様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月 １５ 日現在

機関番号：３６１０２
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ０６４０５

研究課題名（和文）メモリスタ技術を用いた組込み用低電力ニューラルネットワーク・アーキテクチャの研究

研究課題名（英文）Research on low power neural network architecture using memristor technology for embedded systems

研究代表者
河合　浩行（KAWAI, Hiroyuki）

徳島文理大学・理工学部・教授

研究者番号：２０６４３１５９

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：情報を抵抗値として保持する記憶素子を用いた組込み用低電力ニューラルネットワークのハードウエアプラットフォーム基盤技術の確立に向けた技術について検討した。本研究では、不揮発素子を用いたニューラルネットワーク・アーキテクチャ及びそれに適した学習手法と低電力化手法を提案した。また、提案技術を評価環境上に疑似的に構築し、アプリケーションに適用して提案技術の効果を検証した結果、全結合ニューラルネットワーク部ノードのうち53%を削減でき、同じく重み個数のうち65.1%に当たる個数を0値化できた。本結果より、提案技術が組込み用ニューラルネットワークの小型化・低電力化に有用であることを確認できた。

研究成果の学術的意義や社会的意義
本研究の成果は、身近な製品システムに搭載可能な小型・低電力なニューラルネットワークを実装可能とする基盤技術である。今回提案した技術を用いることで、消費電力およびコストへの要求が厳しい製品システムにも、ニューラルネットワークを実装可能としうるものである。これにより、センサ情報を用いてユーザーの使用状況を学びとり、個々のユーザーに適した機能を提供しうる製品システムの開発に寄与することが期待できる。

研究成果の概要（英文）：We studied technologies for establishing a hardware platform of embedded neural networks with low-power consumption. In this research, the neural network is characterized by use of a non-volatile memory device that holds information as resistance values. We proposed a low-power neural network architecture using the non-volatile memory, and a learning scheme and a power reduction technique suitable for it. The proposed technologies were applied to a sample application. The results of this evaluation show that 53% of fully-connected neural network nodes was reduced.Also, 65.1 % of the weight parameters in the network were replaced into zero. Therefore, the proposed technologies is useful for the reduction of both size and the power consumption of the embedded neural network.

研究分野：電気電子工学

キーワード：ニューラルネットワーク　ノーマリーオフコンピューティング　学習

１．研究開始当初の背景

IoT(Internet of Things)の普及が進んでおり、今後多種多様なセンシングデータを手に入れる環境が整うと期待される。一方、多種多様なビッグデータから有為な知見を得る方法として機械学習が脚光を浴びている。同時に、深層学習の処理時間を短縮すべく、ハードウエアプラットフォームとしてのニューラルネットワークの検討も行われている[1][2]。現状の研究は記憶素子に SRAM を用い、かつメモリ容量とデータ転送帯域を低減するために、応用分野を画像等に限定して情報データ精度を犠牲にしている。今後コグニティブ・コンピューティング技術が普及し身近な有用なものとなるためには、高い汎用性を備え、消費電力の少ないニューラルネットワークのハードウエアプラットフォームが必要である。

２．研究の目的

本研究は、メモリスタを用いた組み込み用低電力ニューラルネットワークのハードウエアプラットフォーム基盤技術の確立を目的とする。本ハードウエアプラットフォーム基盤技術の確立に向けて、本研究では、組み込み用低電力ニューラルネットワーク・アーキテクチャ、ノーマリーオフ・コンピューティング技術応用による低電力化技術を検討する。さらに、提案アーキテクチャに適した学習手法・学習結果格納方法を検討する。また、提案技術を評価環境上に疑似的に構築し、アプリケーションに適用した場合の効果を検証する。

３．研究の方法

本研究では、メモリスタ対応組込み用ニューラルネットワークのハードウエアプラットフォーム基盤技術として、

(1) 組込み用ニューラルネットワーク・アーキテクチャ
(2) 学習手法
(3) 低電力化技術

について、研究代表者と研究分担者で分担・併行して研究を進める。アーキテクチャ検討のために必要となるデバイス情報については研究協力者から提供を受けて本研究を進める。

