Tokyo Tech

2022 年 2 月 19 日

報道機関各位

東京工業大学

隠れニューラルネットワーク理論を具現化した AI チップを世界で初めて開発

－消費電力削減・推論精度向上により
走る AI・飛ぶ AI の実現に大きく前進－

【要点】

○最新の「隠れニューラルネットワーク（Hidden Neural Network）」理論に基づく世界で初めての推論アクセラレータ LSI
○世界トップレベルの電力効率（30 TOPS/W 以上）と推論精度の両立を実証
○自動運転車や自律航行ドローン等への省エネルギー・高精度 AI の搭載に期待

【概要】

東京工業大学 科学技術創成研究院の劉載勲准教授、本村真人教授らは、末端（エッジ）機器での AI 応用の発展に向けて今後さらに重要となる、深層ニューラルネットワーク（DNN）の高効率な推論を実現する新規アクセラレータ **LSI**（用語 1）を開発した。

これまで DNN 推論アクセラレータ LSI を自動車などのエッジ機器に搭載・応用するとき、外部メモリアクセス時の電力消費の大きさが課題とされてきた。本研究では DNN の**重み係数**（用語 2）が乱数で固定されている「**隠れニューラルネットワーク（Hidden Neural Network）**（用語 3）」理論と呼ばれる新たな DNN 理論に着目し、推論に必要な情報量と計算時の消費電力を大幅に低減する新規ハードウェア・アーキテクチャ「ヒデナイト（Hiddenite）」を世界で初めて考案した。新提案のオンチップモデル構築（On-chip Model Construction）技術によって、外部メモリアクセスを大幅に削減できることが大きな特徴である。さらに、本アーキテクチャに基づくアクセラレータ LSI を製作し、世界最高レベルの高計算効率（最大 34.8 **TOPS/W**（用語 4））と、省モデルサイズ DNN として最高水準の推論精度（**ImageNet**（用語 5）70.1%）の両立ができることを実測結果で示している。

研究成果の詳細は米国時間 2 月 20 日からオンラインで開催される「ISSCC2022（国際固体素子回路会議）」にて発表される（発表者：廣瀬一俊 博士後期課程 3 年）。ISSCC は集積回路に関する最難関国際会議である。

●背景

人工知能（AI）技術のひとつである深層学習（ディープラーニング）を応用した画像認識技術の実装が広がりを見せており、自動運転車の飛び出し検知機能や、ロボットの作業制御、顔認証等のセキュリティ分野等での活用が期待されている。カメラで捉えた画像・映像を瞬時に処理するために、自動車やロボットといったエッジ機器にコンピュータを搭載し、その場で計算や AI による推論・判断を行うことが求められるが、計算量の爆発的増化とそれに伴う消費電力の増大が課題となっている。特に自動車やドローンといった移動機器においては常に外部から給電することは難しいため、推論精度を保ちつつ、なるだけ少ない電力でコンピュータが駆動することが必要となる。

ディープラーニングは深層ニューラルネットワーク（DNN）と呼ばれる人間の脳を模した情報処理モデルによって、画像や映像などの情報から状況を判断する。このネットワークの構造が複雑・巨大になることが膨大な計算量の要因であり、特に DNN モデルの「重み」といった計算パラメータを外部メモリから読み込む際に多大な電力を消費する。

軽量な DNN モデルを実現する画期的な技術として、2020 年に「DNN 全体の一部分だけを用いても推論精度が劣化しない」ような部分ネットワークを発見する「隠れニューラルネットワーク（Hidden Neural Network）」理論が新たに発表された（図 1）。従来のディープラーニング手法とは異なり、この理論では重みを学習せずに乱数の初期値のまま固定する。代わりに、ネットワークの各結合の重要度を表す「スコア」を学習し、スコアが上位の k%（k は任意の数）だけの部分ネットワークを用いることで、全体の k%のサイズの軽量 DNN モデルを構築する。具体的には、選択対象のスコア上位 k%の結合に値 1 を対応させ、それ以外には値 0 を対応させた「スーパーマスク」を生成し、乱数初期値の重みとスーパーマスクの論理積をとることで上位 k%の部分ネットワークを発掘することができる。

