News Release
2021.8.23

国立研究開発法人新エネルギー・産業技術総合開発機構
国立大学法人東京工業大学

スマホやロボットなどで高効率なAI処理を行うプロセッサーアーキテクチャーを開発
―試作LSIで世界トップレベルの実効効率最大26.5TOPS／Wを達成―

NEDO が進める「高効率・高速処理を可能とする AI チップ・次世代コンピューティングの技術開発」で、東京工業大学はエッジ機器で高効率な畳み込みニューラルネットワーク（CNN）推論処理を行うプロセッサーアーキテクチャーを開発し、大規模集積回路（LSI）を試作しました。

今後、本技術の活用により、例えばスマートフォンにおける先進的な拡張現実（AR）アプリケーションやロボットにおける柔軟な動作制御など、電力供給量などの制約が厳しいエッジ機器でも高度なリアルタイム AI 処理の単独での実行が期待できます。

従来の深く枝刈りされた CNN の推論処理では、メモリへのアクセスが不規則になるため計算効率が低下するという課題がありました。本研究では既存の CNN モデルを変形して高精度で高効率な処理ができる形式に変換するアルゴリズムを開発しました。さらに、このアルゴリズムを効率的に処理するための、入力データの平面シフトを扱う整形機構と直積型並列演算アレイを中核としたアーキテクチャーを提案しました。これにより試作 LSI による実測で、最大 26.5TOPS／W という世界トップレベルの実効効率を達成しました。

なお、東京工業大学は、現在オンラインで開催中（2021年8月22日から24日まで）のプロセッサーLSIの主要国際会議「Hot Chips 33」のポスターセッションで本成果を発表する予定です。

1. 概要

アプリケーションの高度化に伴い大規模化した畳み込みニューラルネットワーク（Convolutional Neural Network; CNN）※1は、応用範囲が急速に広がる一方で、特に電力・面積的制約の大きな端末（エッジ）機器で利用する場合に、要求される計算量とメモリ容量の肥大化が問題となっています。これに対処するため、アルゴリズム面から提案されてきたのが、CNNの枝刈り（プルーニング）※2に代表される、CNNモデルの冗長性を利用し、認識精度を保ったままモデル規模を縮小（スパース化）する手法です。しかしこの手法では、計算量とメモリ容量の削減はできるものの、メモリへのアクセスが不規則となり、データ再利用性と演算器の稼働率が低下するため、並列処理の計算効率が低下することが課題とされてきました。

こうした背景のもと、国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）と国立大学法人東京工業大学（東工大）科学技術創成研究院の本村真人教授らのチームは、「高効率・高速処理を可能とするAIチップ・次世代コンピューティングの技術開発※3」で、エッジ機器での高効率なCNN推論処理を行うプロセッサーアーキテクチャーを開発し、大規模集積回路（LSI）を試作しました。本研究で、深く枝刈りされたことでメモリへのアクセスが不規則になったCNNの推論処理では、チャネル間畳み込みと入力データのチャネル内平面シフトを利用して畳み込みの積和演算を行うことが効率的であることを見いだし、そのため並列演算アレイとデータ整形機構を中核としたアーキテクチャーを提案しました。さらにこれによる試作LSIによる実測で、最大26.5TOPS／W※4という世界トップレベルの実効効率を達成しました。これは、例えば従来、スマートフォンやウエアラブル端末では電力供給量や処理能力が限られているため大規模なCNNモデルによる高度なAIアプリケーションはクラウド側で実行していましたが、本開発によって高度なリアルタイムAI

処理をエッジ側で実行できます。これにより、AIサービスにおけるプライバシーの確保やクラウドへの通信量の削減などが期待できます。

なおこの結果は、現在オンラインで開催中のプロセッサーLSIの主要国際会議であるHot Chips 33※5（期間：2021年8月22日～8月24日、オンライン開催）に採択され、東工大がポスターセッションで発表する予定です。

