Vision AI の大規模化・複雑化を高い電力効率で対応できるミッドレンジ MPU RZ/V2N

ルネサスエレクトロニクス エンベデッドプロセッシングプロダクトグループ
エンベデッドプロセッシング事業部 プロダクトマネジメント第二部 **戸井崇雄**、**下別府正行**、**野瀬浩一**

ルネサスエレクトロニクス ソフトウェア&デジタライゼーショングループ
ソフトウェア開発統括部 システムソリューション開発第一部 **三上顕太郎**

概要

エンドポイントデバイス内で高度な AI(Artificial Intelligence)を行うシステムは、特にリアルタイム性が求められるロボット、AI カメラ、ドローンなどの用途を中心に需要が高まっています。また、より高度かつ多様なタスクを AI で処理可能にするため、処理性能や電力効率の高度化がますます期待されています。

ルネサスは、こうした需要に応えるため、高効率 AI アクセラレータ(DRP-AI)を搭載した Vision AI 向け MPU(Micro Processing Unit)である RZ/V シリーズを提供しています。本ホワイトペーパーでは、RZ/V シリーズの最新製品の概要や、ハードウェア・ソフトウェア協調による AI 処理の最適化技術などを紹介します。

エンドポイント AI のトレンドと課題

エンドポイント AI は近年急速に進化しており、さまざまなアプリケーションでその利用が広がっています。クラウドでの AI 処理には膨大な通信量や通信時間が伴うため、特にリアルタイム性が求められるタスクにおいて、特に Vision AI の分野では、クラウドからエッジへの移行が進んでいます。

エンドポイント AI のトレンド

リアルタイム性向上: 従来のクラウド AI では、大量のデータと豊富な計算リソースをフルに活用していましたが、エンドポイント AI では、データをローカルで処理することで通信遅延を最小限に抑え、リアルタイム性を向上させることができます。これにより、センサーデータをリアルタイムで処理し、即座にフィードバックを提供することが可能となります(図 1)。

AI タスクの多様化: 犬猫の識別から物体認識といった画像認識の世界から、静止画から 3D の奥行を予測するなど、さまざまな予測タスクにも対応するようになっています。特に近年は、認識 AI の精度向上に加え、計画・判断する AI（たとえばロボットの最適動作を計画する AI や、環境や人の行動を基にエネルギー消費を最適化するスマートビルディング）など、多岐にわたる応用開発が進んでいます。

モデルの軽量化: AI モデルの特性を考慮した軽量化技術や、メモリアクセスの効率化などが進められており、これによりエンドポイントデバイスでの高効率な AI 処理が可能となっています。

CNN モデルからトランスフォーマモデルへの発展: CNN(Convolutional Neural Network)は画像認識に特化したモデルですが、トランスフォーマモデルは長距離依存関係を効率的に処理できる自己注意機構(Self-Attention)により、画像認識の認識精度向上に加え、時系列情報や自然言語処理など広範なタスクに対応可能になります。

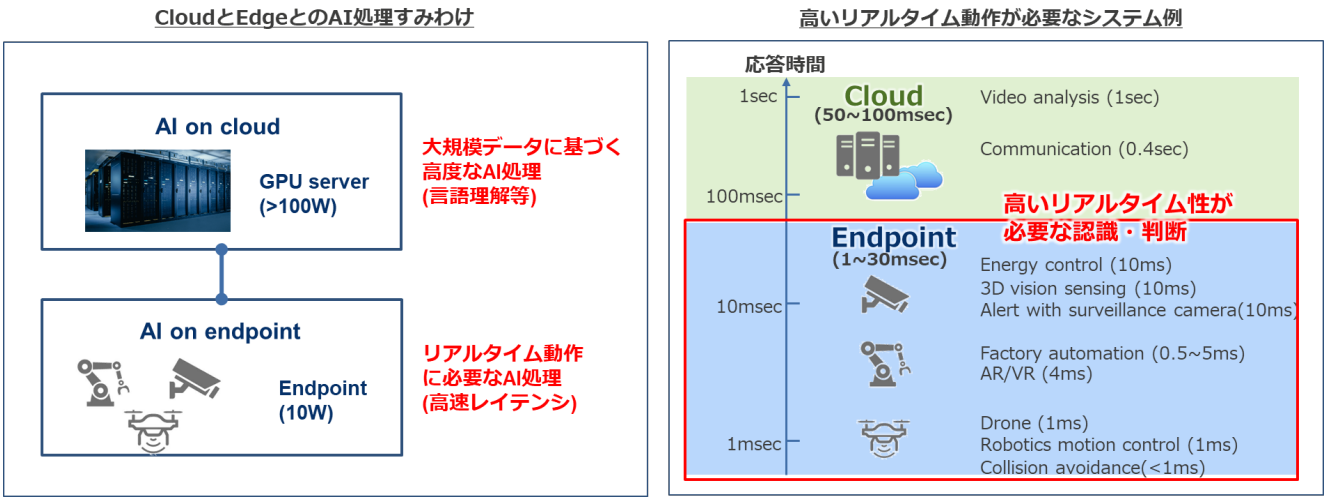


図 1: エンドポイント AI の位置づけ

従来のエンドポイント AI ソリューションの課題

1）AI モデルの大規模化：もともと、画像 AI は従来の画像処理アルゴリズムに比べて 2 桁から 3 桁多い積和演算が必要とされており、高度な AI 処理性能を持つ専用の AI チップや AI アクセラレータが開発されてきました。しかしながら、年々、より高い認識精度を出すために AI モデルの大規模化が進んでおり、その分、画像 1 枚当たりに求められる演算量も増加しています(図 2)。

2）AI モデルの複雑化：AI モデルは年々複雑化しており、特にトランスフォーマモデルのような高度なアーキテクチャが登場しています。これにより、モデルのトレーニングや推論に必要な計算リソースが増加し、エンドポイントデバイスでの実装が難しくなっています。さらに、複雑なモデルはメモリ使用量も増加し、エンドポイントデバイスの限られたリソースでは対応が難しい場合があります。

3）AI ツールの難しさ：特にエンドポイントデバイス向けの最適化には高度な専門知識が必要なため、AI ツールは専門的で複雑化している傾向にあります。一方で、組み込み機器のシステムやアプリを開発するユーザの AI 専門性や要望は多種多様であり、幅広いユーザに対応するための AI ツール環境の整備が不可欠です。また、AI モデルの選定や最適化、デバイス実装には多くの手間がかかり、異なるツール間の互換性の問題や、特定のハードウェアに最適化されたツールの不足も課題となっています。

モデルの大規模化・複雑化により消費電力のさらなる増大が懸念される中で、軽量化されたモデルを高い電力効率で処理可能な DRP-AI3 を搭載する RZ/V2N は、課題解決が可能な新たな製品として、エンドポイントでの AI 実装に大きく寄与できます。