本研究ではこのような最新の隠れニューラルネットワーク理論に着目し、①重みを学習しない、②スーパーマスクで部分ネットワークを発掘する、という既存の DNN にはなかったその新たな 2 つの特徴を生かすことができる DNN 推論アクセラレータを実現した。

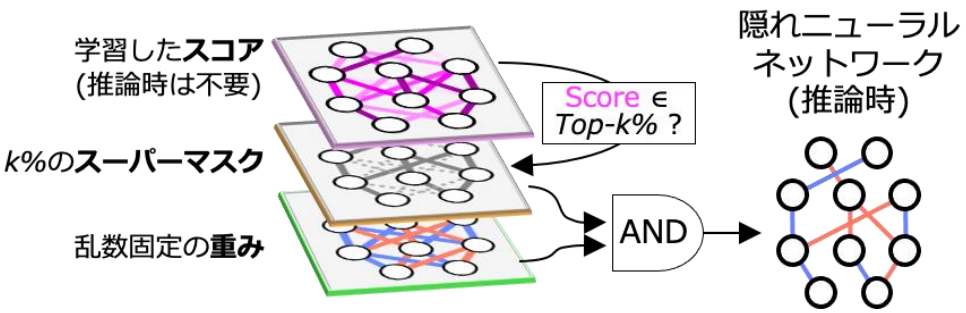

図 1 隠れニューラルネットワーク（Hidden Neural Network）理論の概念図

●研究成果

本研究では、最新の注目 DNN 技術であるこの「隠れニューラルネットワーク」理論を効率的に扱うことができるハードウェア・アーキテクチャを世界で初めて考案し、これを「ヒデナイト（Hiddenite)」アーキテクチャと命名した。そして、そのヒデナイト・アーキテクチャを実際にハードウェア化した LSI チップを試作し、具体的な性能を実測評価した。

【ヒデナイト（Hiddenite）アーキテクチャ】

隠れニューラルネットワークの重みは乱数固定であるため、従来の DNN とは違い、学習時と同じ乱数生成器と乱数シード値があれば再生成することができ、重みの値を記憶しておく必要がなくなる。そこでハードウェアとして乱数生成器を配置し、乱数生成のためのシード値自体も実行時の内部パラメータから生成することで、重みに関するデータの外部転送量を完全にゼロにできることを提案した。さらに、乱数重みについても、**Xorshift**（用語 6）のような非常にハードウェア量が小さい乱数生成回路で生成しても推論精度が低下しないことを新たに発見した。また、スーパーマスクについては、上位 30%程度使用の場合が最も精度が高いため、0 の値が多いことを利用した圧縮が可能であることを見出した。スーパーマスクを事前に圧縮してそれを LSI チップ上で展開することにすれば、外部からの転送量を抑えられる。

このように、重みとスーパーマスクからなる隠れニューラルネットワークのモデルに対して、重みの乱数生成回路とスーパーマスクの展開回路によって LSI チップ上でモデル情報を生成すれば、記憶したり転送したりする情報量を大幅に削減できる。新たに提案したこの「オンチップモデル構築（On-chip Model Construction)」技術が、本研究の根幹となる技術である。

従来型 DNN の最少モデルである**二値化ネットワーク**（用語 7）は、重みの正と負の割合が等しいため圧縮はできず、そのままチップに重みを転送する分の電力を消費する（図 2)。一方で隠れニューラルネットワーク理論に基づく本研究の場合は、二値の重みとスーパーマスクを必要とするため、一見するとモデルサイズは二値化ネットワークの 2 倍となるが、オンチップモデル構築技術により、転送量を二値化ネットワークのおよそ半分に削減することができる。モデル構築による消費電力はモデル情報の転送の消費電力に比べて桁違いに小さいため、全体として電力を大幅に削減できる。

