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, is a constant trade-off between data compression and 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.

A promising development in this area has been neural compression, which leverages deep learning models to learn efficient data representations that retain essential features for ML tasks while achieving high compression ratios [Yang et al., 2022]. Several studies have demonstrated the potential of neural compression techniques to outperform traditional methods in preserving ML model performance at lower bitrates. In one notable study [Löhdefink et al., 2019] show that although GAN-based compression may score worse in PSNR/SSIM, it can yield better semantic segmentation (mIoU) at very low bitrates compared to JPEG2000, when the segmentation model is trained on reconstructions from that codec. Similar findings are reported when looking at time series data. [Zheng and Zhang, 2023] for example, demonstrate that the neurol compression method used, outperforms state-of-the-art benchmarking models in terms of lower reconstruction errors with the same compression ratio. It is important to note, that they do not evaluate downstream ML tasks directly, just reconstruction.

Two major challenges remain across all the examples shown above. First, the rate-utility trade-off. The above mentioned trade-off between compression rate and ML utility is oftern referred to as rate-utility trade-off. Numerous papers like [Bao et al., 2025] and [Zhao et al., 2025] adress this as an ongoing challange.

Another limitation most of these papers fail to address is the computational constraints of in-vehicle embedded systems. The mentioned papers, with the exception of [Bao et al., 2025] which uses dataset distillation, which comes with its own problems [Yang et al., 2024], primarily focus on achieving high compression rates while maintaining model performance, therefore often choosing computationally heavy neural network architectures like recurrent neural networks (RNNs) or transformers.

So while modern 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 simons 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

The goal of compression, in its simplest terms, is to find a reduced representation of data that preserves the information relevant to a task. In natural language processing tasks, this is often achieved through tokenization [Schmidt et al., 2024]. Tokenization is tradionally 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 neural compression: 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. We propose, that this idea can be translated to time series data. Instead of compressing raw sensor values, we 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.

To achieve this we define the following goals and challenges:

- **Main goal:** Develop and evaluate a *task-aware tokenization framework* for automotive data that balances computational efficiency, compression rate, and ML utility.
- **Sub-goals:**
  - Quantify the loss in predictive utility when training ML models on uncompressed, tokenized and compressed data.
  - ...
- **Challenges:**
  - Agree on a downstream ML task or task type (e.g., predictive maintenance, anomaly detection).
  - Define how computational efficiency will be measured (e.g., inference time, model size).
  - Agree on a subset of the available automotive data.

[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 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.

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.

Youneng Bao, Yiping Liu, Zhuo Chen, Yongsheng Liang, Mu Li, and Kede Ma. Dataset distillation as data compression: A rate-utility perspective, 2025. URL `https://arxiv.org/abs/2507.17221`.

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`.

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.

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.

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.

Craig W. Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. Tokenization is more than compression. *arXiv preprint*, pages 678–702, 2024. doi: 10.48550/arxiv.2402.18376.

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.

William Yang, Ye Zhu, Zhiwei Deng, and Olga Russakovsky. What is dataset distillation learning?, 2024. URL `https://arxiv.org/abs/2406.04284`.

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.

Xiaobo Zhao, Aaron Hurst, Panagiotis Karras, and D. Lucani. dreamlearning: Data compression assisted machine learning. *ArXiv*, abs/2506.22190, 2025. doi: 10.48550/arxiv.2506.22190.

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`.