2. 今回の成果

まず、CNNの並列演算の計算効率の改善について検討しました。CNNの畳み込み層における畳み込み演算は、複数チャネルからなる入力活性とカーネル群から複数チャネルの出力データを生成する積和演算の繰り返し※6からなっています。互いに依存関係のない入力チャネル内のピクセル位置と出力チャネル方向の座標軸を選ぶと、積和演算を各行・各列で独立の積（直積）と和の計算に分離でき、データ再利用性を高く保つ並列計算が可能になり、計算効率が高くなります（図1）。この方式は、カーネルサイズ1×1のチャネル間畳み込み（いわゆるPoint-wise畳み込み※7）において、特に計算効率が高くなります。

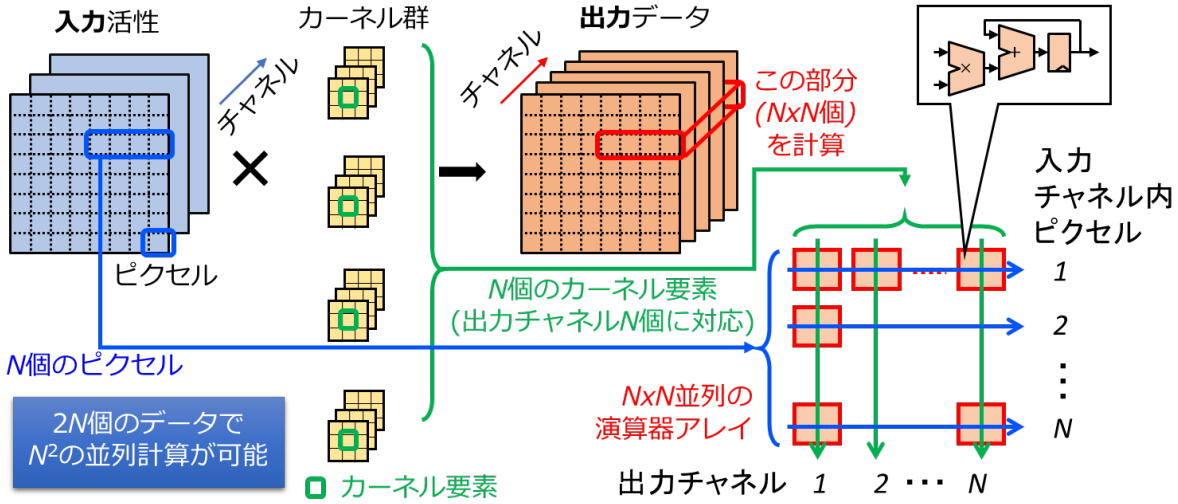

図1　直積型並列演算の概念図

次に、計算効率のアルゴリズムの改善について検討しました。ここでは一般に、あらゆる畳み込み演算は入力データのチャネル内平面シフトとカーネルサイズ1×1のチャネル間（Point-wise）畳み込みの組み合わせに分解できることに着目しました（図2a）。特に、チャネル内とチャネル間にカーネルを分離することで軽量化された畳み込み演算に対しては、入力データのチャネル内平面シフトを扱う整形機構によりチャネル内の畳み込み演算を処理し、直積型並列演算器アレイによりチャネル間の畳み込み演算を処理することで、高いデータ再利用率で処理できることを見いだしました。さらに、畳み込みカーネルをカーネル要素の座標に合わせて枝刈りすると、存在しないカーネル要素に対応する計算をスキップでき、メモリアクセス・演算処理を省けます（図2b）。これにより、深く枝刈りされたCNNの計算効率をより向上させることができました。

