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　６ 年　 ６ 月 １７ 日現在

機関番号：１４６０３
研究種目：基盤研究(C)（一般）
研究期間：2021～2023
課題番号：２１Ｋ１１８０９

研究課題名（和文）Time-Space Re-configurable Flash Computations

研究課題名（英文）Time-Space Re-configurable Flash Computations

研究代表者
ＺＨＡＮＧ　Ｒｅｎｙｕａｎ（ZHANG, Renyuan）

奈良先端科学技術大学院大学・先端科学技術研究科・准教授

研究者番号：００７０９１３１

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）：本研究では、時間または空間領域で再構成可能な近似計算基盤を開発する。提案技術を用いて、ハードウェア（HW）コストを大幅に削減し、適切な計算精度を実現する。時間的な再構成可能なANNのために、独創な"DiaNet "に基づくニューロモーフィック計算基盤を提案し、検証を行った。様々な検証の結果、提案したアーキテクチャは、従来と同等のサービス品質で、HWリソースの使用量を最大95%削減することができた。空間的に再構成のために、非同期確率計算（ASC）手法を提案し、実装し、様々な算術計算により検証した。ASC回路は、同程度の精度で、ハードウェア効率と速度において同期型SCより優位性を確認した。

研究成果の学術的意義や社会的意義
The technologies developed in this project are found as promising candidates of post-Moore soft computing trends for accelerating the artificial intelligence tasks. This work explores the up limit of approximate computing and reasonable scenarios for it by cutting off a great processing energy.

研究成果の概要（英文）：The approximate computing architectures are developed in this project, which are re-configurable in temporal or spatial domain. By using the proposed technologies, the hardware (HW) costs are greatly reduced with reasonable computing accuracy. For temporal re-configurability, a series of neuromorphic computing platforms on the basis of our original topology named “DiaNet” are proposed and verified for artificial neural networks (ANNs). From various validations, the proposed architectures reduce the use of HW resources up to 95% with similar quality of service as conventional works. For spatial reconfigurable computing architectures, the asynchronous stochastic computing (ASC) methodology is proposed, implemented, and validated by various arithmetic calculations. The ASC circuits are found superior to synchronous SC on hardware efficiency and speed with similar accuracy. Moreover, the ASC platform offers rich re-configurability to trade off performance and cost post silicon.

研究分野：高性能近似計算技術

キーワード：approximate computing　Neuromorphic circuits　stochastic computing　low power　artificial intelligence

１．研究開始当初の背景

AI のアプリケーション側とコンピューター・アーキテクチャ側におけるさまざまなトレンドの間にはギャップがある。このプロジェクトはこのギャップを埋めるものではなく、両端をカバーする全く異なるソリューションを提供するものである。

(1) 【応用動向】最近人気の AI チップ製品は、すべてスパイク符号化またはスパイク・ニューラル・ネットワーク（SNN）を搭載していることがわかった。例えば、Loihi@Intel、Akida@BrainChip、Tianjic@Tsinghua Univ.などである。 バイナリの世界から脱却したスパイク実装は、超低消費電力の特徴を備えている。しかし、NN の実装では、その応用分野は限られているように思われる。そのギャップは、NN をユビキタスコンピューティングアーキテクチャにどのように拡張するかということである。もう一つのトレンドは統計に基づく計算である。有名な例のひとつに、推論ラベルだけでなく信頼確率も提供するベイズ NN の成長がある。数十年前、コンピュータアーキテクチャーの重要な課題は、高密度／疎なテンソル計算への対応であったが、これは GPGPU や TPU によって解決された。次の課題は、上記 2 つのトレンドにいかに対応し、同時に超高速（1 サイクル）を実現するかである。

