杉原堅也*
Kenya Sugihara
山本　亮†
Ryo Yamamoto

コンパクトなハードウェアAI

Compact Hardware Artificial Intelligence

要　旨

三菱電機のAI(Artificial Intelligence)技術“Maisart”には全てのモノを賢くする“コンパクト化”という特長がある。機械学習の一つであり，また今日の第３次AIブームのきっかけになった“ディープラーニング(深層学習)”は膨大な演算量を必要とすることから小型機器への搭載が困難であった。これを解決するため，推論に用いるネットワーク構造と計算方法に着目して分析し，推論精度を維持したまま，推論処理の演算量・使用メモリ量を削減し，コンパクト化することに成功した。

コンパクト化を実現する多層ニューラルネットワークの枝刈り技術は，演算量を削減できる一方で，不規則な枝刈りを行う場合，FPGA(Field-Programmable Gate Array)では枝刈り前と後で大幅な回路規模削減が見込めず演算量削減の効果を発揮できない。これは，同一処理を行う並列化した積和演算回路で，枝刈り箇所をスキップする場合とスキップしない場合が混在してしまい，その結果，回路の動作率が低下してしまうためである。そこで，FPGA内の回路動作率を向上させるため，枝の接続に規則性を設け，並列化した回路の動作率を向上させた。これによって，道路標識の画像認識では処理時間が1/10に削減可能になった。また，処理時間を削減しない場合には回路規模を1/10にすることも可能である。

この技術によって，リアルタイム性が求められる分野や，これまでコスト面で人工知能の適用が難しかった分野にも，今後適用範囲を広げていく。

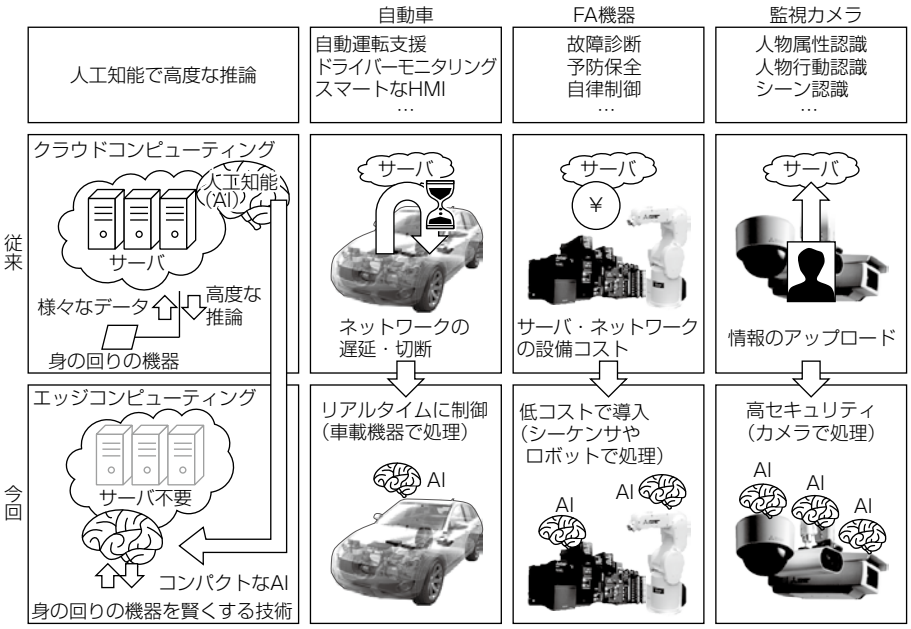
Maisartの“コンパクトな人工知能”の適用分野

当社のAI技術Maisartの一つである“コンパクトな人工知能”の計算順序の効率化と回路構成の最適化によって，小規模なFPGAにも実装できる“コンパクトなハードウェアAI”を開発した。リアルタイム性の向上と低コスト化を実現したことで，家電，エレベーターや自動車など人工知能の適用分野拡大に貢献していく。

*情報技術総合研究所(博士(情報学))　†同研究所

1. まえがき

当社のAI技術Maisartでは全てのモノを賢くする“コンパクト化”という特長を掲げている。当社では，機械学習の一つであり，また今日の第３次AIブームのきっかけになった“ディープラーニング”が膨大な演算量を必要とすることから機器への搭載が困難であるのに対し，これを解決するために，推論に用いるネットワーク構造と計算方法に着目して分析し，推論精度を維持したまま，推論処理の演算量・使用メモリ量を削減し，コンパクト化することに成功した。これによって，従来はクラウドコンピューティングなどでしか実行することが難しかった人工知能による高度な推論が**図１**のようにエッジコンピューティングでも可能になり，適用範囲が広がり，活用の幅が大きく広がった。