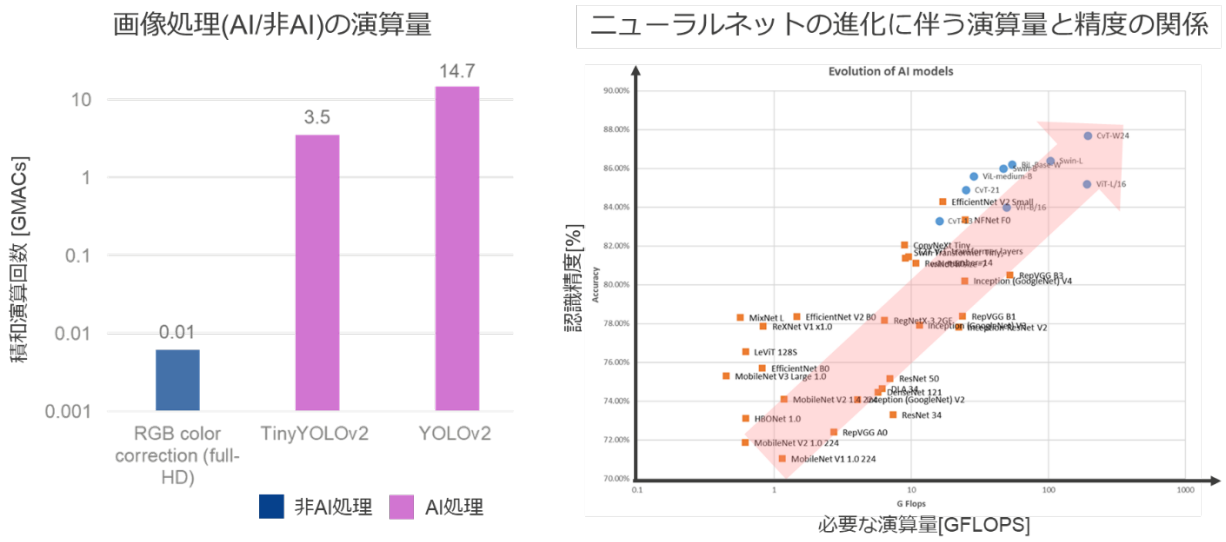


図 2: エンドポイント AI ソリューションの課題

Vision AI に最適な RZ/V シリーズの特徴

エンドポイント Vision AI アプリケーション向け MPU ファミリー

RZ/V シリーズは、ルネサスオリジナルの高効率 AI アクセラレータ（DRP-AI）を採用した Vision AI 向け MPU です。ルネサスの動的再構成プロセッサ（DRP: Dynamically Reconfigurable Processor）を用いることで、多様な AI 演算を効率よく処理することができます。最新の AI アクセラレータである DRP-AI3 を搭載した製品としては、最大 80TOPS のハイエンド向け AI MPU である RZ/V2H に加え、2025 年 3 月からは最大 15TOPS の性能を持つ RZ/V2N の量産も開始し、さまざまなエンドポイント AI アプリケーションに対応するための AI 性能スケーラビリティを提供します。

また RZ/V シリーズは、組み込み製品における AI の電力効率を最大限に引き出す設計がされているため、発熱も低く抑えることができます。たとえば RZ/V2H はファンを使用せずとも、大型ファンを使った組み込み GPU とほぼ同じ温度に抑えられているため、設置サイズや発熱の制限が厳しい組み込み機器にも適用可能になります。電力効率は最大で約 10TOPS/W と、他社を大きく上回る効率を実現しています。

ターゲットアプリケーション

RZ/V シリーズは、エンドポイント AI 市場の中でもハイレンジ（数 10TOPS）からミッドレンジ（1TOPS から 10TOPS）の市場を主なターゲットとしています(図 3)。ハイレンジ向けの RZ/V2H は、主に協働ロボット、AGV、ドローンといった高いリアルタイム性や高度な AI が必要な用途向け、ミッドレンジ向けの RZ/V2N は、主に DMS(Driver Monitoring System)、モバイルロボット、AI カメラといった用途向けです。 コスト効率の高い RZ/V2N は、手頃な価格で高い AI 性能を求めるミッドレンジ AI 市場アプリケーションに最適です。

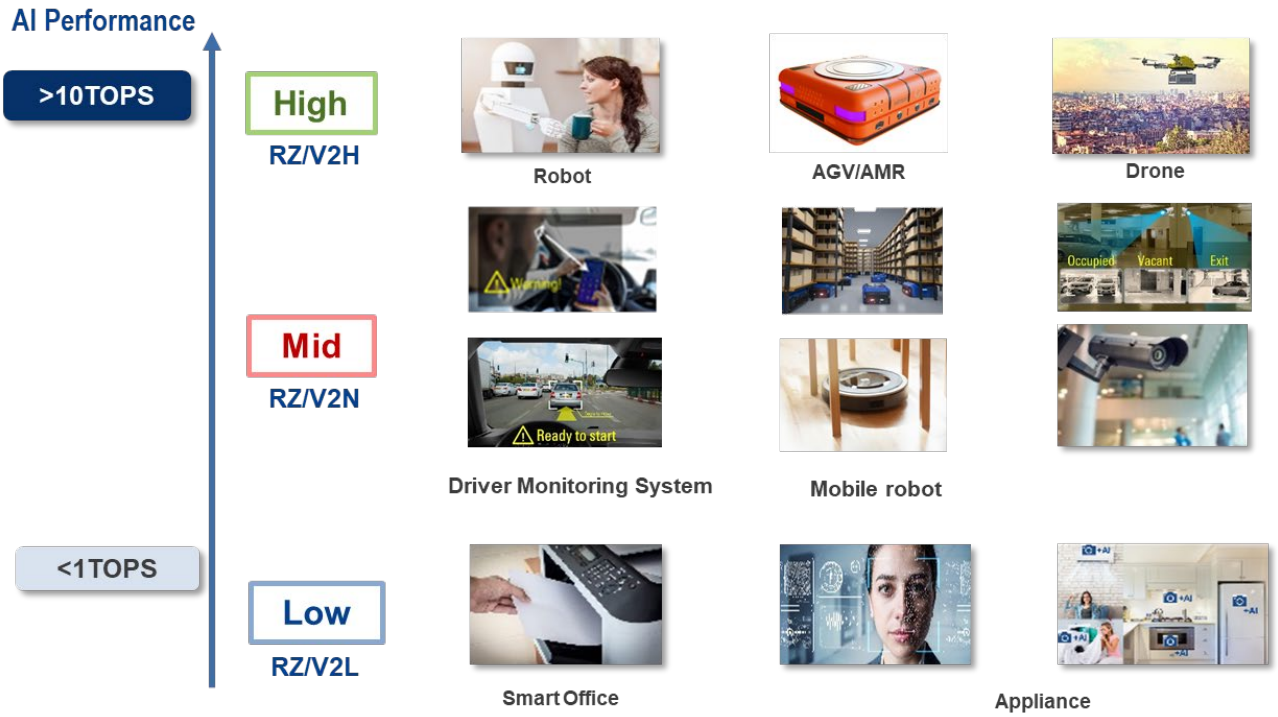


図 3: RZ/V シリーズのターゲットアプリケーション

AI アクセラレータ(DRP-AI)の特徴

DRP-AI は、前述の RZ/V シリーズに搭載されているルネサスオリジナルの AI アクセラレータであり、下記のような技術により高性能と柔軟性の両立を実現します。

幅広い AI 処理に柔軟に対応：画像認識 AI（CNN）処理向けに最適化された AI 積和演算ユニット（AI-MAC）と動的再構成プロセッサ（DRP)との協調動作(図 4)により、大規模化・多様化する Vision AI 処理における高性能と柔軟性の両立を実現します。さらに、AI 処理前の画像処理を DRP で実行することで、システム全体の高速化を実現できます。

AI モデル軽量化に対応：認識精度に影響がない演算を削減した枝刈りモデルを高速処理するルネサス独自の軽量化対応技術などにより、突出した電力効率(最大で約 10TOPS/W)を実現します。

ハードウェア・ソフトウェア協調最適化：ルネサスが提供する専用 DRP-AI ツールにより、DRP-AI 向けに処理を最適化した実行データを生成できます。

本ホワイトペーパーでは、主に DRP-AI ツール関連技術を中心に紹介します。DRP-AI ハードウェア技術の詳細については、DRP-AI のホワイトペーパーおよび ISSCC2024 の論文[1]を参照ください。

