PRESS RELEASE

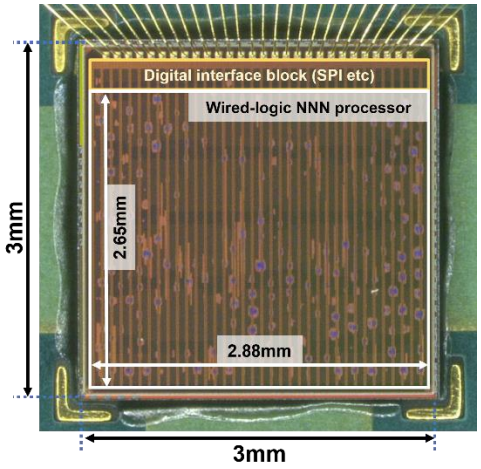
2023 年 6 月 9 日
東京大学
科学技術振興機構（JST）

音声コマンド認識 AI の電力を 3 桁削減、新方式 AI プロセッサを開発

――乾電池 1 本で 2 年以上連続動作、ドローンやロボットへの応用に期待――

発表のポイント

- ◆ 音声コマンド認識 AI の消費電力を 3 桁削減可能な、布線論理型 AI プロセッサを開発。
- ◆ 布線論理型 AI プロセッサの課題は膨大な実装面積。そこで、チップ面積と電力を削減するため新たなアルゴリズムと回路の協調最適化手法を開発し、16 層の深層ニューラルネットワークにおける全てのニューロンとシナプスを 1 チップに実装。
- ◆ 今後は音声コマンド認識に限らず、マシンビジョン、設備点検自動化、ドローン、AR/VR など多くのエッジ AI アプリケーションへの応用に期待。



音声コマンド認識 AI の電力を 3 桁削減可能な、新方式布線論理型 AI プロセッサのチップ写真

発表概要

東京大学大学院工学系研究科の小菅敦丈 講師、澄川玲維 大学院生、濱田基嗣 特任教授、黒田忠広 教授らによる研究グループは、JST 戦略的創造研究推進事業の助成のもと、35 種の音声コマンド認識（注 1）AI を題材に、既存の AI プロセッサと比較し 3 桁以上低電力化できる新方式の布線論理型 AI プロセッサ（注 2）を開発しました。

音声コマンド認識 AI は新たなマシンインターフェースとして急速に発展しています。一方で、認識可能なコマンド数が増え AI モデルが複雑化するほど、消費電力が急増するという課題がありました。これは、深層ニューラルネットワーク（注 3）の処理量が飛躍的に増えてしまうためです。識別可能なコマンド数が 4 種程度であれば 0.1mW 未満での推論が可能な一方、コマンド数が 35 種にもなると 390mW 程度の電力が必要となっていました。

本研究では低電力化のため、人の大脳を真似た布線論理型の新規 AI プロセッサを開発しました（図 1）。省ニューロン省シナプスなアルゴリズム技術と、省面積回路実装技術を新たに開発し、1 チップで 16 層の深層ニューラルネットワークを布線論理型 AI プロセッサで実装する

ことに成功しました。これにより、消費電力の大きかったメモリとの通信を完全になくし、152.8μW での推論を実現しました。この新規 AI プロセッサは、35 種の音声コマンドを識別可能な AI を、乾電池 1 本で 2.2 年にわたり連続動作させることが可能です。今後は、スマートフォン、ドローン、自動車内エンタメ機器制御、AR/VR 機器への応用が期待されます。