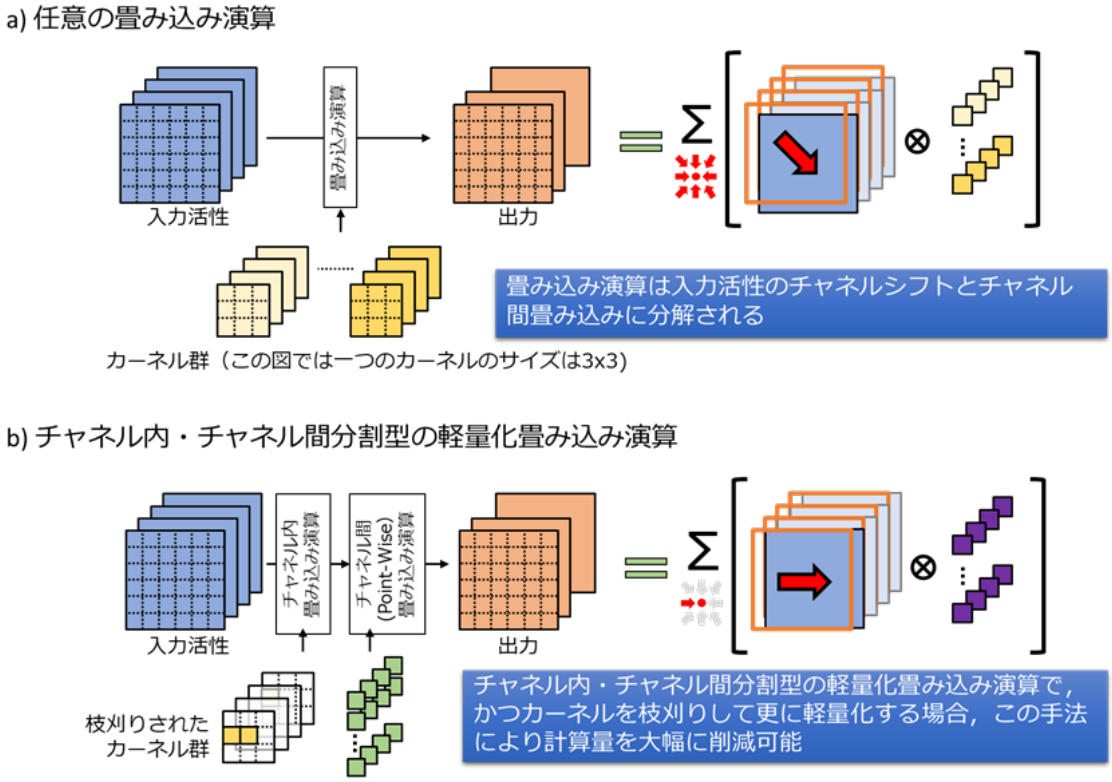

図2　畳み込み演算の分割と枝刈りされたカーネルにおける処理効率化

続いて、既存のCNNモデルを変形し、同一座標のカーネル要素を優先して枝刈りする学習アルゴリズムを構築しました（図3a）。モデルの処理時間・効率と認識精度は、枝刈り後に残存するカーネル要素数（スパース率の逆数）によってトレードオフの関係となります。本アルゴリズムは、既存の学習済みモデルに対して連続的にカーネル要素数を減らしながら再学習を行うもので、求める処理時間・効率と認識精度から、スパース率のトレードオフ点を任意に選択することが可能になりました（図3b）。