HMI：Human Machine Interface

図１．AIコンパクト化

また，この技術を発展させて，さらに計算順序の効率化と回路構成を最適化することによって，小規模なFPGAにも実装できる“コンパクトなハードウェアAI”を開発した。これによって，MaisartをFPGAに実装することで，低コスト化やリアルタイム性を向上させた。

2. コンパクトなAI

現在のAIブームの要因になった“ディープラーニング”技術は，従来のサポートベクタマシンなどによる認識技術と比べて高度な推論が可能になる一方，複雑なモデルを用いて計算するため，学習及び推論処理に必要な演算量・メモリ使用量が膨大になる課題があった。この章ではこれらの課題解決の手法について述べる。

2.1 ディープラーニングの課題

“ディープラーニング”とは，人間が自然に行うタスクをコンピュータに学習させる機械学習の手法の一つである。現在のAIの急速な発展を支える技術であり，その進歩によって様々な分野への実用化が進んでおり，現在のAIの中核的技術とも言える。

ディープラーニングはニューラルネットワーク（Neural Network）という機械学習の手法をベースにしたものであるが，長い間解決されていなかったニューラルネットワーク特有の課題を“多層（ディープ）化”するという工夫で解決している（**図２**）。

この多層化によって，例えば画像認識で言えば，それまでの認識精度が70％台であったものが，ディープラーニングの登場によって一気に80％を超え[1]，2016年以降は90％を超え[2]，極めて高精度な推論が可能になった。しかし，一方で複雑なモデルを用いて計算し，非線形に識別することから，学習用データが大量に必要になり，学習及び推論処理に必要な演算量・メモリ使用量が膨大になるという課題もあった。

図２に示すように“ディープラーニング”は各層間のノード同士が全て枝で接続されており，密なネットワーク構造になっている。各枝では学習処理で定まる枝重みと枝の入力値とで積算を行うため，枝の本数が増加するほど演算量は増大することになる。また，“ディープラーニング”は多層構造を持ち，演算量が更に大きくなる。そのため，大規模サーバや組み込み機器の場合でも，演算を高速化するためのGPU（Graphic Processing Unit）等を用いないと，必要とされる要求仕様を満たせない場合が多く，エッジや機器に搭載する上での大きな課題になっていた。

2.2 ディープラーニングの演算量削減技術

ニューラルネットワークは以前から誤差逆伝播（でんぱ）法と呼ばれる方法で学習を行っていたが，中間層が２層以上になると学習が収束しにくいという問題があった。これに対し，ヒントンらがこの問題の解決のために自己符号化器を用いることを提案した[3]。学習処理を，枝重みの初期化を行う“プレトレーニング”と，教師あり学習で全体を

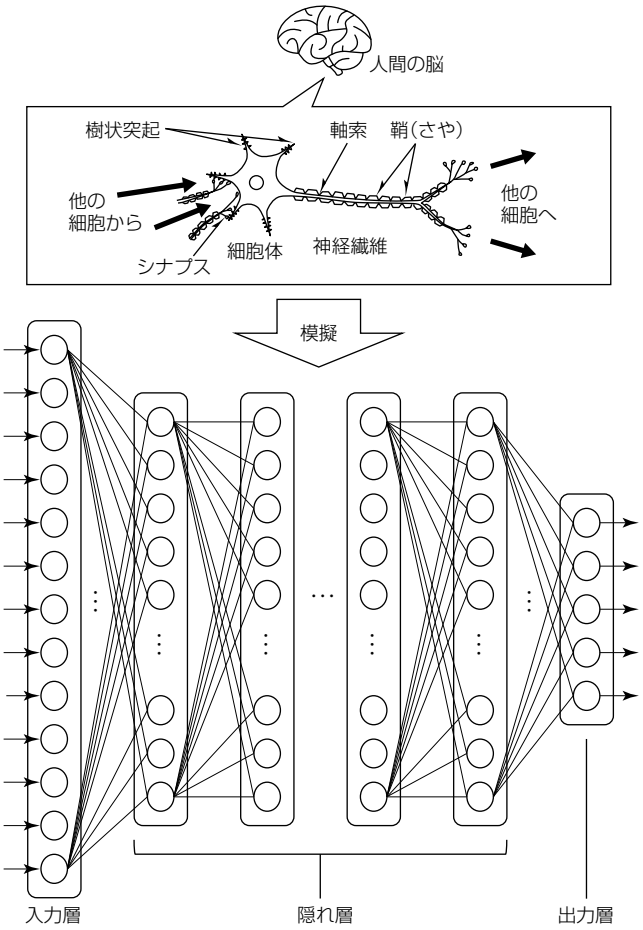


図2．多層化されたニューラルネットワーク

最適化する“ファインチューニング”の2段階に分け，“プレトレーニング”に自己符号化器を用いて一層ずつ教師なし学習で初期化する。自己符号化器は隠れ層に対する入力層をそのまま出力層に用い，次元圧縮した中間層で入力層のデータを出力層で復元できるように学習する手法である。この自己符号化器を用いることによって，学習の収束性が向上した。