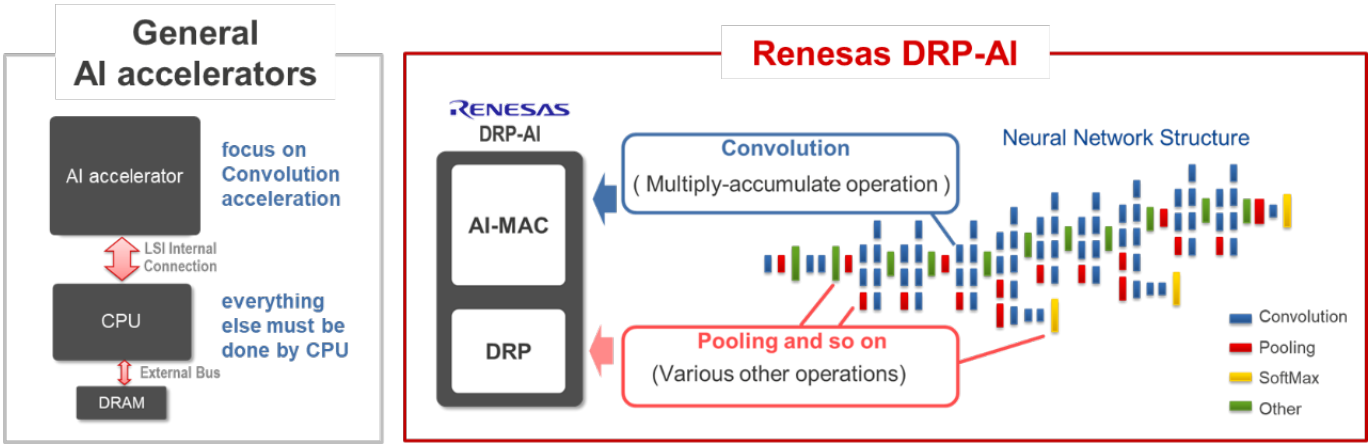


図 4: DRP(Dynamically Reconfigurable Processor)を搭載した AI アクセラレータ(DRP-AI)

DRP-AI ツール構成

ルネサスは顧客の AI アプリケーションの開発を支援するため、AI 初心者からエキスパートまで幅広いユーザが利用できる DRP-AI 向けの AI 開発ツールを提供しています。本章では、これらの開発ツールの全体構成や特徴を紹介します。また後半では、本ツールの特徴であるモデル軽量化を行う際のテクニックなどを紹介します。

DRP-AI ツールの特徴

ルネサスは、DRP-AI アプリケーションの開発環境として、図 5 で示した 2 つの AI ツールの利用形態を提供しています。これにより AI 初心者からエキスパートまで、幅広いユーザが簡単に利用可能になっています。

無料の AI アプリケーションを利用するユーザ: 事前にトレーニングされたモデルを使用した多様な AI アプリケーションをオープンソースで提供しています。ほとんどのアプリケーションは AI モデルの実行ファイルを含んでおり、そのままデバイス上で動作させることができます(図 6 の AS IS or Custom Application)。また、ユーザのデータセットを使用してモデルを再トレーニングする Transfer learning tool (TLT)を提供しており、ユーザは利用シーンに合わせてモデルをカスタマイズすることもできます（図 6 の Re-train RZ/V AI Apps）。詳細は AI applications の Github (https://renesas-rz.github.io/rzv_ai_sdk/latest/) をご参照ください。

独自の AI モデルを利用するユーザ: ユーザのカスタムモデルの実装もサポートしています(図 6 の Custom Model)。DRP-AI 向けに最適化された AI コンパイラ（DRP-AI TVM）を使用することで、ユーザのカスタムモデルを RZ/V 上で高効率に動作する実行ファイルを生成することが可能になります。さらに、カスタムモデルの軽量化をサポートするツール(DRP-AI Extension Pack)も再学習向けに提供しています。

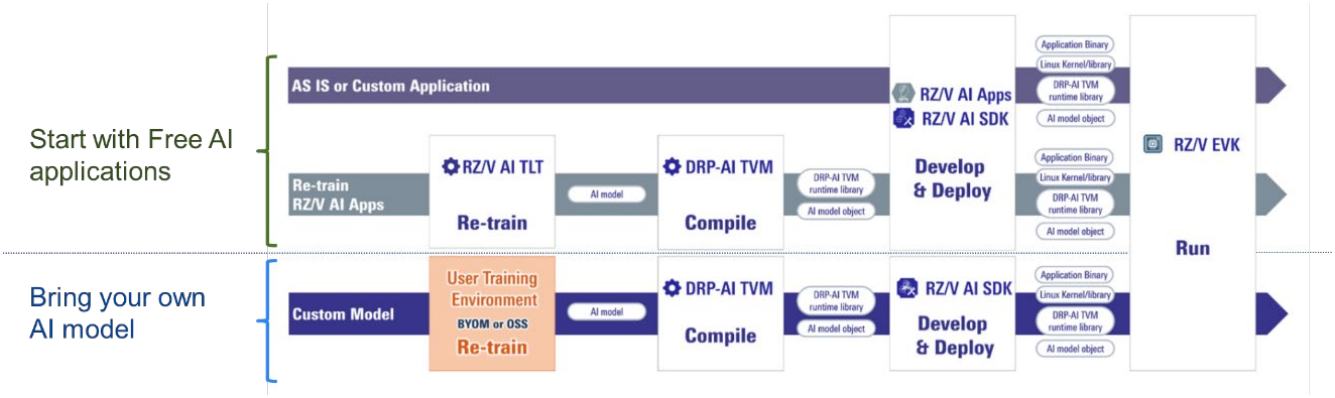


図 5: ユーザ用途ごとの DRP-AI ツール構成

ユーザ独自 AI モデル(BYOM)の実装フロー

BYOM 向けのモデル実装フロー(図 6)は、以下の特徴があります。

多様な Vision AI モデルに対応: 進化の著しい AI モデルを広く活用できるよう、デバイス実装ツールや軽量化ツールは幅広い AI モデルやタスクに対応できる構成となっています。

多種 Framework 対応: PyTorch や TensorFlow など、さまざまな AI フレームワークに対応しています。

RZ/V 製品間で統一の API: RZ/V シリーズの製品間で統一された API を提供し、ユーザの AI アプリケーション開発の一貫性を確保します。

具体的な処理ステップは以下になります。

1) DRP-AI Extension Pack によって DRP-AI サポートが追加された AI フレームワーク(Pytorch あるいは Tensorflow)上で再学習を行い、軽量化（枝刈り）されたモデルを作成します。また、モデルの量子化（32 ビットを 8 ビットに削減）によって低下する認識精度を復元するための再学習(QAT: Quantization Aware Training)にも対応するようになりました（枝刈りモデルや QAT 対応は、RZ/V2H および V2N 向けのオプション）。
2) DRP-AI TVM に、モデルとキャリブレーションデータ(ユースケースに合わせた画像セット)を入力します。
3) DRP-AI TVM は、DRP-AI のハードウェア構成に最適な処理フローやメモリアクセスの方法などを見積もり、デバイス上での実行ファイル（ランタイム）データを生成します。
4) DRP-AI TVM 内で自動的に、キャリブレーションデータを使ったモデルの量子化（32 ビット浮動小数点型から 8 ビット整数型に変換）を行います。軽量化されたモデルはホスト PC 上でシミュレーションを行う事で、元の 32 ビットモデルでの推論結果と比較確認することも可能です（インタプリタモード）。

