

Quantised Neural Network Accelerators for Low-Power IDS in Automotive Networks

Shashwat Khandelwal, Anneliese Walsh & Shanker Shreejith
Reconfigurable Computing Systems Lab, Electronic & Electrical Engineering
Trinity College Dublin, Ireland
Email: {khandels, walsh61, shankers}@tcd.ie

Abstract—In this paper, we explore low-power custom quantised Multi-Layer Perceptrons (MLPs) as an Intrusion Detection System (IDS) for automotive controller area network (CAN). We utilise the FINN framework from AMD/Xilinx to quantise, train and generate hardware IP of our MLP to detect denial of service (DoS) and fuzzing attacks on CAN network, using ZCU104 (XCZU7EV) FPGA as our target ECU architecture with integrated IDS capabilities. Our approach achieves significant improvements in latency (0.12 ms per-message processing latency) and inference energy consumption (0.25 mJ per inference) while achieving similar classification performance as state-of-the-art approaches in the literature.

Index Terms—Controller Area Network, Intrusion Detection System, Machine Learning, Field Programmable Gate Arrays

I. INTRODUCTION & PROPOSED ARCHITECTURE

Controller Area Network (CAN) continues to be the most widely used network architecture for in-vehicle networks owing to its cost-effective nature and ease of use in control applications. However, CAN (and other automotive networks like FlexRay, LIN) does not offer any inherent security and authentication mechanisms, which has been exploited in numerous attacks on in-vehicle systems [1]. Intrusion detection approaches are aimed at detecting such threats on the network and allowing critical systems to enter into a ‘safe working’ mode when such threats are detected. Machine Learning (ML) approaches have shown significant improvement in detection accuracy of such threats and ability to adapt to newer attack vectors [2]–[4]; however, integrating them within a vehicular network is still challenging due to dedicated ECU integration (higher power, space, wiring) with accelerators like GPUs (to meet inference latency).

In this work, we investigate a custom quantised multi-layer perceptron (MLP) model as an IDS and its integration as a coupled accelerator on an ECU. We have explored quantised Convolutional Neural Network (CNN) models as IDSs in our previous works but found lower classification performance when the whole frame is used as the input feature [5], [6]. The proposed architecture, shown in fig. 1, uses a AMD/Xilinx Zynq Ultrascale+ device acting as a standard ECU, with our IDS model integrated as a custom IP on the programmable logic. In our approach, CAN packets received in the interface are handled as usual by the ECU to perform its task; additionally, the packet is copied into a FIFO style buffer capturing a time-series of messages, which is examined by our IDS IP for threat signatures. This integration uses the

ECU to control and manage the IDS accelerator through the software APIs, making them similar to accelerators enabled through AUTOSAR compliant abstraction layers. The MLP is defined using PyTorch and undergoes quantisation aware training using AMD/Xilinx’s FINN framework [7]. We utilise the openly available CAR Hacking dataset [4] for training and validation, which provides a labelled set of real CAN messages and multiple attack vectors captured from the OBD port of an actual vehicle. We train two models, to detect DoS attacks and Fuzzing attacks, using the Brevitas quantisation aware training library from AMD/Xilinx [8]. Design space exploration is performed to arrive at the quantisation level to reduce the resource consumption and computational complexity without compromising on the detection accuracy. From our experiments, we observed that 4-bit uniform quantisation achieved best performance in both DoS and Fuzzing attacks, and hence was chosen for deployment. Streaming layer optimisations and partitioning were chosen during FINN compilation flow to optimise the hardware IP generated from the trained MLP model. Subsequently, the generated IP is integrated into a standard Zynq design flow as a slave memory-mapped peripheral device. The FINN flow also generates low-level drivers for interfacing our IP from the ARM cores on the Zynq device over the AXI interface. We use a Linux operating system (from the PYNQ image) with low-level Xilinx run-time tools integrated to use these APIs for testing our approach. For our tests, we integrate a single IP core (DoS or Fuzzing) to test their performance in isolation.

