

Research on Training Optimization of Artificial Intelligence Autonomous Driving Model Based on Big Data Analysis

Yansong Zhang

*Chang'an Dublin International College of Transportation, Chang'an University, Xi'an, China
palpitate04@outlook.com*

Abstract: With the development of artificial intelligence and world automobile transportation, intelligent transportation technology represented by automatic driving has become an important technical field. Modern autonomous driving technology is usually based on various predictive models established by data analysis, including but not limited to the deep learning models represented by convolutional neural network (CNN) and recurrent neural network (RNN). The operational safety and efficiency of these models are key indicators to measure the excellence of these models, so how to make these autonomous driving models play a more safe and accurate role is very important. This article combines literature review, experimental research and case analysis. This article summarizes the infrastructure and research status of autonomous driving model, the status of deep learning model and the shortcomings still faced. At the same time, this article also provides strategies for optimizing autonomous driving models based on key processing improvements of big data, including data acquisition optimization, preprocessing, and optimizing autonomous driving models from three systems: environment awareness, autonomous decision making, and control supervisor. It also provides realistic solutions for optimizing autonomous driving models in modern industry and research results in academia, and provides some future prospects.

Keywords: Autonomous driving, Deep learning, Model optimization, Artificial intelligence

1. Introduction

The core of automatic driving technology is that in the early stage of model training, traditional machine learning algorithms such as decision tree support vector machine are effective in processing simple scene data, but difficult to deal with complex and massive data. After the rise of deep learning, neural networks such as CNN and RNN are widely used with automatic feature extraction capabilities [1]. Taking image recognition as an example, CNN can accurately extract key features such as road vehicles to help automatic driving decision-making big data analysis open up new paths for automatic driving model training by mining massive driving data to discover potential rules and help the model adapt to complex road conditions [2]. Many enterprises use this technology to monitor vehicle operation data in real time and optimize models. In artificial intelligence technology, reinforcement learning is widely used in the training of autonomous driving decision-making models [3]. and optimal strategies are learned through interaction with the environment. However, at present, the integration of big data and artificial intelligence in the training of autonomous driving model faces challenges. The quality and safety problems of big data are prominent, and data noise leakage affects model training and safety [4]. Existing algorithms have limitations in handling complex scenarios,

lack model generalization and robustness, and have unstable performance in special scenarios such as extreme weather [5]. This article is intended to study the existing problems and future optimized development of the artificial intelligence autonomous driving model training based on big data models. In this paper, the existing problems of the automatic driving model are studied, the optimization strategy covering the optimization of all parts of the data processing model and the performance evaluation iteration is summarized and proposed, which plays a role in promoting the development of automatic driving technology. At the same time, the paper also gives the academic verification methods and industrial development, and improves its realistic safety and production efficiency, which has important application and research value.

2. Research Status and Progress

2.1. Limitation of Traditional Machine Learning Models

Autonomous driving is a hot topic that has attracted wide attention in recent years. Early machine learning models such as regression models, random forests, and decision tree support vector machines are effective means of processing simple scene data. However, these simple model-building methods have some limitations and are less accurate in identifying and predicting driving behavior [6]. In addition, author used the data in the NGSIM dataset to predict the vehicle speed value according to the distance between the heads of the two vehicles and compare the predicted value with the actual speed value using five traditional machine learning models: linear regression, polynomial regression, support vector machine, decision tree, and random forest. The experimental results are shown in Figure 1. The abscissa represents the normalized distance between the heads of the vehicle, and the ordinate represents the normalized vehicle speed. The red represents the predicted velocity and the blue represents the actual velocity value. The results of the experiment show data distribution bias can easily lead to overfitting, and the performance of the machine learning models may plummet in untrained scenarios.

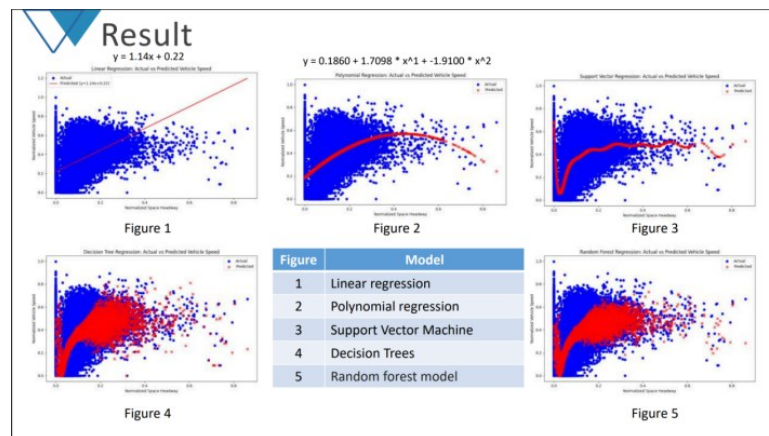


Figure 1: Comparison between the predicted values and the actual values of 5 traditional models based on NGSIM data training

