高精度かつ高度なADASの構築に貢献する車載向け低消費電力DCNNハードウェアアクセラレーター

Low Power Consumption Deep Convolutional Neural Network Accelerator for ADAS

石垣　雄太朗 ISHIGAKI Yutaro　田辺　健 TANABE Ken　田邉　靖貴 TANABE Yasuki

自動車の安全な運転操作を実現する先進運転支援システム（ADAS）では，深層学習によるAI技術の導入，特に深層畳み込みニューラルネットワーク（DCNN）処理を適用した認識・識別技術が注目されている。しかし，DCNN処理は，膨大な演算量とデータ入出力量を必要とするため，これまで，消費電力への要求が厳しい車載システムへの採用は困難であった。

東芝デバイス&ストレージ（株）は，メモリーへのアクセス回数とデータ入出力量を削減することにより，DCNN処理を効率的に実行する車載向けハードウェアアクセラレーター（HWA）を開発し，リアルタイムかつ低消費電力でDCNN処理を実現できることを確認した。

Attention has been increasingly focused on the introduction of artificial intelligence (AI) technologies, particularly recognition and classification technologies applying deep convolutional neural networks (DCNNs), that can play a key role in realizing the safety of automobile driving operations using advanced driver assistance systems (ADAS). However, as DCNN processing involves a huge number of calculations and massive volumes of data, it is difficult to apply it to automotive systems with limited power consumption.

To overcome this problem, Toshiba Electronic Devices & Storage Corporation has developed a real-time hardware accelerator (HWA) with low power consumption for automotive applications. This HWA makes it possible to efficiently implement DCNN processing by achieving reductions in the number of memory accesses and the volume of data.

1. まえがき

自動車の安全な運転操作を実現するため，衝突被害軽減ブレーキといったADASを搭載した車両が増えている。人間が主に視覚情報に基づいて自動車を運転しているのと同じように，カメラ画像を入力とした画像認識情報の利用は，ADASの機能拡充に欠かせない。

近年，深層学習によるAI技術の進展，特に画像認識に特化したDCNN処理を適用することで，画像認識の性能が飛躍的に向上しており，高精度かつ高度なADASを構築するためにDCNN処理の適用が注目されている。しかし，その処理には膨大な演算量とデータ入出力量が必要で，演算システムの消費電力が大きくなるという問題があった。例えば，DCNN処理の研究に広く利用されている汎用GPU（Graphic Processing Unit）の消費電力は，100 W以上にも及ぶ。一方，車載システムには高い信頼性が求められ，厳しい温度条件下でも，故障率の高い冷却ファンを用いずに動作できる半導体製品が要求される。このため，発熱を抑えつつ，DCNN処理を低消費電力で実現することが求められる。

東芝デバイス&ストレージ（株）は，この要求に応えるため，消費電力の増大につながるメモリーへのアクセス回数とデータ入出力量を削減し，DCNN推論処理演算を効率的に実行できるHWA（以下，DCNN HWAと略記）と，その専用ツールを開発した。これにより，リアルタイムでのDCNN処理を，冷却ファンなしかつ低消費電力で実現した。ここでは，今回開発したDCNN HWAと専用ツールの概要について述べる。

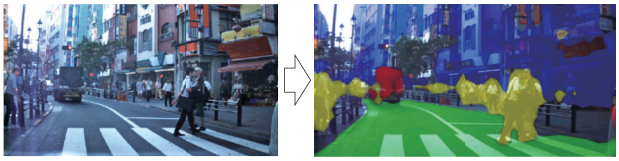

図1. DCNN処理による画像セグメンテーションの例

DCNN処理による高精度かつ高度な画像認識により，道路両端の路側帯を示す白線がない道路でも走路の検出が可能になった。

Example of image segmentation using DCNN

2. ADASへのDCNN処理の適用

ADASへのDCNN処理の適用例の一つとして，ピクセルレベルで画像の分類を行う画像セグメンテーションを**図1**に示す。この例では，入力カメラ画像の各ピクセルについて，

