組込み向けディープラーニングの高速化およびリアルタイム性能保証

High-speed and Real-time Processing for Deep Learning on Embedded System

西　村　　隆*
Takashi Nishimura

ヘテロジニアスなマルチコアアーキテクチャによる複数アルゴリズムの同時動作と，帯域保証機構を用いたメモリアクセスの競合抑制によるリアルタイム性能保証により，ディープラーニングを用いたSoC（System-on-a-Chip）上の画像認識処理の高速化を実現した.

We achieved the speedup of image recognition processing using Deep Learning on System-on-a-Chip (SoC) by simultaneous processing of multiple algorithms with a heterogeneous multi-core architecture and real-time processing with memory access bandwidth control that suppress the conflict.

1. ディープラーニングの組込み実装

ディープラーニングを用いた画像認識はノイズや遮蔽，物体の傾きなどの複雑なシーンを高いロバスト性で高精度に認識できるとともに，学習モデルの更新により認識対象を追加可能なスケーラビリティをもつ技術である[1]. ADAS（Advanced Driver Assistance Systems）をはじめとする車載システムへの組込み実装では，通信途絶の発生し得る環境下の利用を考慮し，組込み機器の限られたリソース上での低遅延応答が要求される．一方でディープラーニングを用いた信号処理は膨大な演算量とメモリアクセスが発生するため[2]，高速化およびリアルタイム性能保証が課題となる.

この課題に対し，ディープラーニング処理をターゲットとするAIプロセッサを搭載したSoC上のソフトウェアとハードウェアの協調設計により以下の対応を行う.

- ヘテロジニアスなマルチコアアーキテクチャによる複数アルゴリズムの同時動作と高速化
- 帯域保証機構を用いたメモリアクセスの競合抑制によるリアルタイム性能保証

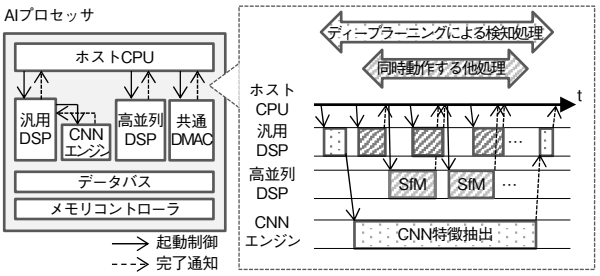
AIプロセッサを用いた協調設計は，リソース競合を抑制しつつハードウェアの稼働率を上げることで，GPU（Graphics Processing Unit）ベースのSoCへの実装では困難であった演算効率の向上を実現させる[3].

2. ヘテロジニアスコアを用いた高速化

ディープラーニングによる画像認識は，いずれもCNN（Convolutional neural network）による特徴抽出が演算の中核となり，分類処理や枠統合などの逐次的な信号処理と組み合わせて機能を実現する．**第1図**に示すAIプロセッサは2種のDSP（Digital Signal Processor）と専用ハードウェアであるCNNエンジンを搭載したヘテロジニアスなマルチコアアーキテクチャとすることで，負荷の高いCNN処理を高速かつ効率良く稼働させつつ，ソフトウェアによる柔軟性を提供する.

このアーキテクチャは，各演算コアがマスタとなり，直接もしくはDMAC（Direct Memory Access Controller）を介してメモリアクセスを行う自走可能なハードウェアモデルとすることで，複数のアプリケーションを同時に動作させることを可能とする.

ホストCPUのシーケンス制御の例を**第1図**に示す. ディープラーニングによる検知処理は，汎用DSPとCNNエンジンがマスタとなりホストCPUの介在なしにフレーム内処理を進める．一方で同時動作するSfM（Structure from Motion）処理は, 1フレームの画像を複数ブロックに分割したタイル単位でホストCPUが汎用DSPや高並列DSPを制御し，パイプライン化することで高スループット性能を実現する.



