Authors: Volvo Group, Embedl
Date: 20241108
Project within FFI – Accelerate Startup Partnership

Content

Brief about FFI

FFI is a collaboration between the state and the automotive industry to jointly finance research and innovation activities with a focus on Climate & Environment and Traffic Safety. The initiative involves activities of approximately 1 billion SEK per year, of which public funds amount to over 400 million SEK.

Read more at www.vinnova.se/ffi.

1. Summary

The purpose of the project was to optimize two of Volvo's AI-powered use cases to run on Volvo's existing as well as future hardware platforms. This would be done with the help of Embedl's state-of-the-art optimization methods.

The project consists of two main components:
- Optimize deep neural networks in the following use case areas to run on Volvo's embedded HW platforms (a) computer vision, (b) signal processing.
- Analysis of future embedded hardware for deep learning.
    - Trends in networks and deep learning accelerators.

Compare different HW architecture options for DNN inference.

3. Background

Volvo Groups AI team wanted to find answers to some broad questions. E.g.:

- How do we optimize chosen deep neural networks for efficient inference on Volvo embedded HW platforms?
- How powerful deep learning solutions can we deploy while still meeting the minimum requirements for chosen applications?

Why do we need efficient models?
- DL models need to run on embedded hardware with constrained resources
- Use shared resources such as memory and compute as efficiently as possible.
- Build a complete and robust solution.

Neural network compression is one technology that allows networks to be run on our current ECUs and Embedl is one company that offers solutions to compress networks.

With this solution, various domains can deploy AI-based functionality on both current and future platforms. Further, the backward/forward compatibility that the Embedl toolchain provides, allows us to be hardware agnostic.

4. Purpose, Research Questions and Method

This project will determine if two of Volvo´s AI-powered use cases can be deployed to Volvo´s existing as well as future hardware platforms.

The project will show potential performance improvements of the selected Volvo use cases on target ECU hardware using Embedl´s state-of-the-art optimization methods.

Show how Embedl potentially contributes to improving productivity and agility. The project will show what state-of-the-art methods can achieve on already existing ECUs, and future compatibility in analyzing what will be technically possible in future hardware.

Method:

Use state-of-the-art optimization methods in Embedl Optimization SDK, such as pruning and quantization.

5. Objectives

- Integrate Embedl Optimization SDK at Volvo Group.
- Use Embedl SDK to optimize chosen models for efficient inference on target HW.
- Determine whether it is viable to reach low inference latency for both use cases.
- Analyze future deep learning accelerators analyzed in written report
- Share knowledge in model optimization tools and methods to reach real-time requirements.

6. Result and Goal Achievement

The results in the project have been positive:

- Both target use cases reach low inference latency on selected HW targets.
- DNN inference capabilities of selected HW targets have been evaluated to satisfactory levels.
- We established a process for evaluating next generation of hardware for DNN inference.

7. Dissemination and Publication

7.1 Knowledge and Result dissemination

How is/planned the project result to be used and disseminated?	Mark with X	Comment
Increase knowledge within the area	X	Volvo Group has been introduced to Embedl's deep neural network inference optimization tools. This is a useful addition to Volvo's toolkit
Carry forward to other advanced technical development projects	X	Post-project there have been introduction meetings to a number of other teams/departments where potential needs for model optimization technology have surfaced
Carry forward to product development projects	X	For future product development projects involving deep learning in embedded systems (edge AI) the solution will be considered.
Introduced to the market		
Used in investigations/regulations/permit cases/political decisions		

7.2 Publications

8. Conclusions and Further Research

o Model optimization can address the need to fit rapidly growing networks on constrained hardware.
o This project has clarified open points for in-vehicle AI.
o Many promising HW SoC options, tools, and runtimes for neural network inference are rapidly becoming available.
o The Embedl toolkit abstracts the use of many of them.

9. Participating Parties

Volvo Group

Embedl:

- Andreas Ask
- Olle Friman
- Daniel Ödman
- Adel Hasic
- Axel Jarenfors
- Ola Tiverman
- Hans Salomonsson