(2) 【ハードウェア実装動向】最近、ディープ CNN のような AI アルゴリズムの実行に並列コンピューティングユニットを適用することが流行している。リコンフィギュラブル・アーキテクチャ（CGRA、特にシストリック・リング : IMAX@NAIST 参照）の多くは、ALU やメモリを特定の機能に整理することで、要求されるタスクにできるだけ適合させようとしている。しかし、ALU の心臓部（機能、構造、オペランド#、容量）は常に一定である。早稲田大学の多階層アーキテクチャ OSCAR でさえ、ALU 内部の柔軟性を提供できていない。従って、本提案が追求する、任意のタスクを完全に並列化することは困難である。コンピュータの性能を向上させるために、新しい材料や物理の基礎が広く研究されている。例えば、低消費電力を目指した薄膜 IGZO 半導体（京大）、BRein の電源に利用されたメモリスター（北大）、ボディエリア SoC に開発された有機 IC（東大）、パターン認識に利用されたスピントルク発振器（東北大）などである。 これらの技術は、ポスト・ムーア時代の多くの問題を何とか解決しているが、特定の物理デバイスに依存しない、設計者に優しいアーキテクチャの誕生が待ち望まれている。

２．研究の目的

新しいコンピューティング・アーキテクチャの開発は、図 1 に示すように、物理学と数学の両側からプレッシャーを受けている。主なユーザーである AI アプリケーションは、確率論的（ベイズモデルなど）とスパイク駆動型（スパイク・ニューラル・ネットワークなど）の 2 つの傾向を示している。もう一方では、フロンティア材料とデバイスがハードウェアの技術革新を推し進める。フロントエンド／バックエンドの進化は、アーキテクチャ層の取り組みを強くリード／推進する。し

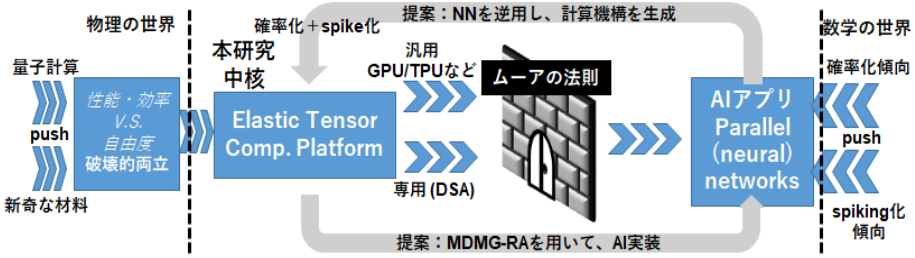

図 1 物理学と数学からの圧力。(MDMG-RA= マルチドメイン・マルチグレイン・リコンフィギュラブル加速器)

かし、我々の専門知識はそのどちらにも近くないため、アーキテクチャ設計者が両方の圧力を素早く解放することは難しい。ユビキタス・プラットフォームは、「再構成性」を最大化するために必要である。このプロジェクトの目的は、時間的・空間的に再構成できるテンソル計算ソリューションを提供することである。同時に、スパイクや確率的な解釈にも対応する。中核となる技術は、①マルチドメイン・マルチグレイン再構成可能ネットワーク・オン・チップ(NoC)の構築、②この巨大で再構成できる NoC を任意のテンソル計算アレイに切断・分割、③この NoC を DiaNet と名付けた新しいバイセクション・ニューラル・ネットワークで組織化、④必要な計算はすべて「フラッシュ・コンピューティング」と名付けた革新的な基本技術で行うが、任意に時間・精度制約を加えることができる。

３．研究の方法

本研究の全体像では、DiaNet（空間的再構成性に対応）とフラッシュ・コンピューティング（時間的再構成性に対応）の 2 つの重要な技術を統合したニューラルネットワークの逆利用によって実施される。

(1)［空間的再構成］DiaNet の進化

図 2 に示すように、全並列 NN のバイセクショントポロジーを考案した。この方法では、チップ上のネットワーク全体が、冗長性のない菱形の小片（DiaNet として知られる）に分割され、再構成される。DiaNet は、従来のフルコネクション・ニューラルネットワーク（FC-NN）の動作を移行する。例えば、ALU 関数はコンパクトな DiaNet によって回帰によって検索することができる。ベクトル計算やパターン認識など、より複雑なタスクも実現可能である。つまり、計算コアを、細粒度（関数）、中粒度（特徴量/オペランド数）、粗粒度（コアの構成）、さらには大域的な粒度（NN 全体のマッピング）に、ポストシリコンで再構成することが可能である。 ダイヤネットのトポロジーはウルトラスパース NN であるため、各処理要素（PE）における計算は非常に単純で、どこでも対称的である。したがって、データ表現と計算の基本は、理論的にはアーキテクチャとは無関係である。従来のデジタル、アナログ、スパイク、確率、量子など、どのようなメカニズムも適用可能である。ダイヤネットの本質は LEGO 遊びである。問題は、LEGO のピース（ダイヤネットでは PE）をどのように設計し、それらのピースを組み立てるかである。私たちはいくつかのバージョンのダイヤネットを開発したが、どのようなトポロジーが広範なアプリケーションに最適かという一般的な結論にはまだほど遠いものであった。テンソル計算やパターン認識のような複雑なタスクでは、コストが大幅に上昇し、DiaNet のメリットを食いつぶしてしまう。また、高度なシリーズになると効率性の問題も出てくる。実用的なプロトタイプの作成が期待される。分割された計算コアをスパイクまたはストキャスティクスで実装する。