DCNN処理で走路，人，建造物，及びそれら以外を識別し，色でラベル付けをしている。従来の画像技術での走路検出には，道路両端の路側帯を示す白線が必要であったが，DCNN処理を適用することで，白線がない道路でも走路を検出できるようになった。このように，DCNN処理は，従来よりも高精度かつ高度な画像認識を実現できることから，ADASの機能拡充を図るために適用が切望されている。

3. DCNN HWA

DCNN処理は，**図2**に示すように多数の処理層の組み合わせで構成される。特に，畳み込み層と全結合層には膨大な積和演算が必要になる。また，この積和演算は大量のデータ（入力画像データ，層間の中間データ，及び重みデータ）を扱うので，膨大なメモリーバンド幅も必要になる。例えば，画像認識用ネットワークであるVGG-16[(1)]の推論処理では，約15.5×10^9回の積和演算と約138×10^6個の重みデータが，認識対象の画像領域ごとに必要となる。

そこで，このようなDCNN処理をリアルタイムかつ低消費電力で実行するために，今回，DCNN HWAとDCNN HWAコンフィグレーション生成ツールを開発した。

3.1 DCNN HWAの構成

開発したDCNN HWAの構成を**図3**に示す。このDCNN HWAは，①LSI外部に接続されているDRAMに出力データを書き込んだり，DRAMから入力データを読み込んだりするDMA（Direct Memory Access）ユニット，②入出力データや中間データを保持するための1 Mi（メビ（2^{20}））バイトの内部メモリー，③DCNN処理の主要な演算を行うための実行ユニット，④全体の制御を行う制御ユニット，及び⑤処理内容の設定を保持する設定バッファー，から構成される。

DMAユニットは，実行ユニットが演算処理を行っている間も，並行してDRAMと内部メモリー間のデータ転送を行う。この並列動作で，実行ユニットのデータ待ち状態を削減し，実行ユニットの稼働率を向上させてDCNN処理を効率的に実行することを可能にした。

内部メモリーは，DRAMよりもアクセス時の消費電力が小さいSRAM（Static RAM）で構成されており，DCNN HWA上の演算は，この内部メモリーにアクセスしつつ実行される。しかし，アクセス回数が多いとその分だけ消費電力は増大するため，内部メモリーについても効率化は重要である。

そこで，実行ユニットは，内部メモリーへのアクセス回数を削減するように設計した。実行ユニットは，**図4**に示すように，四つのPE（Processing Element）で4チャネル分のデータを並列処理する。畳み込み層における各チャネルでのデータ演算時の入力データは同一なので，内部メモリーから読み込んだ入力データは4PE間で共用し，内部メモリーへの総アクセス回数を削減した。また，各PE内では，

