2A-03

センサ端末上でのニューラルネットワーク処理ハードウェア構成

有川 勇輝† Huy Cu Ngo† 岸野 泰恵‡ 坂本 健†

日本電信電話株式会社 NTT 先端集積デバイス研究所†
日本電信電話株式会社 NTT コミュニケーション科学基礎研究所‡

1. 研究背景

Internet of Things（IoT）技術の進展とともに、様々なセンサデータの収集と利用が可能となった。特に、時系列センサデータをニューラルネットワーク（NN）処理することで、有益な情報が抽出できることが知られている[1, 2]。実ユースケースとして、ゴミ収集車に搭載した加速度センサとジャイロスコープのデータを NN 処理することで、ゴミ量を推定する基本フレームワークが考案されている[3, 4]。

このような NN を用いたセンサデータの解析では、大量の演算が必要になるため、計算リソースが十分確保できるクラウドサーバにて行われる。この場合、通信ネットワークを介して大量の時系列センサデータを送信する必要があり、通信コストが増加する。これに対し、上記フレームワークはセンサ端末上で NN 処理する手法を採用している。図 1 に示すように、本アプローチは、センサ端末上でセンサデータを NN 処理し、データ量の小さな出力を得た後、クラウドサーバへ送信する。

このように、センサ端末上で行列積和演算を大量に行う NN 処理を行うために、センサ端末には高性能な演算処理が求められる。しかしながら、センサ端末上に搭載される汎用プロセッサはシステム制御には適しているものの、NN 処理のような大量の演算には向かない。そのため、本研究では、センサ端末に小型 FPGA（Field Programmable Gate Arrays）[5]を搭載し、NN 処理を FPGA 上に構成した専用回路で実行する。本稿では、センサ端末で NN 処理を行うためのハードウェア構成を提案するとともに、その性能評価結果を議論する。

2. 提案ハードウェア構成

本検討では、時系列センサデータに対して、内部状態を考慮に入れる再帰型ニューラルネットワーク（RNN）を用いた処理を行い、その時刻に発生したイベントを分類する。

図 2 に RNN の基本構成と LSTM（Long short-term memory）[6]構成を、図 3 に RNN 処理ハードウェア構成を示す。システム全体は、Processing System（PS）と Programmable Logic（PL）から構成される。PS はシステム全体の制御を行い、行列演算は PL に構成した専用回路にて行う。図 3（a）に示す積和演算のみを PL で実行する構成では、演算過程で PS から PL へのデータ移動が生じる。そのため、データ移動が高頻度に生じ、処理時間のオーバヘッドとなる。それに対し、図 3（b）に示す提案構成は、PS と

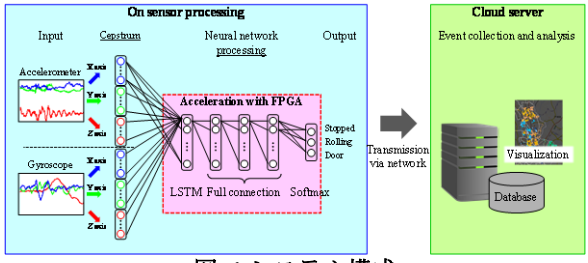


図. 1 システム構成

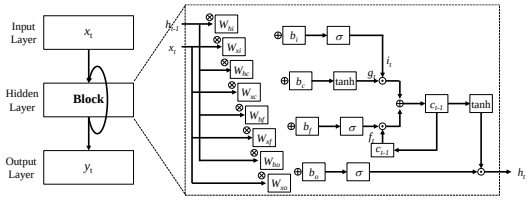
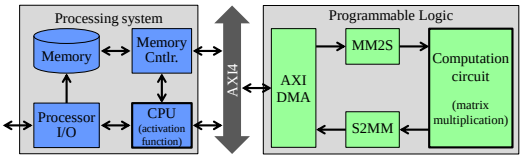
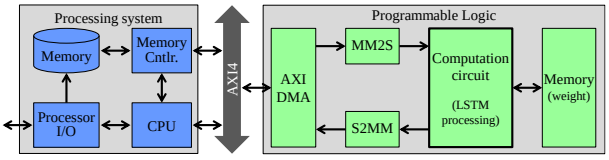


図 2. RNN の基本構成と LSTM



（a）積和演算をPLで処理する構成（従来）



（b）データ移動の頻度を低減した構成（提案）

図 3. RNN 処理ハードウェア構成

Hardware Architecture for On-sensor Neural Network Computing
†NTT Device Technology Labs, NTT Corporation
‡NTT Communication Science Labs, NTT Corporation

