Master Thesis Project Proposal

Neural Compression for Efficient Model Serving

Tim Boleslawsky, gusbolesti@student.gu.se

Emrik Dunwald, gusbolesti@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

2 Problem and Context

2.1 Problem Statement

When discussing the down-stream machine learning capabilities of modern automotive-systems, there is an inherent trade-off between bandwidth/storage and machine learning utility. For example, event-triggered logging is efficient but introduces bias into logged data, possibly omitting predictive precursor signals (gregstanleyandassociates). Moreover, even advanced learned or neural compression approaches exhibit similar trade-offs between compression efficiency and task-relevant information fidelity (Löhdefink).

Let's shed a little more light on where we are coming from and why the above mentioned trade-off is important.

2.2 Why do we use Event-triggered Logging in Automotive Systems?

"Early in-vehicle networking architectures (Classical CAN at 1 Mbit/s; later FlexRay) were sufficient for control loops but not for the sustained high-throughput streams produced by cameras, radar, and LiDAR, as well as growing observability concerns, prompting selective data acquisition strategies." "To prevent bus saturation, manufacturers adopted event-triggered and threshold-based diagnostic logging, which—while reducing communication load—introduced maintenance overhead and diminished holistic observability."

There is a lot of data generated from modern vehicles! Bello et al. [2019] points out a that, as foreseen by Intel, the amount of data generated will increase dramatically: from an average of 1.5 GB of traffic data per Internet user today, we will move toward 4000 GB of data generated per day by an AD car including technical data, personal data, crowd-sourced data, and societal data.

2.3 Limitations of Event-triggered Logging and the Data Quantity Pressure

2.4 Relevance of Downstream Machine Learning Tasks in Modern Automotive Systems

2.5 Traditional Compression Methods

Why can't we use traditional compression methods?

- For video/image (JPEG, MP3, ...): optimized for human perception (e.g., visual quality) rather than machine learning tasks or efficient downstream data use.
- For time series data (algorithmic approaches like CHIMP or Gorilla): Dependence on manually chosen parameters like window size & Sensitivity to data characteristics (entropy, signal variability).

2.6 Neural/ Learned Compression as a Possible Solution

Here is what achievements have been made in the field of neural compression:

- There have been numerous advances in learned compression methods for images, videos, and time series data that outperform traditional methods in terms of rate-distortion performance. Barakat et al. [2025], for example, investigate the use of deep learning based techniques for fisheye image compression in automotive applications. The authors find that learned compression methods can achieve better compression performance particularly at low bitrates crucial for automotive applications, which as the authors point out is crucial for automotive applications.
-

3 Goals and Challenges

3.1 Possibility A

- Select one or more learned compression methods (for image/video, maybe telemetry streams) that perform well in lab/benchmark settings (e.g., in academic datasets).
- Adapt them to the constraints of automotive deployment: e.g., limited compute resources on vehicle ECUs, low latency requirements, energy/power budgets, heterogeneous data (camera + LiDAR + telemetry).
- Evaluate them "in practice" – meaning: on realistic vehicle sensor data (or representative subsamples), under scenarios closer to production: continuous streaming, varying environmental conditions, real-world noise/interference, embedded hardware.
- Measure metrics beyond pure rate-distortion: latency, throughput, memory footprint, energy consumption, robustness, interoperability with downstream ML tasks (e.g., perception, predictive maintenance) and the integration into the vehicle's data pipeline (on-board, bus, cloud).
- Possibly propose modifications (network architecture trimming / quantization / hybrid learned + conventional method) to meet these constraints.

→ Problem: We would need to write C# code and the focus is very much on computer science and not "Data Science/ Machine Learning"!

3.2 Possibility B

- Task specific compression with Signet as the consumer. The ultimate goal is to increase the data quality for downstream data consumers.
- Maybe taking ideas of tokenization to inform compression.

4 Approach

References

Bilal Barakat, Islam Sobh, Chi Cheung Wong, and Others. A comprehensive study of fisheye image compression and perception for autonomous driving. *Neural Computing and Applications*, 37:5765–5780, 2025. doi: 10.1007/s00521-024-10831-w. Early access / Online first if not yet assigned issue.

Lucia Lo Bello, Riccardo Mariani, Saad Mubeen, and Sergio Saponara. Recent advances and trends in on-board embedded and networked automotive systems. *IEEE Transactions on Industrial Informatics*, 15 (2):1038–1051, 2019. doi: 10.1109/TII.2018.2879544.