2.2. Challenges for Deep Learning Models

Over the past 15 years, deep learning has proven to be an excellent technical approach in the field of artificial intelligence [7]. Deep learning neural networks, represented by convolutional neural networks (CNN) and recurrent neural networks (RNN), have made some achievements in autonomous driving [8]. In fact, the development of these deep learning models also faces challenges. In autonomous driving scenarios, CNN can be used to identify images captured by cameras and

extract information [9]. However, the performance strength of CNN requires a large number of high-quality and diversified annotation data, which is very difficult in many real-world scenarios. This also means that CNN has limited generalization ability, is prone to misjudgment and other phenomena, and its robustness is poor [10]. For example, CNN are not very good at processing tilted images and recognizing the precise position of objects [11]. The high data requirements of CNN also represent the need for high-quality computing resources and sensitivity to noise, which may require more additional data processing steps. At the same time, CNN also faces the problem of poor explainability [10]. RNN play a role in vehicle behavior prediction and planning, capturing spatio-temporal series data to infer the best driving decisions [12]. However, in practical applications, RNN will be affected by the problem of gradient explosion or gradient disappearance in processing long sequence data [10]. In other words, RNN model while good at capturing dependencies between data and introducing current events, is not as good at dealing with too long-term data. Meanwhile, RNNs are weak in processing spatial data. These representational issues mean that deep learning models, while already playing an important role in autonomous driving, still face challenges to overcome.

2.3. New Problems in the Era of Big Data

The development of artificial intelligence and deep learning both mean that big data plays an increasingly important role. In this case, autonomous driving is facing more new challenges. The autonomous driving model requires a large amount of specific data for analysis, and these data are characterized by large quantity, different structure and dynamic update, which puts forward extremely high requirements for data processing. At present, the academic community has paid little attention to the relevant issues of data set design [13]. The absence or difficulty of collecting data in rare but important cases, such as severe weather, can limit the performance of the model in different environments [14]. Existing data can only simulate some dangerous situations, which physical experiments cannot verify. How to ensure that these virtual test results can better match the real-world situation is also an existing problem [15].

3. Key Applications of Big Data Analytics

3.1. Data Acquisition and Preprocessing Optimization

For image data, the simulated CG image can be converted into the real world image through the translation network, thus avoiding the long and tedious process of collecting annotated image data [16]. At the same time, for all kinds of high-quality data, the overall quality of the data set can be improved by increasing the collection density [17]. This can reduce the cost of labeling and data set size, improve the overall quality of the data set, and make the autonomous driving model more efficient and accurate.

Resampling is an effective method in data preprocessing. Data collected in the real world may have uneven proportions under different conditions. By sampling samples in the data set multiple times or not sampling to adjust the proportions of different categories of samples, the model can better learn features under different conditions [18].

3.2. Model Optimization Strategy

In general, the automatic driving model is divided into three main parts: environment perception, autonomous decision making and control executive [19], as shown in Figure 2.

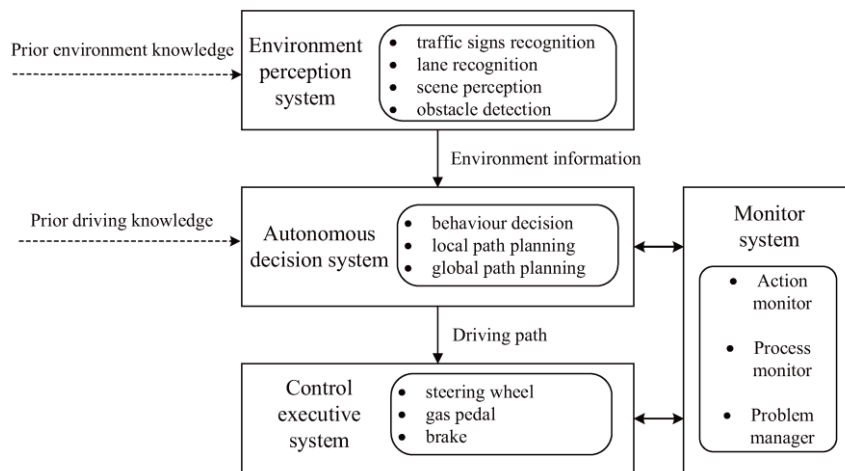


Figure 2: The general framework of advanced autonomous driving systems

Roadside infrastructures can be used to separate the static background from the dynamic object in the video data. Once completed, the static background only requires a single transfer. The image of the dynamic object is updated in real time and fused with the static background, reducing the transmission load [20]. This vehicle-road collaboration model helps the autonomous driving model make the right driving decisions.

