Proceedings of CONF-MLA 2024 Workshop: Securing the Future: Empowering Cyber Defense with Machine Learning and Deep Learning
DOI: 10.54254/2755-2721/94/2024MELB0082

Exploring the Role of Transformer Models in Vision, Multi-Modal, and Orbital Data for Autonomous Driving

Yuxiang Gong

Faculty of Science and Technology, Beijing Normal University - Hong Kong Baptist University United International College, Zhuhai, Guangdong, 519087, China

t330034009@mail.uic.edu.cn

Abstract. Autonomous driving aims to reduce human error in driving, improve traffic efficiency, and provide a more comfortable driving experience. The integration of computer vision, advanced sensors, and machine learning has been pivotal in this advancement. The introduction of Transformer models has particularly revolutionized the field by offering a novel approach to processing data through attention mechanisms, which is crucial for tasks involving complex relationships between data elements. The paper categorizes research into three main approaches based on input data types: camera-based perception, multi-modal data fusion, and orbital data integration. As autonomous driving technology progresses towards higher levels of autonomy, with L2+ systems becoming standard, challenges remain in accurately interpreting complex environments, handling edge cases, and navigating legal and regulatory landscapes. The paper concludes that while Artificial Intelligence (AI) and deep learning advancements have brought autonomous driving closer to full realization, further research is necessary to address current limitations and ensure safe and reliable autonomous vehicle operation.

Keywords: Autonomous driving, Transformer, Computer vision.

1. Introduction

Autonomous driving represents an innovative method of vehicle control that employs computer vision (CV), advanced sensors, and machine learning algorithms, enabling vehicles to move automatically without driver, or with minimal human intervention [1,2]. This cutting-edge technology holds the promise of substantially cutting down on traffic accidents attributable to human mistakes, optimizing the efficiency of traffic flow, and lightening the cognitive and physical load of driving for people. The progression of deep learning has been instrumental in pushing the development of autonomous driving technology forward. Initially, Recurrent Neural Networks (RNNs) were used to process sequential data and recognize patterns in time-series data, which was beneficial for tasks like speech recognition. However, the capability of RNNs to manage extended sequences was constrained by the challenge of the vanishing gradient issue. Convolutional Neural Networks (CNNs) emerged as a fundamental component in image processing tasks, owing to their proficiency in detecting spatial hierarchies within image data. CNNs have been pivotal in enhancing the precision of object detection and recognition, capabilities essential for autonomous vehicles to comprehend their environmental context. The advent of the Transformer architecture represented a major paradigm shift within the deep learning domain [3,4]. Unlike RNNs and CNNs, which are primarily designed for sequential and grid data respectively,

DOI: 10.54254/2755-2721/94/2024MELB0082

Transformers are based on the concept of attention mechanisms. They enable the model to dynamically assess the significance of various input data segments, rendering them exceptionally suited for processing tasks where the interconnections between elements outweigh their sequential or spatial arrangement.

As autonomous driving technology matures, its applications will extend beyond personal vehicles to transformative uses in the taxi and logistics industries, potentially revolutionizing urban transportation and supply chain management. Incorporating Artificial Intelligence (AI) and deep learning, especially with the innovations introduced by Transformer models, is poised to be crucial for unlocking the complete potential of autonomous driving technology.

2. Principle of Transformer

Since its introduction by Vaswani et al. in 2017, the Transformer model has emerged as a foundational technology for natural language processing (NLP) and a variety of sequence-to-sequence applications [5]. Its primary advantage lies in the Self-Attention mechanism, enabling the model to create direct connections between any two positions within a sequence, effectively managing long-range dependencies. This mechanism is more efficient in processing sequence data than traditional RNN and CNN, because it can parallelly process all elements without the need for gradual iteration. The fundamental structure of the Transformer model has encoders and decoders, which are stacked with multiple identical layers. Each layer is equipped with self-attention mechanisms and feed-forward neural networks, and optimizes the training process through residual connections and layer normalization. The encoder transforms the input sequence into a succession of hidden states, and the decoder, relying on these states along with its previous outputs, constructs the target sequence. Its advantages are mainly reflected in three aspects: First, the self-attention mechanism empowers the model to process every element of the sequence simultaneously, significantly enhancing computational efficiency; Secondly, the model has powerful representation ability, which effectively leverages the global information of the input data, and is especially suitable for complex natural language processing tasks. Finally, it can adapt to long sequence data, overcoming the problem of gradient disappearance or explosion encountered when processing long sequences leveraging RNNs. Nevertheless, the Transformer model also faces challenges in practical applications, such as its self-attention mechanism leading to high demands on compute and memory resources, and resource consumption increases linearly as the sequence length increases. In addition, it is highly sensitive to input, and the stability of the output may be affected by small changes in the input.

