（西暦）　　２０２４年　１２月　２日

電気・電子情報工学専攻	学籍番号	M213262
申請者氏名	別役 拓哉	

指導教員氏名	市川　周一

論　文　要　旨(修士)

論文題目	確率的計算を用いたGate Recurrent Unitの検討

GRU (Gate Recurrent Unit)は，RNN (Recurrent Neural Network)の一種であり，長期的な依存関係を学習しやすいように改良されたモデルである．主な用途として株価予想や天気予測などの時系列データの処理，翻訳や感情分析などの自然言語処理などに使用される．GRU は計算が複雑であるため，実装時にハードウェア資源や消費電力が大きくなる可能性がある．IoT やモバイル/エッジコンピューティングの発展に伴い，ハードウェア資源や消費電力の制約は大きな問題になっている．

近年，ストカスティック・コンピューティング (Stochastic Computing：SC)がニューラルネットワークの計算に適用され，電力効率の向上が報告されている．主に算術演算 (加算や乗算)を簡素化し，ハードウェアコストと消費電力の削減に成功している．Maor ら(2019) は RTL 記述を用いて RNN の一種である LSTM を SC 化することで消費電力の削減に成功している．しかし，レイテンシーについての評価はされておらず，その実用性は明らかになっていない．

本研究では，C++言語記述と高位合成を用いて GRU の SC 化を行う．SC には 32 bit LFSR を使用し，ユニポーラ形式の実装を行って，LSTM と消費電力やハードウェアコストを比較した．

評価には，MNIST テストデータである手書き文字画像を使用する．Python ライブラリである Keras で学習済の重みとバイアスを保存し，C++に取り込むことで推論を可能とした．Keras と自作 C++コードの精度を比較し，同等の精度であることを確認した．また GRU のユニット数による精度変化も調査した．さらに Vitis HLS と Vivado を使用し，LUT や FF のリソース使用量やレイテンシー，消費電力の計測を行った．比較のため LSTM についても計測を行った．GRU・LSTM 共に最大 97%の精度を達成した．リソース使用量は GRU の方が LSTM より約 2.8%少なく，レイテンシーは LSTM が GRU より 50%ほど少ない．消費電力は，GRU が LSTM より 7%小さい．

次に SC (乗算)と SC(乗算+加算)の 2 種類を GRU に実装し，精度やリソース使用量，レイテンシー，消費電力を調査した．SN ビット長 32～1024 で SN ビット長と精度の関係をした結果，SC (乗算)のみでは SN ビット長が 512 以上，SC (乗算+加算)では SN ビット長が 1024 で GRU と同等の精度になった．リソース使用量は SC (乗算)のみでは 13%の増加，SC (乗算+加算)では 36%の増加となった．レイテンシーは SN ビット長が増加するにつれて増加した．消費電力は SC 乗算のみでは 24%の削減，SC (乗算+加算)は 3%の削減となった． SC (加算)のスケーリング演算に乗算器や除算器が必要であるため SC (乗算)より消費電力が増加する結果となった．電力 C/P (cost/performance)と資源 C/P は本研究の SC 化手法では LSTM・GRU 共に悪くなる結果となった．しかし，すべての演算器を SC 化できていないため GRU が SC と相性が悪いとは断言できない．

今後の課題は，活性化関数の SC 化やハードウェアの最適化，SN ビット長によるレイテンシーの削減，高位合成におけるスケーリング手法，バイポーラ形式の SC 導入の 5 つである．