TOKYO UNIVERSITY OF SCIENCE
1-3 KAGURAZAKA, SHINJUKU-KU,
TOKYO 162-8601, JAPAN
Phone: +81-3-5228-8107

２０１８年　５月　７日

報道関係各位

モノ側で情報の高度な利活用を行う人工知能(ＡＩ)処理型リコンフィギュラブル半導体
～認識精度を向上させつつ半導体実装で姿を消せる賢いゼロを活用した
スパースターナリーニューラルネットワークＡＩチップ～

東京理科大学

研究の要旨

東京理科大学工学部電気工学科河原尊之教授は、小面積ながら高い認識精度を達成できる新しいＡＩ[1]コンピュータ用半導体(ＡＩチップ)の実現手法を開発しました。１、０、－１の３値を用いてニューラルネットワークを構成して高い認識精度を実現し、これをＦＰＧＡ[2]半導体チップに実装する時には０を配線の省略によって表現することで素子数と配線数を低減するという簡便ながら効果的な手法です。その結果、実数値を使う理想方式とほぼ変わらない高い認識精度を実現すると共に、精度は低いが面積は最小と言われている２値を使う方式よりも小面積とできる見通しを得ました。これにより電力や搭載サイズの制限が厳しいモノ側(エッジ)に適したＡＩ組込みコンピュータ半導体素子の実現が期待され、情報が急増するＩｏＴ[3]社会の持続可能な発展に寄与できます。

本研究成果は、2018 年 5 月 7-9 日に台北市にて開催される The IEEE International Symposium on Next-Generation Electronics (ISNE 2018)にて発表致します。

【研究の背景】

ＩｏＴ[3]社会の持続可能な発展には、社会インフラや環境といった実世界(物理世界)の情報(データ)から有意なデジタルビット情報を賢く取り出す技術が重要です。すなわち、センサ素子が高性能化する中で、これを搭載した多くの”モノ”が直接情報網に繋がるのでは実世界からのソースデータで溢れることになり、通信回線やクラウドに大きな負荷が生じてしまいます。電力効率も悪くなります。よって、これを解決するに、モノ側にてこの急増する膨大なデータをその場で処理し、有用な情報のみを取り出す必要があります。この処理は一般的に従来のノイマン型情報処理では極めて大きな処理量となる種類のものです。このためＡＩ[1]を活用した、新しい情報処理装置が期待されています。この分野は、また、日本が得意な組込みマイコンがＡＩと結びつき大きく発展していく分野ともなります。

このＡＩ処理を実現するためにニューラルネットワーク（ＮＮ）技術の開発が盛んです。従来のコンピューティングと比較して、ＮＮでは学習において、内部の結合重み・バイアスなどのパラメータの最適値を求めます。ひとたびそれらの最適値が決定されれば、データをＮＮに入力することで高速・高精度な認識が可能となります。モノ側でのデータの利活用のための処理とは、学習結果に基づいたこの推論と呼ばれる処理が主となります。課題として

は、理想的には実数として扱うパラメータは実装時にも８ビット程度の精度は必要とされており、またＮＮの出力を求める際には積和計算を多用するため、回路規模が大きくなり消費電力が増大してしまうことです。モノ側(エッジ)においては、サーバやデータセンタ側と比較して電力制限は非常に厳しいものとなります。このため、パラメータを究極的に２値化（バイナリ）することにより回路規模やメモリを削減する研究が進められています。しかしながら、この場合、認識の精度は必ずしも高くないという問題がありました。

また、実世界データの種類は多く、その活用時の処理は多様です。ＮＮは学習によって広範囲な情報に対応できますが、モノ側で推論を行うためには、学習結果に従ってモノ側においてＮＮとしての再構成ができることが必要となります。これを実現し、推論を低電力かつ高性能に行うには、製造後に半導体回路を自在に再構成でき柔軟性に優れるＦＰＧＡ[2]のようなリコンフィギュラブル半導体が適しています。

このように次世代ＩｏＴ向けに、電力制限が厳しいモノ側(エッジ)において、高度なセンサ情報の認識を高精度かつ低電力にて行なえ、広い範囲の応用に対応できる、推論を主としたＡＩエッジコンピュータ技術（ＡＩチップ技術）の開発が必要となってきておりました。

【研究成果の概要】

上記背景のもと、今回、ＦＰＧＡを用いたスパースターナリーニューラルネットワークＡＩチップ技術を独自に開発しました。具体的な技術内容は以下となります。

１）1, 0, -1 の３値で重みを表現する３値化（ターナリ）ＮＮを用い、かつ、重みの値が０の部分はＦＰＧＡ実装の時に配線の省略によって表現します。従来のバイナリＮＮでは重みと入出力を２値化することによりＸＮＯＲ回路でＮＮが構成できました。しかしながら、図１(a)に示すようにニューロン同士に必要な配線は多いままであり、また、認識率の向上には問題に応じた工夫が必要でした。今回、図１(b)に示すように３値で重みを表現することにより認識率の向上を図り、０の機能を重み０の配線は実装しないことで実現して小型化しました。重みは実効的には３値を取りながらも入出力は２値となります。

２）図２(a)に示すように、データセットＭＮＩＳＴ[4]を用いた評価では認識率は 93～96%と高く、理想的な実数重みの場合との精度の差は 0.3～0.6%と小さな結果となりました。また、図２(b)に示すように、ＦＰＧＡ実装の試行では重み０の配線は実装しないため、２値の場合と比べてＮＮの中間層－出力層間の重み配線数は 46%削減でき、メモリ使用面積も30%削減できました。このように、理想的な実数重みの場合と同等の精度を実現し、かつ、精度は低いが面積は最小と言われている２値重みの実装よりも小面積化できる見通しを得ました。