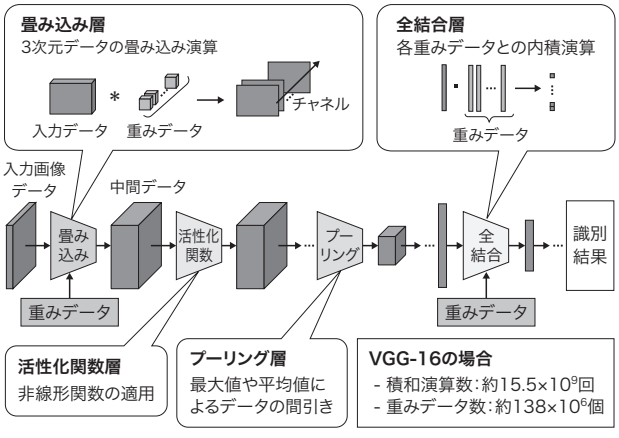

図2. 典型的なDCNNの構造

DCNN処理は，多数の処理層の組み合わせで構成され，膨大なデータと積和演算が必要になる。

Typical architecture of DCNN

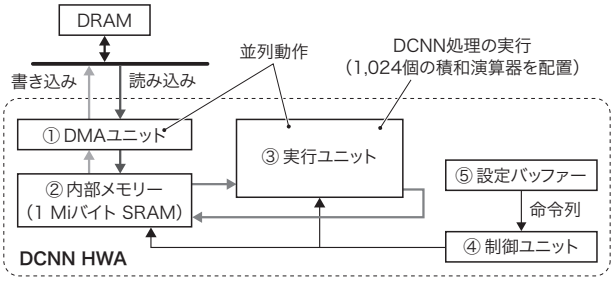

図3. DCNN HWAの構成

実行ユニットとDMAユニットの並列動作によって，効率良くDCNN処理を実行できる。

Architecture of HWA for DCNN processing

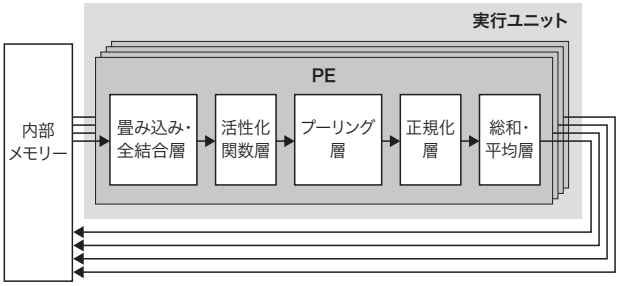

図4. 実行ユニットの構成

DCNN処理で主要な演算を行う回路をカスケード接続し，内部メモリーへのアクセス回数を削減した。

Architecture of execution unit

DCNN処理において頻出する各層の処理を行う演算回路をカスケード接続し，カスケード接続された層間では，中間データの内部メモリーでの書き込み／読み込みを不要にして内部メモリーへのアクセス回数を削減した。更に，PE内の畳み込み・全結合層ユニットには，畳み込み層と全結合層の膨大な積和演算をリアルタイムに処理できるように，四つのPE合計で1,024個（一つのPE当たり256個）の16ビット浮動小数点精度の積和演算器を搭載した。

3.2 DCNN HWAコンフィグレーション生成ツール

パソコン（PC）上で学習したモデルのDCNN HWAへの移植作業を効率化するため，学習済みモデルの変換を行うDCNN HWAコンフィグレーション生成ツールを開発した。開発したツールを用いたDCNN処理フローを**図5**に示す。このツールは，オープンソースの深層学習フレームワークであるCaffe[2]のネットワーク定義と重みデータを入力とし，DCNN HWA用の実行バイナリー（DCNN HWAコンフィグレーションバイナリー）を生成する。これは，DCNN HWAを制御する設定と，DCNN HWA用に変換された重みデータから構成される。DCNN HWAは，このDCNN HWAコンフィグレーションバイナリーと入力データである画像データを受け取り，DCNN処理を実行して推論結果を出力する。

このツールは，(1)DCNN HWAが持つ1 Miバイトの内部メモリーに，演算に必要なデータが収まるようにネットワークを分割し，(2)複数の層（演算）を統合（部分ネットワーク統合）して実行することで，DRAMアクセスのデータ量を削減する最適化を行っている。以下でそれぞれについて説明する。

(1) ネットワーク分割　DCNN処理で必要となる膨大なデータの全てを，DCNN HWAの1 Miバイトの内部メモリーに，同時に配置することはできない。そこで，このツールがネットワークを自動分割する。分割方法としては，DCNNの中間データを水平・垂直方向にタイル状に分割する“タイル分割”と，重みデータを出力チャネルごとに分割する“重み分割”を行う。**図6**にタイル分割の例を示す。最適化前ネットワーク（図6(a)）の演算bの実行に必要なデータAと実行結果のデータBの合計が1 Miバイトを超える場合は，内部メモリーに同時に配置することはできない（図6(b)）。すなわち，このままではDCNN HWAで処理できない。そこで，ネットワークを分割し，内部メモリーに配置できるようにする。図6(c)の例では，水平方向にタイルを2分割する。そして，データAの一部であるA1をDMAユニットにより内部メモリー上に読み込み（図6(d)の①），実行ユニットにより演算b1を実行し，データB1を得て，B1をDMAユニットでDRAM上に書き込む（図6(d)の②）。同様に，データA2を演算し，得られたB2をDRAMに書き込む（図6(d)の③，④）ことで，タイル分割前の

