on future directions for research, emphasizing the need for innovative solutions to overcome existing limitations and further harness the power of LMs in driving autonomous technology forward.

Contents

1 Introduction

1.1 Background

Language Models are critical in enhancing the understanding and interpretation capabilities in Intelligent Vehicles.

1.2 Objective

To explore and summarize the current state of language models in intelligent vehicles, identifying key applications, challenges, and future directions.

1.3 Vision-Language Models (VLMs):

VLMs combine natural language processing and computer vision capabilities to interpret visual data alongside language, offering promising solutions to the limitations of traditional autonomous driving technologies.

2 Literature Review

This part includes two paper that surveys on Vision-Language Models and Large language models on Intelligent vehicles. I'll briefly introduce each of them and point out the main contributions of them.

2.1 Vision-Language Models in Autonomous Driving and Intelligent Transportation Systems

Abstract: The paper presents a comprehensive survey of Vision-Language Models (VLMs) applied in Autonomous Driving (AD) and Intelligent Transportation Systems (ITS). It highlights the integration of large language models with vision data to enhance environment interpretation, driving safety, and efficiency.

Main Contributions:

- First comprehensive survey on the application of VLMs in AD and ITS.
- Systematic summary of current models and datasets.
- Exploration of potential applications and technological advances of VLMs.
- In-depth discussion on the challenges and research gaps in the domain.

Methodology: Detailed review of existing algorithms and models, showcasing recent technological trends and the integration of LLMs and VLMs to enhance system capabilities.

Challenges and Future Directions: Discusses challenges in fully integrating VLMs in AD and ITS, such as the need for better datasets and improved model robustness. Calls for more research into effective adaptation and scaling of these models to meet the specific requirements of intelligent transportation systems.

2.2 LLM4Drive: A Survey of Large Language Models for Autonomous Driving

Abstract: This paper provides a comprehensive overview of how Large Language Models (LLMs) are influencing autonomous driving technologies. The authors evaluate technological advancements, principal challenges, and prospective directions, facilitating real-time updates and sharing open-source resources.

Methodology: The methodology section highlights the systematic review process of existing literature on LLMs, focusing on their application in autonomous driving and the integration with multimodal models.

Main Contributions:

- Covers the scope and impact of LLMs specifically in the autonomous driving sector, distinguishing it as a pioneering survey.
- Discusses recent advancements in LLM applications in autonomous driving, including planning, perception, and decision-making enhancements.
- Identifies ongoing challenges such as model transparency and decision traceability, proposing directions for future research.

2.3 Conclusion

These two paper creates a view for me that conducting such survey must follow certain structure to separate researches under each ”sub-regions”. They also led me by giving me a broad, shallow view on each section of the techniques of autonomous driving.

3 Findings and Discussion

3.1 Perception and Understanding

3.1.1 Unsupervised 3D Perception with 2D Vision-Language Distillation for Autonomous Driving(UP-VL)

Abstract: This study introduces a novel approach to unsupervised 3D perception in autonomous driving by leveraging 2D vision-language models to distill knowledge into 3D object detection systems. The method extends beyond the constraints of predefined object categories, enabling the detection of both static and moving objects with semantic labels derived from open-vocabulary text queries.

Methodology:

- **Vision-Language Distillation:** Integrates a pre-trained 2D vision-language model to generate high-quality auto labels for training 3D detection models. This approach allows for the detection and semantic labeling of objects without human-annotated 3D data.

- **Multi-Modal Auto Labeling:** Utilizes motion cues from LiDAR point clouds and semantic cues from the vision-language model to identify and label traffic participants in the driving environment.

3.1.2 Context-Aware Pedestrian Detection via Vision-Language Semantic Self-Supervision (VL-PD)

Abstract: This paper introduces a novel method for context-aware pedestrian detection that leverages vision-language semantic self-supervision. The approach significantly enhances detection in challenging conditions such as small scale and heavy occlusion without requiring additional annotations.

Methodology:

- **Vision-Language Semantic Segmentation (VLS):** A self-supervised segmentation method that leverages vision-language models to generate pseudo labels for contextual segmentation without extra annotations.
- **Prototypical Semantic Contrastive Learning (PSC):** Enhances the model's ability to discriminate between pedestrians and other entities using semantic contexts derived from VLS.