第1図　ヘテロジニアスコアのシーケンスモデル
Fig. 1　Sequence model of heterogeneous architecture

3. メモリアクセス競合抑制による性能保証

ディープラーニングを用いた画像認識を含め，複数のアルゴリズムを同時動作させるADASシステムでは，データ転送路やメモリアクセス帯域など共有リソースの競合による性能オーバヘッドが課題となる．特にディープラーニン

* オートモーティブ社　開発本部
R&D Div., Automotive Company

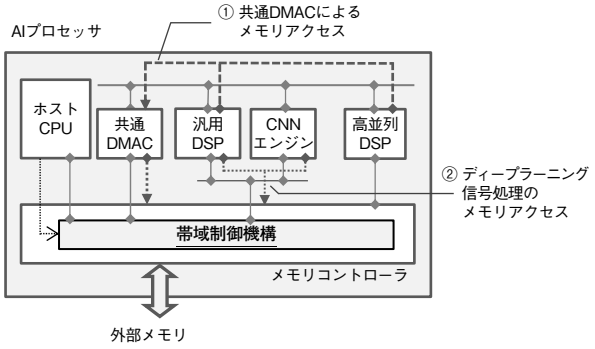
グを用いた信号処理は，膨大かつ細粒度の外部メモリアクセスを発生させる特性をもつため，システム性能への影響が大きい．

本章ではメモリアクセス競合を抑制するためのデータ転送フロー設計とメモリコントローラによる帯域制御を述べ，リアルタイム性能保証のための性能設計の応用例を示す．

3.1 データ転送フロー設計

第2図はAIプロセッサをメモリアクセスフローの観点で示したものである．複数コア間のデータ受け渡しをホストCPUが制御する信号処理では，複数の論理チャネルをもつ共通DMACによりメモリアクセスを一元管理し（**第2図①**），パイプライン処理の効率化をはかる．

データ転送量が多く細粒度なデータアクセスが頻発するディープラーニングの信号処理では，転送コマンドの独占により他コアの転送待ちに影響が及ぶのを防ぐため，汎用DSPもしくはCNNエンジンに内蔵したDMACを用いて独立したバス経路（**第2図②**）でメモリアクセスを行い，データ転送の輻輳（ふくそう）を抑制する．



第2図　データ転送フローと帯域制御機構
Fig. 2　Data flow and bandwidth control mechanism

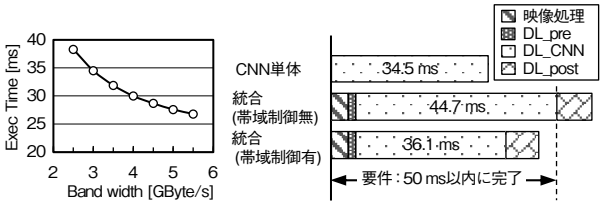
3.2 メモリコントローラによる帯域制御

ディープラーニングを用いた信号処理は，ニューラルネットワークの設計パラメータにより，メモリアクセス帯域の局所ピークの値や間隔がそれぞれ異なる．**第2図**に示すAIプロセッサのメモリコントローラは，各コアと外部メモリ間のデータ転送帯域を独立に制御する機構を備え，あらかじめ設定した値に従いデータ転送を調停することでメモリアクセスの性能設計を可能とする．CNNエンジンから一時的に大量のメモリアクセスが発生した場合は，帯域設定を超過する転送コマンドに対して優先度を下げることで他コアのデータ転送を保証する．

3.3 性能設計の応用例

ADAS向け検知アルゴリズム実装を例として，ディープラーニング信号処理のメモリアクセス帯域と処理速度のトレードオフ評価結果を**第3図（a）**に示した．この評価は性能保証のための帯域値や，演算とメモリアクセスのボトルネック解析の手がかりとすることができる．

第3図（b）はCNNエンジン単体動作時の処理性能とシステム統合時の処理性能を比較した結果である．帯域制御機構を活用することによりシステム統合へ移行した際の性能劣化を5 %未満に抑制することで要件とする時間内に処理を完了している．



(a) 帯域と処理速度トレードオフ　　(b) 単体性能と統合システム性能

第3図　ディープラーニング信号処理の速度評価
Fig. 3　Processing performance evaluation of deep learning

4. 動向と展望

車載分野でのディープラーニングを用いた画像認識技術の適用は年々増加しており，カメラ入力から車両駆動までの低遅延応答や処理フレームレート向上の要求はますます高まっている．本稿で述べたソフトウェアとハードウェアの協調設計はSoCに依存せずに適用可能な手法であり，組込み機器全般での処理高速化やリアルタイム性能保証の実現へ貢献できると考える．

参 考 文 献

[1] A. Krizhevsky et al. “Imagenet classification with deep convolutional neural networks”. Proc. NIPS, 2012.

[2] A. Canziani, et al. “An Analysis of Deep Neural Network Models for Practical Applications”, Proc. ICLR, 2017

[3] John L. Hennessy et al., Computer Architecture: A Quantitative Approach, Sixth Edition, Morgan Kaufmann Publishers, Burlington, 2017, Chapter 7.