４．研究成果

(1) 組込み用ニューラルネットワーク・アーキテクチャ

通過電流履歴を抵抗値として保持するメモリスタ技術を使いニューラルネットワークの基本素子回路を実現する。具体的には、学習結果情報であるシナプス結合ごとの結合重み情報を従来のようにデジタル値で保持せずにメモリスタの抵抗値として書き込み／保持する。図 1 にメモリスタを用いたニューロンの基本回路案を示す。また、メモリスタベースの組込み用ニューラルネットワークのアーキテクチャを図 2 に示す。本提案ニューラルネットワーク回路では、重み情報はメモリスタデバイスの抵抗値として記憶する。ニューロン出力は、閾値を使いコンパレータで生成する。

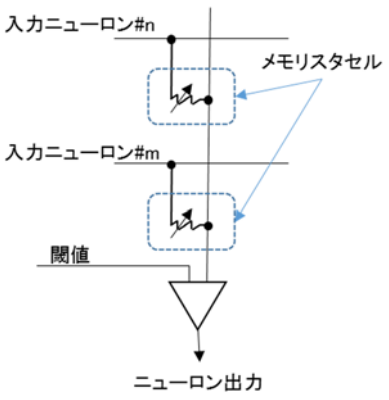

図 1　ニューロン基本回路

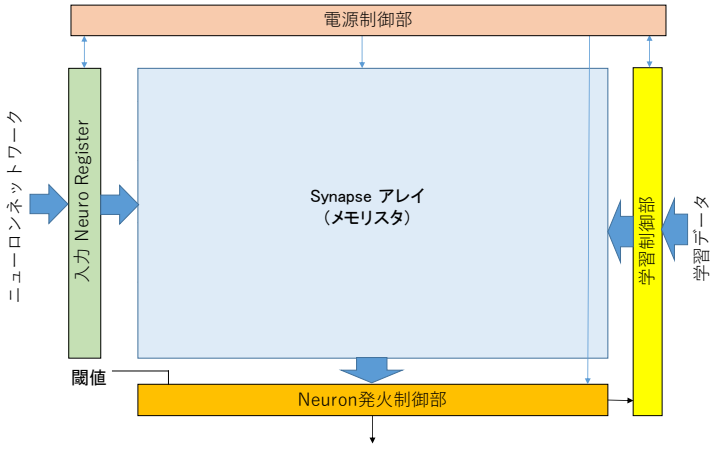

図 2　ニューラルネットワークアーキテクチャ

(2) 学習手法

組込み応用では、搭載可能なメモリ容量に制約がある。本研究では、メモリスタを用いた場

合に予想される精度起因の認識率低下抑止方法を検討した。また、格納するメモリ容量を削減する手法として、重み情報の 0 値化手法として分散補正型面積相殺手法を提案した。さらに、非 0 重みをも持つノード削除手法として、出力寄与度を用いたニューロン削除手法を提案し、その有効性を調べた。

本研究における学習方法検討では、冗長性を持たせた全結合 3 層ニューラルネットワーク（入力層：784、隠れ層 1：1000 ノード、隠れ層 2：1000 ノード、出力層：10 ノード）を用いて技術検討を行った。学習・テストデータには手書き文字認識データベースである MNIST を使用した。

①量子化影響低減手法

図 1 のメモリスタベース基本回路では、学習済み重み値を抵抗値として記憶する際の精度が課題となる。そこで、疑似的に活性化関数の出力を量子化し、精度が認識率に与える影響をシミュレーションにより評価した。図 3 には、活性化関数 Tanh、Sigmoid と ReLU の各場合において、活性化関数出力の精度（区間数）を 3～15 まで変えた場合に得られるテストデータ認識率を示している。なお、ReLU1 は隠れ層毎にそれぞれの最大値で区間を決めた場合の結果であり、ReLU2 は両隠れ層の最大値で両層の区間を決めた場合の結果である。本結果から、特に ReLU は Tanh と Sigmoid に比べて精度の影響を受けやすいことが確認できた。また、ReLU 使用時には、量子化により生じる認識率低下には隠れ層毎の活性化関数のダイナミックレンジを考慮した量子化が有効である。

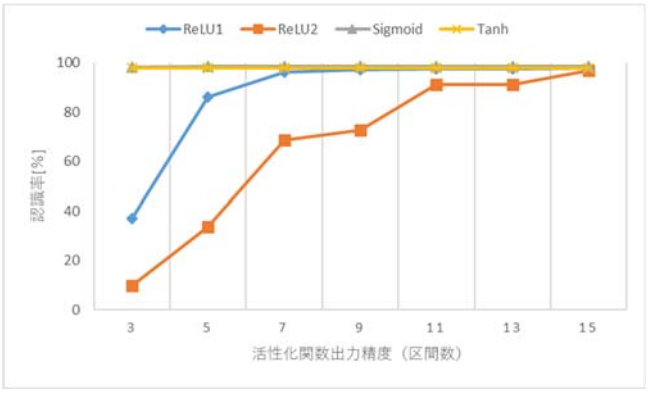

図 3　活性化関数毎の量子化影響

②分散補正型面積相殺手法[3]