3. Transformer-based autonomous driving using vision data

Vision sensor-based autonomous driving mainly uses cameras as the primary perceptual input device to understand the environment around the vehicle through image processing and computer vision technology. This includes but is not limited to monocular vision, binocular vision, and RGB-D vision.

By creating Bird's-Eye-View (BEV) representations without relying on depth information, BEV Transformer (BEVFormer) model, is proposed as a groundbreaking approach, allowing for adaptive learning of BEV features. It stands out through three main design aspects: first, it adeptly integrates spatial and temporal characteristics through the application of attention mechanisms within grid-like BEV queries; second, it incorporates a spatial cross-attention module designed to synthesize spatial features from images obtained by various cameras.; and third, it features a temporal self-attention module that captures temporal information from previous BEV features. In terms of performance, the BEVFormer model has demonstrated remarkable results on the testing set, achieving a 56.9% nuScenes detection score (NDS) under similar parameters and computational requirements, surpassing the previous models. Additionally, it has shown superior performance in the map segmentation task, outperforming Lift-Splat by more than 5.0 points, especially in the most challenging lane segmentation category [6].

Moreover, an innovative end-to-end monocular 3D lane detection model, characterized by the integration of a Transformer-based module for spatial feature transformation. The architecture leverages

DOI: 10.54254/2755-2721/94/2024MELB0082

Transformer-based mechanisms to achieve spatial transformation of features. It incorporates a deformable attention mechanism that significantly reduces computational and memory requirements. The model simultaneously unifies 2D and 3D lane detection. This mechanism dynamically adjusts to capture prominent features within the local region through cross-attention, offering more representative and robust feature extraction compared to traditional Inverse Perspective Mapping (IPM) transformations [7].

Compare to other sensors such as laser radar, cameras cost less and it can provide rich visual information to assist understand complex traffic scenes. The development of deep learning has driven rapid advances in visual perception techniques, especially in object detection and segmentation. But it is also unstable, the performance of the visual system is greatly affected by light and weather. Another disadvantage is visual processing models based on deep learning usually require high computational resources.

4. Transformer-based autonomous driving using multi-modal data fusion

In autonomous driving, multimodal data fusion entails the amalgamation of data from various sensors to achieve a more holistic and resilient environmental perception. This includes but not limited to image data, point cloud data, radar signals, etc., which can be integrated to improve the ability to identify, classify, and locate the vehicle's surroundings.

Among representative works, the unified yet efficient multi-modal transformer (UniTR) is designed specifically for 3D outdoor perception tasks, which often involve understanding the environment using data from LiDAR sensors, for providing 3D sparse point clouds, and cameras, for providing 2D multi-view dense images. As a multi-modal Transformer, UniTR can process these different types of data in parallel, which means it can handle both point cloud data and image data simultaneously. The goal of UniTR is to learn a BEV representation that is particularly useful for applications like autonomous driving. This representation helps in perceiving the 3D structure of the outdoor environment more effectively. In the nuScenes benchmark, UniTR achieved 3D object detection advancements of +1.1 NDS and +12.0 mean Intersection over Union (mIoU) in BEV segmentation [8].

The TransFuser is another multi-modal fusion model that focuses on combining various input modalities at a deeper level to enhance feature extraction. It aims to incorporate global context into the model's understanding by considering the overall scene or data, as well as pair-wise interactions, which might refer to the relationships between different elements within the data. In this manner, TransFuser is capable of generating a more detailed and nuanced input representation, which proves beneficial for tasks that demand an understanding of the complex relationships and interactions within the data. In the urban driving simulator, TransFuser improved driving performance, resulting in a 76% reduction in collisions compared to the geometry-based fusion approach. In addition, TransFuser showed a distinct advantage in processing high-density dynamic objects and complex scenes, especially in situations that require global context inference, such as processing multiple traffic at uncontrolled intersections. In the Car Learning to Act (CARLA) simulator, TransFuser reduced the number of collisions per kilometer by an average of 48%, demonstrating its great potential to improve the safety of automated driving. These data demonstrate TransFuser's robust performance and broad application potential in automated driving simulation and rendering tools [9].