表 1. 評価条件、実装結果および性能評価結果

FPGA ボード	FPGA ボード搭載の CPU		FPGA				処理時間		見積り消費電力
	デバイス	クロック周波数	クロック周波数	FPGA リソース使用率			ソフトウェアベース処理 (Baseline)	ハードウェアベース処理*1	
				BRAM	FF	LUT			
ZCU104	Arm Cortex-A53	1.2 GHz	400 MHz	124 (19 %)	73174 (15 %)	147782 (64 %)	322 ms	0.710 ms (x 454)	5.4 W
ZC706	Arm Cortex-A9	1.0 GHz	200 MHz	124 (11 %)	74039 (16 %)	146492 (67 %)	455 ms	1.03 ms (x 441)	3.5 W
Zynq7010	Arm Cortex-A9	866 MHz	200 MHz	120 (Available)	35200 (Available)	17600 (Available)	Approx. 4000 ms	Approx. 500 ms (x 8)	1.8 W

*1: 8 bit 固定小数点（整数部：2 bit、小数部：6 bit）

PL 間のデータ移動頻度を低減することに着目しており、NN の重みパラメータを PL で保持することで、PS から PL への重みパラメータのデータ移動頻度を低減した。提案構成は、PL のメモリリソースを消費するが、データ移動頻度を低減できるため、処理を高速化できる。

3. 評価

提案ハードウェア構成を市販 FPGA ボード(Xilinx ZCU104、ZC706)と、Xilinx Zynq7010 を搭載したカスタム FPGA ボードに実装し、処理性能を評価した。表 1 に評価条件、実装結果および性能評価結果を示す。ハードウェアベース処理で用いる演算精度は、8bit 固定小数点とした。[3, 4]に示す実ユースケースを用いた評価では、32bit 浮動小数点と比較し、推論精度の劣化は 2%程度であった。比較として、FPGA に搭載されている CPU を用いてソフトウェアベース処理を行った場合の処理時間を測定した。消費電力は設計ツール（Xilinx SDSoC）を用いて見積った。

表 1 に示すように、推論精度の劣化を 2%程度に抑えつつも、提案ハードウェア構成はソフトウェアベースの処理に対して、最大 454 倍程度の高速化を確認した。消費電力の観点では、より小型な FPGA の利用やクロック周波数を 200 MHz に抑えることが望ましい。消費電力の増加が許容できるのであれば、クロック周波数を 400 MHz に上げることで、NN 処理を 1 ミリ秒以下に抑えられる。以上より、提案ハードウェア構成を用いることで、数ワット程度の消費電力で 1 ミリ秒以下の処理時間で時系列センサデータを処理できることを示した。今後、PL に構成する専用回路の並列度やデータ再利用などデータ移動の更なる効率化を図ることで、低電力化・高性能化を目指す。

4. まとめ

本稿では、センサ端末上で NN 処理を行うためのハードウェア構成を議論した。PS と PL 間のデータ移動を低減するハードウェア構成を提案し、数ワット程度の消費電力および 1 ミリ秒以下の処理時間で時系列センサデータを処理できることを示した。

参考文献

[1]M. Mohammadi, A. Al-Fuqaha, S. Sorour and M. Guizani. 2018. Deep Learning for IoT Big Data and Streaming Analytics: A Survey. IEEE Communications Surveys & Tutorials 20, 4 (Fourthquarter 2018), 2923-2960.

[2]Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, Lisha Hu. 2019. Deep learning for sensor-based activity recognition: A survey. Pattern Recognition Letters 119, (February 2018), 3-11.

[3]Yasue Kishino, Yoshinari Shirai, Koh Takeuchi, Futoshi Naya, Naonori Ueda, Yin Chen, Takuro Yonezawa, Jin Nakazawa. 2017. Detecting Garbage Collection Duration Using Motion Sensors Mounted on Garbage Trucks Toward SmartWaste Management. In Proceedings of the Third International Conference on Smart Portable, Wearable, Implantable and Disability-oriented Devices and Systems (SPWID 2017). Venice, Italy, 1-4.

[4]Yasue Kishino, Koh Takeuchi, Yoshinari Shirai, Futoshi Naya and Naonori Ueda. 2017. Datafying city: Detecting and accumulating spatio-temporal events by vehicle-mounted sensors. In Proceedings of IEEE International Conference on Big Data (Big Data 2017). Boston, MA, 4098-4104.

[5]Xilinx. 2016. Zynq UltraScale+ MPSoC Product Brief. https://www.xilinx.com/support/documentation/product-briefs/zynq-ultrascale-plus-product-brief.pdf

[6]Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural Computation 9, 8, 1735-1780.