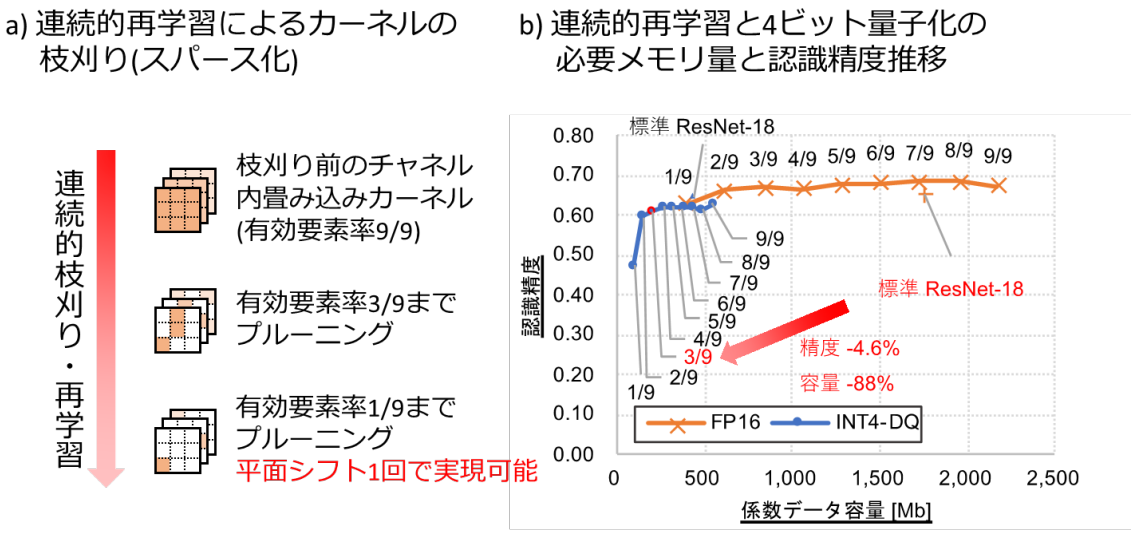

図3　連続的枝刈り・再学習技術とその効果

本アーキテクチャーの試作チップを台湾積体電路製造（TSMC社）の40nmプロセスにて製作しました（図4）。並列演算アレイサイズを32×32とし、活性値・係数値には4ビット固定小数点（INT4）量子化を採用しました。その結果、最大534MHz 1.1Vにおいて400mW以内の電力消費を実測しました。これは、カーネル要素数を9分の1まで枝刈りした後のスパース化した不要カーネル要素の省略を考慮すると、スマートフォンなどのエッジ機器向けCNN推論プロセッサーとしては世界トップレベルの実効効率26.5TOPS／Wに相当します。

本成果は、低電力・高速での大規模CNNモデルの推論処理を可能にし、従来クラウド処理が前提であった大規模CNNモデルによる高度なAIアプリケーションを、電力や処理能力、外部通信データ量の制約が厳しいスマートフォンやロボットなどのエッジ機器でも使用できるようにします。高度なAIアプリケーションとは、例えば、スマートフォンにおけるTPO（Time、Place、Occasion）に応じた賢いAR（Augmented Reality: 拡張現実）アプリケーションや、ロボットにおける柔軟で適応的な動作制御・姿勢制御などです。また、これらのAIアプリケーションにおいて、クラウド処理の課題であったプライバシーやネットワークの不通・遅延などの問題を回避し、スマートデバイスでのAI応用の可能性を広げることが期待できます。

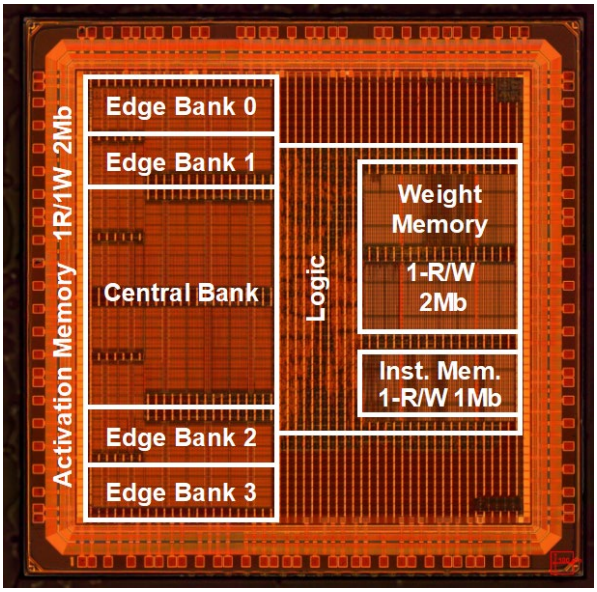

図4　試作されたLSIの顕微鏡写真

3. 今後の予定

エッジ機器で利用されるニューラルネット推論チップは、さらに高度な枝刈りや量子化による小型化が進むものと予想されます。NEDOと東工大の研究チームは、本研究の試作チップで実証した技術をさらに発展させ、枝刈り後の精度向上のための学習技術や、RISC-Vプロセッサーなどとのシステムレベル統合技術の開発など、より高精度・高効率なニューラルネット推論チップの実現を目指し、スマートフォンやロボットなどのエッジ機器での高度なAIアプリケーションの実現を目指します。