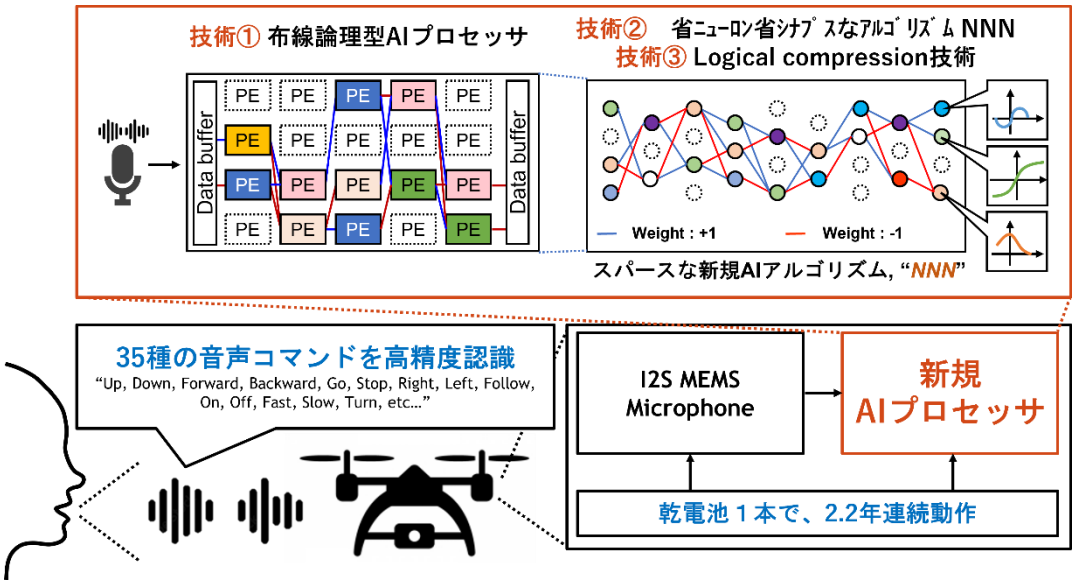


図 1：開発した新規 AI プロセッサの概要

35 種の音声コマンドを高精度認識可能な AI を、152.8μW の電力で連続動作可能。
乾電池 1 本で 2.2 年間の連続動作できる計算になる。

本研究成果は、2023 年 6 月 9 日（日本時間）に国際会議 2023 Symposium on VLSI Technology and Circuits で発行される「Technical Digest」に掲載されました。

発表内容

〈研究の背景〉

AI 技術は多くの産業に技術革新をもたらし、日常生活を変革すると期待されています。膨大な数のニューロンとシナプスを持つ深層ニューラルネットワークが技術の中核であり、シナプス接続を学習により最適化することでさまざまな能力を獲得しています。

AI 技術の課題は、極めて大きな消費電力です。ニューロンとシナプス数を増やすほど多種多様なタスクを高精度に処理でき、高性能な AI を実現できることが知られています。一方、巨大な AI モデルであるほど計算処理量が増え、コンピュータが消費する電力も膨大なものとなります。

本研究グループでは AI 処理の低電力化を実現するため、人の大脳を真似た布線論理型新規 AI プロセッサを開発してきました。1 チップ上にすべてのニューロンとシナプスを展開実装することで、消費電力の大きいメモリアクセスをなくし低電力化を実現してきました。これまでに、画像分類タスクにおいて、GPU（注 4）に比べて 4 桁以上電力を削減できることを実証し、プロセッサ分野における最高峰の学会の 1 つである 2022 IEEE Hot Chips 34 Symposium で発表しました。一方で、実際の応用に対してどのくらいの性能を発揮できるか、高い AI 計算性能を実現するためにどのようなシステム上の工夫が必要であるか、という点に関しては検証できていませんでした。

〈研究の内容〉

新たなマシンインターフェースとして期待される音声コマンド認識 AI を題材とし、高い精度、少ないチップ実装面積、低い消費電力すべてを同時に実現するため、省ニューロン省シナプスなアルゴリズム技術と、省面積回路実装技術を新たに開発しました。1 チップで 16 層の深層ニューラルネットワークを布線論理型 AI プロセッサとして実装することに成功し、152.8μW での推論を実現しました。従来の AI プロセッサ（ISSCC’22）（注 5）と比較し、消費電力を 1/2552 以下にできました（図 2）。これにより、35 種の音声コマンドを識別可能な AI を、乾電池 1 本で 2.2 年にわたり連続動作させることが可能です。

