Technische Universität Dortmund

Fakultät Informatik
Design Automation for Embedded Systems Group

TODO Titel

Fachprojekt

Jack Diep, Florian Köhler, Yannick Naumann

Betreuung:
M.Sc. Mikail Yayla

26. August 2021

Inhaltsverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einleitung

Inhalt dieses ersten Kapitels: Motivation, Hintergrund, Aufbau der Arbeit etc.

1.1 weitere Unterkapitel

⋮

Kapitel 2

Das Netzwerk

2.1 Aktivierungsfunktion

2.2 Binärer Linear-Layer

2.3 Verluste durch binäre Linear-Layer

Kapitel 3

Binarisierung

Kapitel 4

Export

Nachdem das Netzwerk nun trainiert wurde, müssen die Ergebnisse, die Kantengewichte und Schwellwerte der Neuronen, nun exportiert werden. Im Folgenden sollen diese dann in den BNN-Beschleuniger Baustein importiert und verwendet werden. Da der Import in VHDL stattfindet, eignen sich hier simple Formate, sprich eine einfache Textdatei. Diese kann dann, Zeichen nach Zeichen, von dem Import-Buffer eingelesen und in einer Matrix gespeichert werden.

4.1 Export der Kantengewichte

Da es sich bei unserem Netzwerk um ein *FullyConnected Neural Network* handelt, ist insbesondere jedes Neuron mit jedem Neuron der Folgenden Schicht verbunden. Bei unserem BNN ergibt sich also folgende Kantenanzahl

$$784 \cdot 500 + 500 \cdot 1024 + 1024 \cdot 1024 = 1.952.576$$

Diese Gewichte müssen alle, mit möglichst wenig Mehrkosten, in die Datei geschrieben werden. Da es sich bei den Gewichten lediglich um binäre Werte, Einsen und Nullen, handelt, ist kein Trennzeichen zwischen den Gewichten notwendig. Die Gewichte sind außerdem, trivialer Weise, präfixfrei und können fortlaufend in die Datei geschrieben werden.
Um die Gewichte zu extrahieren, wird zuerst über jeden *Layer* iteriert. In jedem *Layer* wird nun jedes Neuron abgelaufen. Jedes dieser Neuronen hat nun jeweils eine Kante

zu jedem Neuron in der nachfolgenden Schicht. Hier wird ebenfalls über alle Kanten iteriert und das jeweilige Gewicht wird hinten an eine Variabel an gehangen. Ist nun ein *Layer* fertig, wird der Inhalt der Variable, welche als Zwischenspeicher dient, in die Datei *weights.txt* geschrieben. So wird für jede Schicht ein IO-Zugriff gemacht.

4.2 Export der Schwellwerte

4.3 Verbesserungen beim Export

Wie in Kapitel 4.1 beschrieben, wird für jeden *Layer* ein IO-Zugriff gemacht. Die ursprüngliche Implementierung sah hier vor, dass jedes Kantengewicht direkt der Datei angehangen wird. Diese ursprüngliche Implementierung sorgte jedoch für erhebliche Performance-Einbußen.

Literaturverzeichnis

[1] Beutelspacher, Albrecht: Das ist o.B.d.A. trivial!. Tipps und Tricks zur Formulierung mathematischer Gedanken. Vieweg Verlag, Braunschweig und Wiesbaden, 2004.