5) ランタイムデータを RZ/V2H に実装（デプロイ）することで、デバイス上で AI の推論を実行できます。

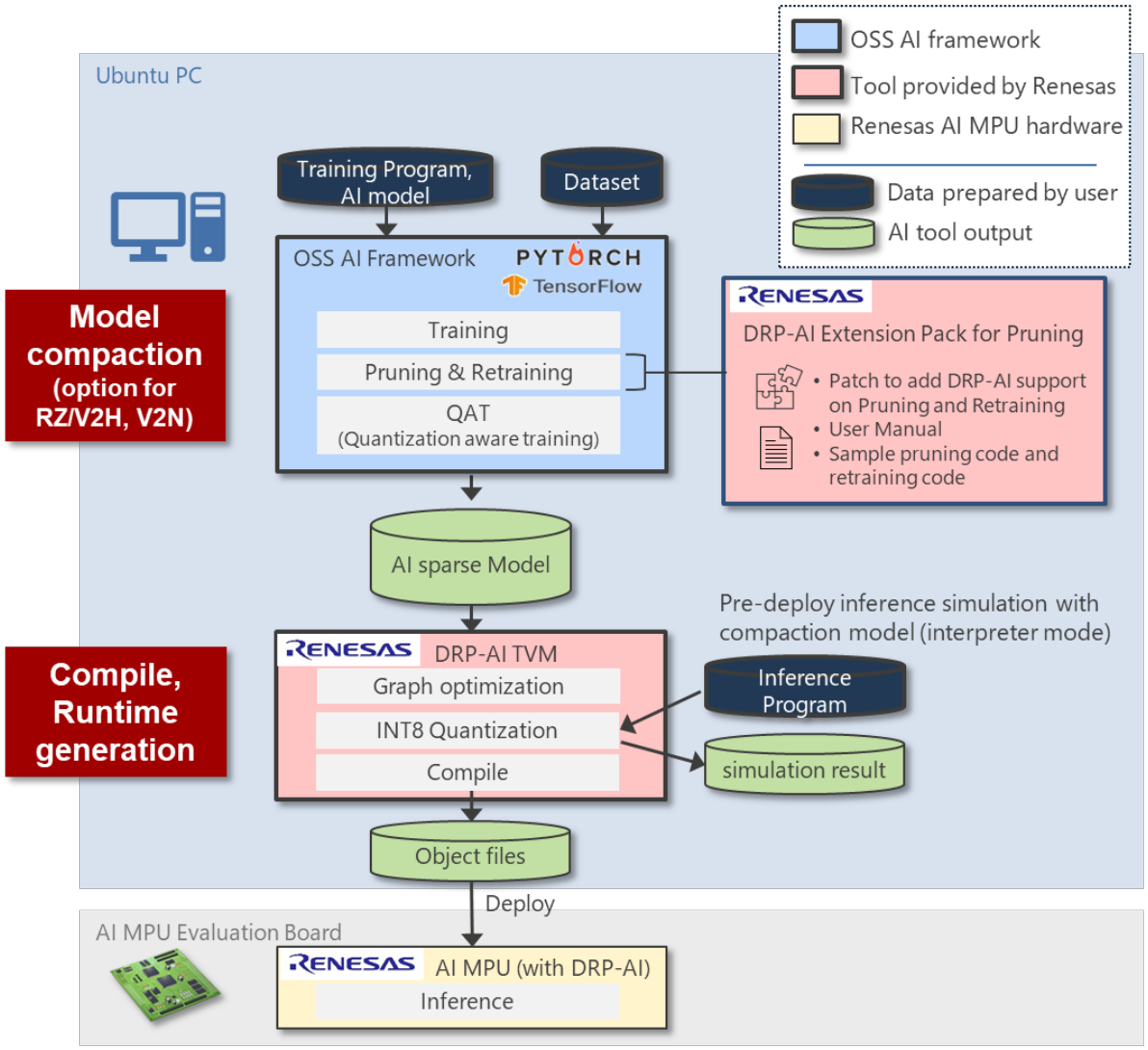


図 6: BYOM (Bring Your Own Model)のツールフロー

DRP-AI Extension Pack (枝刈りツール)

DRP-AI Extension Pack (枝刈りツール)による AI モデル軽量化

RZ/V2H, V2N に搭載されている最新の DRP-AI(DRP-AI3)には、図 7 に示す枝刈りと呼ばれる技術によって軽量化された AI モデルを高速・低電力に処理するための独自の技術が搭載されています。枝刈りは、ニューラルネットワーク内のノード間の重みを削除することで、パラメータの数を減少させる手法です。これにより、ハードウェアの消費電力を削減し、推論プロセスを高速化します。

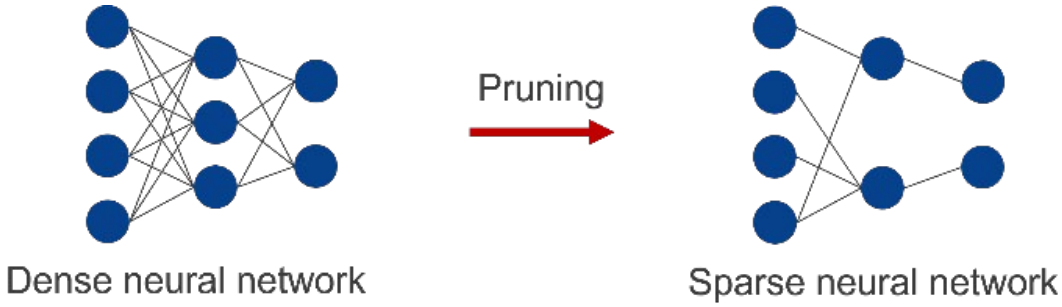


図 7: 枝刈りによるモデルの軽量化

DRP-AI Extension Pack は、ユーザの PyTorch や TensorFlow トレーニングプログラムと組み合わせることで、DRP-AI に最適化された枝刈り機能を提供します。DRP-AI Extension Pack は以下の特徴があり、多様なユーザ要望に対応しつつ、軽量化技術になじみがないユーザでも容易に開発が可能となっています。

1）ユーザは枝刈り率を設定するだけで、ハードウェアに適切な軽量化モデルを自動的に生成できます。

2）学習プログラムがあれば、任意の CNN モデルに適用できます。

DRP-AI Extension Pack を用いた AI モデルの圧縮（枝刈り）フロー

DRP-AI Extension Pack を用いたユーザ独自の AI モデル(BYOM)の枝刈りフローを図 8 に示します。

- **学習プログラムとデータセットの準備**: AI モデルを学習した際の学習プログラムおよび学習に用いるデータセットの準備が必要となります。（対応するフレームワークは、Pytorch および Tensorflow です）
- **パッチの適用**: 学習プログラムを部分的に書き換え、DRP-AI Extension Pack のパッチを適用します。
- **枝刈りおよび再学習の実行：**再学習は、Pytorch あるいは Tensorflow がインストールされているサーバあるいは PC にて実行可能です。
- **枝刈り結果の分析**: DRP-AI TVM によるコンパイルおよび量子化後に、ボード実装評価あるいは後述のインタプリタモードで枝刈りモデルの認識精度を分析し、最適な枝刈り率を決定します。