II. RESULTS & CONCLUSION

To evaluate the proposed approach as an integrated IDS accelerator, we evaluate the classification performance of the model in identifying DoS and Fuzzy attacks from the Car Hacking dataset and compare them against state of the art techniques in the literature. Table I compares the performance of both our models (4-bit QMLP) against other results from the literature, in terms of inference precision, recall, F1 score and false negative rate. Subsequently, we quantify the per-message processing latency of the model, starting from the arrival of the CAN message at the interface to determine the detection delay incurred by the approach. Table II compares our result against other approaches in literature, which utilise different platforms (GPUs, Jetson edge accelerators, Raspberry Pi) and approaches (block of CAN messages v/s individual messages). It should

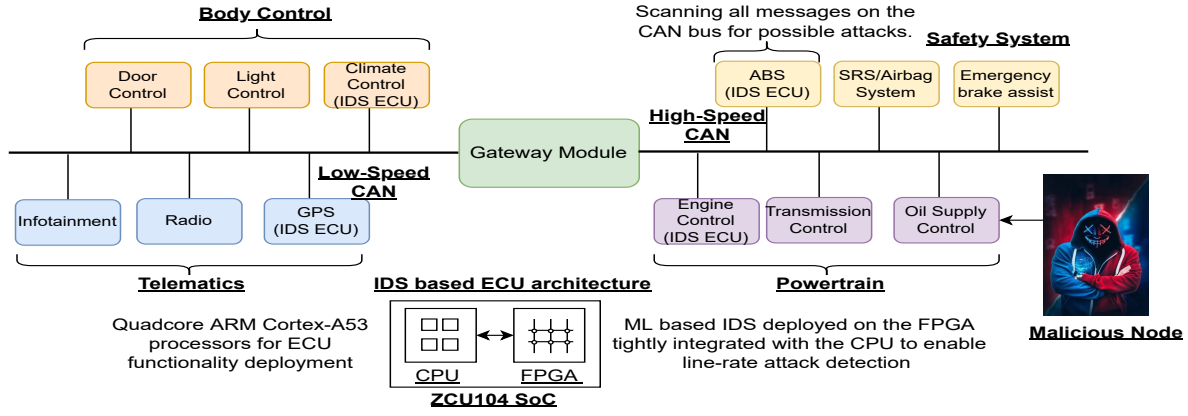


Fig. 1: System diagram of the CAN network (featuring components connected to high/low speed CAN) incorporating ECUs capable of intrusion detection.

TABLE I: Accuracy metric comparison (%) of our quantised FPGA accelerators against the IDSs in reported literature.

Attack	Model	Precision	Recall	F1	FNR
DoS	DCNN [4]	100	99.89	99.95	0.13
	MLIDS [3]	99.9	100	99.9	-
	NovelADS [10]	99.97	99.91	99.94	-
	TCAN-IDS [11]	100	99.97	99.98	-
	GRU [2]	99.93	99.91	99.92	-
	4-bit-QMLP	99.99	99.99	99.99	0.01
Fuzzy	DCNN [4]	99.95	99.65	99.80	0.5
	MLIDS [3]	99.9	99.9	99.9	-
	NovelADS [10]	99.99	100	100	-
	TCAN-IDS [11]	99.96	99.89	99.22	-
	GRU [2]	99.32	99.13	99.22	-
	4-bit-QMLP	99.68	99.93	99.80	0.07

be noted that the latency metric in the case of block-based IDS does not consider the delay in acquiring the number of CAN messages required, and hence could result in potentially larger delays in attack detection. Our approach achieves 0.12 ms per CAN frame, which is a $4.8\times$ improvement over the dedicated line-rate MTH-IDS approach [9]. In terms of throughput, our QMLP coupled ECU can process over 8300 messages per second at highest payload capacity, achieving near-line-rate detection on high-speed critical CAN networks. We observe that our model consumed 2.09 W when measured directly from the device's power rails (using the PYNQ-PMBus package) while performing inference and other tasks on the ECU (with Linux OS), thus consuming 0.25 mJ of energy per inference, significantly better than other approaches that use GPUs (9.12 J per inference for our 8-bit quantised MLP model on an A6000 for reference) and/or dedicated CPUs for IDS. The single model deployed consumes less than 4% of resources on the device, allowing multiple models to be executed simultaneously for a comprehensive IDS integration at slightly higher energy consumption than what we observed, making it an ideal deployment framework for line-rate IDS in automotive networks.