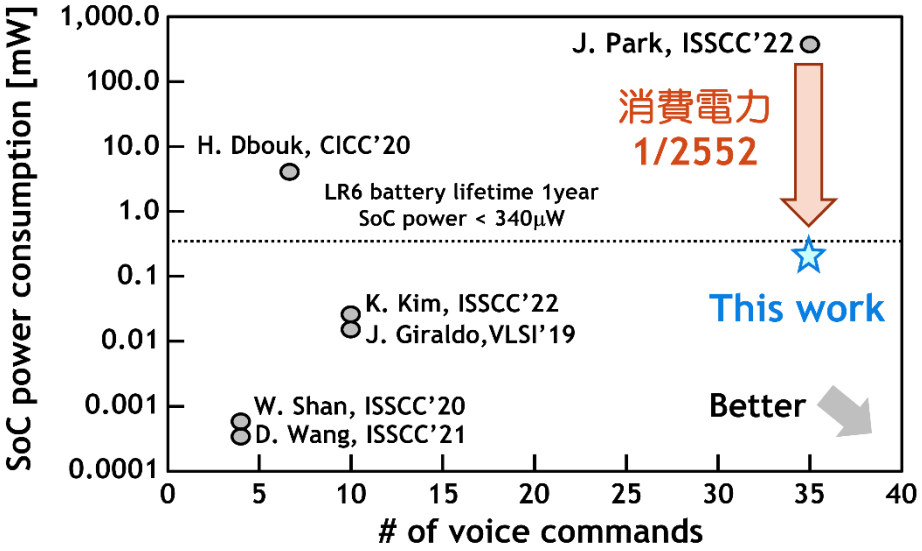


図 2：これまでの音声コマンド認識プロセッサとの性能比較

従来の AI プロセッサ（ISSCC’22）と比較し、1/2552 以下に電力消費を削減。

従来の音声コマンド認識 AI では、識別可能なコマンド数が 4 種程度であれば単純な AI 処理で完結するため 0.1mW 未満での推論が可能でした。一方、コマンド数が 35 種にもなると、16 層もの深層ニューラルネットワークが必要になり、390mW 程度の電力が必要となっていました。

布線論理型 AI プロセッサは人間の大脳を真似た方式であり、ニューラルネットワークを構成するニューロンとシナプスすべてをチップ上に並列実装しています。頻繁なデータ移動やメモリとの通信をなくすことができ、低消費電力化を実現しています。一方、16 層もの深層ニューラルネットワークを布線論理型 AI プロセッサとして実装しようとすると、大きな実装面積が必要でした。特に音声コマンド認識 AI 用途では長いビット幅が必要であり、個々の回路規模は大きくなります。試算では、30 チップ以上にも上る実装面積が必要でした。チップ間通信の電力消費が大きいことに加え、チップ枚数も多いことから、実装にかかる巨額のコストと巨大な面積が課題でした。

本研究グループは、布線論理型 AI プロセッサの実装面積を削減するため、深層ニューラルネットワークを簡素化し必要なニューロンとシナプス数を大幅に削減する“非線形ニューラルネットワーク（Non-linear Neural Network（NNN））”技術をこれまで提案してきました。ニューロンの非線形関数を個々に最適化することでニューラルネットワークの表現能力を高め、従来の深層ニューラルネットワークに比べて 2 桁少ないニューロン数とシナプス数で、複雑な AI タスクを実現する技術です。さらに音声コマンド認識向けにビット幅を削減し、ニューロンを省

面積な回路として実装しやすい形に変換する“Logical Compression”技術も新たに開発しました（図 3）。また、認識精度の劣化を抑えるため、ニューロン回路を AI のモデルとして再度取り込み AI モデルを再度最適化する、“Logical Compression Aware Re-Training”技術も併せて開発。結果、音声コマンド認識の精度を保ったまま、回路面積を 1/497 に削減することに成功しました。

