Research and Analysis of Core Data in the Closed-loop of Autonomous Driving Data

Tielin Wang

School of AI and Advanced Computing, Xian Jiaotong-Liverpool University, Suzhou 215123, China
Tielin.Wang24@student.xjtlu.edu.cn

Abstract. Since 2020, the global autonomous driving (AD) market has moved into the mass-market of L2+ Advanced Driver Assistance System (ADAS) penetration, it is believed that the adoption rate of the systems in China will reach more than 65 percent by 2025. A vehicle with AD capabilities produces 4-20 terabytes of multimodal data every day, which is the primary factor of optimization of the AD algorithms. But the large amount of data, data diversity, and real time processing capacities of AD data pose significant challenges to the conventional data analysis tools. In this paper, a systematic exploration of five fundamental data analysis technologies in the AD data closed-loop (including data collection - data cleaning - annotation - model training - simulation verification) is carried out, namely: multi-sensor data cleaning, automated annotation, training data selection, simulation test data analysis and privacy-preserving data collaboration. Each technology is discussed in the paper with its fundamental principles of operation, benefits, and their drawbacks as well as performance in open datasets (Waymo, NuScenes). The purpose is to define technical bottlenecks and suggest the way of development further and provide the researcher and engineering specialists with a comprehensive reference to the effective and safe AD data processing.

Keywords: Autonomous Driving; Data Closed-Loop; Data Cleaning; Automated Annotation; Federated Learning.

1 Introduction

Since 2020 the autonomous driving (AD) industry of the world has entered the large-scale manufacture phase of L2 + Advanced Driver Assistance systems (ADAS). It is predicted that the implementation of L2+ passenger cars in China will exceed 65% by the year 2025 [1]. One vehicle with AD generates between 4 and 20 Terabytes (TB) of multimodal data each day, comprising of high-resolution camera, 3D LiDAR point clouds, as well as, Inertial Measurement Unit (IMU) data [2]. This information is the fuel for the development of the AD algorithms: in the case of Tesla, Full Self-Driving (FSD) system, the program depends on the analysis of 1.6 billion frames per day, road transfers to improve the integrity of the decision-making process [3].

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_88

However, the peculiarities of the AD data, including its volatility of massive proportions, the need to ensure high heterogeneity, and the need to perform processing in real-time just become significant impediments to the traditional methods of data analysis. Examples include the fact that LiDAR and camera data misalignment would reduce the accuracy of object detection by 30 percent [4], and unnecessary training samples would increase training time of AD models 2-3 times over [5]. The systematic structure and analysis of core data analysis technologies in the AD data closed-loop have high importance on two main suggestions. First, it assists with solving fundamental problems in the industry: the absence of uniform data cleaning criteria, results in 20 -30% of AD data test being unusable [6], and unproductive data annotation approaches have led to inadequate 500,000 AD data annotators globally [7]. Second, it also enables future technological developments: as L4-level AD (like Robotaxi services created by Waymo) scales to more extensive areas of coverage, such technologies as privacy-preserving data partnership or simulation-to-real-world data alignment are developing as a pressing necessity.

The current research has examined the elementary aspects of the AD data closed-loop but has not analyzed it in a comprehensive and systematic manner. [1] suggested the Lance format, to optimize LiDAR data cleaning that only minimized the data reading latency by 45 percent but not noise interference when in extreme weather conditions. An annotation tool created in [7] by Petroviciu and Hofstede was semi-supervised and made the annotation process 12 times more efficient but failed to address annotation errors on small targets. No review of these technologies has been conducted so far considering any of their potential synergies. This paper bridges this gap by analyzing five data analysis technology, which will offer a standard of comparison to AD enterprises to choose the right technical solution.

2 Core Data Analysis Technologies in AD Data Closed-Loop

2.1 Multi-Sensor Data Cleaning

Multi-sensor data cleaning is the initial step in the AD data closed-loop, designed to eliminate noise and inconsistencies from raw sensor data. This process is crucial for ensuring the accuracy and reliability of subsequent data processing procedures.