TABLE II: Per-message latency comparison against other state-of-the-art IDSs reported in literature.

Models	Latency	Frames	Platform
GRU [2]	890 ms	5000 CAN frames	Jetson Xavier NX
MLIDS [3]	275 ms	per CAN frame	GTX Titan X
NovelADS [10]	128.7 ms	100 CAN frames	Jetson Nano
DCNN [4]	5 ms	29 CAN frames	Tesla K80
TCAN-IDS [11]	3.4 ms	64 CAN frames	Jetson AGX
MTH-IDS [9]	0.574 ms	per CAN frame	Raspberry Pi 3
4-bit-QMLP (ours)	0.12 ms	per CAN frame	Zynq Ultrascale+

REFERENCES

- [1] C. Miller and C. Valasek, "Remote exploitation of an unaltered passenger vehicle," *Black Hat USA*, vol. 2015, no. S 91, 2015.
- [2] H. Ma, J. Cao, B. Mi, D. Huang, Y. Liu, and S. Li, "A GRU-Based Lightweight System for CAN Intrusion Detection in Real Time," *Security and Communication Networks*, vol. 2022, 2022.
- [3] A. K. Desta, S. Ohira, I. Arai, and K. Fujikawa, "MLIDS: Handling Raw High-Dimensional CAN Bus Data Using Long Short-Term Memory Networks for Intrusion Detection in In-Vehicle Networks," in *Proc. Intl. Telecommunication Networks and Applications Conference (ITNAC)*, pp. 1–7, IEEE, 2020.
- [4] H. M. Song, J. Woo, and H. K. Kim, "In-vehicle network intrusion detection using deep convolutional neural network," *Vehicular Communications*, vol. 21, p. 100198, 2020.
- [5] S. Khandelwal, E. Wadhwa, and S. Shreejith, "Deep learning-based embedded intrusion detection system for automotive can," in *2022 IEEE 33rd International Conference on Application-specific Systems, Architectures and Processors (ASAP)*, pp. 88–92, IEEE, 2022.
- [6] S. Khandelwal and S. Shreejith, "A Lightweight Multi-Attack CAN Intrusion Detection System on Hybrid FPGAs," in *Proc. Intl. Conf. on Field Programmable Logic and Applications (FPL)*, pp. 425–429, IEEE, 2022.
- [7] Y. Umuroglu, N. J. Fraser, G. Gambardella, M. Blott, P. Leong, M. Jahre, and K. Vissers, "Finn: A framework for fast, scalable binarized neural network inference," in *Proc. Intl. Symposium on Field-Programmable Gate Arrays (FPGA)*, pp. 65–74, 2017.
- [8] A. Pappalardo, "Xilinx/brevitas," 2021.
- [9] L. Yang, A. Moubayed, and A. Shami, "MTH-IDS: A Multitiered Hybrid Intrusion Detection System for Internet of Vehicles," *IEEE Internet of Things Journal*, vol. 9, no. 1, pp. 616–632, 2021.
- [10] K. Agrawal, T. Alladi, A. Agrawal, V. Chamola, and A. Benslimane, "NovelADS: A Novel Anomaly Detection System for Intra-Vehicular Networks," *IEEE Transactions on Intelligent Transportation Systems*, 2022.
- [11] P. Cheng, K. Xu, S. Li, and M. Han, "TCAN-IDS: Intrusion Detection System for Internet of Vehicle Using Temporal Convolutional Attention Network," *Symmetry*, vol. 14, no. 2, p. 310, 2022.