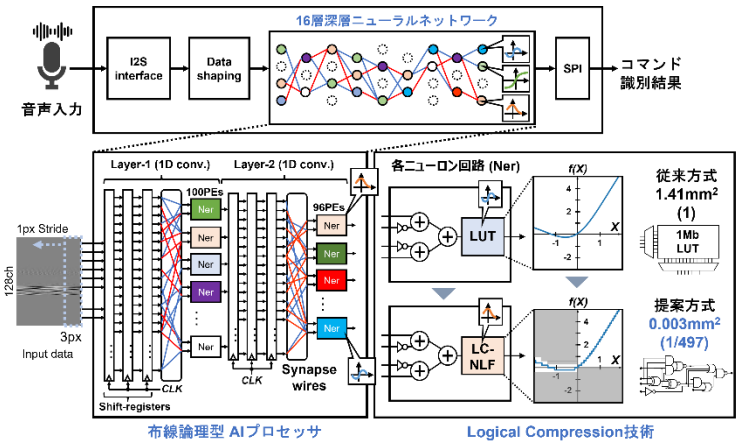


図 3：音声コマンド認識に向けた布線論理型 AI プロセッサ

チップ実装面積削減のための Logical Compression 技術を開発。

さらに開発した再学習アルゴリズムと組み合わせることで、認識精度を保ちながら面積を 1/497 に削減。

本研究では 16 層もの深層ニューラルネットワークを、40nm プロセスで製造された 3mm×3mm の 1 チップに布線論理型 AI プロセッサとして実装することができました（図 4）。これにより、152.8μW 消費電力での推論が可能になりました。半導体回路設計分野で最も権威ある学会である ISSCC、VLSI シンポジウムにて 2019 年以降発表された論文と比較したところ、同程度の消費電力で 3.5 倍以上のコマンド数を認識できました。コマンド数が増えると深層ニューラルネットワークの規模が増え、一般に大幅に電力が増大しますが、同程度の 100μW 台の消費電力に抑えることに成功しました。従来 AI プロセッサ（ISSCC’22）と比較し、1/2552 もの消費電力削減を実現しています（図 2）。

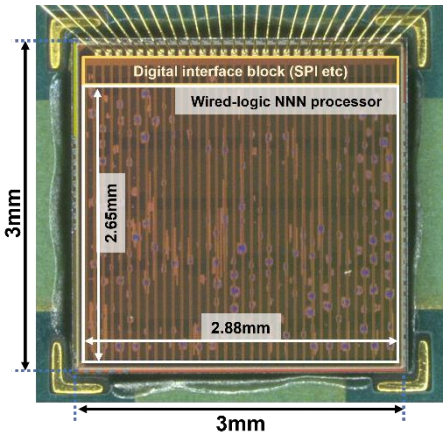


図 4：試作した音声コマンド認識向け布線論理型 AI プロセッサ

40nm プロセスで開発。16 層の深層ニューラルネットワークを 1 チップに実装。

〈今後の展望〉

開発した布線論理型 AI プロセッサはすべてデジタル回路で構成され、Python などの高位プログラミング言語から、短い設計期間で AI プロセッサの製造図面にまで変換できることが特徴です。このため、短期間に機能更新を繰り返す AI アプリケーションに最適といえます。今後は音声コマンド認識に限らず、マシンビジョン、設備点検自動化、物流倉庫、無人店舗など、カメラやドローンなどの端末に直接 AI を搭載したエッジ AI アプリケーションへ展開することを目指しています。

〈関連の記事〉

「システムデザイン研究センター 小菅敦丈 講師が「MIT Technology Review Japan Innovators Under 35」を受賞されました」(2021/12/22)
https://www.t.u-tokyo.ac.jp/topics/foe/topics/setnws_20211221112322095760 1664.html