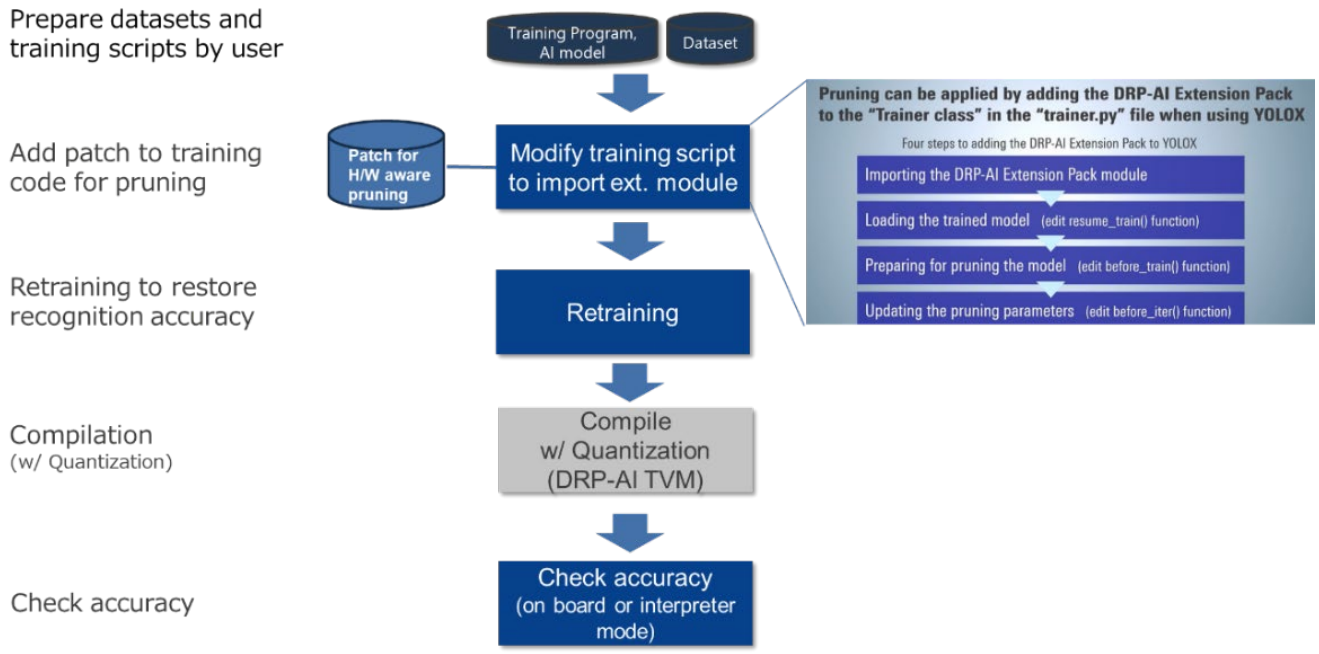


図 8: DRP-AI Extension pack によるモデル圧縮フロー

DRP-AI TVM の github に、学習プログラムの修正方法を示すビデオやフローの詳細を説明したウェブガイド、サンプルスクリプト等を提供しています。（https://renesas-rz.github.io/rzv_drp-ai_tvm/pruning.html）

認識精度を考慮した、自由度の高い枝刈り設定

枝刈りによる DRP-AI での高速化効果を最大化するには、DRP-AI のハード構成に適した枝刈りノードの選定が重要となります。この選定ルールについては、DRP-AI Extension Pack 内のパッチに組み込まれており、ユーザは選定ルールを意識せずに適切な枝刈りモデルを作成することができます。このパッチを使うためには、学習スクリプトを準備し、その一部のコードの修正が必要となります。その一方、枝刈りできる CNN モデルに制約が無いことが特徴で、広いタスクに活用可能となっています。

ノードの削減率（枝刈り率）は任意の値（70%、80%、90%など）に設定することが可能であるため、認識精度と速度とのトレードオフをきめ細かく調整可能です。また、枝刈りすると認識精度に影響を与えるレイヤを、枝刈り対象レイヤから自動的に除外する処理も行っております。これにより、高い枝刈り率設定の際にも認識精度の劣化を抑えています。現在は、除外レイヤ以外は一律の枝刈り率となりますが、今後自由度を高める改良を行う予定です。

枝刈り時の認識精度低下に対する対策

近年の CNN モデルはスケーラブルモデルが主流であり、必要な認識精度に応じて S、M、L といったサイズが選べるようになっています。一般的に、モデルサイズが大きい方が、枝刈り前のベースモデルに比べて枝刈り時の認識精度の低下が少ない傾向にあります(図 9)。よって、特定の AI モデルにおいて、枝刈り

時の認識精度が必要以上に低下してしまった場合の対策としては、１）枝刈り率を低めに設定する、という方法のほかに、２）一段大きいサイズのモデルを選択し枝刈り率を高めに設定する方法も有効です。

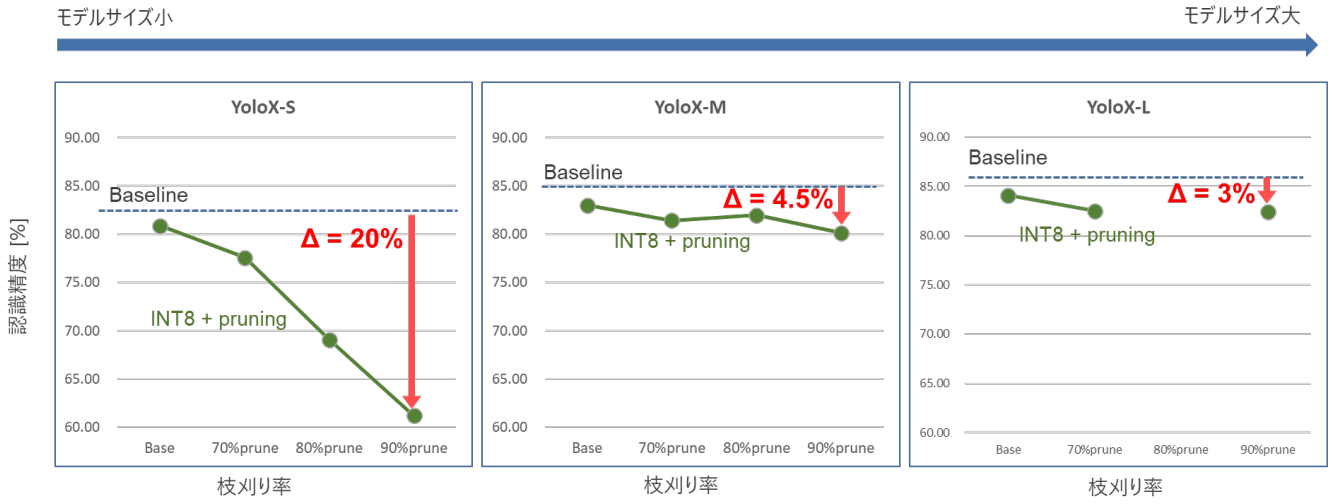


図 9: モデルサイズ・枝刈り率と認識精度との関係

枝刈り時の推論速度の事前見積手法

枝刈りによる推論速度向上と認識精度維持とのバランスを取るためには、枝刈り率を変えて再学習を行い、推論速度および認識精度を確認するループを繰り返して調整する必要があります。この際、より少ないループ回数で調整するには、時間がかかる再学習を行う前に枝刈り時の推論速度を見積り、ターゲットとなる枝刈り率を先に定めておくことが有効です。DRP-AI TVM では、時間のかかる再学習をする前に、枝刈り前後での推論時間を見積もるフローが提供されています。具体的には、枝刈り率を設定した後に仮の枝刈りモデルを作成後、推論速度を実機上で測定する方法になります。これにより、枝刈りモデルの導入判断に役立てることができます。詳細については"DRP-AI Extension Pack (Pruning Tool) Sparse Model Processing Speed Check Guide "をご確認ください。

AI コンパイラ（DRP-AI TVM）

DRP-AI TVM の特徴

DRP-AI TVM は、オープンソースの AI コンパイラとして広く活用されている Apache TVM をベースに DRP-AI 向けに最適化したコンパイラです(図 10)。DRP-AI TVM は以下の特徴を持っており、最新モデルへの柔軟性、最適な処理性能、製品間のスケーラビリティを実現しています。