The development status of multimodal data fusion technology in autonomous driving is positive, but it also faces some challenges. The advantages of different sensors can complement each other, improving the robustness and accuracy of perception. Via integrating multiple sensor data, the systematic safety can be increased. But at the same time, data synchronization and calibration between different sensors is a critical issue that requires a deep understanding of the performance and characteristics of the various sensors. With the increase of sensor data, enhancing the computational efficiency and real-time capabilities of the fusion algorithm thus emerges as a significant challenge.

DOI: 10.54254/2755-2721/94/2024MELB0082

5. Transformer-based autonomous driving using orbital data

Orbital data in the context of autonomous driving refers to the use of data from satellites and other high-altitude sources to aid in navigation and decision-making processes. This can include Global Positioning System (GPS) data, satellite imagery, and other geospatial information that can be integrated into the vehicle's systems to enhance its understanding of the environment.

AutoBots represents a framework specifically crafted for predicting the movements of multiple agents, a critical component for autonomous driving systems. It utilizes a Transformer-based architecture to capture the collective probability distribution of forthcoming trajectories for every agent within a given scenario. The framework harnesses latent variables and sequence modeling via Transformers to craft trajectories that are coherent with the scene. Its encoder integrates temporal and social multi-head self-attention (MHSA) modules that enable the processing of information across both time and social contexts. The decoder, on the other hand, employs learnable seed parameters in conjunction with MHSA modules to deduce the entire forthcoming scene in a single pass, efficiently. This methodology enables AutoBots to generate either a solitary trajectory for an ego-agent or a spectrum of potential future trajectories for all agents involved. Validated using the NuScenes dataset for motion prediction tasks, AutoBots has demonstrated impressive real-time inference capabilities, rendering it an apt solution for autonomous driving scenarios [10].

The Scene Transformer serves as a cohesive framework for forecasting the trajectories of multiple agents within autonomous driving contexts. It incorporates a masking technique to adeptly manage diverse forecasting challenges, including motion prediction, conditional motion prediction, and goal-oriented prediction. Engineered to seize the intricate dynamics and interconnections among various agents within a scene, this model deploys a succession of Transformer layers to achieve this. The encoder is tasked with digesting input data to distill key features, whereas the decoder is responsible for producing the anticipated trajectories. Studies have demonstrated that the Scene Transformer proficiently encapsulates the multi-agent dynamics and their interactions, offering a formidable approach to trajectory forecasting essential for autonomous driving technologies [11].

6. Discussion

Autonomous driving technology has seen rapid development over the past year, L2+ assisted driving has gradually become the standard, at present driveless cars as an important development direction of the future automotive industry, there is no substitute but there is still a certain distance from the full promotion of the market.

In self-driving research, Google is doing well, in 2010, Google announced in an official blog post that it was developing an autonomous system. Its driverless cars were licensed in 2012, and total distance driven has exceeded 483,000 kilometers with almost 0 accidents. The core of Google driveless car’s external device are 64 laser rangefinders on the roof of the car, they can provide accurate 3D map data in 200 feet, autonomous system combines laser data with high-resolution maps to make different types of data models to avoid obstacles and follow traffic laws during autonomous driving. It also collects data by cameras and radar sensor to assist driving system. Regarding the developmental landscape, the Society of Automotive Engineers (SAE) categorizes autonomous driving into levels ranging from L0 to L5. L0 signifies the absence of driving automation, whereas L2 denotes partial automation, characterized by advanced driver assistance systems. The transition from L2 to L3 represents a significant technological leap; at Level 3, vehicles possess environmental awareness and can make decisions based on their surroundings, marking the beginning of conditional driving automation. A pivotal distinction between Levels 3 and 4 is that Level 4 vehicles are capable of self-intervention in the case of malfunctions or issues, thus eliminating the need for human intervention in most driving scenarios. Level 5 signifies full driving automation without the need for human oversight. Presently, L2+ (Level 2 Plus) systems are gaining popularity, incorporating features such as adaptive cruise control, lane keeping assist, and traffic jam assist. The industry is actively progressing towards achieving Level 4 (full autonomy under specific conditions) and Level 5 (full autonomy under all conditions). However, these levels are still in the testing and development phase, with limited real-world deployment. Autonomous