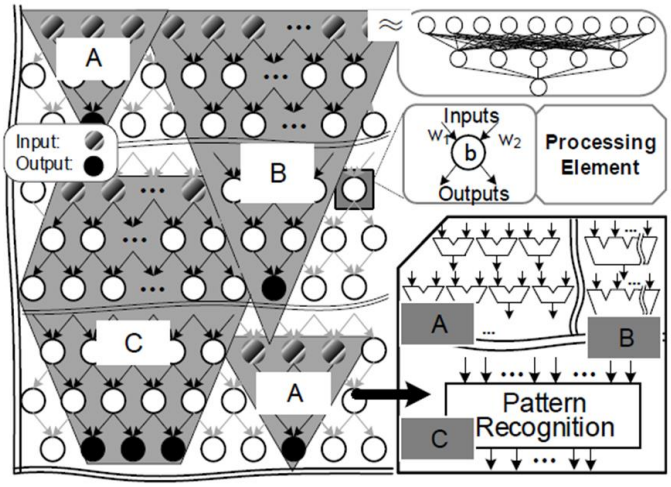

図 2 DiaNet トポロジーとそれによるテンソル・コンピュータ

(2)［時間的再構成］ 高効率 stochastic 計算＝Flash computing

フラッシュ・コンピューティングの概念を図３に示す。フラッシュ・コンピューティングは、新しい符号化機構とデータ表現によって動作を実現する。ストキャスティック・コンピューティング（SC）やスパイク・ニューラル・ネットワークとは異なり、データは任意のタイミングにおけるスパイクの確率によって運ばれる。最も単純な論理ゲートとキャパシタだけで、ニューラルネットワーク全体を実行することができる。スナップショットは高速推論を提供し、累積はベイズのような統計機能を提供する。この意味で、操作は超高速かつ効率的に行うことができる。一方、精度と速度は自由に調整可能である。ストキャスティック・コンピューティングやスパイク・ニューラル・ネットワークなど、近隣には多くの証拠があるが、スナップショット符号化技術は存在しない。ゼロからのスタートである。現在のところ、シリコンのアナログやデジタルの努力でスナップショットの確率的動作をエミュレートすることは可能である。おもちゃのような NN も、フラッシュ・コンピューティングのメカニズムによって検証された。しかし、より深く研究する前に、基礎的な理論を確立する必要がある。このプロジェクトでは、テンソル計算のための確率計算回路の設計と検証を行っている。