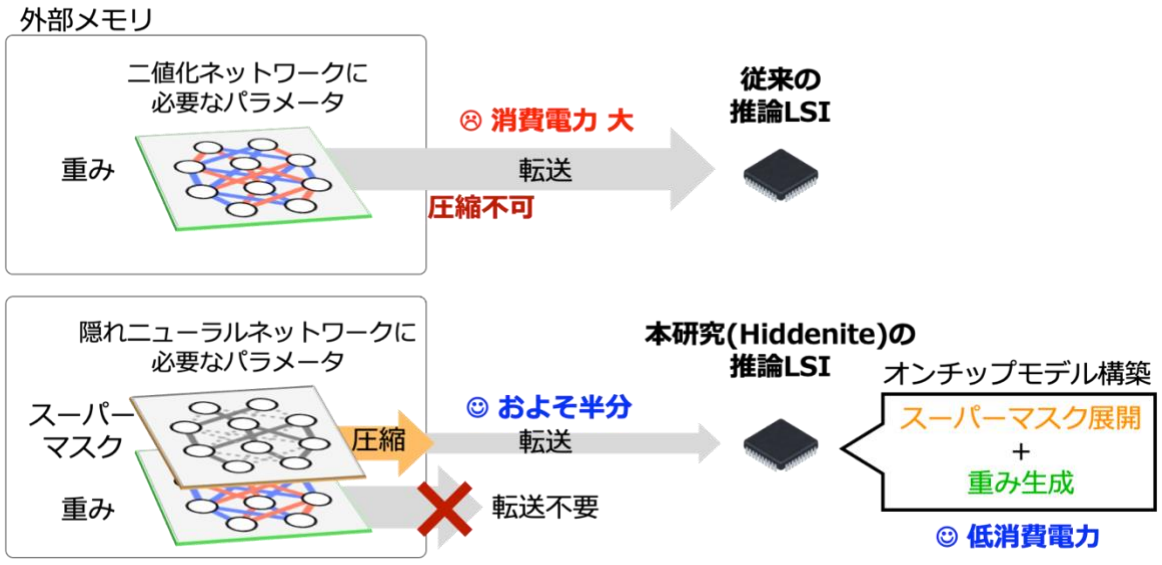

図 2 従来技術と本研究（Hiddenite）による推論 LSI のデータ転送量の比較

このオンチップモデル構築技術を備えた DNN 推論アクセラレータ全体のアーキテクチャを「ヒデナイト（Hiddenite; Hidden Neural Network Inference Tensor Engine）」と呼ぶ（図 3）。重みとスーパーマスクがそれぞれ 1 ビットであるため乗算器を必要とせず、演算要素が非常にコンパクトになり、エッジ機器への応用用途を想定した限られたハードウェア量の場合でも、ニューロン演算器を多数配置して並列演算を行うことが可能である。その並列演算能力を余すことなく生かすため、畳み込みニューラルネットワークの 4 次元（入力チャネル数・特徴マップの高さ・特徴マップの幅・出力チャネル数）方向にバランスよく割り当てる工夫も施している。