前処理として、学習後重み値のうち閾値範囲に含まれる 0 近傍重み値を 0 値化する（重み値 A）。閾値が±0.15 の場合、認識率は 96.7%、0 値化重み数は 1.52M 個であった。さらに重み情報削減（0 値化）を行う手法として、図 4 に示す重み面積相殺手法を提案した。本手法では、各ノードが持つ重み値について正負側各々で合計を求め、それらの絶対値を比較し、小さい側の重み値をすべて 0 値化する。他方、大きい側の重み値については、小さい側の総和を超えない範囲で絶対値の小さい側から 0 値化する（重み値 B）。この面積相殺は、大幅な重み値数削減効果を期待できるが、認識率低下を引き起こしかねない。そこで、重み値の重要度判別指標として、各ノードについて重み値 A 時の積和演算値から重み値 B 時の値を引いたものの分散を用いる。このノード毎の分散値が閾値を超えていた場合、重み値 B を破棄して重み値 A に戻すことで認識率低下を回避する。

図 5 に、隠れ 1 層の積和演算結果の分散分布を示す。また、図 6 に本手法を適用した結果を示す。面積相殺のみを行った場合の認識率（10.32%）に比べて、分散補正を行った結果 95.06%まで回復し、前処理に比べて 99K 個多く重みを 0 値化でき、本手法が有効であることを確認できた。

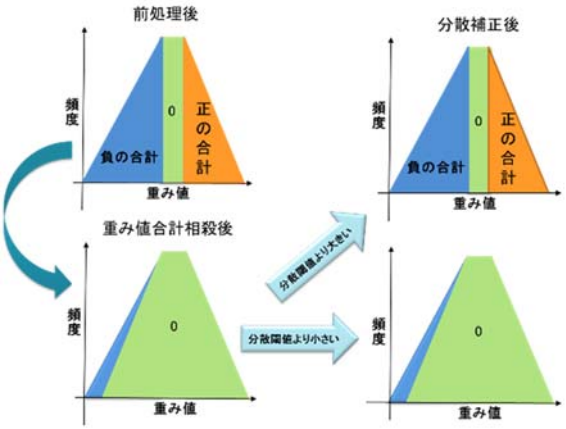

図 4　分散補正型面積相殺手法

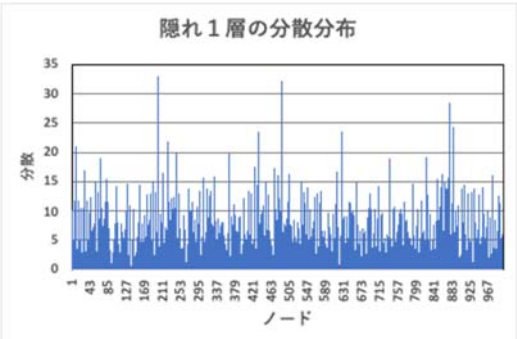

図 5　積和演算結果の分散分布

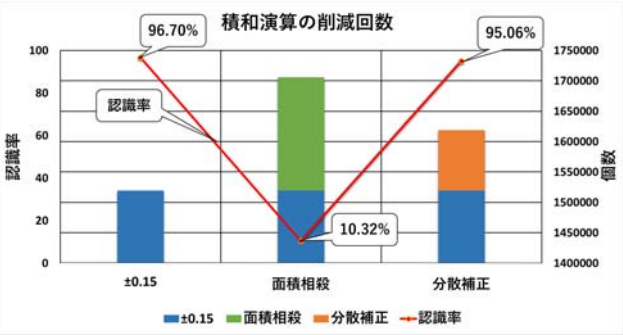

図 6　分散補正適用効果

③出力寄与度を用いたノード削除手法[4]