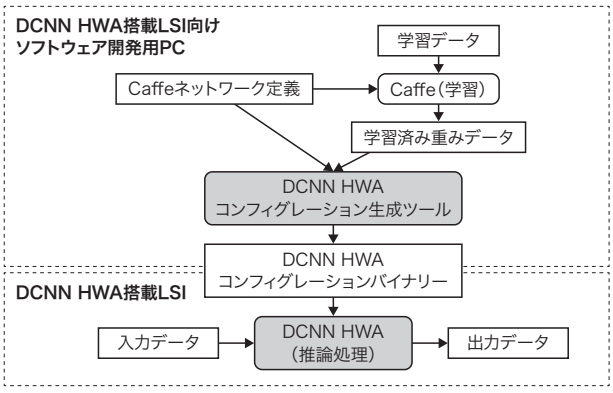

図5. DCNN HWAコンフィグレーション生成ツール

開発用PC上で学習したネットワーク定義と重みデータを，DCNN HWAコンフィグレーションバイナリーに変換する。

DCNN HWA configuration generator tool

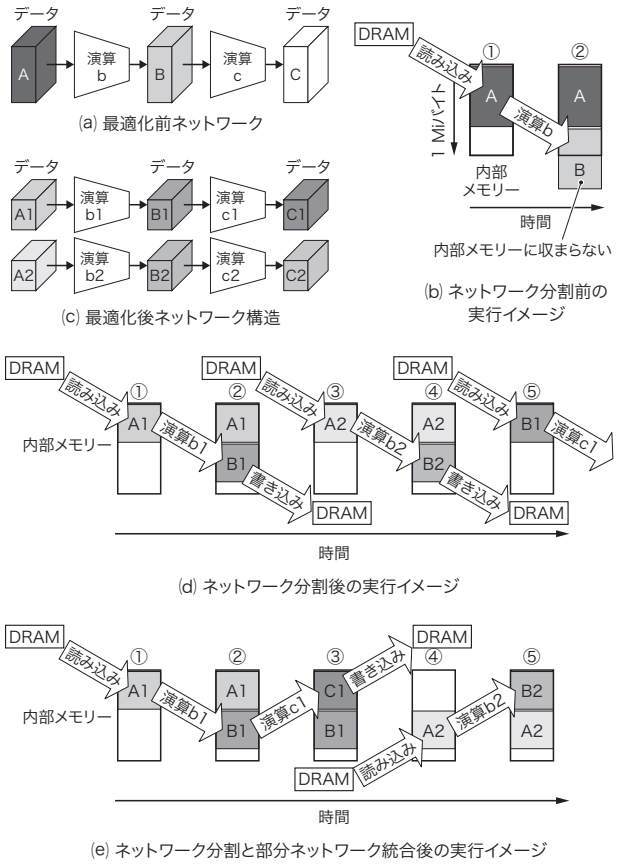

図6. ツールによるデータ量削減のための最適化例

ネットワーク分割とネットワーク統合により，膨大なデータが必要となるDCNN処理を，DCNN HWA上で効率良く実行できるようにした。

Example of optimization for reduction in amount of data using DCNN HWA configuration generator tool

データBを得る。

(2)　部分ネットワーク統合　　複数の層間の中間データを演算の都度DRAMに書き込むと，DRAMアクセスに伴うデータ量が増え，消費電力が増大する。そこで，このツールは，(1)で分割したネットワークで複数の演算を統合し，中間データをDRAMに書き込まず，内部メモリーに配置したまま処理を継続できるように最適化する。図6(d)では，演算b1の結果であるデータB1は演算c1で利用される。しかし，演算b1，演算c1ともにタイル分割されており，データB1をDRAMに書き込み（図6(d)の②），再度DRAMから読み込む（図6(d)の⑤）必要がある。そこで，図6(e)に示すように，演算b1と演算c1で部分ネットワークを構成し，演算c1は演算b1によって得られた内部メモリー上のB1を利用して演算するように統合する（図6(e)の③）。その結果，データB1とデータB2は，DRAMでの書き込み／読み込みが省略され，DRAMアクセスに伴うデータ入出力量が削減される。

