Master Thesis Project Proposal

Tokenization as a Neural Compression Strategy in Automotive Embedded Systems

Tim Boleslawsky, gusbolesti@student.gu.se

Emrik Dunvald, gusdunvem@student.gu.se

September 2025

Suggested Supervisor at CSE: Yinan Yu

Suggested Supervisor at Company: Dhasarathy Parthasarathy

Relevant completed courses student 1:

- ...

Relevant completed courses student 2:

- ...

1 Introduction

- Modern vehicles generate vast amounts of data from multi-modal sensors such as cameras, radar, LiDAR, and in-vehicle networks (IVNs) like CAN and LIN networks.
- Legacy IVNs such as Classical CAN (1 Mbit/s) and LIN (20 Kbit/s) were never designed for continuous high-bandwidth streams.
- To avoid overload, event-triggered or selective logging schemes are used.
- These reduce bandwidth but limit observability and introduce sampling bias, degrading downstream machine learning (ML) performance.
- The proliferation of ADAS and intelligent systems further multiplies data quantity and complexity.
- Hence, there is a pressing need for adaptive and ML-aware logging frameworks that preserve informational value while respecting resource constraints.

[aa dummy, 9999]

2 Problem

Constructing downstream ML models for automotive systems, or in fact Internet-of-Things (IoT) systems in general, is a constant trade-off between compression large quantities of data and maximizing model performance [Muniz-Cuza et al., 2024]. 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. Another approach to reducing the data volume is event based logging which offers a very lightweight solution for the problem but similarly risks loosing useful data.

The use of task-aware compression modules or neural compression are two potential solutions to this problem which are believed to have room for further exploration as explored in literature surveys such as Liu et al. [2023]. Here it is mentioned that the use of machine learning for producing compressed representations of the data on edge devices is largely underexplored due to the high cost of these models and the restraints of embedded systems. This presents an interesting gap in the literature for further exploration.

Neural compression models leverage deep learning techniques to learn efficient data representations to compress data [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].

Task-aware compression techniques, on the other hand, focus on optimizing compression algorithms to retain information that is most relevant for specific ML tasks []. 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.

When combining these two techniques task-aware neural compression models, have shown promise in reducing the loss of utility associated with higher compression ratios, commonly referred to as the rate-utility trade-off. These models are specifically designed to retain essential features for ML tasks while achieving high compression ratios [Yang et al., 2022]. Studies such as [Kawawa-Beaudan et al., 2022] and Chowdhury et al. [2022] have empirically evaluated task-aware neural compression models and shown that this approach have potential for optimizing two parameters. In [Kawawa-Beaudan et al., 2022] they use a hierarchical autoencoder-based compression network together with a recognition model and implement two hyperparameters to trade off between distortion, bitrate, and recognition performance, while in Chowdhury et al. [2022] the ability to reconstruct the to perform prediction tasks.

There are two major limitations with the examples discussed above. First, while there exist some exploration of task-aware neural compression techniques for image and video data [Kawawa-Beaudan et al., 2022], there is a notable lack of research focusing on time series data, which is the predominant data type in automotive and IoT applications.

The second limitation, that most of these papers fail to address, is the computational constraints of in-vehicle embedded or IoT systems. The mentioned papers, if they use neural compression, primarily focus on achieving high compression rates while maintaining model performance. Because of this, computationally heavy neural network architectures like recurrent neural networks (RNNs) or transformers were chosen [Zheng and Zhang, 2023, Löhdefink et al., 2019, Kawawa-Beaudan et al., 2022].

So while modern task-aware compression techniques, like neural compression, have shown promising advancements in balancing the compression and model performance trade-off, there remains a significant gap in systematically understanding and optimizing the rate-utility trade-off, specifically in vehicular contexts, where computational resources and bandwidth are often constrained.

3 Context

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, Unknown, 2019].

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, Unknown, 2019]. The design of in-vehicle embedded systems must address strict requirements for reliability, safety, real-time performance, and increasingly, cybersecurity, as these systems are critical to both vehicle operation and passenger safety [Bello et al., 2019, Navet and Simonot-Lion, 2017, Mun et al., 2020].

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. 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, complicating the detection of incipient faults and comprehensive condition monitoring [Nunes et al., 2023, 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].

Two developments in recent years further underline the shortcomings of event-triggered logging in automotive systems: the massive increase in signal-based data in the in-vehicle network and the growing relevance of downstream ML tasks.