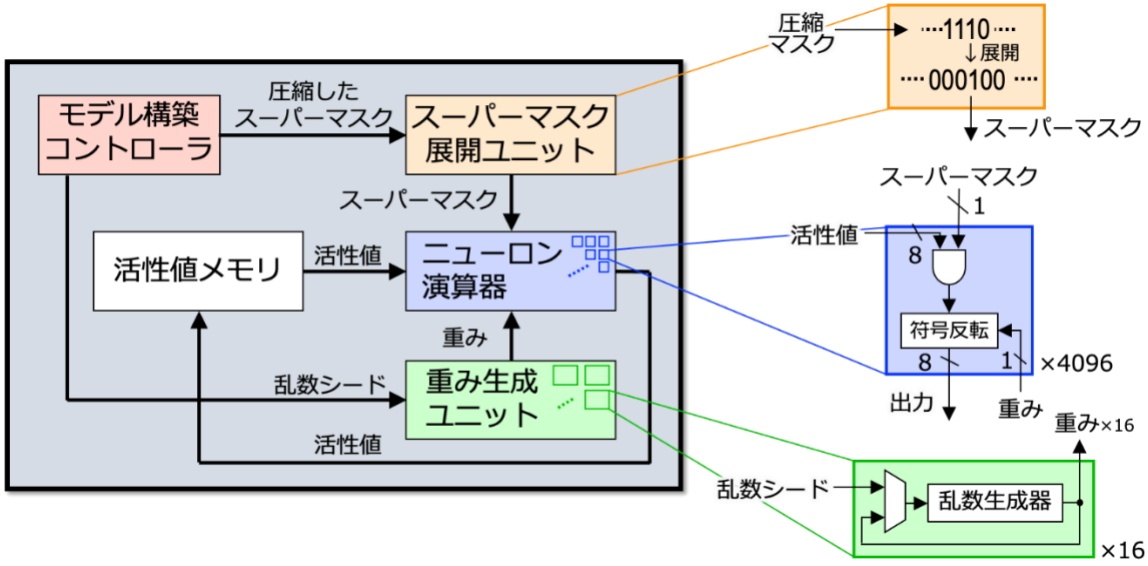

図 3 ヒデナイト（Hiddenite）アーキテクチャの概略 とその LSI 実現

【LSI の試作と性能評価】

このヒデナイト・アーキテクチャに基づいて、TSMC 社の 40 nm プロセスでプロトタイプチップを設計・試作した（図 4）。わずか 3 mm×3 mm の大きさで 4,096 個ものニューロン演算器を並列動作させ、4 次元の並列性を活かした高速推論処理が可能である。本チップは、DNN モデルの転送量を二値化ネットワークの半分に抑えながらも（図 2 で説明）、最大 34.8 TOPS/W という世界トップレベルの演算効率を達成している（この時の消費電力は 40 mW 程度と非常に小さい）。また、推論精度については、従来と同等の値を達成することができている（図 5）。

