LSTM-RNN 用アクセラレータ回路の負荷割当法の検討

1180386　山 﨑　尚 之　【 コンピュータ構成学研究室 】

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の負荷割り当てについて最適な方法の検討を行う．

2　負荷割り当て方法

負荷割り当ての方法は図 1 のフロー図に示す 3 方向の軸に沿った方法が考えられる．

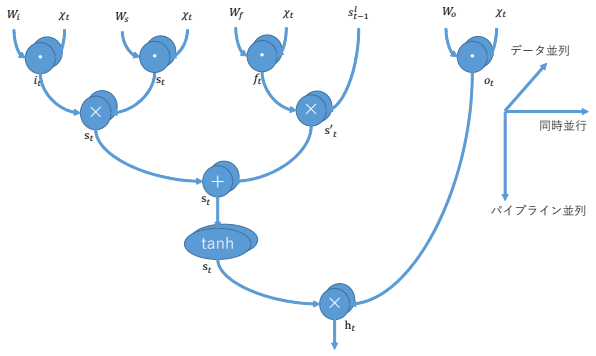

図 1　LSTM の計算フロー図

同時実行並列方向とパイプライン並列方向の負荷割り当ては，図 1 の演算命令を左右と上下にそれぞれ分け，各コアに割り振られた命令を実行する方法である．データ並列方向の負荷割り当ては全命令を各コアで実行するが，各演算に必要な行列やベクトルの要素を半分に分ける方法である．データ並列方向の負荷割当ではデータの重複を限りなく少なくできるが，演算に必要なデータを半分にして演算を行うため，図 1 にはない足し合わせの処理が必要となる．

3　性能の見積もり

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，1 回の演算処理にかかる時間，1 回の通信にかかる時間，1 要素のデータを保存するためのメモリ領域をそれぞれ 1 とし，実行時間と回路面積を算出し比較を行った結果を表 1 に示す．実行時間は演算時間と通信時間を足し合わせた値である．また，性能の指標として実行時間と面積の積をとる．この値が低い値であるほど 1 コアあたりの負荷が低く性能が高いと言える．

結果として，データ並列方向に負荷割り当てを行った場合，二番目に性能の良い同時実行並列方向の割り当てに比べ，18%低負荷での実行が可能であるという見積もり結果が得られた．

また，通信方法として，Bus，Interconnection Network，Crossbar の 3 種類が主な手法として考えられが，どの方法を選択したとしても，同程度負荷を軽減させることが可能であるという結果が得られた．

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	実行時間	回路面積	性能
同時実行並列	202250	5150	$1.24*10^{10}$
パイプライン並列	301650	5150	$1.81*10^{10}$
データ並列	203050	4169	$8.92*10^{9}$

次に，最も性能の高い結果が得られた，データ並列方向の負荷割り当てについて稼働率 (実行命令数/(実行命令数＋ NOP 命令実行数)) を求める．マルチコアで動作させる場合，通信回路が他のコアとの通信に使われていて，通信が行えない時などに，処理を行わない NOP 命令が実行される．この NOP 命令が行われているとき，演算処理は進まないためそのコアは止まった状態となり稼働率が低下する．LSTM を実行した場合，1 タイムステップで 1 コアあたり実行される NOP 命令は〇〇回であり，稼働率は△△%であった．

4　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，実行時間，回路面積の観点から最も性能が高くなるという見積もり結果が得られた．今後，この見積もりをもとに実装を行い，実際の性能評価を行う．

参考文献

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