# LSTM-RNN 用アクセラレータ回路の相互結合網の検討

1180386　山崎　尚之　【 岩田研究室 】

## 1　はじめに

言語処理，音声認識の分野で長期短期記憶LSTM(Long Short-term Memory) が注目されており，組込みシステムでリアルタイム翻訳などに用いる場合アクセラレータ回路を使用し，高速に処理する必要がある．LSTM-RNN(Recurrent Neural Network) は一般的な人工ニューラルネットワーク ANN(Artificial Neural Network) に比べ，3 種のゲートと自状態の重みが必要で演算量が大きくなる．そのため，LSTM-RNN 用アクセラレータの要求が求められるようになっている．

先行研究として，LSTM を含む微分可能ニューラルコンピュータ DNC(Differentiable Neural Computer) 用単一コアの提案がされている [1]．しかし，シングルコアでは処理性能に限界があるためマルチコア化が求められる．

本研究では，LSTM をマルチコアで動作させる場合に必要な相互結合網回路での負荷の割り当て法について，負荷を見積もり検討を行う．また，見積もった性能が達成できているか FPGA 上に実装し，性能を評価する．

## 2　負荷割り当て方法の検討

負荷割り当ての方法を考える上で重要な LSTM の計算フローと並列処理の軸を図 1 に示す．

図 1　LSTM の計算フロー図

並列処理の方法は以下の 3 方向にそった負荷割り当てが考えられる．

- 同時実行並列方向の負荷割り当て
- パイプライン並列方向の負荷割り当て
- データ並列方向の負荷割り当て

この 3 つの負荷割り当て方法をコア数 50，中間層のニューロン数 200，入力数を 50 とし，同時実行可能な時間の割合 p，1 対多の通信が可能な割合 q，各コアに分割保存可能なデータの割合 r を，負荷の割り当てをしない場合とそれぞれ比較した結果を表 1 に示す．

表 1　各負荷割り当て方法の負荷見積もり

| 負荷割り当て方法 | p | q | r |
|---|---|---|---|
| 同時実行並列 | 99.3% | 5% | 5% |
| パイプライン並列 | 99.3% | 4% | 4% |
| データ並列 | 99.9% | 8% | 8% |

表 1 より，すべての項目でデータ並列が高い負荷割り当てを実現していることが分かる．ここで，q，r が低い値となっているのは，分割数の変化に関係しないデータの割合が多くなるためと考えられる．

次に、1 番性能の高かったデータ並列方向の負荷割り当てに対し，ニューロン数と入力数，コア数を変化させた場合の各コアのローカルメモリの削減割合を算出した結果を図 2 に示す．

図 2　コア 50 個の時のメモリ削減割合

図 2 より LSTM のネットワーク規模が大きくなるにつれて，メモリの削減割合が減少していることが分かる．これは負荷割り当てが分割数に関係しないデータが他のデータに比べて，中間層のニューロン数の総数に大きく影響されるためと考えられる．

## 3　まとめ

負荷の割り当て方法として，データ並列方向の負荷割り当てが 3 つの割り当て方法の中では最も高い性能が得られることが分かった．しかし，ネットワーク規模が大きくなると負荷の削減割合が減少し性能が落ちることが分かった．今回算出した結果は入出力にかかる処理時間などを考慮していない時の結果である．よって今後よ

り精密な見積もりを行い最適な負荷割り当て法を提案し，それに適した相互結合網回路を設計し実装を行う．

## 参考文献

[1] Akane Saito, Yuki Umezaki, and Makoto Iwata, “Hardware Accelerator for Differentiable Neural Computer and Its FPGA Implementation, ” Proceedings of the 2017 International Conference on Parallel and Distributed Processing Techniques and Applications(PDPTA’17), pp. 232-238, July 2017.