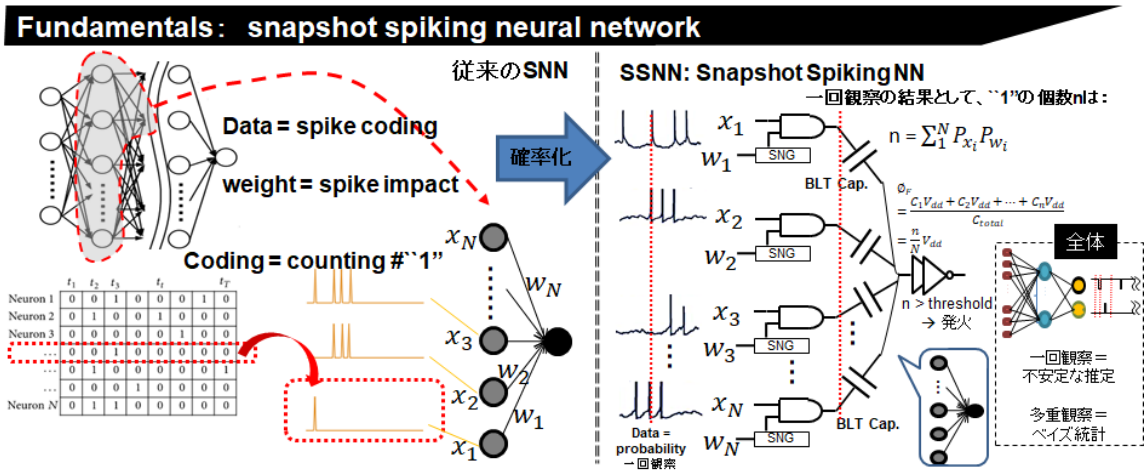

図 3 Flash computing の概念

４． 研究成果

本研究では、時間または空間領域で再構成可能な近似計算基盤を開発する。提案技術を用いて、ハードウェア（HW）コストを大幅に削減し、適切な計算精度を実現する。時間的な再構成可能な ANN のために、独創的"DiaNet "に基づくニューロモーフィック計算基盤を提案し、検証を行った。様々な検証の結果、提案したアーキテクチャは、従来と同等のサービス品質で、HW リソースの使用量を最大 95%削減することができた。空間的に再構成のために、非同期確率計算（ASC）手法を提案し、実装し、様々な算術計算により検証した。ASC 回路は、同程度の精度で、ハードウェア効率と速度において同期型 SC より優位性を確認した。

(1)［空間的再構成］ DiaNet の進化

DiaNet1.0 のプロトタイプを公開し、回帰のような単純な機能を実現することが可能であり、効率的であることを証明した。しかし、高次元パターン認識のような複雑な動作はまだ開発中である。単純なプロトタイプでは、様々な分野の要求に応えることは困難であった。これまで、我々はいくつかの DiaNets を進化させてきた。深さの爆発や勾配の消失問題を防ぐために、I/O 層の統合やスキップ接続など、様々な進化を提案した（DiaNet2.0、3.0 と命名）。実験により、本手

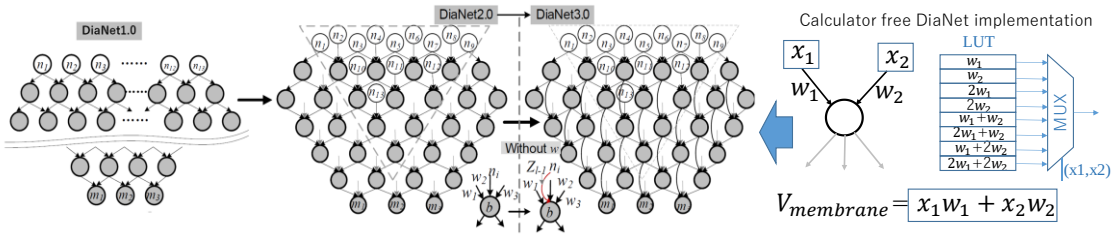

図 4 DiaNet の進化

法の有効性が様々なアプリケーションで実証された。MNIST 認識では、層数を DiaNet プロトタイプの 8.8%に削減しつつ、精度を 98.41%に向上させた。さらに、最新の LeNet5 と比較して、進化型 DiaNet は 90.86%のパラメータ削減を達成した。

(2)［時間的再構成］高効率 stochastic 計算

非同期ストキャスティック・コンピューティング（SC）のプロトタイプを提案し、様々な機能検証を行った。異なる周波数の複数のクロック信号を結合することにより、マルチプレクサ・チェーンが実装され、連続カオスの様式で非同期確率数（ASNG）を生成する。マッシブ・オペランド・コンピューティング・プラットフォームの場合、位相シフト・トポロジーが一連のインバータによって行われる。各 ASNG は 48 個の MOS トランジスタで構成される回路によって生成され、最新技術（SOTA）よりもストリーム長が短く、精度は同等以上である。唯一の ASNG、SC 乗算、マルチ・オペランド演算機能は、7nm CMOS 技術による非同期 SC プロトタイプに基づいて実装され、検証されている。SOTA と同程度の精度を持つトランジスタレベルのシミュレーション結果から、提案回路は、乗算、有限状態マシンに基づく SC 近似、多項式展開において、全 SOTA の中で最も優れた項目に対して、それぞれ 13.9 倍/1.46 倍/51.9 倍、22.4 倍/14.9 倍/8.68 倍、19.7 倍/12.4 倍/7.34 倍の電力/速度/トランジスタ数を獲得した。さらに、柔軟な電力-精度-速度のトレードオフを検証し、分析した。既存の SC 計算技術に対する比較を表 1 に示す。