なお、ターナリーニューラルネットワーク実現のための学習についても、重みの正則化、閾値や学習係数の設定、適した活性化関数の選択などの手法を独自に開発しました。

【今後の展望】

・デモシステムの構築による小型化と精度の実機全体検証

・実応用での検証

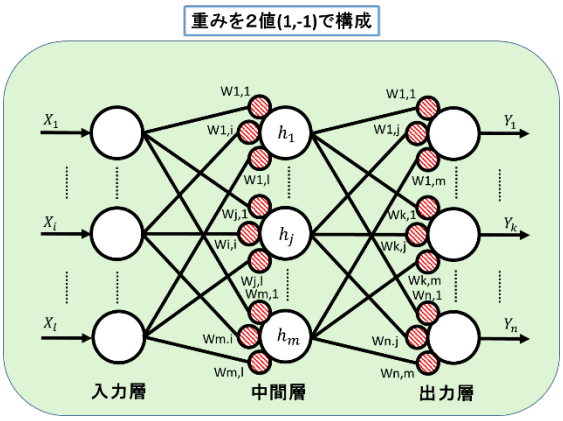

図　1（a）従来方式

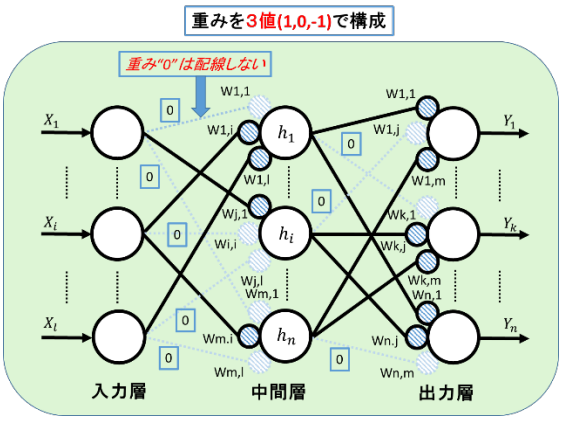

図　1（b）提案方式

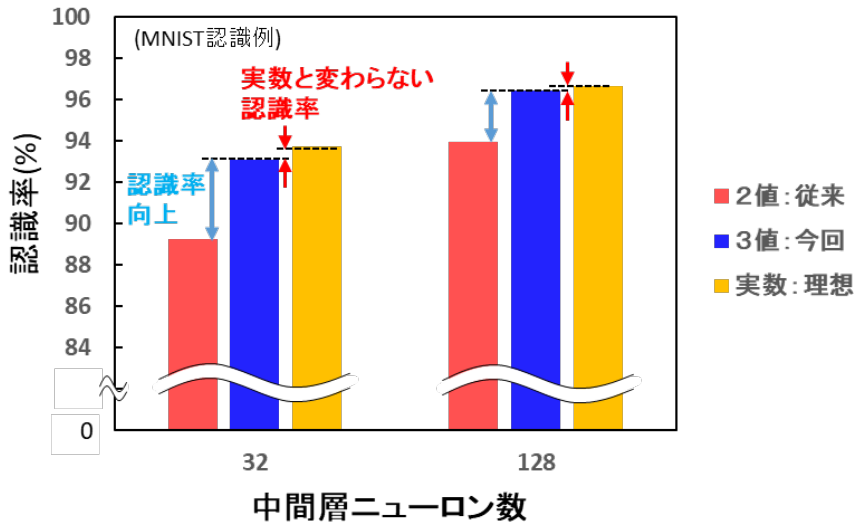

図　2（a）認識率向上

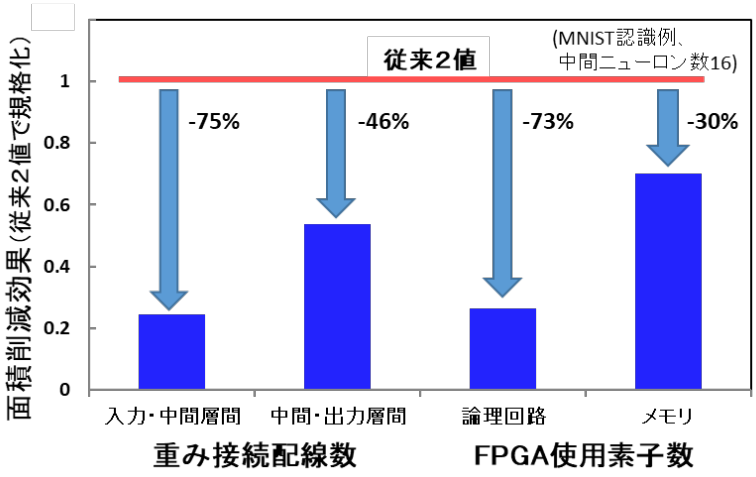

図　2　（b）チップ面積低減効果

用語

1･･･AI：Artificial Intelligence、人工知能。

2･･･FPGA：Field Programmable Gate Array。予め用意された論理回路の組み合わせを製造後に変えることができ、所望の機能を実現できるＬＳＩ。

3･･･IoT：Internet of Things、モノのインターネット化。従来は直接接続されていなかった"モノ"がインターネットを介して情報をやり取りする能力を備えていくこと。

4･･･MNIST：0~9 の 10 種類の手書きの数字が用意されたデータセット。各数字に 6000 種類の手書きデータが用意されている。ＮＮの学習や推論の評価に使用される。

～本件に関するお問い合わせ～
東京理科大学　研究戦略・産学連携センター
〒162-8601　東京都新宿区神楽坂 1-3
TEL：03-5228-7440　FAX：03-5228-7441