Master Thesis Planning Report

Tokenization as a Neural Compression Strategy in Automotive Embedded Systems

Tim Boleslawsky, gusbolesti@student.gu.se

Emrik Dunvald, gusdunvem@student.gu.se

February 2026

Suggested Supervisor at CSE: Yinan Yu

Suggested Supervisor at Company: Dhasarathy Parthasarathy

1 Background

The In-Vehicle Embedded System: An in-vehicle embedded system is a specialized computer system integrated within a vehicle to perform dedicated functions, often in real-time, and is essential for controlling, monitoring, and enhancing various automotive operations. These systems typically consist of both hardware and software components, such as electronic control units (ECUs), sensors, actuators, and communication interfaces, which are responsible for tasks like engine management, safety features, infotainment, and advanced driver assistance systems [Navet and Simonot-Lion, 2017, Fairley, 2019].

In-Vehicle Networks and Event-Triggered Logging: Modern vehicles may contain dozens or even hundreds of these embedded systems, interconnected through in-vehicle networks (e.g., CAN, LIN, FlexRay, Ethernet), enabling efficient communication and coordination among different vehicle subsystems [Bello et al., 2019, Navet and Simonot-Lion, 2017, Fairley, 2019]. Event-triggered logging and diagnostic frameworks, which record data only when anomalies or threshold crossings occur, are often adopted to reduce data transmission and avoid bus saturation in complex systems, such as the in-vehicle embedded system. However, this selective approach can reduce holistic visibility of system health, as it may miss subtle degradation patterns or early warning signs that do not cross predefined thresholds. This complicates the detection of emerging faults and comprehensive condition monitoring [Nunes et al., 2023, Montero Jiménez et al., 2020, Azar et al., 2022]. Additionally, the need to carefully tune event thresholds and diagnostic criteria introduces maintenance challenges, as improper settings can lead to missed events or excessive false positives, further complicating system upkeep and reliability [Nunes et al., 2023, Azar et al., 2022].

Traditional Compression: Compression, as originated in information theory by Shannon [1948], is the process of encoding information using fewer bits than the original representation. Compression techniques can be broadly categorized into lossless and lossy methods. Lossless compression is based on two principles: distribution modeling, sometimes called entropy modeling, and entropy coding. Entropy modeling involves creating a probabilistic representation of the data, while entropy coding assigns shorter codes to more frequent symbols based on their probabilities, thereby minimizing the average code length. Lossy compression allows for some loss of information in exchange for higher compression ratios. This is typically achieved through techniques such as transform coding and quantization [Sayood, 2018]. For the purpose of this project, the focus will be on lossy compression as this is more suitable for common downstream ML tasks where some loss of fidelity is acceptable as long as the relevant information for the task is preserved.

Traditional compression methods, based on these information theory principles, often fall short in automotive applications, especially as a precursor for downstream ML tasks. For video/image compression, traditional methods like JPEG or MP3 are optimized for human perception (e.g., visual quality) rather than ML tasks or efficient downstream data use [Ma et al., 2019]. For time series data, algorithmic approaches like CHIMP or Gorilla depend on manually chosen parameters like window size and are sensitive to data characteristics such as entropy and signal variability. This limits their effectiveness in capturing the nuances required for accurate ML model performance in automotive contexts [Johnsson, 2025]. These algorithmic approaches were investigated by Johnsson [2025] in a previous Master's thesis project. This work builds upon this thesis by exploring an alternative approach to compressing vehicle telemetry data.

Rate-Utility Trade-off and Related Research: Constructing downstream ML models for automotive systems, or in fact Internet-of-Things (IoT) systems in general, is a constant trade-off between handling large quantities of data and maximizing model performance. Traditional compression techniques can reduce data volume, but often at the cost of losing critical information necessary for accurate ML tasks such as predictive maintenance, anomaly detection, and fleet analytics. The impact of this trade-off is well-documented in the literature. Muniz-Cuza et al. [2024], for example, study the impact of lossy compression techniques on time series forecasting tasks and observe a constant trade-off between compression ratio and forecasting accuracy.

Existing research approaches these challenges from three different angles: utility-aware adaptive telemetry, neural compression, and task-aware compression.

- First, utility-aware adaptive telemetry methods aim to employ policy learning methods to dynamically adjust telemetry parameters to reduce maintenance costs while preserving data utility for downstream tasks. Although this approach is still emerging, recent research has demonstrated promising results

[Zhang et al., 2023].

