LSTM-RNN 用マルチコアアクセラレータの負荷割当法の検討

1180386　山崎　尚之　【 コンピュータ構成学研究室 】

1　はじめに

近年，再帰型ニューラルネットワーク RNN(Recurrent Neural Network) が言語処理や音声認識の分野で注目され，リアルタイム翻訳機などの組込みシステムに搭載することが期待されている．

先行研究 [1] では，長期短期記憶 LSTM(Long Short-term Memory) を含む微分可能ニューラルコンピュータ DNC(Differentiable Neural Computer) 用アクセラレータ回路が提案されている．しかし，単一コアでは十分な性能が得られないためマルチコア化が必須である．

本研究では，大規模な LSTM 型 RNN を対象として，上記アクセラレータをマルチコア化した回路上で効果的な並列処理を行うための負荷割当法を検討した．

2　負荷割当法

本研究で対象とするアクセラレータ回路は，２つのベクトルの積和演算と線形関数を 1 命令で実行可能な 5 段命令パイプライン構成となっている．ベクトルポインタをレジスタファイルでフェッチして，2 個の SRAM モジュールから連続的にベクトルを読み出し、結果をアキュムレータに保存する回路から構成されている．その後，必要に応じて，活性化関数を LUT で参照する．

図 1 に示す多数の LSTM ニューロンの計算をマルチコアアクセラレータの上で実行する場合，各コアで独立して部分的に実行可能な分割法が重要になる．図 1 から分かるように，原理的には，同時並行，パイプライン実行，データ並列の軸が考えられる．本研究では，各コアに必要なローカルメモリ量が最も少ないデータ並列に着目して，負荷割当法を検討した．この場合，図中の積和演算を複数コアで部分的に実行して，最後に集約する処理が必要になる．

データ並列の負荷割当法では，入力ベクトルの全データが揃っていなくとも，割り当てられたデータが全て揃ったコアから順次先行評価できる．よって，演算命令より通信命令を常に優先的に実行し，各コアのアイドル時間を極力少なくすると性能向上が図れる．また，上述の集約処理の遅滞が全体の処理時間に影響するため，集約に必要なデータの通信命令の実行を最優先する．これらの必要条件を満たした上で，図 1 の計算を随時データ駆動的に実行する．

3　性能見積もり

提案したデータ並列方向の負荷割当法について，アクセラレータ内のパイプラインストールを考慮したクロックサイクルベースの性能評価を行った．コア数 32 の場合について，中間層のニューロン数に対するコアの平均稼働率の変化を図 2 に示す．この結果からニューロン数の増加に伴い稼働率が上昇する．これは，通信命令に対する演算命令の比が増加し，コア間通信網での遅延が他の演算命令の代替実行で隠蔽できる確率が向上するためである．

4　まとめ

大規模 LSTM 型 RNN をマルチコア上で並列実行するための負荷割当法として，データ並列方向の負荷割当と命令スケジューリング法を提案した．性能見積もりの結果，RNN の規模が増えるほど，各コアの稼働率を向上できることが分かった．今後，コア間通信網の回路を設計して，具体的な通信コストを加味した性能を評価する必要がある．

参考文献

[1] A. Saito, Y. Umezaki, and M. Iwata, “Hardware Accelerator for Differentiable Neural Computer and Its FPGA Implementation,” PDPTA’17, pp. 232-238, July 2017.

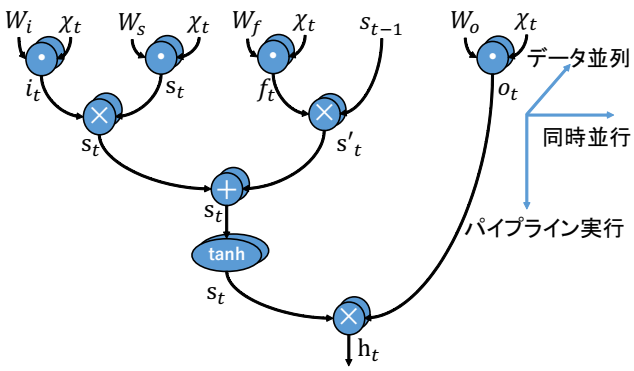

図 1　複数の LSTM ニューロンの計算グラフ

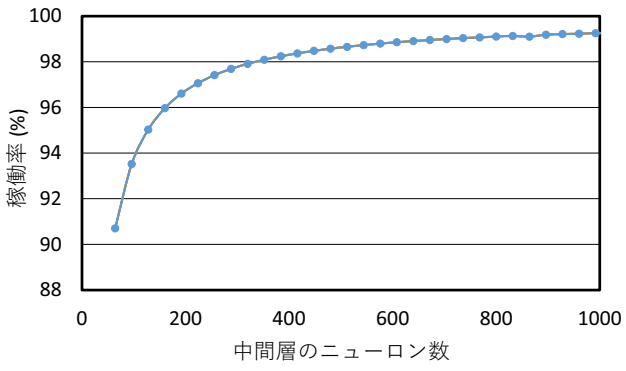

図 2　全コアの平均稼働率（コア数 32 の場合）