Kurzfassung des Schlussberichts

Verbundprojekt:
EMDRIVE - Konzeption und RT-kompatible Erweiterung von Zentralrechenplattformen und Embedded Compute Netzwerken für zukünftige hochautomatisierte Fahrzeuge EMDRIVE

Teilvorhaben:
Entwurfsmethodiken für effiziente, adaptive und optimierte Beschleunigerarchitekturen im Automotivekontext

Zuwendungsempfänger:
Karlsruher Institut für Technologie (KIT)

Förderkennzeichen:
16ME0454

Aufgabenstellung und Stand der Technik

Zukünftige autonome Fahrzeuge sind durch eine zunehmende Zentralisierung der verwendeten Elektrik/Elektronik-Architektur (E/E-Architektur) geprägt. Daraus folgt ein Anstieg in der Komplexität des Designprozesses dieser Architekturen. Zusätzlich steigt die benötigte Performance um Funktonen des autonomen Fahrens realisieren zu können.

Basierend auf diesen Trends war es das Ziel dieses Teilvorhabens neue Methodiken des Hardware/Software Co-Designs für die effiziente Ausführung von KI Algorithmen in eingebetteten Systemen zu entwickeln. Dabei wurde auf mehrere aktuelle Entwicklungen speziell im Bereich KI aufgebaut. Durch die zunehmende Zentralisierung wird nur ein Teil der Datenverarbeitung direkt auf dem Sensor ausgeführt, komplexere Operationen, wie z.B. Object Detection oder Segmentation, werden von einer zentralen Recheneinheit ausgeführt. Es erfolgt also eine verteilte Ausführung der Aufgaben auf heterogenen Plattformen. Gleichzeitig ergeben sich durch die Verbreitung von KI Algorithmen neue Herausforderungen. So ist es notwendig die verwendeten Netzwerke während dem Einsatz im Feld weiterhin zu trainieren, um auch auf neue Situationen reagieren zu können. Zudem müssen die Netze abgesichert werden damit Out-of-Distribution Fehler, die zu fehlerhaften Einschätzungen der Fahrsituation führen können, vermieden werden.

Insgesamt soll damit durch die gemeinsame Betrachtung von Hardware- und Softwarekomponenten die Lücke im Stand der Technik geschlossen werden, indem Methodiken für das Design zukünftiger eingebetteter Systeme im Automotive geschaffen werden.

Ablauf und Ergebnisse des Teilvorhabens

Während der Projektlaufzeit vom 01.02.2022 bis zum 31.01.2025 hat der Zuwendungsempfänger an allen sieben Arbeitspakten mitgewirkt.

Zu Beginn des Projektes wurde in enger Abstimmung mit den Konsortialpartnern ein Anforderungskatalog entwickelt um damit benötigte Komponenten, Funktionen und Schnittstellen festzulegen. Diese konnten über die Laufzeit des Projektes als Grundlage für die weitere Arbeit

verwendet werden. Weiterhin wurden Use-Cases sowie, darauf aufbauend, Möglichkeiten zur Demonstration der entwickelten Systemkomponenten definiert. Insgesamt konnte durch diese Vorarbeiten eine möglichst praxisnahe Arbeit sichergestellt werden.

Aufbauend auf den Vorarbeiten konnte anschließend die Hardware/Software Co-Design Methodik entwickelt werden. Hierfür wurden verschiedene Komponenten entwickelt, die zusammen genutzt werden können, um den Designprozess eingebetteter Systeme zu vereinfachen.

Zunächst wurde ein Framework entwickelt, um die Partitionierung neuronaler Netze auf heterogene Recheneinheiten zu untersuchen. Dabei kommen analytische Modelle zum Einsatz, die für die Modellierung von Performance und Energie KI Beschleunigern genutzt werden können. Ein solches Modell wurde für einen spezifischen Beschleuniger ebenfalls im Projekt entwickelt. Als weitere Verbesserung wurde eine Komponente integriert, die die automatisierte Design Space Exploration für KI Beschleuniger ermöglicht. Diese konnte genutzt werden um die Architektur des Beschleunigers für die von der Universität zu Lübeck sowie Infineon entwickelten Netzwerke zu optimieren. Durch die Integration der beiden Module kann somit die Co-Optimierung der verteilten Ausführung sowie der verwendeten Hardware automatisiert werden.

Im nächsten Schritt wurden Methoden zur Kompilierung neuronaler Netze und, damit verbunden, Möglichkeiten zur automatisierten Einbindung von Komponenten für das Online Learning untersucht. Hier wurde ein auf Apache TVM basierter Compiler entwickelt, der für das Deployment quantisierter Netzwerke und für das Online Learning genutzt werden kann. Dieser fügt, während das Modell kompiliert wird, automatisch zusätzliche Komponenten in das Netzwerk ein, die das Training ermögliche. Damit wird sichergestellt, dass die verwendeten Netzwerke auf im Felde weiter optimiert werden können um auch neue, zuvor unbekannt Objekte detektieren zu können.

Weiterhin wurden Methodiken entwickelt, um die Anomaliedetektion bei der Ausführung neuronaler Netze zu ermöglichen. Dafür wurde zunächst ein Tool entwickelt welches, basierend auf Inferenz-Traces, eine Auswahl passender Methoden zur Anomaliedetektion durchführt. Das Tool erzeugt die benötigten Traces automatisiert aus den Feature-Maps vorher spezifizierter Layer und trainiert damit einen Klassifikator, der anschließend verwendet werden kann. In einem weiteren Schritt wurde eine Methodik hinzugefügt, um zur Laufzeit auf erkannte Anomalien zu reagieren. Dazu werden mehrere Early Exits in das Netzwerk eingefügt und mit einem vorher trainierten Anomaliedetektionsmodul versehen.

Abschließend wurden Demonstratoren entwickelt, um die entwickelten Komponenten anschaulich vorzustellen. Als Grundlage dient hierbei der Use-Case der Radardatenverarbeitung im Automobilbereich, wobei speziell die Verwendung von neuronalen Netzen für diesen Schritt untersucht wurde. Hierfür wurde eine Smart Sensor Platform bestehend aus den vorher beschriebenen Komponenten entwickelt um diese praxisnah evaluieren zu können.