Recent industry and research reports indicate that the data quantity generated by ADAS (Advanced Driver Assistance Systems) sensors in vehicles is growing at an extremely rapid pace. According to a 2023 technical paper referencing McKinsey's 2021 automotive electronics report, by 2030, about 95 % of new vehicles will be connected, up from around 50 % today, and a single car can generate up to 1 terabyte (TB) of data per hour from its sensors 1 [Bertoncello et al., 2021, Samantaray, 2023]. This explosive growth is driven by the increasing number and sophistication of sensors—such as cameras, radars, and lidars—required for advanced safety and autonomous driving features, with the complexity and volume of data presenting significant challenges for storage, processing, and transmission within embedded automotive systems 14 [Samantaray, 2023].

Modern vehicles increasingly rely on data-driven intelligence to enhance safety, reliability and efficiency. Beyond perception and control, downstream ML tasks — those leveraging collected vehicle and sensor data for offline analysis, optimization and predictive functions — have become central to automotive-system design. These tasks include predictive maintenance [Theissler et al., 2021], anomaly and intrusion detection [Övgü Özdemir et al., 2024], and fleet-level analytics like fuel consumption or maintenance scheduling [Chen et al., 2025].

Recent reviews highlight that while event-triggered and anomaly-based data collection can optimize resource use, they often result in fragmented or incomplete datasets, making it harder to implement robust predictive maintenance strategies and limiting the effectiveness of ML models that rely on continuous, high-resolution data streams [Nunes et al., 2023, Jiménez et al., 2020]. Multi-model and hybrid approaches are being explored to address these limitations, but the trade-off between data reduction and diagnostic completeness remains a significant challenge in both industrial and automotive contexts [Jiménez et al., 2020, Azar et al., 2022].

Now, given the need for efficient data handling in the context of downstream ML tasks, and the shortcomings of event-triggered logging, one might look to traditional compression methods. Unfortunately, these methods often fall short in automotive applications. 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 (TBC, maybe cite Simon's thesis that covered exactly this last year).

Existing research approaches these challenges from two different angles. 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 (TBC). Other research focuses on neural compression techniques that 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]. Here task-aware approaches have shown especially promising results as discussed in section 2.

4 Goals and Challenges

In this paper, the goal will be to determine weather using tokenization as a transformation step can allow the subsequent entropy modeling to be done using a smaller architecture while still achieving similar results to traditional neural compression implementations. Tokenization is traditionally understood as the mapping of high-dimensional, continuous inputs into a sequence of discrete symbols drawn from a finite vocabulary [Grefenstette, 1999]. Tokenization therefore serves as a form of simplified representation of the data; it reduces dimensionality, constrains representations to a compact code space, and can be made task-aware so that the retained tokens are maximally useful for prediction or classification. Instead of compressing raw sensor values, this approach would aim to learn a discrete vocabulary of prototypical temporal patterns that are maximally informative for downstream tasks. This approach is expected to give us two distinct advantages:

- better computational efficiency compared to RNN and transformer based neural compression methods.
- an interpretable intermediate layer of tokens instead of continuous values.

4.1 Goal

The main goal of this paper will be to develop a compression framework for time-series data using a learned tokenizer and lightweight entropy modeler. This paper is expected to present the difference in predictive utility between data compressed using a neural compression with tokenized inputs, traditional neural compression and no compression. In addition, the paper also aims to present the difference in computational cost and number of parameters for each approach. The expectation is that the approach which utilizes a small tokenizing module will have a small memory footprint and low latency while still producing a compressed representation which offers comparable predictive utility.

4.2 Sub-Goals

- Produce a lightweight task-aware tokenization framework for time series data.
- Quantify the loss in predictive utility when training ML models on uncompressed, tokenized and compressed data.
- Measure the computational cost of the system in terms of latency and peak memory usage.
- Measure the rate-utility tradeoff of the implementation in regards to the relevant ML task.

4.3 Challenges

- Lack of established learned compressors for time-series data
- Difficulty optimizing rate-distortion trade-off
- Generalizing the results over heterogenous sensors

4.4 Approach

- **Dataset:** Use available automotive sensor and telemetry test-fleet data supporting tasks such as predictive maintenance and anomaly detection.
- **Task 1:** Train downstream ML models on uncompressed data to quantify loss in predictive utility.
- **Task 2:** Implement established neural compression methods such as CompressAI as baselines, measuring rate-utility trade-offs.
- **Task 3:** Develop a learnable tokenization module that discretizes data into semantically meaningful units optimized for downstream tasks.
  - Design tokenization schemes for automotive sensor data (time series).
  - Define ML-aware utility metrics that correlate compression rate with downstream model performance (e.g., accuracy, F1-score).
- **Task 4:** Evaluate and compare the methods.
  - Measure rate-utility curves across the methods.
  - Evaluate trade-offs between computational efficiency.