当社はこの自己符号化器が持つ機能を実現するための条件を導出し，演算量に大きな影響を与える枝の本数を削減することが可能かどうかを検討した。その結果，大幅な演算量削減にもかかわらず，正解率はほとんど劣化しないことが分かった。すなわち，当社が提案した手法はネットワーク構造と計算方法を効率化するアルゴリズムである。圧縮センシング理論を活用し，疎である事前分布の情報を伝播させるために最適な疎なネットワーク構造とそれに対する計算アルゴリズムを開発した[4][5]。

この手法によって，従来型の“ディープラーニング”と比べて認識精度を維持したまま，低演算量化することが可能になった。演算量をどの程度まで削減できるかについては推論対象にも依存するが，画像認識の場合に，学習・推論の演算量及びメモリ使用量を従来比で1/10まで削減可能なケースが報告されている[6]。低演算量化することによって，処理能力の低い組み込み機器や小型コンピュータ等に搭載することが容易になる。

3．コンパクトなハードウェアAI

一般に多層ニューラルネットワークは，大規模な演算を要するため，エッジコンピューティングで要求されるリアルタイム性を達成するためには，アクセラレータが利用される。アクセラレータの種類は，GPU，専用ASIC（Application Specific Integrated Circuit），FPGAがある。用途などにも依存するが，エッジコンピューティングでは低消費電力であり，高いスループットが出せるFPGAが候補になる。

大規模なディープニューラルネットワークによる推論をFPGAで実現する場合，回路規模が大きくなり，FPGAが高コストになってしまうという問題がある。演算量を削減するコンパクト化を行えば，一般的に回路規模は削減傾向にはなるが，実装対象のアクセラレータによって，その効果は異なる。

この章ではFPGA向けのコンパクト化について述べる。

3.1　FPGA向け量子化技術

GPUなどは，浮動小数点演算専用のハードウェアを搭載しているため，問題にはならないが，FPGAの場合は，浮動小数点演算は，回路規模増大の原因になる。そこで，ニューラルネットワークの演算対象となるデータの量子化を行うことで，回路規模を削減する。重みデータを作成する際に，FPGA側の量子化処理と同じ処理を学習時に利用することで，推論精度を高めることができる。

3.2　FPGA向け枝刈り技術

実装先であるFPGAを意識しない不規則な枝刈りを行った場合，演算量自体は削減されても，FPGAでその効果を発揮できない。そのため，枝刈り前と後で，大幅な回路規模削減が見込めない。これは，同一処理を行う並列化した積和演算回路で，枝刈り箇所をスキップする場合とスキップしない場合が混在してしまい，その結果回路の動作率が低下してしまうためである（**図3**）。

そこで，FPGAの回路の動作率を向上させるため，枝の接続に規則性を設け，並列化した回路の動作率を向上させた（**図4**）。動作率の向上によって，必要な回路並列数を削減できる。