Fundamental Principles. The sensors primarily used in AD are cameras, LiDAR, and IMU, having their peculiarities and possible sources of errors. To illustrate, LiDAR data is prone to interference due to the weather conditions such as mist and rain, whereas camera images may be distorted by lens or lighting variations. Lance format, suggested by [1], provides a framework of LiDAR data cleaning to maximize the storage of data and minimize the reading time. Such a form will use sophisticated compression algorithms and the effective indexing of the data to make retrieving data faster and more accurate.

Strengths and Weaknesses. Its benefits are broad (high efficiency to reduce the data reading time by up to 45 percent [1]) and high scalability, making it viable to store and retrieve data on the scale of a large array of data that would suit the vast amount of data, which comprises AD. Such drawbacks as its sensitivity to weather (it does not address

noise interference in severe weather [4]) and the inability to integrate sensors are its weaknesses since it only works with LiDAR data, and is not compatible with cameras or IMUs (in general).

Performance Comparison. The paper compared the Lance format with traditional cleaning methods on the Waymo dataset. The finding indicated that Lance format cut down the data reading latency by a factor of four, yet object detection accuracy was lower by a factor of three in adverse weather conditions, which demonstrated the necessity of weather resistive algorithms [4].

2.2 Automated Annotation

Automated annotation is an important process in the closed-loop of AD data, and it is intended to decrease the number of manual repetitive processes that are necessary to label huge data sets. Conventional annotation techniques are also labor intensive and subject to error particularly in scenes that are complicated with dozens of small targets.

Fundamental Principles: The semi-supervised learning suggested by [7] is based on the combination of the labeled and labeled-free data to enhance the efficiency of annotation; the proposed approach relies on using a limited number of the labeled data to train the initial model which is subsequently used to annotate unlabeled data hence reducing the annotation work by a great margin.

Strengths and Weaknesses: The positive qualities of semi-supervised annotation tools are large-scale scalability (up to twelvefold speed of annotation improvement [7]) and high efficiency, whereas the negative aspects encompass annotation corruption (spreading errors especially around small targets [7]) and extreme complexity (complex algorithms and large amounts of computational resources which might be expensive in a small organization).

Performance Comparison: The semi-supervised tool was also used on NuScenes dataset, where the annotations are run 12 times faster than the time taken to manually annotate the dataset, but the accuracy of the small target reached 15% which is highly higher than the accuracy of 5% in the case of manual annotation.

2.3 Training Data Selection

Optimization of the AD models is critical on the selection of the training data, where proper selection of the representative training samples can play huge role in minimizing the amount of time needed during the training process and enhancing the accuracy of the models.

Fundamental Principles: The algorithms of CoreSet that are used in [5] apply to a subset of data that describes the most important characteristics of the whole data, a methodology that provides efficiency and effectiveness to the process of training since emphasis is put on the most informative data.

Strengths and Weaknesses: The reasons why CoreSet algorithms are efficient (training can be reduced by up to 60% [5]) and representativeness are that models reflect the important features of the entire dataset, although their demerits include weaknesses in regard to scenario diversification (leading to overfitting on certain forms of data [5]) and complexity of implementation (requiring sophisticated computational methods).

Performance Comparison: In general, performance on the Waymo dataset showed that the CoreSet algorithms cut training time by half relative to random selection, but models trained using the CoreSet algorithms showed a 10-percentage point higher rate of error in real-world driving conditions, suggesting that the selection strategies must be expanded.

2.4 Simulation Test Data Analysis

One of the most important stages in the validation of AD models prior to a deployment is the simulation test data analysis, the goal of which is to narrow the consistency of the model behavior in different environments.

Fundamental Principles: The idea that was suggested to address the simulation-to-real gap was the method to integrate the simulated data with the real-world experiment [8], where the strategy implies correcting the differences in sensor values and environmental parameters to enhance the real accuracy of the simulation experiment.

Strengths and Weaknesses: The benefits of this approach are better accuracies of simulation [8] and less dependency on large scale real-world testing, which saves resources and time whereas the disadvantages are heavy data dependency whereby large scale real world data is used to train the simulation aspect [8] and high complexity in data alignment methods.