[aa dummy, 9999]

5 Approach

- **Dataset:** Use available automotive sensor and telemetry test-fleet data supporting tasks such as predictive maintenance and anomaly detection.
- **Task 1:** Train downstream ML models (NeuralCompression or CompressAI) on uncompressed data to quantify loss in predictive utility.
- **Task 2:** Implement established neural compression methods (TBC) as baselines, measuring rate-utility trade-offs.
- **Task 3:** Develop a learnable tokenization module that discretizes data into semantically meaningful units optimized for downstream tasks.
  - Design tokenization schemes for automotive sensor data (time series).
  - Define ML-aware utility metrics that correlate compression rate with downstream model performance (e.g., accuracy, F1-score).

- **Task 4:** Evaluate and compare the methods.
  - Measure rate-utility curves across the methods.
  - Evaluate trade-offs between computational efficiency.
- **Optional Task 5:** Evaluate the use of the tokenization framework as a precursor to neural compression methods, to further improve rate-utility trade-off.
- **Expected Outcome:** Demonstrate that task-aware tokenization achieves comparable rate-utility trade-off to established neural compression approaches, while increasing computational efficiency.

[aa dummy, 9999]

References

aa dummy, 9999.

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.

Michele Bertoncello, Christopher Martens, Timo Möller, and Tobias Schneiderbauer. Unlocking the full life-cycle value from connected-car data. Technical report, McKinsey & Company, McKinsey Center for Future Mobility, Feb 2021. URL `https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/unlocking-the-full-life-cycle-value-from-connected-car-data`. White paper.

Fanghua Chen, Hong Jia, and Wei Zhou. Vehicle maintenance demand prediction: A survey. *Applied Sciences*, 15(20), 2025. ISSN 2076-3417. doi: 10.3390/app152011095. URL `https://www.mdpi.com/2076-3417/15/20/11095`.

Ranak Roy Chowdhury, Xiyuan Zhang, Jingbo Shang, Rajesh K. Gupta, and Dezhi Hong. Tarnet: Task-aware reconstruction for time-series transformer. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 212–220, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539329. URL `https://doi.org/10.1145/3534678.3539329`.

Gregory Grefenstette. *Tokenization*, pages 117–133. Springer Netherlands, Dordrecht, 1999. ISBN 978-94-015-9273-4. doi: 10.1007/978-94-015-9273-4_9. URL `https://doi.org/10.1007/978-94-015-9273-4_9`.

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.

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`.

Yuxin Liu, Xiaoyu Wang, Yuchen Zhang, Jian Li, and Wei Chen. Machine learning-based models for predicting patients' no-show in medical appointment. *Journal of Healthcare Engineering*, 2023:5025255, 2023. doi: 10.1155/2023/5025255. URL `https://doi.org/10.1155/2023/5025255`.

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.

Hyeran Mun, Kyusuk Han, and Dong Hoon Lee. Ensuring safety and security in can-based automotive embedded systems: A combination of design optimization and secure communication. *IEEE Transactions on Vehicular Technology*, 69:7078–7091, 2020. doi: 10.1109/tvt.2020.2989808.

{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.

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.

Andreas Theissler, Judith Pérez-Velázquez, Marcel Kettelgerdes, and Gordon Elger. Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry. *Reliability Engineering & System Safety*, 215, 2021. ISSN 0951-8320. doi: https://doi.org/10.1016/j.ress.2021.107864. URL https://www.sciencedirect.com/science/article/pii/S0951832021003835.

Unknown. Automobile embedded real-time systems. In *Systems Engineering of Software-Enabled Systems*. Wiley, 2019. doi: 10.1002/9781119535041.app2.

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.

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.

Övgü Özdemir, M. Tuğberk İşyapar, Pınar Karagöz, Klaus Werner Schmidt, Demet Demir, and N. Alpay Karagöz. A survey of anomaly detection in in-vehicle networks, 2024. URL https://arxiv.org/abs/2409.07505.