ヘテロジニアス構成に対応: 多様化する AI モデルに対応するため、DRP-AI で対応していないレイヤ構造（オペレータ）でも CPU に割り振り、DRP-AI と CPU 間で協調動作させることが可能です。各レイヤの DRP-AI と CPU への割り振りは、DRP-AI TVM が入力されたモデル構造を理解し、自動的に行われます。

複数の AI フレームワークのサポート：DRP-AI TVM は、ONNX、PyTorch、TensorFlow など、主要な AI フレームワークをサポートしています。

DRP-CPU 統合ランタイムの生成: AI アクセラレータ(DRP-AI)と CPU 間での協調動作を効率よく行うため、データ転送と演算の効率的なスケジューリング（たとえば、CPU が扱う linux 仮想メモリ空間と DRP-AI が扱う物理メモリ空間双方のメモリ管理、重みデータの再利用、バースト転送による外部 DRAM トラフィックの最小化など）を計画し、統合ランタイムを自動生成する機能を有しています。

DRP-AI 向け軽量化モデル対応とコンパイル: DRP-AI TVM 内で、入力された AI モデルを DRP-AI 向けに量子化し、コンパイルします。具体的には、一般的な AI モデルのデータタイプである 32 ビット浮動小数点形式（FP32）から 16 ビット(FP16)への量子化(RZ/V2L, V2M, V2MA) あるいは INT8 への量子化（RZ/V2H, V2N）を行います。DRP-AI Extension Pack で軽量化した枝刈りモデルを入力すると、自動的に DRP-AI の枝刈り機能が適用されるよう最適化やコンパイルを行います。

RZ/V シリーズ間の互換性：DRP-AI TVM は、複数の製品間で同一のコンパイラおよび API を活用することができるため、DRP-AI TVM で動作する AI アプリケーションは、製品間で共通に活用できます（一部設定ファイルなどの変更が必要になる場合があります）。

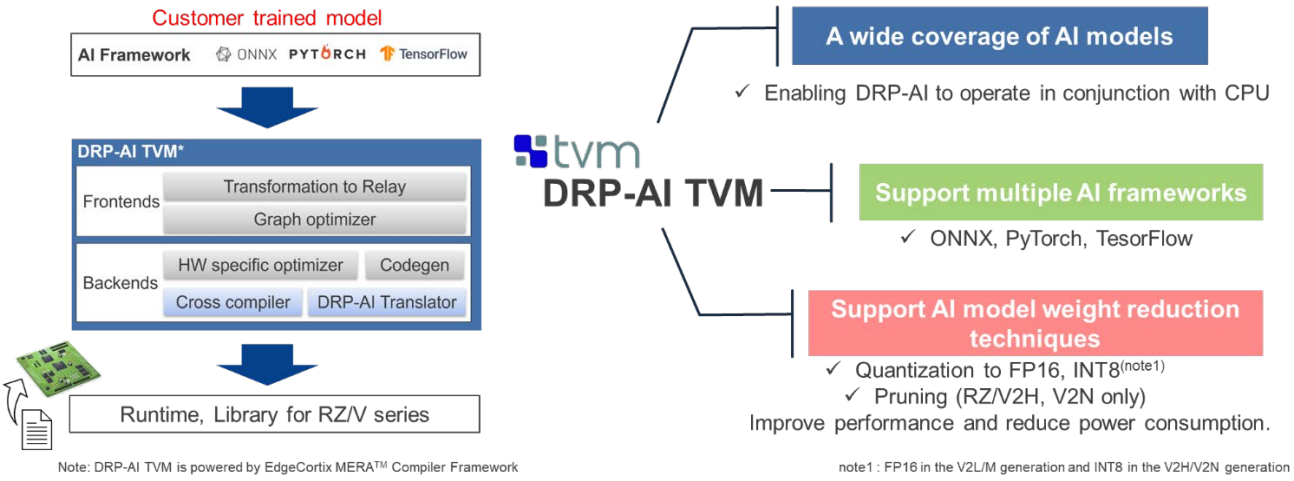


図 10: DRP-AI 向け AI コンパイラ (DRP-AI TVM)

DRP-AI 向け量子化

DRP-AI の AI モデル・アプリ実装フローでは、再学習が不要な量子化手法である PTQ（Post-Training Quantization）は DRP-AI TVM 内で実行します。その他の方法として、再学習によって量子化モデルの認識精度を改善する QAT（Quantization-Aware Training）にも対応しており、これは DRP-AI TVM の入力前にユーザで実行するオプションのフローとなります。QAT の具体的なフローについては、DRP-AI TVM の GitHub ページを参照ください。（https://github.com/renesas-rz/rzv_drp-ai_tvm/tree/main/QAT）

DRP-AI 向け枝刈りモデル対応

DRP-AI Extension Pack で軽量化した枝刈りモデルを使用すると、自動的に DRP-AI の枝刈り機能が適用されます。コンパイラは枝刈りされた重みデータを自動的に解析し、DRP-AI の枝刈り機能を最大限に有効化するコードを生成します。枝刈り済みモデルを使用することで、画像 1 枚当たりの AI 処理量や演算量メモリと DRP-AI 間での重みデータの転送量を大きく削減できるため、推論時間の短縮や高い電力効率での推論が可能となります。

DRP のパイプライン処理による前処理の高速化

DRP-AI を用いた CNN の推論を高速化する際、AI 処理向けに画像調整する前処理の時間が相対的に大きくなります。このアーキテクチャでは、組み込み CPU の代わりに DRP を利用することで、全体の性能を向上させることができます。DRP-AI テストチップでの例では、物体認識アプリ(YOLOv2)では、前処理を含めた全体の推論時間は、組み込み CPU で処理した場合に比べて 6.5 倍速くなります[1]。 さらに、DRP に最適化された画像処理ライブラリを順次拡充しております。これらを実装することで、CPU 動作に比べて約 1 桁の速度向上が可能となります。

RZ/V2H, V2N の CNN モデル推論性能評価

CNN モデルの推論性能

図 11 は、代表的な CNN モデルについて、DRP-AI 搭載製品の世代間での性能比較をした結果です。DRP-AI3 を搭載した RZ/V2N(最大性能 4TOPS(枝刈り無し), 15TOPS(枝刈りあり))は、前世代の DRP-AI を実装した RZ/V2M（最大性能 1TOPS）に比べ、枝刈り無しモデルで約 4 倍、枝刈り率 90%のモデルで最大 10 倍以上の性能向上を実現しています。AI モデル実行時の性能向上は、AI-MAC 数の増強によるピーク性能強化以外にも、量子化等による軽量化、内蔵メモリも含めたメモリ管理などの技術も貢献しています。

