Lehrstuhl für Integrierte Digitale Systeme und Schaltungsentwurf

RWTH AACHEN

Masterarbeit

Entwurf eines FPGA Beschleunigers zur quantitativen Bewertung von Deep-Learning Algorithmen

Deep Neural Networks (DNN) bestehen aus einer Vielzahl von trainierbaren Einzelschichten, die zu einem komplexen neuronalen Netz verschaltet werden. Das Training der Einzelschichten erfolgt aufgabenabhängig und – wie in neuronalen Netzen üblich - datengetrieben mit Hilfe von sog. Deep-Learning-Algorithmen realisiert. Typischerweise finden hier Verfahren Anwendung, die z.B. auf dem *steepest-gradient-descent–Algorithmus* basieren und Fehlerinformation zur Anpassung von Trainingsgewichten geeignet in tiefer gelegene Schichten transportieren (*error backpropagation*). Während die eigentlich Trainingsverfahren zur Anpassung von Trainingsgewichten theoretisch gut untersucht sind, ergeben sich hinsichtlich der zugrunde gelegten Arithmetik eine Vielzahl offener Fragestellungen. Speziell mit Blick auf energieeffiziente digitale Hardwareimplementierungen sind Fragen der erforderlichen Wortbreiten zur Repräsentation von Konstanten, Fehlersignalen, Gewichten, Trainingsdaten und Adaptionssignalen weitgehend offen. Allgemein zeigt sich zunächst empirisch, dass aufgrund der erforderlichen graduellen Anpassung von Trainingsgewichten in der Trainingsphase einerseits höhere Wortbreiten zu Abbildung von Trainingssignalen zur Vermeidung von Fehlerfortpflanzungseffekten erforderlich sind, andererseits im Vergleich relativ geringe Wortbreiten bei der Inferenz („Musterabruf“) benötigt werden.

In dieser Arbeit sollen anhand eines beispielhaften Neuronalen Netzes zur Ziffernerkennung („LeNet“) Untersuchungen hinsichtlich der Auswirkungen limitierter Wortbreiten in dedizierten Hardwareimplementierung vorgenommen werden. Für die Durchführbarkeit umfassender Studien wird zusätzlich die Realisierung einer Simulationsbeschleunigungshardware auf Basis von FPGAs angestrebt.

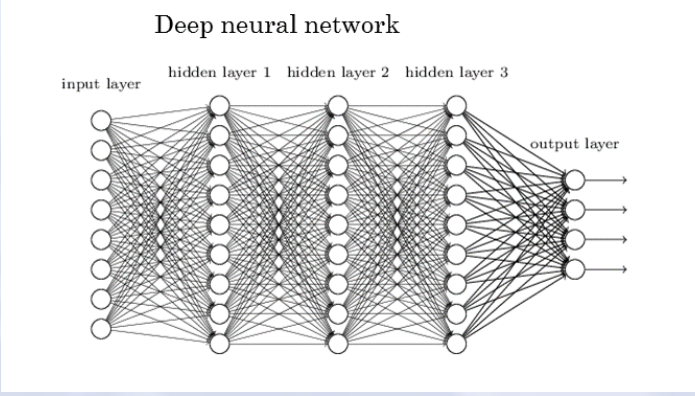

Bild 1: Prinzipskizze eines „Deep Neural Network“ mit einer Vielzahl von Schichten

Aufgaben

1. Spezifikation und Beschreibung einer rekonfigurierbaren Hardwarearchitektur mit zunächst dediziert getrennten Signalpfaden für „Inferenz“ und „Error-Backpropagation“ zur Implementierung von „LeNet“
2. Quantifizierung der Auswirkungen limitierter Wortbreiten in digitalen Arithmetikeinheiten hinsichtlich der Erkennungsrate; „False Positive“ vs. „true Negatives“.
3. Quantifizierung der Auswirkungen limitierter Wortbreiten in digitalen Arithmetikeinheiten hinsichtlich der Lernparameter; schichten-und epochen-abhängige Lernraten, Anzahl erforderlicher Lernepochen
4. Verschmelzung beider Signalpfade, Realisierung von konfigurierbaren Arithmetik-einheiten und Speicherschnittstellen, Benchmarking
5. Abbildung fundamentaler Arithmetikeinheiten auf ein FPGA zur Simulationsbeschleunigung

Voraussetzungen

- selbstständiges Arbeiten
- Kenntnisse in Matlab, C++, VHDL/Verilog, FPGA Design sind von Vorteil

Kontakt

Prof.Dr.-Ing. Tobias Gemmeke
gemmeke@ids.rwth-aachen.de
+49 241 80 97600

Dr.-Ing. Arne Heittmann
heittmann@ids.rwth-aachen.de
+49 241 80 97591