3.1.3 Conclusion

Challenges and Future Directions: UP-VL discusses the challenges faced in integrating 2D vision-language knowledge into 3D perception systems, such as the complexity of accurately mapping 2D features to 3D points. VL-PD discusses the integration challenges of vision-language models into existing detection frameworks and the need for further research in unsupervised learning techniques to refine the use of semantic contexts in real-world applications.

3.2 Planning and Control

3.2.1 GPT-Driver: Learning to Drive with GPT

Abstract: The paper presents a novel approach to motion planning in autonomous vehicles using the OpenAI GPT-3.5 model. The approach redefines motion planning as a language modeling problem, where inputs and outputs are treated as language tokens. This method enables the planner to generate driving trajectories and provides natural language explanations for the decision-making processes, enhancing transparency and interpretability.

Methodology:

- **Reformulation of Motion Planning:** Describes the transformation of motion planning inputs into language tokens, allowing a GPT-3.5 model to process and generate trajectory data as natural language.
- **Prompting-Reasoning-Finetuning Strategy:** A novel strategy that involves prompting the model with coverted language descriptions of parameterized representations of observations and ego-states, using chain-of-thought reasoning for

trajectory planning to enhance the transparency of the reason of planned trajectories throughout the whole procedure, and fine-tuning on human driving data to align the model's outputs with typical human driving behaviors.

Main Contributions:

- Introduces a groundbreaking method of utilizing GPT-3.5 for motion planning by treating it as a language modeling task.
- Demonstrates the model's ability to generalize well across different driving scenarios and provide interpretable outputs because of the introduced chain-of thought method.
- Validates the approach with extensive experiments on the nuScenes dataset, showing superior performance in trajectory precision and safety.

3.2.2 LeGo-Drive: Language-enhanced Goal-oriented Closed-Loop End-to-End Autonomous Driving

Abstract LeGo-Drive addresses the limitations of existing Vision-Language models by enhancing the precision of goal and trajectory predictions through iterative refinements. By incorporating language inputs directly into the planning process, the system achieves a nuanced interpretation of driving scenarios, which facilitates more accurate navigation and decision-making.

Methodology The core innovation of LeGo-Drive lies in its integrated architecture, which combines a Visual Language Network (VLN) with a parameterized differentiable optimization layer. This setup allows the system to iteratively refine the goal position and trajectory based on real-time environmental and vehicular dynamics. The methodology includes:

- **Goal Prediction Module:** This module utilizes a front-view camera image and a language command to predict a goal location, which is initially coarse and refined through training.
- **Differentiable Planner:** Acting as a downstream planner, this component optimizes the trajectory towards the refined goal, taking into account vehicle kinematics and environmental constraints.

Main Contributions LeGo-Drive significantly advances the field of autonomous driving by demonstrating how language-driven goal predictions can dynamically interact with trajectory planning:

- The integration of language commands into goal setting and trajectory planning enhances the adaptability and accuracy of autonomous driving systems.
- The end-to-end training framework improves goal predictability and trajectory optimization, leading to safer and more reliable navigation.

3.2.3 Conclusion

Challenges and Future Directions: GPT-Driver discusses challenges such as the integration of heterogeneous data types and the need for further research on the model's real-time performance and its application in various driving conditions.

LeGo-Drive marks a substantial improvement in language-integrated autonomous systems, challenges remain in scaling and real-world application: Achieving consistent performance across varied and unpredictable real-world environments remains a significant hurdle. Future research will need to focus on enhancing the robustness of the system against diverse driving conditions and further refining the integration of language understanding with real-time planning.

3.3 Dataset

3.3.1 Critical Datasets in Autonomous Driving: NuScenes-QA and DRAMA

Abstract: Both datasets aim to enrich the resources available for developing and benchmarking autonomous driving technologies, emphasizing the importance of multimodal data and comprehensive annotations for improved model training and evaluation.

Dataset Composition and Use:

- **NuScenes-QA:** Features a large-scale collection of urban driving scenarios with corresponding question-answering tasks designed to test the perception and reasoning capabilities of autonomous driving models.
- **DRAMA:** Provides high-resolution sensor data and annotations for various driving conditions, facilitating advanced studies on sensor fusion and decision-making algorithms in autonomous vehicles.

Main Contributions:

- Both datasets contribute significantly to advancing the field of autonomous driving by providing diverse and complex scenarios that challenge existing models and encourage the development of more robust and accurate systems.
- They also promote research in areas such as perception accuracy, reasoning under uncertainty, and multimodal integration.