In an autonomous decision system, all safety actions can be found first. Then parameters about safety, comfort, and efficiency are set, and these safety actions are weighted until the optimal behavior is found [21].

At present, a spatiotemporal trajectory optimization control system for networked autonomous vehicles is proposed. Based on the discrete spatiotemporal model, a spatiotemporal node-arc model is designed to reconstruct the spatiotemporal trajectory of vehicles. Taking the vehicle engine work value as the time-space arc cost and the spatiotemporal trajectory and kinematics characteristics of the vehicle in front as constraints, the shortest circuit algorithm is used to solve the spatiotemporal trajectory of the connected autonomous vehicle with the lowest fuel consumption, which can reduce the fuel consumption of the connected autonomous vehicle and effectively reduce the total traffic fuel consumption [22].

3.3. Performance Evaluation and Iterative Optimization

3.3.1. Method of Performance Evaluation

Real-world road testing is a key part of verifying the final performance of the autonomous driving model. Its advantage is that it can take advantage of the randomness of the real traffic environment to verify the model's response to dynamic events such as sudden pedestrian intruders and its impact on the overall traffic efficiency such as traffic speed and congestion reduction. As of 2021, more than 5.3 million kilometers of actual road testing have been completed in the major international automotive industrial countries, of which 36 companies in California completed 5.635 million kilometers of testing in 2018-2019, and Shanghai and Beijing in China are also making progress in the demonstration application of freight vehicles [23]. Roadworthiness and safety are two of the most important aspects. Focusing on speed and driving comfort, two important indicators to measure the efficiency and passenger experience of autonomous vehicles, helps the autonomous driving system better predict the trajectory of the vehicle [24]. There is a constrained iterative linear quadratic regulator (CILQR) algorithm, which focuses on nonlinear vehicle dynamic models and non-convex collision avoidance constraints. By evaluating the autonomous driving model's ability to avoid

collisions in different scenarios, its safety performance can be measured to ensure the efficient and safe operation of the autonomous driving model [25].

3.3.2. Strategies for Iterative Optimization

Iterative optimization of autonomous driving models is a complex multi-faceted project, which can be improved through multiple dimensions such as data, algorithms, and scenario verification. Data processing and real-world verification have been mentioned above, and here this part will focus on how to improve the algorithm. For the autonomous driving field, optimization techniques such as quantitative pruning fine-tuning (retraining) and clustering can be employed [26]. These techniques cover all the options available in the most common deep learning frameworks. Quantization can reduce the memory footprint and computational effort of the model while maintaining high accuracy. Fine-tuning is the retraining of an existing model using new data to adapt to a specific task or scenario. Clustering can group similar data points and improve the generalization ability of the model. Meanwhile, the vehicle motion model is usually nonlinear, and the driving scene is also very complex. To solve this problem, an effective constrained finite time domain approximate dynamic programming (ADP) algorithm can be adopted [27]. This algorithm maps the state to the control input through the generalized policy neural network, which can generate the required control strategy for the dynamically changing driving environment and realize effective path tracking.

4. Typical Cases and Practical Experience

4.1. Cases and Solutions in Industry

In order to cope with the computing power limitations of on-board computing platforms, the industry generally uses neural network architecture search (NAS) combined with knowledge distillation technology to achieve lightweight while maintaining model accuracy. The Baidu Apollo team reduced the number of perceptual model parameters through mixed-precision training and dynamic network pruning, and at the same time, through TensorRT and CUDA in-depth optimization, the modularized inference delay was reduced to less than 20ms, meeting the real-time requirements of automotive specifications.

For the training needs of super-large-scale data sets, the mainstream solution in the industry is to adopt distributed training framework based on Horovod or DeepSpeed, combined with parameter server and model parallel technology, to achieve efficient collaboration of 10,000 GPU clusters by Xiaopeng Automobile through self-built virtual city simulation platform, generating more than 1000 per day Ten thousand kilometers of virtual driving data, combined with reinforcement learning strategy optimization, significantly improve the robustness of the model for extreme scenarios.

