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

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，音声認識の分野で注目されており，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．

先行研究として，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) 用単一コアの提案 [1] がされているが，単一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．

LSTM に代表されるような大規模で複雑なネットワークは今後更に増えると予想され，これらをマルチコアアクセラレータ回路で動作させると，負荷はネットワーク規模に従って大きくなる．よって，ネットワークに合わせた高効率な負荷割り当て方法を検討することが必要となる．

本研究では，LSTM を一例として取り上げ，マルチコアで動作させる場合の負荷割り当てについて，今後現れる可能性の高い，より大きなネットワーク規模の深層ニューラルネットワーク DNN(Deep Neural Network) にも対応可能な，負荷割り当て方法の検討を行う．

2　負荷割り当て方法

単一 LSTM アクセラレータ回路の構成は，5 段のパイプラインと 2 つのデータメモリ，1 つのアキュムレータを備えた構造となっている．各データメモリから演算に必要な対応するデータを取得することで積和演算や積算を行う．必要に応じて，LUT にあらかじめ登録した活性化関数を参照することで出力を求める．この一連の流れを繰り返し実行することで，LSTM を行うことが可能である．このような機能を備えたコアをマルチコアで，並列処理させる時の負荷割り当て方法について，負荷の割り当てとデータ並列の負荷割り当てを実現するスケジューリングについて検討する．

負荷割り当ての方法は図 1 に示す 3 方向の軸に沿った方法が考えられる．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り当てではデータの重複を最大限少なくできるが，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．

また，データ並列の負荷割り当てを行う際のスケジューリングは，全ての出力が出揃うまでに部分的に次のタイムステップの演算が可能であるという特徴から，演算命令より通信命令を優先して行う．演算命令の途中で通信命令が実行可能になった場合も，演算命令の実行を止め通信命令を優先して実行し，通信命令の実行が終わり次第続きから再開する．また，同じタイミングで複数コアの入力命令が実行可能な状態になった場合は，足し合わせを行うコアのペアで入力の時間差を少なくするため，ペアの片方のコアに入力が行われると次のクロックサイクルでは，ペアのもう片方のコアを優先して入力が行えるようにする．演算命令は始め入力行列が入力されると 4 つの重み行列との積和演算命令すべてを実行可能状態にする．その後は，各演算命令の終了応じて続きの演算命令が実行可能となる．

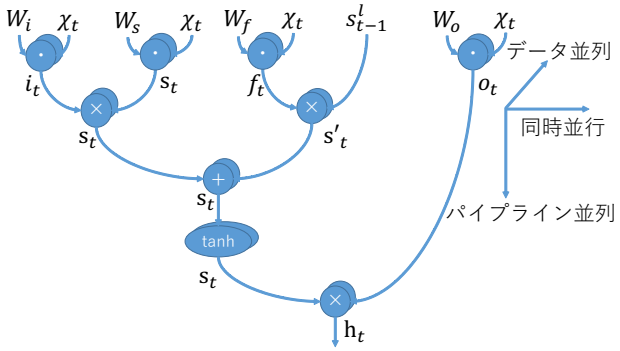

図 1　LSTM の計算フロー図

3　性能の見積もり

コア数 32，中間層のニューロン数 256，入力数 128 とした場合の性能について見積もりを行った．各割り当て方法の実行時間と回路面積を 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 として概算し比較を行った．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，7.2%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．

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，ストールの発生回数を調べ稼働率を求める．

コア数を 32 で固定し中間層のニューロン数を増加させていった際の稼働率の変化を図 2 に示す．ニューロン数の増加に伴い稼働率も上昇していることが見てわかる．これは，ニューロン数を増加させた場合，通信命令より演算命令の増加率が高いため，他コアの通信により

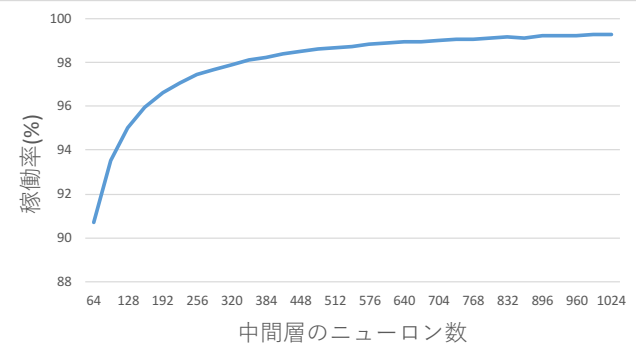

図 2　稼働率

通信網が占有されている時に演算命令を代行できる確率が上がるからであると考えられる．

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．また，稼働率がネットワーク規模の大きさに伴い上昇することから，より複雑で演算回数の多いネットワーク構成の DNN であっても，高稼働率での実行が可能であるといえる．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.