3.3　FPGAへの実装試行

量子化と枝刈りは，精度と回路規模に影響を与える。量

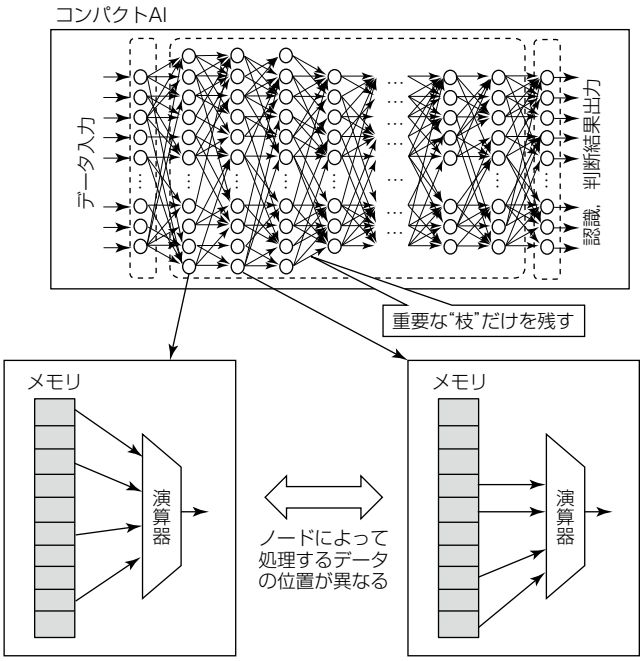


図３．演算量削減技術の課題

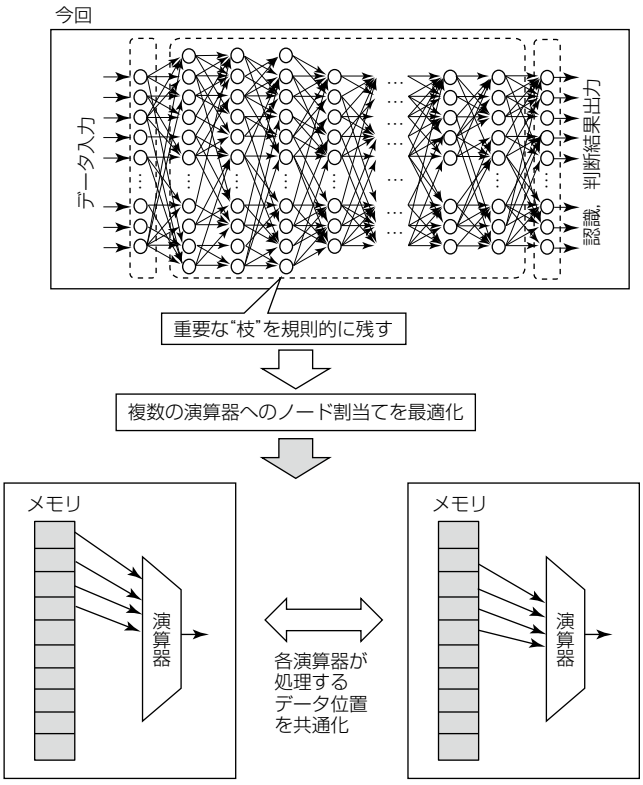


図４．コンパクトなハードウェアAI

子化によるビット数削減が大きい，又は枝刈りする量が多い場合，回路規模は小さくなるが，精度が劣化する。又は量子化によるビット数削減が小さい，又は枝刈りする量が少ない場合，精度は高いが，回路規模が大きくなってしまう。そこで，精度と回路規模のトレードオフをとりながら，要求となる精度を出すことができる，最小限の量子化と枝刈りに対応した学習を実施する。

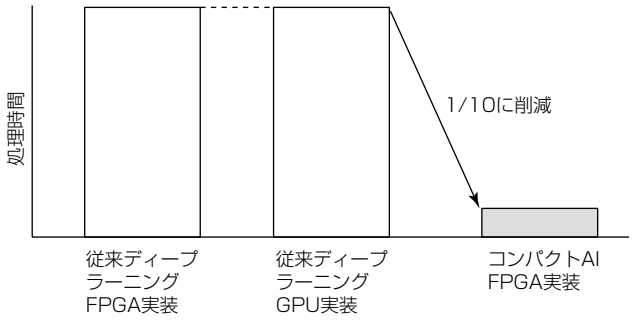


図５．従来手法との処理時間比較

この技術を道路標識の画像認識用アルゴリズムに適用試行した。その結果，従来のディープラーニングアルゴリズムと比較して，推論精度を維持しながら推論処理にかかる演算時間を1/10に短縮し，リアルタイム性の向上を図った(**図５**)。また推論処理の演算時間が同等の場合には，従来に比べて回路規模を1/10に削減した。

4. むすび

当社のAI技術Maisartの一つである“コンパクトな人工知能”の演算順序効率化と回路構成最適化と，それによって小規模なFPGAにも実装できるディープラーニングのアルゴリズムを開発した。ディープラーニングは高度な推論が可能だが，多層のニューラルネットワーク構造を用いるため，推論に必要な演算量とメモリ量が膨大になる。ニューラルネットの重要な枝だけを残す“コンパクトな人工知能”で，推論処理の演算量を減らして省メモリ化を図り，組み込みCPUへの実装を可能にした。しかしながら，この手法では，枝の接続が不規則であったため，FPGAへの実装では並列処理を行う構造をとれず，回路規模を削減できないという課題があった。今回，枝の接続に規則性を持たせるために演算順序を効率化し，演算器へのノード割当てを最適化することで，小規模なFPGAへの実装を可能にした。この技術によって，リアルタイム性が求められる分野や，これまでコスト面で人工知能の適用が難しかった分野にも，適用範囲を広げていく。

参考文献

(1) Krizhevsky，A.，et al.：ImageNet classification with deep convolutional neural networks，NIPS，1097～1105 (2012)
(2) He，K.，et al.：Deep residual learning for image recognition，Proc. IEEE CVPR，770～778 (2016)
(3) Hinton，G.E.，et al.：Reducing the dimensionality of data with neural networks，Science 313(5786)，504～507 (2006)
(4) 松本　渉：深層学習での演算量削減技術，三菱電機技報，**91**，No.6，361～364 (2017)
(5) Matsumoto，W.，et al.：A deep neural network architecture using dementionality reduction with sparse matrices，ICONIP，397～404 (2016)
(6) 中尾尭理，ほか：疎なネットワーク構造を持つDeep Learningを用いた映像分析システム，情報処理学会第79回全国大会，6B-07 (2017)