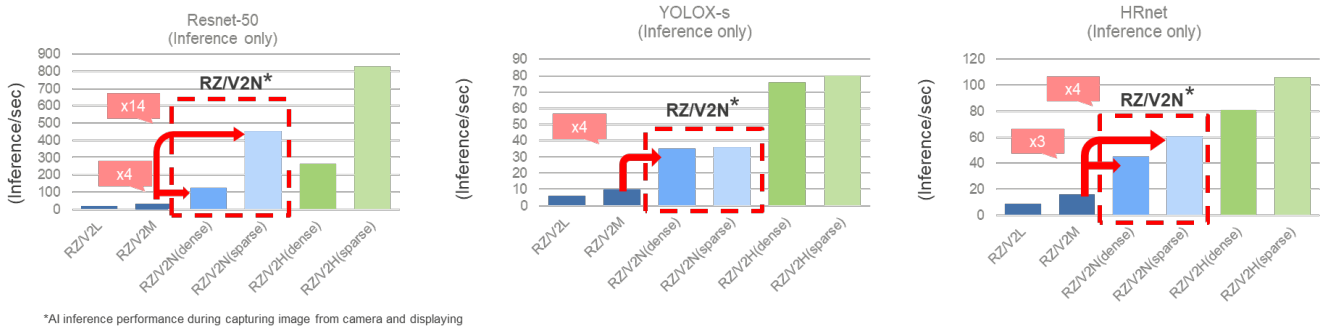


図 11: RZ/V シリーズ間での CNN 推論速度比較

DRP-AI3 による速度向上効果の変動要因

最大性能の強化や枝刈り対応技術による速度向上の効果が、モデルによって異なる理由を以下に示します。

まず、最大処理性能（ピーク性能）は AI-MAC 内の積和演算器の搭載数の強化などで実現しております。しかしながら、積和演算器の数を増やしても、使う AI モデルのサイズやメモリ帯域の制約により演算器を常時使いきれない場合があるため、必ずしもピーク性能に比例した性能向上をしない場合があります。また、RZ/V2N はメモリ帯域の制限が RZ/V2H より厳しいため、AI 以外の高負荷なアプリケーションを同時に実行する際に、RZ/V2N の方が AI 処理性能の変動が大きい可能性があります。よって、AI 処理だけでなく、AI 以外の処理負荷を考慮したデバイス選定が重要となります。

次に、枝刈り効果のモデル依存性について示します。DRP-AI による枝刈りモデル高速化技術は、枝刈りによってニューラルネットワークのノード数（つまり演算回数）を減らし、画像 1 枚当たりの推論時間および消費電力を削減する機能です。消費電力については、枝刈りによって積和演算の回数が減らせるため、その分電力効率が向上します。一方、推論時間に関しては、使用する AI モデル等によって効果が変わります。その理由として、枝刈りによって積和演算時間およびノード情報（weight）の通信量(図 12(A))を削減できるため 1 画像当たりの推論時間の短縮が可能になりますが、レイヤごとの演算結果（feature map）は枝刈りでも削減されないため（図 12(B)）、枝刈り高速化するにつれて feature map の通信量が増えてしまい、メモリ通信量（帯域）の制限により速度が頭打ちになる傾向にあるためです。しかしながら GPU での枝刈りでの速度向上は 10%程度[2]であることに比べると、ルネサスの枝刈りは 20%～300%程度の性能向上が得られるため、性能面でも有効性がある技術となっています。

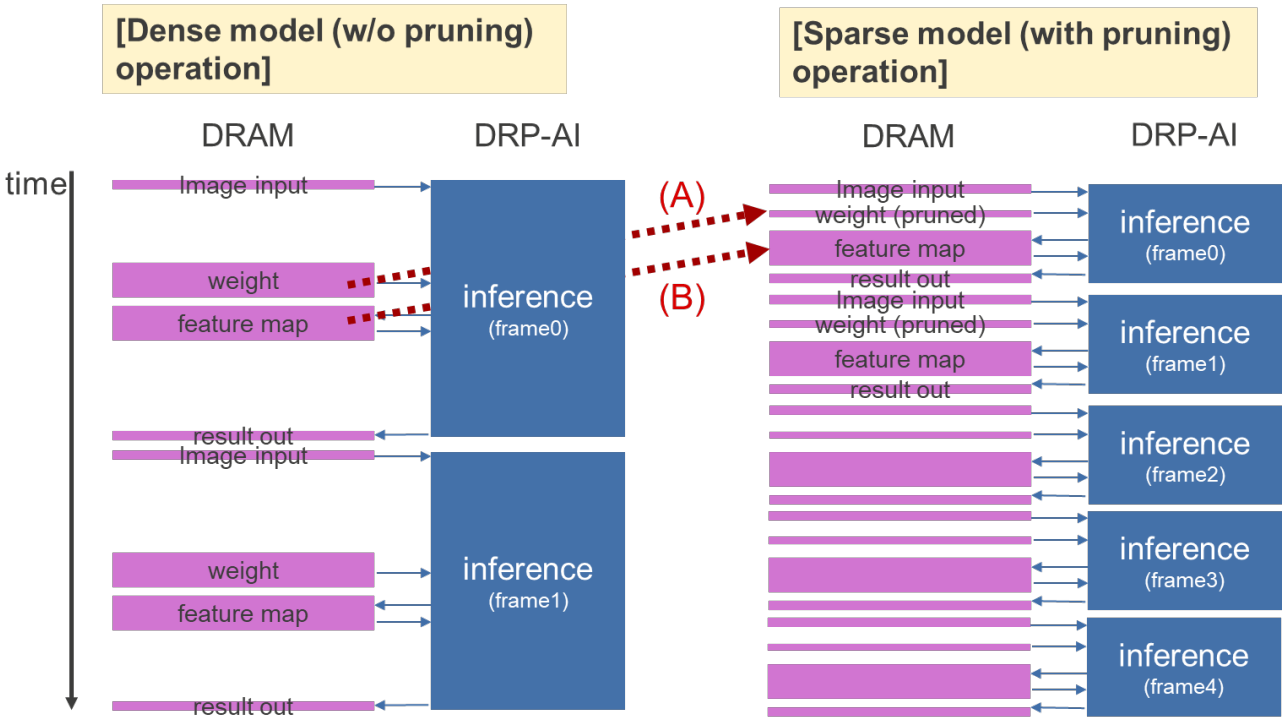


図 12: 枝刈り前モデル(Dense)と枝刈り後モデル(Sparse)とのメモリアクセス比較（イメージ）

トランスフォーマモデルのエンドポイント導入

近年、トランスフォーマモデルはその高い性能と多用途性から、エンドポイントデバイスへの導入が進んでいます。特に、ChatGPT のような会話モデルに加え、ViT[3]などの Vision Transformer モデル、SayCan[4]/RT-2[5]のようなロボットの行動生成モデルが注目されています。この新しいトレンドに対応するため、高い柔軟性を持つ DRP-AI ツールをトランスフォーマ向けに改良し、RZ/V2H, V2N でのトランスフォーマモデルによる推論を可能にしました。

一方、トランスフォーマは CNN モデルとは異なる演算種類が多用されることや、CNN よりも必要なメモリサイズや演算量が大きくなるといった課題があります。ルネサスでは、DRP-AI ツールのさらなる改良に加え、トランスフォーマモデルの演算効率向上と、枝刈り等のモデル小型化技術を強化した次世代 DRP-AI の開発を推進していく予定です。

トランスフォーマモデルの組み込み実装トレンドと課題

トランスフォーマモデル活用の拡大

Vision AI は、近年の技術進化により、現在の主要要素である CNN（Convolutional Neural Network）からトランスフォーマモデルへと発展しています。この背景にはいくつかの重要な要因があります(図 13)。

まず、トランスフォーマモデルは CNN に比べて認識精度が高いことが挙げられます。これは、トランスフォーマが自己注意機構（Self-Attention）を利用することで、画像内で離れた距離の情報の依存関係も効率的に処理できるためです。これにより、画像認識だけでなく、時系列予測や自然言語処理などの広範なタスクにも対応できるようになりました。さらに、トランスフォーマモデルはマルチモーダルなデータ処理にも適しています。例えば、画像とテキストの同時処理や、音声と映像の統合など、複数のデータ形式を組み合わせたタスクにおいても優れた性能を発揮します。また、トランスフォーマモデルの軽量化技術の開発も進んできたことで、高度なトランスフォーマモデルを低電力かつリアルタイムに処理することが可能になってきました。