- Second, neural compression techniques learn data representations optimized for both compression efficiency and ML task performance. This research is heavily inspired by deep generative models like GANs, VAEs, and autoregressive models, but focuses on compressing the data, instead of generating realistic data samples [Yang et al., 2022]. Neural compression techniques extend the introduced lossy compression principles in two key ways. First, they offer an alternative to traditional distribution modeling by leveraging deep neural networks to learn complex data distributions directly from the data, capturing intricate patterns and dependencies that traditional statistical models may miss. Second, they substitute traditional approaches to transform coding and quantization with learned representations [Yang et al., 2022]. Studies as early as 2019 have shown that neural compression methods can outperform traditional compression techniques for image and video data, especially at low bitrates [Löhdefink et al., 2019]. The same has been shown for time series data [Zheng and Zhang, 2023, Liu et al., 2024].
- Lastly, task-aware compression techniques focus on optimizing compression algorithms to retain information that is most relevant for specific tasks [Yang et al., 2022]. This idea has shown promise in handling time-series data more efficiently in IoT systems. Azar et al. [2020] and Sun et al. [2025] for example explore task-aware compression algorithms that adaptively prioritize data features based on their relevance to downstream tasks, demonstrating improved performance in resource-constrained environments.

2 Aim

The aim of this project will be to investigate the use of neural compression techniques, specifically tokenization-based approaches, for compressing automotive time series data. The goal is to develop a compression framework that effectively balances the trade-off between compression ratio and utility for downstream machine learning (ML) tasks, such as predictive maintenance and anomaly detection.

The motivation behind using tokenization as a neural compression strategy for automotive time series data is to produce discrete latent representations that simplify the entropy modeling task within the compression pipeline. By constraining the data to a finite set of tokens, the complexity of modeling the underlying data distribution is reduced, enabling the use of lightweight entropy models that are computationally efficient. This is particularly advantageous in automotive applications where computational resources are limited, and real-time processing is often required.

The goal will be to design, implement and evaluate a tokenization-based neural compression framework tailored for time-series data. Evaluation will be conducted by comparing relevant parameters between the proposed framework and traditional compression methods.

3 Problem Formulation

The problem to be solved with this thesis is that of efficiently compressing vehicle telemetry data for downstream machine learning tasks. The issue is that vehicles generate a vast amount of telemetry data from various sensors and systems, which can be very challenging to store and transmit due to constraints on edge devices and communication bandwidth [Samantaray, 2023]. While techniques for data compression do exist, they are either not optimized for the specific characteristics of vehicle telemetry data or do not take into account the requirements of downstream machine learning tasks.

Traditional compression techniques such as event-triggered logging and algorithmic time series compression methods have limitations when applied to vehicle telemetry data. Event-triggered logging can miss important information that does not cross predefined thresholds, leading to incomplete data for machine learning models [Nunes et al., 2023, Montero Jiménez et al., 2020, Azar et al., 2022]. Algorithmic compression methods often rely on manually chosen parameters and may not effectively capture the complex patterns present in vehicle telemetry data, which can negatively impact the performance of downstream machine learning tasks.

More recent approaches such as utility-aware adaptive telemetry and task-aware compression have shown promise in addressing some of these challenges. However, these methods often require complex policy learning or adaptive algorithms that may not be feasible for real-time applications in vehicles [Zheng and Zhang, 2023, Löhdefink et al., 2019, Kawawa-Beaudan et al., 2022, Liu et al., 2024].

One less explored but promising approach to compressing data which could be suitable for vehicle telemetry data is that of neural compression in combination with tokenization. A system which incorporates a lightweight neural encoder and a small tokenizer should in theory be able to produce a lossy low entropy representation of the data which can be efficiently stored and transmitted. The main challenge will be to ensure that the compressed representation retains as much relevant information as possible for downstream machine learning tasks while still maintaining a high compression ratio.

4 Limitations

Due to time constraints and the scope of this thesis, there will be several limitations to the work presented:

- The work will be limited to only testing a single head / type of loss for the neural compression model. While it would be interesting to explore multiple heads, this would be time consuming in testing when running all the model variations.
- The work will only implement the encoder and tokenizer part of the pipeline while forgoing the entropy coding. This is done due to the focus of the thesis being on the neural compression step and while entropy coding is important in a full compression pipeline, it is less relevant for the research question at hand.
- The work will also forgo testing on real hardware as that would be very time consuming and difficult to set up while contributing relatively little to the main research question.
- The work will be limited to only testing regression tasks and anomaly detection as downstream tasks due to time constraints and structure of the available datasets. These two tasks should however be a good representation of common downstream tasks and adding more tasks would likely not drastically change the conclusion of the thesis.
- Due to wanting to keep the compressed data representation as versatile as possible, no task-aware compression techniques will be implemented. While task-aware compression could potentially yield better performance for specific tasks, it would limit the generalizability of the compressed representation and add significant complexity to the work. It also results in a more relevant comparison to other compression techniques which are not task-aware.

5 Methodology

6 Risk Analysis

7 Timeline

References

Joseph Azar, Abdallah Makhoul, Raphaël Couturier, and Jacques Demerjian. Robust iot time series classification with data compression and deep learning. *Neurocomputing*, 398, 02 2020. doi: 10.1016/j.neucom.2020.02.097.

