Lucerne University of
Applied Sciences and Arts
HOCHSCHULE
LUZERN
Technik & Architektur
FH Zentralschweiz

Diplomand **Hinninger Timo**

Dozenten Prof. Dr. Jürgen Wassner

Themengebiet Embedded Systems, Signalverarbeitung & Kommunikation

Projektpartner Hochschule Luzern
Institut Elektrotechnik

Skalierbarer FPGA-Beschleuniger für Neuronale Netze

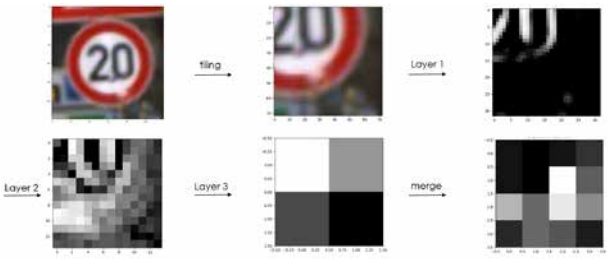

Abb. 1 Architektur des BinArray System mit vier SAs

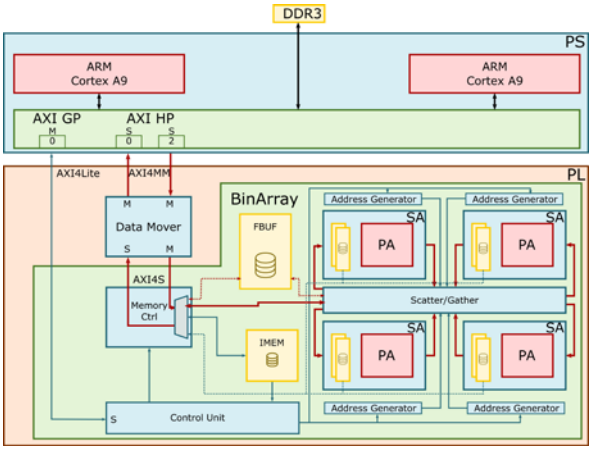

Abb. 2 Ablauf: Kacheln des Eingangsbildes, bearbeiten aller Layer und zusammensetzten der Ausgangsteilbilder anhand eines CNN mit drei Layern angewendet am GTSRB Datensatz.

Problemstellung

Convolutional Neuronal Networks (CNN) gelten aufgrund ihrer überlegenen Genauigkeit als State-of-the-Art für Computer Vision und andere Signalverarbeitungsaufgaben. BinArray, ein Low-Cost CNN Beschleuniger, der am CC ISN entwickelt wurde, arbeitet speziell komprimierte CNN effizient ab und ermöglicht den Gebrauch von CNN in Embedded Edge Computern mit begrenzten Ressourcen. BinArray ist anhand von drei Designparametern bezüglich Durchsatzes und Genauigkeit einstellbar. Der dritte Parameter, welcher die Anzahl an Teilbildern, die parallel bearbeiten werden können, wird in dieser Thesis eingeführt.

Lösungskonzept

Um an verschiedenen Teilbildern arbeiten zu können, muss das Eingangsbild zuerst gekachelt werden. Da bei der Faltungsoperation ohne Erweiterung des Eingangsbildes Daten verloren gehen, wurde ein Konzept entwickelt, welches beim Kacheln des Eingangsbildes die nötigen zusätzlichen Pixel berechnet und diese für die jeweiligen Teilbilder zusätzlich abspeichert. Abbildung 1 zeigt den Vorgang des Kachelns an einem Beispielnetzwerk für vier Teilbilder. Das Beispielnetzwerk hat drei Faltungsschichten. Am Ende werden die Teilausgangsbilder zum gesamten Ausgangsbild zusammensetzt.

Realisierung

Die Teilbilder werden von jeweils einem Sysolic Array (SA) abgearbeitet. Diese SA besitzen jeweils einen eigenen Speicher der die Eingangs- sowie die Ausgangsteilbilder abspeichert. Die SA's können so unabhängig voneinander alle Layer des CNNs abarbeiten. Das Kacheln kann aus dem Programmierbaren Logik (PL) Teil des Systems in welchem sich das BinArray und somit auch die SA's befinden, in den Processing System (CPU) Teil ausgelagert werden (Abb.2).

Ergebnisse

Ergebnisse zeigen, dass durch Erhöhen der Anzahl an SA von einem auf zwei ein erhöhter Durchsatz erreicht werden kann. Der erhöhte Durchsatz geht auf Kosten von zusätzlich benötigten Hardware Ressourcen. Der Anteil an Hardware Ressourcen der SA's entsprechen fast dem gesamten Hardware Gebrauch. Somit verdoppelt sich der Hardware Aufwand beinahe mit der Verdoppelung an SA's. Weiter ist zu erwähnen, dass kacheln des Bildes vor allem effektiv ist, wenn sich die Anzahl an Ausgangspixel der letzten Faltungsschicht durch die Anzahl an SA's teilen lässt.

Ausblick

In der Arbeit wurden weitere Kachelmethoden bzw. Speichermethoden aufgezeigt. Diese können in einer nächsten Projektphase implementiert und verglichen werden.