今回、非 0 値重みをもつノード削除手法として、出力寄与度を用いたノード削除手法を提案し有効性を評価した。本ノード削除手法では、式（1）を用いて各ノード出力の次段ノード出力に対する寄与度C_{ij}を定義した。

$$C_{ij}=\frac{W_{ij}}{\sum_{i=1}^{I} W_{ij}+b_i} \qquad (1)$$

ここで、W_{ij}は第 k 層第 j 番目ノード重み値、b_iはバイアス項である。まず、寄与度閾値より小さい寄与度の重み値を 0 に置換する。次に、図 7 のノード A 出力に対応した重み値が別途設定する閾値以上の 0 個数であれば、ノード A は出力寄与度が低いと判断して削除する。この際、ノード A が持っていた前層からの入力に対する重み値も全て 0 にする。順次入力層に向かって 0 値化効果を伝搬させながら同様の手順でノードの削除を行う。効果検証のために、目標認識率を 95%とし、全層同じ寄与度閾値を用いてノード削除効果を検証した。学習後の認識率は 98.27%であった。寄与度閾値は±0.5 とした（表 1）。隠れ層 2 はノードに対応した重み値の 0 個数が 10 個、隠れ層 1 は 680 個、入力層は 868 個をノード削除閾値とした場合、計 1465 個（全体の 52.43%）のノードを削減でき、全体の 80.72%の積和演算回数を削減できた。このとき、認識率は 95.02%であった。本結果より、本手法はメモリ容量低減に有効であることが確認できた。

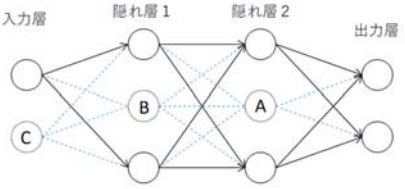

図 7　ノード削除手法

表 1　削除ノード数と 0 の個数

	±0.5	
各層	削除ノード数	0値化した個数
出力層	0	9080
隠れ層 2	417	731703
隠れ層 1	777	707405
入力層	271	0
合計	1465	1448188

（3）低電力化技術

本研究では、重み情報保持手段として不揮発メモリを用いることを前提としている。そこで、低電力化手法として、ノーマリーオフ型メモリ電源制御について検討した。不揮発性素子の場合、電源 ON/OFF に伴うデータ再設定時間オーバーヘッドが不要である。この特徴を活かし、ノーマリーオフ型電源制御により、非動作時にメモリブロックの電源を OFF することにより、スタンバイ電力削減を図る。電源制御方式の概要を図 8 に示す。さらに、動作時電力低減手法として、メモリ電源分割設計による冗長ブロック電源 OFF を行う。上記（2）で提案した手法を適用することにより、重み行列のスパース化および出力寄与度の優先度判定が可能となる。したがって、(2) の手法とメモリブロック単位のノーマリーオフ型電源制御を組み合わせることにより、冗長ブロックへの電力供給 OFF を実現でき、電力低減が可能である。

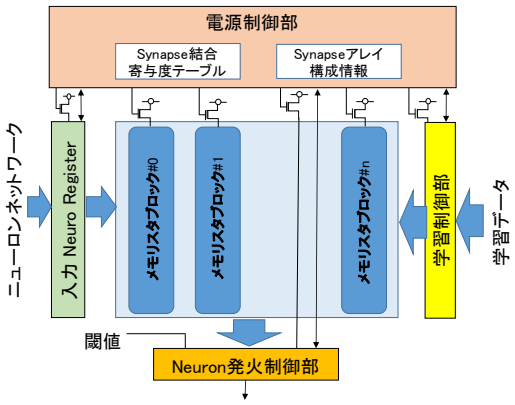

図 8　ノーマリーオフ型電源制御

（4）統合評価

提案技術の統合評価として、MNIST を用いた手書き文字認識と異なるアプリケーションを用いて、提案手法の効果検証を行った。統合評価では、MEMS 触覚センサを用いて、紙の種類の識別用の深層学習ネットワークに（2）の提案手法を適用し、その効果について考察した。今回の検討では、6 種類の紙の識別用深層学習ネットワークを対象とした。表面凹凸および摩擦の各 1000 点の電圧データを組み合わせた 2000 点を入力とし、convolution 層の後に 64 ノードの全結合層を持つ深層学習ネットワークの構成を対象とした。識別数を 6（紙の種類）とし、Softmax 関数により出力が決定される。convolution 層の出力数は 1664 であるため、全結合層（入力側）は