「若手研究者紹介：小菅 敦丈 講師」(2023/5/2)
https://www.t.u-tokyo.ac.jp/topics/tp2023-05-08-069

発表者

東京大学大学院工学系研究科附属システムデザイン研究センター
小菅 敦丈（講師）
濱田 基嗣（特任教授）
黒田 忠広（教授）
澄川 玲維（修士課程）
柴 康太（博士課程：研究当時）
許 耀中（修士課程：研究当時）

論文情報

〈雑誌〉Technical Digest
（国際会議 2023 Symposium on VLSI Technology and Circuits で発行）
〈題名〉A 183.4nJ/inference 152.8μW Single-Chip Fully Synthesizable Wired-Logic DNN Processor for Always-On 35 Voice Commands Recognition Application
〈著者〉Atsutake Kosuge＊, Rei Sumikawa, Yao-Chung Hsu, Kota Shiba, Mototsugu Hamada, Tadahiro Kuroda

研究助成

この研究成果は、主として、以下の事業・研究領域・研究課題によって得られました。
JST 戦略的創造研究推進事業 個人型研究（さきがけ）
研究領域：「情報担体とその集積のための材料・デバイス・システム」(研究総括：若林　整 東京工業大学　工学院　教授)
研究課題：「デバイス・システム協調による超低電圧布線論理型 AI プロセッサ」
研究代表者：小菅 敦丈（東京大学 大学院工学系研究科　講師）

用語解説

（注 1）音声コマンド認識：複数のキーワードを認識することで、スマートフォン、PC、ロボットを制御する音声インタフェースの方式。
“Up”、“Down”、“Move”、“Stop”、“Fast”、“Slow”、“Right”、“Left”などの名詞や動詞を登録することで、細かい機器制御を音声で行うことができる。

（注 2）布線論理型 AI プロセッサ：演算器同士を物理的に結線し、結線を組み替えることで、プログラムの命令を実行する方式。汎用プロセッサと異なり、命令や各種データのメモリへの格納が原則不要であり、高速かつ低消費電力であるという特徴がある。

（注 3）深層ニューラルネットワーク：脳の仕組みを模した AI モデルの 1 つであり、多数のニューロンとシナプスからなる層を多段に重ね、シナプスの係数を計算により最適化することで所望の認知機能を獲得する。

（注 4）GPU：Graphic Processing Unit の略称であり、画像認識に特化した汎用プロセッサを指す。大規模な行列計算を高い電力効率と短い時間で実行できることから、行列計算を多数行う AI 処理に多く採用されている。

（注 5）ISSCC：International Solid-State Circuits Conference の略称であり、米国電気電子学会 固体回路分科会（IEEE Solid-State Circuit Society）が主催する最高峰のフラグシップ学会である。ここでは 2022 年度の ISSCC で発表された AI プロセッサに関する論文を指している。

J. -S. Park *et al.*, “A Multi-Mode 8K-MAC HW-Utilization-Aware Neural Processing Unit with a Unified Multi-Precision Datapath in 4nm Flagship Mobile SoC,” *IEEE International Solid- State Circuits Conference (ISSCC), Dig. Tech. Papers*, pp. 246-248, 2022.

問合せ先

〈研究に関する問合せ〉
東京大学大学院工学系研究科
講師　小菅 敦丈（こすげ　あつたけ）
Tel：03-5841-6655　E-mail：kosuge[at]dlab.t.u-tokyo.ac.jp

〈報道に関する問合せ〉
東京大学大学院工学系研究科 広報室
Tel：03-5841-0235　E-mail：kouhou[at]pr.t.u-tokyo.ac.jp

科学技術振興機構 広報課
Tel：03-5214-8404　E-mail：jstkoho[at]jst.go.jp

〈JST 事業に関する問合せ〉
科学技術振興機構 戦略研究推進部 グリーンイノベーショングループ
安藤 裕輔（あんどう　ゆうすけ）
Tel：03-3512-3526　E-mail：presto[at]jst.go.jp