

Training von Deep Learning Algorithmen für die Anwendung auf einem Mikrocontroller

Diplomand



Dominic Wunderlin

Einleitung: Im Bereich der künstlichen Intelligenz gewinnen neuronale Netze immer mehr an Bedeutung. Daher wird es immer wichtiger, Mikrocontrollern den Zugriff auf solche Netzwerke zu ermöglichen. Mikrocontroller haben ganz offensichtliche Einschränkungen, einerseits Speicherplatz, andererseits Performance. In dieser Arbeit wurden die Machbarkeit und mögliche Grenzen eines solchen Netzwerks ermittelt. Daher wurde ein Netzwerk zur Erkennung von Sprachbefehlen erstellt und optimiert, damit es auf einem STM ARM-Mikrocontroller ausgeführt werden kann.

Vorgehen / Technologien: Für das Erstellen eines Netzwerkes ist TensorFlow als Ausgangsplattform gewählt worden (Version 2.6.0). Ein bekannter, von TensorFlow verwendeter Audio-Datensatz (von Pete Warden) wird für das Trainieren des Netzwerkes verwendet. Die Befehle stop, go, left, right sollen erkannt werden. Der Datensatz ist in drei Pakete unterteilt: Einem Trainingsatz, der für das Trainieren zuständig ist, einem Validierungssatz, der für die Erkennung von Overfitting zuständig ist, und einem Testsatz, der dazu da ist, um das Netzwerk auf die Zuverlässigkeit zu testen. Für ein besseres Training wird der Datensatz einer Datenaugmentation unterzogen.

Bei der Datenaufbereitung und Umwandlung sind gewisse Grenzen und Grössen vom Mikrocontroller vorgegeben. Mit diesen Vorgaben wurden die Wavedateien einem Mel-Spektrogramm Algorithmus von Librosa unterzogen und als Grafik, im Format eines png ausgegeben (siehe Grafik). Für die Netzwerk-Architektur stehen ein fully-connected und ein Convolutional Neuronales Netzwerk (CNN) zur Verfügung. Diese beiden Netzwerkarchitekturen wurden mit einem systematischen Ansatz ermittelt. Der Ansatz: Kleinstmögliche Architektur, höchstmögliche Zuverlässigkeit.

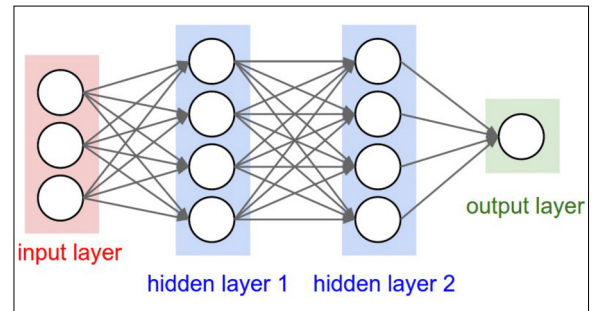
Für eine noch kleinere Architektur stellt TensorFlow das Quantisieren zur Verfügung, die jedoch einen Zuverlässigkeitsverlust beinhaltet. Mit einer geeigneten Anpassung konnte jedoch dieser Verlust so gering wie möglich gehalten werden. Es gibt zum einen die Möglichkeit, mit der post-training-quantization das Netzwerk, nach dem es trainiert wurde, zu quantisieren oder während dem Training mit dem quantized-aware-training (QAT).

Ergebnis: Durch Quantisierung kann die Netzwerkgrösse um das bis zu Vierfache reduziert werden, ohne dass es zu signifikanten Verlusten an Zuverlässigkeit kommt.

In der Grafik sind die Resultate des CNN und des fully-connected Netzwerk aufgelistet, jeweils einmal unquantisiert, einmal mit der post-training-quantization und einmal mit dem QAT. Wie in der Grafik ersichtlich konnten keine signifikanten Unterschiede festgestellt werden.

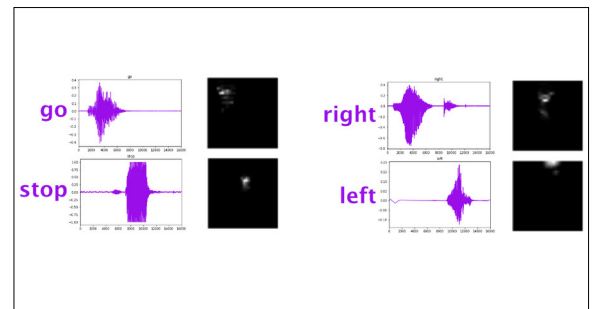
Struktur eines neuronalen Netzwerkes

Website: researchgate.net



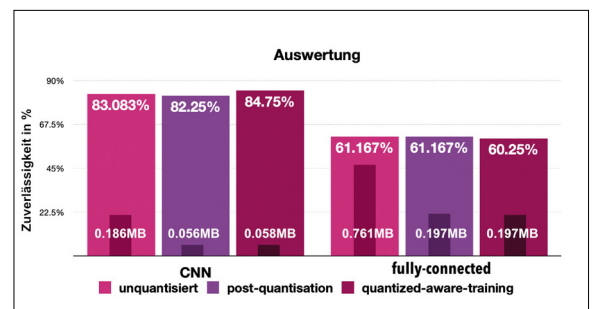
Wavedatei zu Mel-Spektrogramm

Eigene Darstellung



Zuverlässigkeit und Grösse der Netzwerke

Eigene Darstellung



Referent

Hannes Badertscher

Korreferent

Gabriel Sidler, Teamup Solutions AG, Zürich, ZH

Themengebiet

Artificial Intelligence

Projektpartner

ICAI Interdisciplinary Center for Artificial Intelligence, Rapperswil, SG