表 1　様々な SC プロセッサーの HW コスト比較

		Sole/MUL				FSM			Polynomial		
	TVLSI'19	TVLSI'2	TVLSI'22	ICCD'23	this	TVLSI'19	TVLSI'21	this	TETC'1	TC'22	this
Strategy†	Deter.	SCC	AFE	Quan.	Async.	MC	MC	FSM	FC	McL.	Bern.
Tech. (nm)	45	45	FPGA	65	7	40	45	7	65	40	7
MAE/MSE	0.0019/-	-/0.00001	0.008/-	-/0.045	0.005/-	0.035/-	-/-	0.0138/-	0.017/-	-/0.00017	0.019/-
# of Tr.s ‡	25225	2960	21LUT	1350	126	3800	1650	190	9457	6021	820
Time (ns)	- ×1024	- ×1024	2.5 ×56	- ×1024	0.75 ×128	6.26 ×1024	1.4 ×1024	0.75 ×128	3.78 ×1024	1.16 ×1024	0.75 ×128
Power (uW)	-	-	6000	12	0.861	47	77	2.1	-	379.3	19.3
Energy ♯ (pJ)	-	-	840	9.2	0.083	301	110	0.2	-	451	1.82
EDP ♯	-	-	$>10^6$x	887x	1x	$>10^5$x	8213x	1x	-	3066x	1x
ADP ♯	1602x	188x	-	85.7x	1x	1335x	130x	1x	465x	80x	1x

Deter. = deterministic SN generation; SCC = minimum stochastic computing correlation; AFE = attitude and frequency encoding = hybrid of binary and SC over multi-wire; Quan. = stochastic bit quantization; MC = Markov chain; FC = factor combination; McL. = McLaughlin expansion; Bern. = Bernstein expansion.

５．主な発表論文等

〔雑誌論文〕　計7件（うち査読付論文　6件／うち国際共著　1件／うちオープンアクセス　0件）

１．著者名 Chen Olivia、Zhang Renyuan、Luo Wenhui、Wang Yanzhi、Yoshikawa Nobuyuki	４．巻 Early access
２．論文標題 Extremely Energy-Efficient Non-Linear Function Approximation Framework Using Stochastic Superconductor Devices	５．発行年 2024年
３．雑誌名 IEEE Transactions on Emerging Topics in Computing	６．最初と最後の頁 1～12
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TETC.2023.3330979	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 該当する

１．著者名 Oshio Reon、Sugahara Takuya、Sawada Atsushi、Kimura Mutsumi、Zhang Renyuan、Nakashima Yasuhiko	４．巻 44
２．論文標題 A Compressed Spiking Neural Network Onto a Memcapacitive In-Memory Computing Array	５．発行年 2024年
３．雑誌名 IEEE Micro	６．最初と最後の頁 8～16
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/MM.2023.3285529	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 －

１．著者名 Chen Yan、Zhang Renyuan、Kan Yirong、Yang Sa、Nakashima Yasuhiko	４．巻 Early Access
２．論文標題 Bisection Neural Network Toward Reconfigurable Hardware Implementation	５．発行年 2022年
３．雑誌名 IEEE Transactions on Neural Networks and Learning Systems	６．最初と最後の頁 1～11
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TNNLS.2022.3195821	査読の有無 無
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 －

１．著者名 ERLINA Tati、ZHANG Renyuan、NAKASHIMA Yasuhiko	４．巻 E104.C
２．論文標題 A Feasibility Study of Multi-Domain Stochastic Computing Circuit	５．発行年 2021年
３．雑誌名 IEICE Transactions on Electronics	６．最初と最後の頁 153～163
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transele.2020ECP5015	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 －