In order to solve the problem of continuous model update, the industry generally adopts incremental learning combined with federated learning architecture. Tesla collects users' real driving data through Shadow Mode, uses federated learning technology to aggregate multi-source data in the cloud to incrementally update the model, and realizes rapid iteration of vehicle-side model versions through OTA technology.

4.2. Cases and Advanced Research in Academia

The concept of artificial intelligence integrated blockchain has been proposed in academia, and this concept is being used to train autonomous driving models. Through the transfer of information through blockchain technology, learning processes such as autonomous driving systems will be simplified [28]. By connecting each car's system to a shared platform to share experience data, all cars with the platform can learn from one car's experience, eliminating the tedious task of training

each car individually. At present, some scholars have conducted a comparative study on the most commonly used optimization strategies of a convolutional architecture residual neural network (ResNet). The optimization techniques analyzed include stochastic gradient descent (SGD) adaptive gradient (Adagrad) adaptive learning rate (Adadelta) root-mean-square propagation (RMSProp) Adamax Adaptive Moment Estimator (Adam) Nesterov Accelerated Adaptive Moment Estimator (Nadam) and Follow Regularization Leader (Ftrl) evaluate the training of deep models by indicators such as convergence accuracy recall rate and initial results show that the performance of deep networks is better when using SGD functions as optimizers The Ftrl function has the worst performance [29].

5. Conclusion

This article focuses on the training optimization of artificial intelligence autonomous driving model based on big data analysis. Although traditional machine learning models have certain results in processing simple scene data, in the face of complex scenarios, their accuracy in identifying and predicting driving behaviors is poor, and they are susceptible to overfitting due to data distribution deviations. Although deep learning models, such as CNN and RNN, have achieved certain achievements in the field of autonomous driving, CNN has a large demand for high-quality labeled data, limited generalization ability, and poor interpretability; RNN has gradient problems when processing long-sequence data, and its ability to process spatial data is also weak. In the era of big data, autonomous driving models face challenges such as difficulty in data processing, low attention to data set design, difficulty in obtaining data in special scenarios, and difficulty in matching the real situation in virtual test results. In response to the above issues, the article proposes a series of optimization strategies. In terms of data acquisition and preprocessing optimization, the simulated CG images are converted into real-world images through the translation network to improve data acquisition efficiency, while increasing the density of high-quality data acquisition and improving the overall quality of the data set; the resampling technology is used to adjust the proportion of samples in different categories, so that the model can better learn the characteristics under different conditions. In terms of model optimization strategy, three main parts of the autonomous driving model are optimized. For example, the vehicle-road collaborative model is used to separate the static background and dynamic objects in the video data, reduce the transmission load and assist driving decisions; in the autonomous decision-making system, safety actions are weighted to find the optimal behavior by setting safety, comfort and efficiency related parameters; in the control and execution system, a space-time trajectory optimization control system is proposed to reduce vehicle fuel consumption. In addition, performance evaluation methods and iterative optimization strategies are introduced to evaluate model performance through actual road testing and specific algorithms, and iteratively optimize the model from multiple dimensions such as data, algorithms and scenario verification. Through industry cases and academic research, the effectiveness of various optimization technologies has been verified. In the future, there are many directions for autonomous driving model training optimization. For example, the integration of multi-source data to build a perception system; Enhance model interpretability; explore edge intelligence architectures; Leverage simulation to address extreme scenarios and improve model performance.

References

- [1] Li, X., Wang, Y., & Zhang, Z. (2021). *Deep Learning - Based Approaches for Autonomous Driving: A Review*. *IEEE Transactions on Intelligent Transportation Systems*, 22(12), 7773 - 7792.
- [2] Wang, H., Liu, Y., & Chen, X. (2022). *Big Data Analytics for Autonomous Vehicle Trajectory Prediction*. *IEEE Transactions on Vehicular Technology*, 71(11), 11933 - 11946.