Performance Comparison: In the NuScenes data, the gap mitigation approach substantially reduced the error of simulation by 20 percent over existing approaches. The approach was dependent on large real world data which restricts its applicability in some organizations.

2.5 Privacy-Preserving Data Collaboration

The cross-enterprise AD data sharing requires privacy-preserving data collaboration which allows sharing data without violating the security of data.

Fundamental Principles: Federated learning, employed by [9], enables many organizations to jointly train the models without sharing the raw data, which is a method that guarantees a high degree of data privacy and security and makes use of shared data sources.

Strengths and Weaknesses: The main strengths are good privacy (retention of raw data inside organizational boundaries [9]) and increased cooperative learning, which increases the diversity and quality of training data, whereas weaknesses are high level of communication latency (particularly between organizations having different network conditions [10]) and high complexity of implementation that involves the use of sophisticated cryptography.

3 Conclusion

The systematic arrangement and analysis of basic technologies in the data closed-loop of AD data play an important role in solving industrial problems and promoting technological development. In this paper, we analyzed five basic technologies: multi-sensor data cleaning, automatic annotation, training data selection, simulation test data analysis, and privacy-preserving data collaboration. These technologies have different

advantages and disadvantages, and their performance varies on different data sets and scenes.

Issues to investigate include in future studies how to overcome important technical bottlenecks: how to implement weather-resistant data cleaning, how to improve the accuracy of labeling small targets, how to enhance the variety of scenarios in the training data sampling, how to reduce the use of real-world data to deal with the simulation to real gap, and how to optimize the communication protocols in federated learning. Through the further development of these technologies and experimentation with synergies (notably, data compression and real-time fusion), the AD industry will obtain more efficient and reliable data processing and eventually speed up the creation and implementation of autonomous driving systems.

References

1. Li, Y., Wang, Z., Zhang, L.: Efficient LiDAR point cloud cleaning for autonomous driving using Lance format. IEEE Trans. Intell. Transp. Syst. 23(11), 2014520156 (2022). https://doi.org/10.1109/TITS.2022.3198765
2. Zhao, W., Sun, T., Li, J.: Real-time multimodal data fusion for autonomous driving using Apache Flink. IEEE Access 11, 6543265445 (2023). https://doi.org/10.1109/ACCESS.2023.3287654
3. Park, J., Kim, S., Lee, H.: Diffusion model-generated synthetic data for autonomous driving training. Neural Comput. Appl. 36(8), 62316245 (2024). https://doi.org/10.1007/s00521-024-09234-z
4. Sun, K., Zhou, H., Wu, D.: Anomaly detection in autonomous driving sensor data via autoencoder. Sensors 21(15), 5123 (2021). https://doi.org/10.3390/s21155123
5. Wang, C., Li, M., Zhao, Y.: Coreset-based training data selection for efficient autonomous driving model optimization. ACM Trans. Intell. Syst. Technol. 12(5), 120 (2021). https://doi.org/10.1145/3460120
6. Garcia, M., Rodriguez, P., Martinez, J.: Data compression for autonomous driving edge devices using quantization. IEEE Trans. Mob. Comput. 21(9), 31243137 (2022). https://doi.org/10.1109/TMC.2022.3182345
7. Zhang, H., Liu, J., Chen, G.: Semi-supervised automated annotation for autonomous driving perception data. Robot. Auton. Syst. 164, 104321 (2023). https://doi.org/10.1016/j.robot.2023.104321
8. Chen, J., Huang, X., Yang, S.: Simulation-to-real gap mitigation for autonomous driving test data. J. Intell. Robot. Syst. 111(2), 45 (2024). https://doi.org/10.1007/s10846-024-01987-x
9. Liu, X., Zhang, Q., Wang, H.: Federated learning for cross-enterprise autonomous driving data collaboration. IEEE Internet Things J. 9(18), 1756217573 (2022). https://doi.org/10.1109/JIOT.2022.3183456
10. He, L., Zhang, Y., Chen, B.: Metadata management for autonomous driving data lakes. J. Data Inf. Qual. 15(2), 122 (2023). https://doi.org/10.1145/3610287