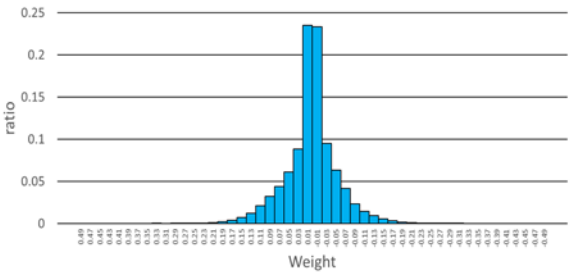

図 9　1000epoch 学習後の全結合層の重み分布

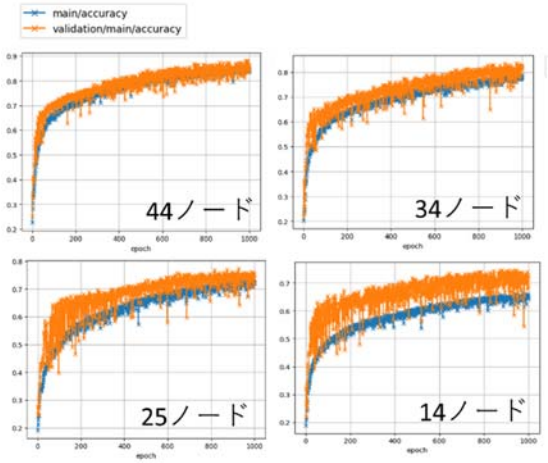

図 10 -0.04～0.04 の範囲の重みを 0 値化した場合の学習曲線

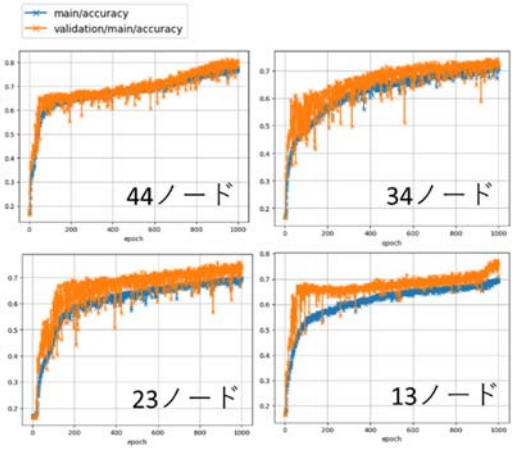

図 11 最初から少ないノード数で学習した場合の学習曲線

1664x64 個の波形情報を持つことになる。今回提案した手法により、1000epoch 学習後のネットワークに対して(2)の手法を適用した。1000epoch 学習後の全結合層の重み分布を図 9 に示す。今回は１）－0.02～0.02、2）－0.04～0.04、3）－0.06～0.06、4）－0.08～0.08 の 4 種類の範囲の重みを 0 にして検討を行った。なお、この処理により 0 になる重みの数の全体数に対する割合は、1)46.9%、2)65.1%、3)77.6%、4)86.1% であった。重みの 0 値化後、全結合層の各ノードについて、重み 0 の入力数が一定数以上のノードを削除する。全結合数 64 のうち、削除ノード数をおおよそ 20、30、40、50 とした 4 パターンについて検討した。同様に、残ったノード数は 44, 34, 25, 14 となった。これらの構成による学習曲線を図 10 に示す。44 ノードの時に識別率 0.8 を超えた。また、34 ノードの時も識別率が 0.8 近くになった。今回提案手法の意義を確認するために、最初からノード数を削減した状態で 1000epoch 学習させた場合の結果を図 11 に示す。44 ノードのとき、識別率は 0.8 程度となり、これは図 10 の 44 ノードの結果とほぼ同程度といえる。一方、34 ノードの時は 0.7 となり、これは図 10 の 34 ノードでの 0.8 に及ばない。以上のことから、元々20 ノード程度は本質的に冗長であったため、0 値化する重み範囲によらず識別率は 0.8 程度となった。一方、さらに 10 ノード削減する場合、識別率を高くするには、0 値化範囲最適化が必要であり、今回提案手法が、所望のメモリ容量に合わせてネットワーク規模縮小する上で有効であることを確認できた。

メモリスタに使われる Emerging 系のメモリは未だ開発途上にあり、デバイスばらつきの影響で粒度が粗い傾向にある（4~8bits 程度）。今回提案の手法は、推論精度の要求が比較的緩いマルチモーダルセンサー応用のアプリケーション(粗粒度画像認識等)に適しており、本研究で目指すメモリスタへの適用に望ましい手法と思われる。学習方法としては、事前学習を実施し、本質的に冗長であったノード数を抽出して、本番深層学習モデルを Fix させることが望ましいことが分かる。本研究で提案する（2）の学習法は現時点、完全に自動化できていないが、完全自動化することで、従来に比べて大幅にコンパクトな学習モデルの設計が可能となり、サーバ等での事前学習とエッジでの本番学習のフロー構築の可能性を示すことができた。また、この事前-本番学習フロー適用方式では、このフローを介して、学習モデルをハードウエア化する場合、粒度の大きなハードウエア構成（メモリ型ニューラルネットワークやリコンフィギャラブルロジック型ニューラルネットワーク等）への適用が可能となる。