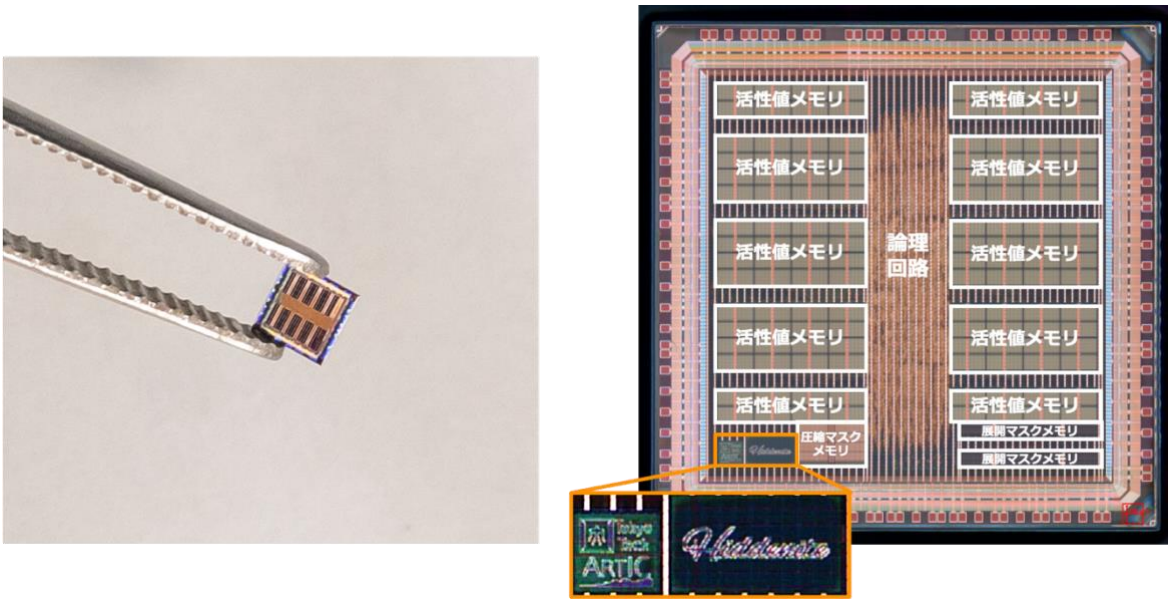

図 4 隠れニューラルネットワーク理論に基づく DNN 推論アクセラレータ LSI：
ヒデナイト（Hiddenite）チップ

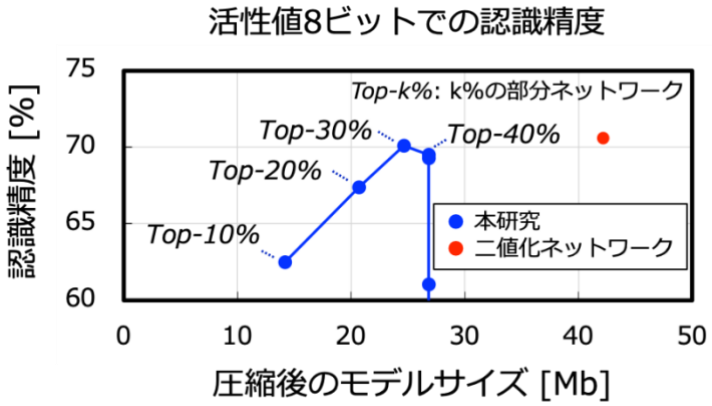

図 5 本研究で実現したモデルサイズと認識精度、及びその従来二値化ネットワークとの比較
（Resnet50、ImageNet、活性値 8 ビットでの比較）

●今後の展開

今回の試作 LSI は小さいチップ面積でのハードウェア化を目指したが、提案したヒデナイト・アーキテクチャ自体は、より大規模な隠れニューラルネットワークのモデルに対しても汎用的に推論可能である。研究チームは、より一層の推論精度向上のための学習アルゴリズムやその動作原理の研究を推し進めるとともに、さらにそれを活かした高精度かつ高効率な DNN アクセラレータの実現を目指している。

【用語説明】

（1） **LSI**：Large-Scale Integration（大規模集積回路）の頭文字で、多数の回路素子を 1 枚の基板に集積したもの。LSI チップないしはチップとも呼ぶ。

（2） **重み係数**：ニューラルネットワークにおいて入力データの重要度を表した値。学習によって更新される。ただし、隠れニューラルネットワークの場合は、学習されずに固定されたままである。

（3） **隠れニューラルネットワーク（Hidden Neural Network）**：V. Ramanujan らによる 2020 年の論文「What's Hidden in a Randomly Weighted Neural Network?」にて発表された新しい DNN 理論。

（4） **TOPS/W**：消費電力 1 W あたりの処理速度（TOPS; tera operations per second）として電力効率を表す単位。この数値が大きいほど、ある問題をある速度で処理する際の消費電力が小さいため、高効率であるといえる。例えば KU Leuven が VLSI2021 で発表した畳み込みニューラルネットワーク推論 LSI「DepFiN」は最大で 15.8 TOPS/W であった。

（5） **ImageNet**：画像認識の研究で用いられる大規模なデータセット。1,400 万枚を超える画像にどのような物体が写っているかが示されている。

（6） **Xorshift**：疑似乱数列生成法のひとつ。排他的論理和とビットシフトのみで構成されるため高速かつ軽量に疑似乱数を生成できる。

（7） **二値化ネットワーク**：1 ビットの重みで{-1, +1}の二値を表現することで算術演算を軽量化し、電力効率を向上させる。

【発表情報】

会議名：International Solid-State Circuits Conference (ISSCC) 2022
論文番号：15.4
論文タイトル：Hiddenite: 4K-PE Hidden Network Inference 4D-Tensor Engine Exploiting On-Chip Model Construction Achieving 34.8-to-16.0TOPS/W for CIFAR-100 and ImageNet
著者：廣瀬一俊（発表者、博士後期課程 3 年）、劉載勲（共同主著、准教授）、安藤洸太（特任助教）、大越康之（修士課程 1 年）、Ángel López García-Arias（博士後期課程 1 年）、鈴木淳之介（修士課程 2 年）、Thiem Van Chu（助教）、川村一志（特任助教）、本村真人（教授）

【問い合わせ先】
東京工業大学 科学技術創成研究院 AI コンピューティング研究ユニット
教授　本村真人
Email: motomura@artic.iir.titech.ac.jp
TEL: 045-924-5654　FAX: 045-924-5654

【取材申し込み先】
東京工業大学 総務部 広報課
Email: media@jim.titech.ac.jp
TEL: 03-5734-2975　FAX: 03-5734-3661