Kamyar Azar, Zohreh Hajiakhondi-Meybodi, and Farnoosh Naderkhani. Semi-supervised clustering-based method for fault diagnosis and prognosis: A case study. *Reliability Engineering & System Safety*, 222: 108405, 2022. doi: 10.1016/j.ress.2022.108405.

L. L. Bello, R. Mariani, S. Mubeen, and S. Saponara. Recent advances and trends in on-board embedded and networked automotive systems. *IEEE Transactions on Industrial Informatics*, 15:1038–1051, 2019. doi: 10.1109/tii.2018.2879544.

Richard E. Fairley. *Automobile Embedded Real-Time Systems*, pages 377–389. Wiley-IEEE Press, 2019. doi: 10.1002/9781119535041.app2.

Simon Johnsson. Large scale efficient data readout for vehicle fleets. Master's thesis, Chalmers University of Technology, 2025.

Maxime Kawawa-Beaudan, Ryan Roggenkemper, and Avideh Zakhor. Recognition-aware learned image compression. *Electronic Imaging*, 34(14):220–1–220–5, January 2022. ISSN 2470-1173. doi: 10.2352/ei.2022.34.14.coimg-220. URL `http://dx.doi.org/10.2352/EI.2022.34.14.COIMG-220`.

Jinxin Liu, Petar Djukic, Michel Kulhandjian, and Burak Kantarci. Deep dict: Deep learning-based lossy time series compressor for iot data, 2024. URL `https://arxiv.org/abs/2401.10396`.

Jonas Löhdefink, Andreas Bär, Nico M. Schmidt, Fabian Hüger, Peter Schlicht, and Tim Fingscheidt. Gan-vs. jpeg2000 image compression for distributed automotive perception: Higher peak snr does not mean better semantic segmentation. *arXiv preprint arXiv:1902.04311*, 2019. doi: arXiv:1902.04311v1.

Siwei Ma, Xinfeng Zhang, Chuanmin Jia, Zhenghui Zhao, Shiqi Wang, and Shanshe Wang. Image and video compression with neural networks: A review. *IEEE Transactions on Circuits and Systems for Video Technology*, 30:1683–1698, 2019. doi: 10.1109/tcsvt.2019.2910119.

Juan José Montero Jiménez, Sébastien Schwartz, R. Vingerhoeds, B. Grabot, and M. Salaün. Towards multi-model approaches to predictive maintenance: A systematic literature survey on diagnostics and prognostics. *Journal of Manufacturing Systems*, 2020. doi: 10.1016/j.jmsy.2020.07.008.

Carlos Enrique Muniz-Cuza, Søren Kejser Jensen, Jonas Brusokas, Nguyen Ho, and Torben Bach Pedersen. Evaluating the impact of error-bounded lossy compression on time series forecasting. In *Advances in Database Technology - EDBT*, number 3 in Advances in Database Technology - EDBT, pages 650–663. OpenProceedings, March 2024. doi: 10.48786/edbt.2024.56.

N. Navet and F. Simonot-Lion. *Automotive Embedded Systems Handbook*. CRC Press, 2017. doi: 10.1201/9780849380273.

P. Nunes, J. Santos, and E. Rocha. Challenges in predictive maintenance – a review. *CIRP Journal of Manufacturing Science and Technology*, 2023. doi: 10.1016/j.cirpj.2022.11.004.

Rojalin Samantaray. Adas sensor data handling in the world of autonomous mobility. *SAE Technical Paper Series*, 2023. doi: 10.4271/2023-01-0993.

Khalid Sayood. *Introduction to Data Compression, Fifth Edition*. Morgan Kaufmann Publishers Inc., 5th edition, 2018. ISBN 978-0-12-809474-7.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. doi: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1948.tb01338.x`.

Guoyou Sun, Panagiotis Karras, and Qi Zhang. Highly efficient direct analytics on semantic-aware time series data compression, 2025. URL https://arxiv.org/abs/2503.13246.

Yibo Yang, S. Mandt, and Lucas Theis. An introduction to neural data compression. *Found. Trends Comput. Graph. Vis.*, 15:113–200, 2022. doi: 10.1561/0600000107.

Penghui Zhang, Hua Zhang, Yibo Pi, Zijian Cao, Jingyu Wang, and Jianxin Liao. Adapint: A flexible and adaptive in-band network telemetry system based on deep reinforcement learning. *IEEE Transactions on Network and Service Management*, 21:5505–5520, 2023. doi: 10.1109/tnsm.2024.3427403.

Zhong Zheng and Zijun Zhang. A temporal convolutional recurrent autoencoder based framework for compressing time series data. *Applied Soft Computing*, 147:110797, 2023. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2023.110797. URL https://www.sciencedirect.com/science/article/pii/S1568494623008153.