【注釈】

※1　畳み込みニューラルネットワーク（Convolutional Neural Network; CNN）

画像などの 2 次元のデータの認識に特によく用いられるニューラルネットワークの一種です。2 次元の平面データを複数チャネルがもつ入力活性に対し、カーネルとの離散畳み込みを適用して特徴を抽出する畳み込み層、入力サイズの縮小と特徴の位置ずれの吸収を行うプーリング層、畳み込み層とプーリング層が抽出した特徴の分類を行う全結合層などから構成されま

す。

※2 枝刈り（プルーニング）

ニューラルネットワークの認識精度を保ちつつ、必要計算量・メモリ容量を削減するために、ニューロン（特徴マップのベクトル次元）やシナプス結合（重み係数の要素）の一部を省略するアルゴリズムです。係数の要素のうち、絶対値の小さいものから一定の割合で 0 に置き換えて再学習するなどの手法が用いられます。

※3 高効率・高速処理を可能とする AI チップ・次世代コンピューティングの技術開発

研究開発項目：革新的 AI エッジコンピューティング技術の開発／動的再構成技術を活用した組み込み AI システムの研究開発

事業期間：2018 年度～2022 年度

※4 TOPS／W

消費電力 1W 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です。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ます。

※5 Hot Chips 33

Hot Chips は毎年 8 月にシリコンバレーで開催される技術シンポジウムです。本年はオンライン開催となります。

公式サイト：https://hotchips.org/

論文タイトル：Edge Inference Engine for Deep & Random Sparse Neural Networks with 4-bit Cartesian-Product MAC Array and Pipelined Activation Aligner

発表者：安藤洸太（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 特任助教）

※6 積和演算の繰り返し

CNN の基本演算である畳み込み演算は、チャネル内平面 XY ピクセル座標とチャネル方向の 3 軸を持つ入力活性の各ピクセルについて、カーネル（重み係数）の XY 座標と入力チャネルの 3 軸で内積計算を行う 6 重ループとなります。このうち、入力 XY 座標と出力チャネルの軸は互いに依存関係にないので、独立に並列化が可能です。

※7 Point-wise畳み込み

畳み込みのうち、カーネル係数の XY サイズが 1×1 であり、チャネル間の畳み込みのみを担当するものは特に Point-wise 畳み込みと呼ばれます。入力活性の XY 平面のピクセル間の相関を考慮しないで、チャネル方向（特徴空間）の変換のみを担います。これとは対照的に、チャネル方向を考慮しないで、チャネル内でのピクセルの変換のみを扱う畳み込みは Depth-wise 畳み込みと呼ばれます。通常の畳み込みを Depth-wise 畳み込みと Point-wise 畳み込みのペアで近似することで、必要メモリ容量の削減を図った CNN モデルとして、これまで MobileNet が提案されています。

4. 問い合わせ先

（本ニュースリリースの内容についての問い合わせ先）

NEDO IoT推進部 担当：広瀬、西山 TEL：044-520-5211

東工大 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

担当：本村真人 教授 TEL：045-924-5654 E-mail：motomura@artic.iir.titech.ac.jp

総務部 広報課 TEL：03-5734-2975 E-mail：media@jim.titech.ac.jp

（その他NEDO事業についての一般的な問い合わせ先）

NEDO 広報部 担当：坂本、橋本、鈴木（美）、根本

TEL：044-520-5151 E-mail：nedo_press@ml.nedo.go.jp