

要旨

微分可能ニューラルコンピュータ向き ハードウェアアクセラレータ回路の検討

齋藤 あかね

近年の深層人工神経回路網 DNN(Deep Neural Network) 研究の進展に伴い, DNN の高速化を目的としたアクセラレータ回路が提案されている. これらは従来の多層 DNN のみを対象としており, 微分可能ニューラルコンピュータ DNC(Differentiable Neural Computer) のような新たに提案されたニューラルネットワークを効率良く実行するには機能が不足している.

DNC は LSTM 型のニューラルネットワークと外部メモリで構成されているため, LSTM ニューロンの計算と外部メモリへの読み書き操作を高速化する必要がある.

本研究では, パイプライン型ハードウェアアクセラレータを用いて, 1 命令で重み付き和とその非線形活性化関数を実行する. このアクセラレータでは, 各 SRAM モジュールに格納されている入力データと重みが順番にフェッチされた後, それぞれが逐次乗算されて累積され, 最終的な結果データが SRAM モジュールに書き戻される. さらに, 命令で記述された即値を使用することでダイレクトアドレッシングモードを使用している.

提案アクセラレータは Intel FPGA チップ Stratix V 向けに設計し, 性能評価を行った. その結果, 回路は 256 ビット × 32 ワードの外部メモリを有する DNC を 0.5M ニューロン/秒でシミュレートできることが確認された.

キーワード 微分可能ニューラルネットワーク, ハードウェアアクセラレータ

Abstract

A Study on Hardware Accelerator for Differentiable Neural Computer

Saito Akane

Along with the recent progress of deep artificial neural network (DNN) research, accelerator circuits aiming at high speed of DNN has been proposed. These are dedicated only for conventional multilayer DNNs so that they are lack of efficiency to execute a newly proposed neural network such as Differentiable Neural Computer (DNC).

Since the DNC is composed of a long-term short-term memory (LSTM) type neural network and an external memory, it is necessary to speed up the calculation of the LSTM neurons and its reading / writing operations to the external memory.

In this research, a pipelined hardware accelerator is studied in order to execute a weighted sum and its non-linear activation function by a single instruction. In this accelerator, all input data and weights stored in each SRAM module are fetched in order, and then each of them is multiplied and accumulated consecutively, and the final resultant data is written back into the SRAM module. Furthermore, the direct addressing mode is allowed by using an immediate value described in the instruction.

The proposed accelerator was designed for Stratix V, Intel FPGA chip, and the performance evaluation was carried out. As a result, the circuit was confirmed to simulate a DNC with 256bit x 32word external memory at 0.5 M neurons per second.

key words Differentiable Neural Computer, Hardware Accelerator