DOI: 10.54254/2755-2721/94/2024MELB0082

systems can sometimes misinterpret their surroundings. This can lead to accidents if the system fails to recognize a hazard or incorrectly identifies a non-hazardous situation as dangerous. Another challenge is autonomous vehicles need to struggle with edge cases that are rare but challenging, such as unusual road conditions, unexpected behaviors of other drivers, or complex traffic scenarios. There are still many legal and regulatory hurdles to overcome before fully autonomous vehicles can be widely deployed. This includes liability issues, safety standards, and the integration of autonomous vehicles into existing traffic systems. Liability for self-driving accidents is a major point of contention. For example, now in Shenzhen, China, there is a rule indicate that as long as the L3 automatic driving system is turned on, if the vehicle has a violation or accident responsibility, the first responsible person is always the driver. This is the focus of attention and controversy caused by this law. Now the concern of the market is that many smart cars can already achieve L3 capabilities, but due to the problem of accident liability identification, the car companies that have the initiative will sell at L2 level positioning, and will launch L2+ products that are close to L3 level indefinitely, and the gimmick of assisted driving is becoming more and more sufficient.

7. Conclusion

The research presented in this paper has explored the advancements in autonomous driving technology, with a focus on the innovative integration of Transformer-based models with camera, multi-modal, and orbital data. The autonomous driving landscape has seen significant advancements, with the BEVFormer model achieving a notable 56.9% NDS, surpassing previous models in both detection and segmentation tasks. Multi-modal fusion models like UniTR and TransFuser have shown promising results in enhancing environmental perception, while orbital data models such as AutoBots have demonstrated potential in multi-agent motion prediction. Despite these strides, the technology faces challenges like misinterpretation of surroundings, which can lead to accidents, and the need to address rare but complex scenarios. The integration of autonomous vehicles into existing traffic systems is also hindered by legal and regulatory uncertainties, particularly regarding liability. Looking ahead, continued research is crucial to refine these systems, ensuring they can handle edge cases and meet rigorous safety standards. The industry anticipates overcoming these hurdles to realize the full potential of autonomous driving, which promises to revolutionize transportation and logistics.

References

[1] Yurtsever, E., Lambert, J., Carballo, A., & Takeda, K. (2020). A survey of autonomous driving: Common practices and emerging technologies. IEEE access, 8, 58443-58469.

[2] Wang, W., Wang, L., Zhang, C., Liu, C., & Sun, L. (2022). Social interactions for autonomous driving: A review and perspectives. Foundations and Trends® in Robotics, 10(3-4), 198-376.

[3] Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., ... & Tao, D. (2022). A survey on vision transformer. IEEE transactions on pattern analysis and machine intelligence, 45(1), 87-110.

[4] Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. AI open, 3, 111-132.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention Is All You Need. arXiv preprint arXiv:1706.03762.

[6] Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., ... & Dai, J. (2022). Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In European conference on computer vision, 1-18.

[7] Chen, L., Sima, C., Li, Y., Zheng, Z., Xu, J., Geng, X., ... & Yan, J. (2022). Persformer: 3d lane detection via perspective transformer and the openlane benchmark. In European Conference on Computer Vision. 550-567.

[8] Wang, H., Tang, H., Shi, S., Li, A., Li, Z., Schiele, B., & Wang, L. (2023). Unitr: A unified and efficient multi-modal transformer for bird's-eye-view representation. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 6792-6802.

DOI: 10.54254/2755-2721/94/2024MELB0082

[9] Chitta, K., Prakash, A., Jaeger, B., Yu, Z., Renz, K., & Geiger, A. (2022). Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(11), 12878-12895.

[10] Girgis, R., Golemo, F., Codevilla, F., Weiss, M., D'Souza, J. A., Kahou, S. E., ... & Pal, C. (2021). Latent variable sequential set transformers for joint multi-agent motion prediction. arXiv preprint arXiv:2104.00563.

[11] Ngiam, J., Caine, B., Vasudevan, V., Zhang, Z., Chiang, H. T. L., Ling, J., ... & Shlens, J. (2021). Scene transformer: A unified architecture for predicting multiple agent trajectories. arXiv preprint arXiv:2106.08417.