＜引用文献＞

[1] 中原啓貴他,” メモリベースに基づく 2 値化深層畳込みニューラルネットワークの実現,“ 信学技報, vol.116, no.210, RECONF2016-37, pp.63-68 (2016.9)
[2] Filipp Akopyan 他, “True North: Design and Tool Flow of a 65mW 1Million Neuron Programmable Neurosynaptic Chip,” IEEE Transactions on Computer-Aided design of integrated circuits and systems, Vol.34, No.10, pp1537-1557, Oct. 2015.
[3] 木村太一,三好寛太,河合浩行,”分散を用いたニューラルネットワークの積和演算回数削減手法の検討“, 令和元年度電気関係学会四国支部連合大会 講演論文集, 1-3(2019.9)
[4] 沖雄太,久保孝敦、河合浩行, “組込み用ニューラルネットワークのバックワード型ノード削減手法に関する検討”, 令和元年度電気関係学会四国支部連合大会 講演論文集, 1-1(2019.9)

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 M. Hayashikoshi, H. Noda, H. Kawai, Y. Murai, S. Otani, K. Nii, Y. Matsuda, H. Kondo	４．巻 4
２．論文標題 Low-Power Multi-Sensor System with Power Management and Nonvolatile Memory Access Control for IoT Applications	５．発行年 2018年
３．雑誌名 IEEE Transactions on Multi-Scale Computing Systems	６．最初と最後の頁 784-792
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計9件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 沖雄太、久保孝敦、河合浩行
２．発表標題 組込み用ニューラルネットワークのバックワード型ノード削減手法に関する検討
３．学会等名 令和元年度電気関係学会四国支部連合大会
４．発表年 2019年

１．発表者名 三好寛太、木村太一、河合浩行
２．発表標題 組込み用ニューラルネットワークの低電力化に向けた検討
３．学会等名 令和元年度電気関係学会四国支部連合大会
４．発表年 2019年

１．発表者名 木村太一、三好寛太、河合浩行
２．発表標題 分散を用いたニューラルネットワークの積和演算回数削減手法の検討
３．学会等名 令和元年度電気関係学会四国支部連合大会
４．発表年 2019年

１．発表者名
久保孝敦、沖雄太、河合浩行

２．発表標題
組込み用ニューラルネットワークのフォワード型ノード削減に関する検討

３．学会等名
令和元年度電気関係学会四国支部連合大会

４．発表年
2019年

１．発表者名
M. Hayashikoshi

２．発表標題
Embedded memory solutions for AI, ML and IoT

３．学会等名
ISSCC2019（国際学会）

４．発表年
2019年

１．発表者名
有本和民、藤井知、山内直樹、木下研作、吉川憲昭

２．発表標題
ドローン高度適応型高精度着陸システム

３．学会等名
電子情報通信学会　無線通信システム研究会

４．発表年
2018年

１．発表者名
M.Hayashikoshi, H.Noda, H.Kawai, K.Nii, H.Kondo

２．発表標題
Low-Power Multi-Sensor System with Task Scheduling and Autonomous Standby Mode Transition Control for IoT Applications

３．学会等名
2017 IEEE Symposium in Low-Power and High-Speed Chips(COOL CHIPS)（国際学会）

４．発表年
2017年

１．発表者名 坂村 賢士，有本 和民，茅野 功，横川 智教
２．発表標題 バッテリ駆動ノーマリオフコンピューティングシステムにおける高効率エネルギ供給の検討
３．学会等名 平成 29 年度(第 68 回)電気・情報関連学会中国支部連合大会
４．発表年 2017年

１．発表者名 坂村賢士，有本和民，茅野 功，横川智教
２．発表標題 スタック回路を用いたノーマリオフコンピューティングの検討
３．学会等名 電子情報通信学会　VLD研究会
４．発表年 2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	有本　和民 (ARIMOTO Kazutami) (10501223)	岡山県立大学・情報工学部・教授 (25301)	
研究協力者	林越　正紀 (HAYASHIKOSHI Masanori)		