4. DCNN処理実行時の消費電力測定

今回開発したDCNN HWAを搭載したLSI（試作チップ）では，DCNN HWA単体の消費電力が約1.5 Wであることを確認した[3]。また，試作チップでのDCNN処理による画像セグメンテーションの実行例を，**図7**に示す。このときの評価用ボード全体の消費電力（AC（交流）アダプターでの損失，及び周辺インターフェースやDRAMチップなどの周辺デバイスでの消費電力を含む）は，約6 Wであった。GPUは，膨大なメモリーバンド幅と大量の積和演算器を備え，かつ高い汎用性を持つことから，消費電力は100 Wクラスになる。一方，開発したDCNN HWAは，DCNN処理に特化した回路構成とメモリーアクセスの効率化で，冷却ファンなしでも動作可能な低消費電力でDCNN処理が実現できる。

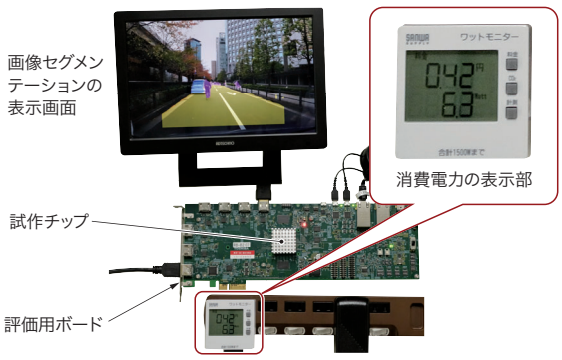

図7. DCNN処理実行時の評価用ボードの消費電力測定
試作チップで画像セグメンテーションを実行し，評価用ボード全体の消費電力が約6 Wであることを確認した。
Measurement of power consumption of evaluation board when running image segmentation on DCNN HWA

5. あとがき

DCNN HWAとその専用ツールを開発した。このDCNN HWAは，DRAMアクセスと演算の並列動作，回路構成，及びツールによって，メモリーへのアクセス回数とデータ入出力量を削減し，DCNN HWA単体で約1.5 W，評価用ボード全体で約6 Wと，ファンレスでのDCNN処理を実現した。このDCNN HWAは，当社の車載向け画像認識プロセッサー Visconti5に搭載予定であり，DCNN処理を利用した高精度かつ高度なADASの構築に貢献していく。また，低消費電力でDCNN処理を実現したいというニーズは車載分野以外でも多いことから，今後，このDCNN HWAを様々な製品向けに展開していく。

文　献

(1) Simonyan, K.; Zisserman, A. "Very Deep Convolutional Networks for Large-Scale Image Recognition". Proceedings of 3rd International Conference on Learning Representations (ICLR 2015). San Diego, CA, 2015-05, ICLR. arXiv.org e-Print archive, 2015, arXiv:1409.1556v6. <https://arxiv.org/pdf/1409.1556.pdf>, (accessed 2019-06-05).

(2) Jia Y. et al. "Caffe: Convolutional Architecture for Fast Feature Embedding". Proceedings of the 22nd ACM international conference on Multimedia (ACM Multimedia 2014). Orlando, FL, 2014-11, Association for Computing Machinery. 2014, p.675–678.

(3) Yamada, Y. et al. "A 20.5TOPS and 217.3GOPS/mm² Multicore SoC with DNN Accelerator and Image Signal Processor Complying with ISO26262 for Automotive Applications". 2019 IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers. San Francisco, CA, 2019-02, IEEE. 2019, p.132–134.

石垣　雄太朗　ISHIGAKI Yutaro
東芝デバイス&ストレージ（株）デバイス&ストレージ研究開発センター エンベデッドコア技術開発部
IEEE・電子情報通信学会会員
Toshiba Electronic Devices & Storage Corp.

田辺　　健　TANABE Ken
東芝デバイス&ストレージ（株）デバイス&ストレージ研究開発センター エンベデッドコア技術開発部
Toshiba Electronic Devices & Storage Corp.

田邉　靖貴　TANABE Yasuki, Ph.D.
東芝デバイス&ストレージ（株）デバイス&ストレージ研究開発センター エンベデッドコア技術開発部
博士（工学）
Toshiba Electronic Devices & Storage Corp.