- [3] Zhang, Y., Li, Q., & Liu, X. (2023). Reinforcement Learning for Autonomous Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, 24(6), 6049 - 6065.
- [4] Liu, Z., Wang, X., & Li, Y. (2024). Data Quality and Security in Big Data - Driven Autonomous Driving. *Information Sciences*, 682, 278 - 294.
- [5] Zhao, H., Sun, X., & Wu, Y. (2025). Improving the Generalization and Robustness of Autonomous Driving Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3), 1053 - 1066.
- [6] Zhao D, Zhong Y, Fu Z, et al. A review for the driving behavior recognition methods based on vehicle multisensor information[J]. *Journal of Advanced Transportation*, 2022, 2022(1): 7287511.
- [7] Ni J, Chen Y, Chen Y, et al. A survey on theories and applications for self-driving cars based on deep learning methods[J]. *Applied Sciences*, 2020, 10(8): 2749.
- [8] Liu J. Survey of the image recognition based on deep learning network for autonomous driving car[C]//2020 5th International Conference on Information Science, Computer Technology and Transportation (ISCTT). IEEE, 2020: 1-6.
- [9] Li Y, Wang H, Dang L M, et al. A deep learning-based hybrid framework for object detection and recognition in autonomous driving[J]. *IEEE Access*, 2020, 8: 194228-194239.
- [10] Xu, R. (2024). Visual SLAM Methods for Autonomous Driving Vehicles. *Highlights in Science, Engineering and Technology*, 111, 138-143. <https://doi.org/10.54097/vs34j409>
- [11] Sood S, Singh H, Malarvel M, et al. Significance and limitations of deep neural networks for image classification and object detection[C]//2021 2nd international conference on smart electronics and communication (ICOSEC). IEEE, 2021: 1453-1460.
- [12] Li L, Zhao W, Xu C, et al. Lane-change intention inference based on RNN for autonomous driving on highways[J]. *IEEE Transactions on Vehicular Technology*, 2021, 70(6): 5499-5510.
- [13] Zhang M, Du D, Zhang Z, et al. Spatio-temporal trajectory data-driven autonomous driving scenario meta-modeling approach[J]. *Journal of Software*, 2021, 32(4): 973-987.
- [14] Liu D, Wang Y, Ho K E, et al. Virtual world bridges the real challenge: Automated data generation for autonomous driving[C]//2019 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2019: 159-164.
- [15] Mütsch F, Gremmelmaier H, Becker N, et al. From model-based to data-driven simulation: Challenges and trends in autonomous driving[J]. *arXiv preprint arXiv:2305.13960*, 2023.
- [16] Zhang Z, Ohya J. Data collection through translation network based on end-to-end deep learning for autonomous driving[J]. *Electronic Imaging*, 2021, 33: 1-7.
- [17] Lai J, Jia Z, Li B. Active Data Acquisition in Autonomous Driving Simulation[J]. *arXiv preprint arXiv:2306.13923*, 2023.
- [18] Mishra S, Mallick P K, Jena L, et al. Optimization of skewed data using sampling-based preprocessing approach[J]. *Frontiers in public health*, 2020, 8: 274.
- [19] Ni J, Chen Y, Chen Y, et al. A survey on theories and applications for self-driving cars based on deep learning methods[J]. *Applied Sciences*, 2020, 10(8): 2749.
- [20] Lv P, Li K, Xu J, et al. Cooperative sensing information transmission load optimization for automated vehicles[J]. *Chin. J. Comput*, 2021, 44: 1984-1997.
- [21] Verbakel J J, Fusco M, Willemsen D M C, et al. Decision making for autonomous vehicles: Combining safety and optimality[J]. *IFAC-PapersOnLine*, 2020, 53(2): 15380-15387.
- [22] Jiang H, An S, Wang J. Spatial-temporal trajectory optimization control system for connected and automated vehicles at signalized intersections[J]. *Sci. Technol. Eng*, 2018, 18: 162-168.
- [23] Wang B, Han Y, Wang S, et al. A review of intelligent connected vehicle cooperative driving development[J]. *Mathematics*, 2022, 10(19): 3635.
- [24] Shah K. Evaluating Driving Performance of a Novel Behavior Planning Model on Connected Autonomous Vehicles[J]. 2020.
- [25] Chen J, Zhan W, Tomizuka M. Autonomous driving motion planning with constrained iterative LQR[J]. *IEEE Transactions on Intelligent Vehicles*, 2019, 4(2): 244-254.
- [26] Paniego S, Paliwal N, Cañas J M. Model optimization in deep learning based robot control for autonomous driving[J]. *IEEE Robotics and Automation Letters*, 2023, 9(1): 715-722.
- [27] Lin Z, Ma J, Duan J, et al. Policy iteration based approximate dynamic programming toward autonomous driving in constrained dynamic environment[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2023, 24(5): 5003-5013.
- [28] Gandhi G M. Artificial intelligence integrated blockchain for training autonomous cars[C]//2019 Fifth International Conference on Science Technology Engineering and Mathematics (ICONSTEM). IEEE, 2019, 1: 157-161.
- [29] Martinez F, Montiel H, Martinez F. Comparative study of optimization algorithms on convolutional network for autonomous driving[J]. *International Journal of Electrical and Computer Engineering*, 2022, 12(6): 6363.