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，音声認識の分野で再帰型ニューラルネットワーク RNN(Recurrent Neural Network) 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．

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る．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，1 コアあたりの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，今後出てくる可能性のある，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 にあらかじめ用意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について検討を行う．

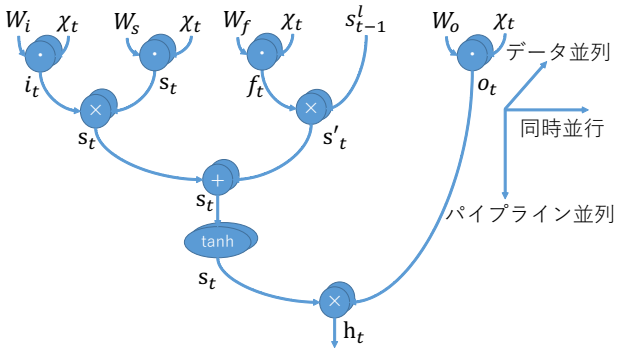

図 1　LSTM の計算フロー図

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当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．

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，NOP 命令の実行回数を調べ稼働率を式 (1) により求める．

$$\text{稼働率} = \frac{\text{実行命令数}}{\text{実行命令数} + NOP\ \text{命令実行数}} \quad (1)$$

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 662 回であり，稼働率は 95%であった．また，コア数を固定し中間層のニューロン数を増加させると，稼働率が上昇することが分かった．

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．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をとる 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.