1．著者名 Kan Yirong、Wu Man、Zhang Renyuan、Nakashima Yasuhiko	4．巻 69
2．論文標題 MuGRA: A Scalable Multi-Grained Reconfigurable Accelerator Powered by Elastic Neural Network	5．発行年 2022年
3．雑誌名 IEEE Transactions on Circuits and Systems I: Regular Papers	6．最初と最後の頁 258～271
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TCSI.2021.3099034	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 －

1．著者名 Wu Man、Kan Yirong、Erlina Tati、Zhang Renyuan、Nakashima Yasuhiko	4．巻 464
2．論文標題 DiaNet: An elastic neural network for effectively re-configurable implementation	5．発行年 2021年
3．雑誌名 Neurocomputing	6．最初と最後の頁 242～251
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.neucom.2021.08.059	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 －

1．著者名 Nguyen Van-Tinh、Trinh Quang-Kien、Zhang Renyuan、Nakashima Yasuhiko	4．巻 9
2．論文標題 STT-BSNN: An In-Memory Deep Binary Spiking Neural Network Based on STT-MRAM	5．発行年 2021年
3．雑誌名 IEEE Access	6．最初と最後の頁 151373～151385
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2021.3125685	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8件（うち招待講演　0件／うち国際学会　7件）

1．発表者名 Haohui Jia, Ziwei Yang, Pei Gao, Man Wu, Chen Li, Yirong Kan, Renyuan Zhang
2．発表標題 Automatic Sleep Staging via Frequency-Wise Spiking Neural Networks
3．学会等名 IEEE International Conference on Bioinformatics and Biomedicine, (BIBM)（国際学会）
4．発表年 2022年

１．発表者名
Zheng Chen, Lingwei Zhu, Ziwei Yang, and Renyuan Zhang

２．発表標題
Multi-Tier Platform for Cognizing Massive Electroencephalogram

３．学会等名
International Joint Conference on Artificial Intelligence, (IJCAI)（国際学会）

４．発表年
2022年

１．発表者名
Guangxian Zhu, Yirong Kan, Renyuan Zhang, Yasuhiko Nakashima

２．発表標題
A Stochastic Coding Method of EEG Signals for Sleep Stage Classification

３．学会等名
IEEE International System-on-Chip Conference, (SOCC)（国際学会）

４．発表年
2022年

１．発表者名
Man Wu, Yirong Kan, Renyuan Zhang, and Yasuhiko Nakashima

２．発表標題
Training Deep Spiking Neural Networks with Ternary Weights

３．学会等名
IEEE International System-on-Chip Conference, (SOCC)（国際学会）

４．発表年
2022年

１．発表者名
Van-Tinh Nguyen, Tieu-Khanh Luong, Emanuel Popovici, Quang-Kien Trinh, Renyuan Zhang, Yasuhiko Nakashima

２．発表標題
An Accurate and Compact Hyperbolic Tangent and Sigmoid Computation Based Stochastic Logic

３．学会等名
IEEE International Midwest Symposium on Circuits and Systems, (MWSCAS)（国際学会）

４．発表年
2022年

１．発表者名
Sugahara Takuya, Renyuan Zhang, and Yasuhiko Nakashima

２．発表標題
Training Low-Latency Spiking Neural Network through Knowledge Distillation

３．学会等名
IEEE Symposium on Low-Power and High-Speed Chips（国際学会）

４．発表年
2021年

１．発表者名
Van Tinh NGUYEN, T. -K. Luong, E. Popovici, Q. -K. Trinh, Renyuan Zhang and Yasuhiko Nakashima

２．発表標題
An Accurate and Compact Hyperbolic Tangent and Sigmoid Computation Based Stochastic Logic

３．学会等名
IEEE International Midwest Symposium on Circuits & Systems（国際学会）

４．発表年
2021年

１．発表者名
Man Wu, Yirong Kan, Van_Tinh Nguyen, Renyuan Zhang, Yasuhiko Nakashima

２．発表標題
Ternarizing Deep Spiking Neural Network

３．学会等名
信学技報

４．発表年
2021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	木村　睦 (Kimura Mutsumi) (60368032)	奈良先端科学技術大学院大学・先端科学技術研究科・客員教授 (14603)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関