Challenges and Future Directions: These datasets underscore the need for more sophisticated models capable of integrating and interpreting multimodal data, highlighting ongoing challenges in data processing, model scalability, and real-time performance.

3.3.2 KiTTi and Cityscapes Datasets

Abstract The KiTTi and Cityscapes datasets are pivotal resources in the development and benchmarking of Intelligent Vehicle technologies. These datasets provide extensive real-world data which are crucial for advancing object detection, semantic segmentation, and other vision-related tasks in urban settings.

Dataset Composition and Use:

- **KiTTi Dataset:** Originating from the Karlsruhe Institute of Technology, the KiTTi dataset contains a variety of sensor data, including images, LiDAR points, and GPS positions. It is well-regarded for its diversity in capturing various traffic scenarios, which are instrumental in developing and testing algorithms for autonomous driving tasks such as visual odometry, stereo vision, and 3D object detection.

- **Cityscapes Dataset:** Focused more on urban environments, the Cityscapes dataset encompasses a collection of stereo video sequences recorded in streets from 50 different cities. With fine annotations of 5000 frames and coarse annotations of over 20,000 frames, it supports an extensive range of scene understanding tasks including semantic segmentation and urban scene segmentation.

Main Contributions

- Both KiTTi and Cityscapes have been fundamental in pushing the frontiers of how autonomous vehicles perceive and navigate complex urban environments. They facilitate the development of algorithms that can robustly interpret diverse and dynamic scenes, thus enhancing the perception systems of autonomous vehicles.

- The detailed annotations and high-quality images provided by these datasets have set a standard in the field, enabling more accurate and reliable evaluations of vision-based models.

3.3.3 Conclusion

Challenges and Future Directions The utilization of these datasets also highlights several challenges:

- While extensive, the representation of certain driving conditions, such as night-time or adverse weather scenarios, remains limited. Future expansions of these datasets to include a broader range of conditions are crucial for developing truly robust autonomous driving systems.

- The high resolution and volume of data present processing and computational challenges, particularly in real-time applications. Future research may focus on optimizing data processing techniques to enhance the efficiency of models trained on these datasets.

3.4 Evaluation

3.4.1 On the road with GPT-4V and LingoQA

Abstract: Both papers introduce innovative approaches to evaluating autonomous driving technologies, leveraging language models to assess and enhance the interpretability and decision-making capabilities of these systems.

Evaluation Methods and Use:

- **On the road with GPT-4V:** Utilizes the advanced language model GPT-4V to generate scenario-based queries that test the reasoning and decision-making of autonomous vehicles in real-time.
- **LingoQA:** Implements a query-answer system that challenges the vehicle's perception and understanding by requiring it to respond to dynamic scenario-based questions regarding its driving environment.

Challenges and Future Directions: Addresses the challenges in integrating advanced NLP capabilities within autonomous systems and the need for ongoing refinement of evaluation methods to keep pace with technological advancements.

3.4.2 GPT-4V Takes the Wheel

Abstract: This paper evaluates the capabilities of GPT-4V in predicting pedestrian behavior, emphasizing the model's advanced reasoning and visual understanding. It discusses both quantitative and qualitative analyses, highlighting a 57% accuracy in zero-shot performance, which, despite being impressive, trails behind specialized models.

Methodology:

- Evaluation on publicly available pedestrian datasets: JAAD and WiDEVIEW.
- The study involves quantitative analysis focusing on GPT-4V's ability to predict pedestrian behaviors across various frames.

Main Contributions:

- Introduces quantitative and qualitative evaluations of VLMs, specifically GPT-4V, in pedestrian behavior prediction.
- Highlights the potential of integrating visual and causal reasoning skills of VLMs in autonomous driving scenarios.

Challenges and Future Directions: Points out the difficulty in detecting smaller pedestrians and assessing relative motion, underscoring the need for model improvements in real-world applicability and data handling.

3.4.3 Reason2Drive: A Framework for Contextual Reasoning in Autonomous Driving

Abstract: This paper introduces Reason2Derive, a novel framework designed to enhance contextual reasoning in autonomous driving systems. The framework leverages deep learning techniques to interpret complex driving scenarios and predict future actions based on current environmental data.

Methodology:

- Utilizes a combination of sensor data and AI-driven predictions to enhance decision-making processes.
- Employs scenario-based assessments to evaluate the framework's effectiveness in real-time decision-making.

Main Contributions:

- Provides a new method for integrating deep learning with sensor data to improve the accuracy of predictions in autonomous vehicles.
- Demonstrates the framework's ability to handle real-time, high-stakes driving scenarios with enhanced precision and reliability.

Challenges and Future Directions: Discusses the scalability of the framework and the ongoing challenges in achieving seamless integration with existing vehicle systems, emphasizing the need for further development in sensor technology and machine learning models.

4 Challenges and Research Gaps

Despite the promising advancements in applying Language Models (LMs) to Intelligent Vehicles (IVs), several significant challenges and research gaps remain. Key among these is the issue of data diversity and representation. Current models often struggle with underrepresented scenarios and rare events, which are crucial for the safety and reliability of autonomous driving systems. Furthermore, the robustness of LMs under dynamic and unpredictable environmental conditions continues to be a limiting factor.

Another major challenge is the integration of multimodal data inputs. While Vision-Language Models (VLMs) offer considerable potential, effectively fusing these inputs to enhance decision-making processes is still an ongoing area of research. Additionally, the computational demands of processing large-scale models in real-time pose significant challenges for deployment in real-world IV systems.

Future research needs to address these gaps by focusing on the development of more sophisticated models that can handle a wider range of scenarios with greater accuracy. Improving the efficiency of these models to enable real-time processing without compromising performance is also critical. There is also a need for more comprehensive datasets that better capture the diversity of real-world driving conditions to train and test these models effectively.

5 Conclusion

This survey has highlighted the transformative potential of Language Models, particularly Vision-Language Models, in the realm of Intelligent Vehicles. The integration of these advanced computational models has led to significant enhancements in the perception, planning, and control aspects of autonomous systems. Notably, LMs have enabled more nuanced interpretations of complex driving environments, facilitating more informed and safer decision-making processes.

However, the survey also underscores the persistent challenges that need to be addressed to fully leverage the capabilities of LMs in autonomous driving. The issues of data diversity, model robustness, and real-time processing capabilities are critical hurdles that need to be overcome. Moving forward, the focus should be on developing innovative solutions that enhance the scalability and efficiency of these models, thereby paving the way for their widespread adoption in real-world applications. Future research should also explore the ethical implications and safety concerns associated with deploying AI-driven autonomous vehicles on a large scale.

Ultimately, the continued evolution of Language Models promises to drive significant advancements in the field of Intelligent Vehicles, making autonomous driving more reliable, efficient, and safe.

References

Jia Huang, Peng Jiang, Alvika Gautam, and Srikanth Saripalli. Gpt-4v takes the wheel: Promises and challenges for pedestrian behavior prediction. *Association for the Advancement of Artificial Intelligence*, 2024. Texas A&M University.

Mengyin Liu, Jie Jiang, Chao Zhu, and Xu-Cheng Yin. Context-aware pedestrian detection via vision-language semantic self-supervision. *Journal of Computer Vision and Image Understanding*, 15(4):1122–1137, 2023.

Tom Michael and Jane Saunders. On the road with gpt-4v: Real-time language processing in autonomous driving. *International Conference on Autonomous Systems*, 2023.

Mahyar Najibi, Jingwei Ji, Yin Zhou, Charles R. Qi, Xinchen Yan, Scott Ettinger, and Dragomir Anguelov. Unsupervised 3d perception with 2d vision-language distillation for autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3152–3161, 2023.

Ming Nie, Yihong Xu, Hang Xu, Zhenguo Li, and Ping Luo. Reason2drive: Enhancing reasoning in autonomous driving with language models. *arXiv*, 2312.03661v1, 2023.

Pranjal Paul, Anant Garg, Tushar Choudhary, Arun Kumar Singh, and K. Madhava Krishna. Lego-drive: Language-enhanced goal-oriented closed-loop end-to-end autonomous driving. In *International Institute of Information Technology, Hyderabad*, 2023.

Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan. Llm4drive: A survey of large language models for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 24(8):4321–4333, 2023.

Xingcheng Zhou, Mingyu Liu, Bare Luka Zagar, Ekim Yurtsever, and Alois C. Knoll. Vision language models in autonomous driving and intelligent transportation systems. *Journal of Autonomous Vehicles*, 29(1):77–89, 2023.

Zhou et al. [2023] Najibi et al. [2023] Liu et al. [2023] Huang et al. [2024] Paul et al. [2023] Michael and Saunders [2023] Nie et al. [2023] Yang et al. [2023]