さらに最近では、CNN と Self-Attention を融合させたモデル(たとえば YOLOv10 や Topformer など)が、従来の画像 AI(CNN モデル)よりも精度と性能とのバランスを高くできるモデルとして注目されています。これらのモデルは、CNN の特徴抽出能力と Self-Attention の長距離依存関係把握能力が組み合うことで、より高い精度と効率を実現しています。

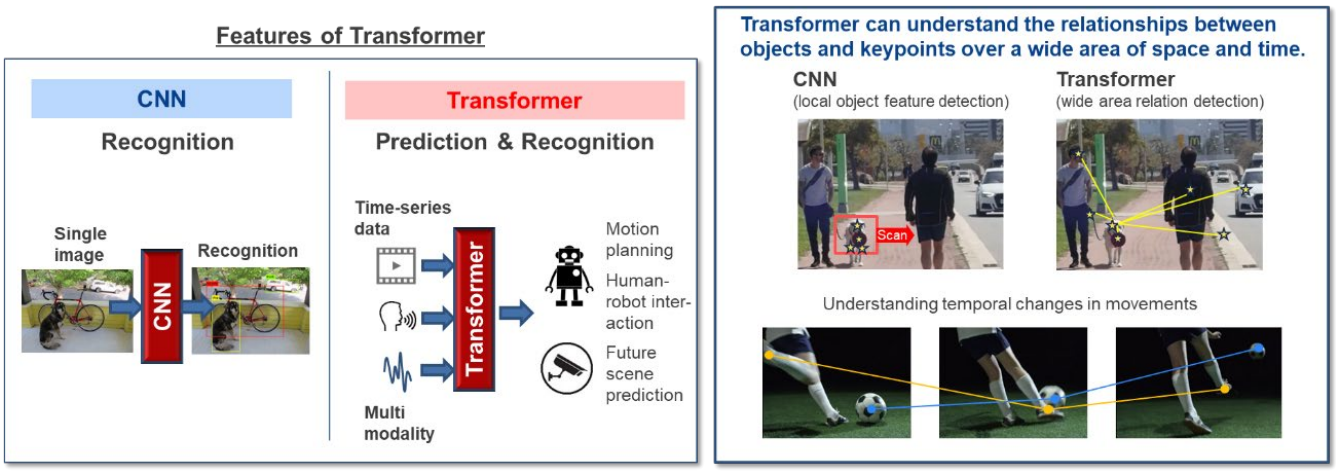


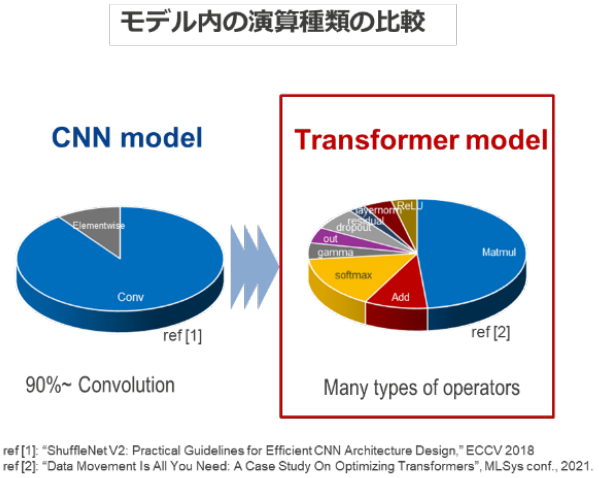
図 13: トランスフォーマモデルの特徴

トランスフォーマのエンドポイント実装に向けた課題

しかしながら、トランスフォーマは CNN モデルとは異なる種類の演算が多用されることや、CNN よりもモデルサイズが大きくなるといった課題があります（図 14）。そのため、トランスフォーマモデルをエンドポイントに実装するには、これらの課題を解決していく必要があります。

演算の複雑性の増大: 90%以上の演算が畳み込み処理(Convolution)で構成される CNN[6]とは異なり、トランスフォーマは重みデータを用いない feature map 同士の積和演算や Softmax、GELU、LayerNorm など、トランスフォーマ特有の複雑性の高い演算が多用される複雑な構成となっています[7]。そのため、AI コンパイラが対応しきれずにデバイスに実装できない場合や、これらの演算が性能ボトルネックになり AI 推論速度が低下するなどの問題があります。

モデルサイズや演算量の増大: トランスフォーマモデルは、従来の CNN モデルよりも認識精度は高くできます。一方で、モデルサイズや演算量が数倍大きくなるという傾向にあるため、メモリや演算リソースが限られるエンドポイントデバイスへの実装の障壁となっています。



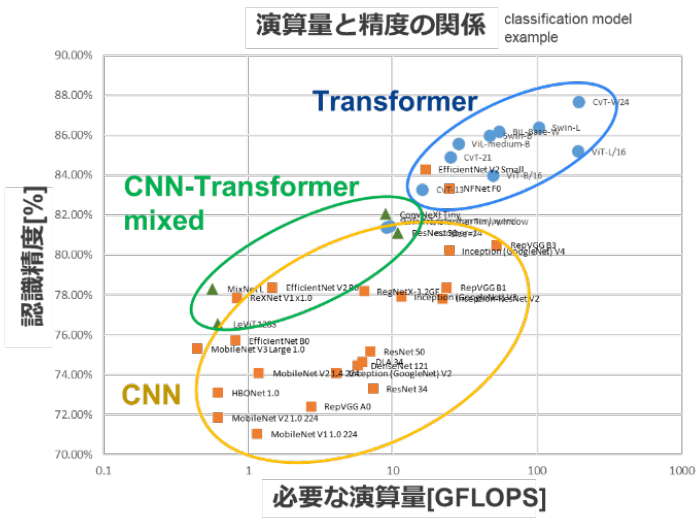


図 14: トランスフォーマモデルのエンドポイント AI 実装課題

DRP-AI3 でのトランスフォーマモデル対応

DRP-AI3 でのトランスフォーマ最適化技術

RZ/V2H, V2N に実装されている DRP-AI3 は、CNN に最適化されたアーキテクチャになっています。ルネサスでは、これらの技術をトランスフォーマにも対応できるよう、AI コンパイラなどの環境を進化させております。具体的には、以下の技術の開発を継続的に推進しております。

多様な Activation の高速化: トランスフォーマの多様かつ大規模な演算処理に対して、動的再構成プロセッサ（DRP）を活用した高速ライブラリの開発を継続中です。

行列演算の最適化: DRP-AI TVM 内部で、トランスフォーマの演算の多くを占める行列演算を CNN 向けの演算（オペレータ）に変換することで、DRP-AI による高速処理が可能になります。

枝刈り・量子化技術の適用：DRP-AI 向けの量子化・枝刈りツールは、CNN だけでなくトランスフォーマにも適用可能です。したがって、CNN と同様、枝刈りモデルによる高速化にも対応可能です。また、Quantization Aware Training(QAT)にも対応予定です。

AI コンパイラ（DRP-AI TVM）のトランスフォーマ対応

ルネサスでは、DRP-AI TVM のトランスフォーマ対応を強化し、2024 年 10 月リリースの Version 2.4 より、一部のトランスフォーマモデルおよびオペレータが RZ/V2H, V2N で実行可能になりました。2025 年 3 月現在では、ViT や Swin[8]などの Vision transformer に加え、トランスフォーマモデルと CNN モデルを組み合わせたモデルなどが対応しております(図 15)。なお、DRP-AI3 はもともと CNN 向けのアーキテクチャであることや、DRP-AI ツールも継続進化中の段階であるため、いくつか制約があります。

・一部のオペレータがサポートされていないため、変換できないモデルがある。
・INT8 量子化による精度低下が大きい場合がある。
・DRP-AI 未対応オペレータは CPU で処理するため、速度向上は CPU 比 10 倍程度までになる。

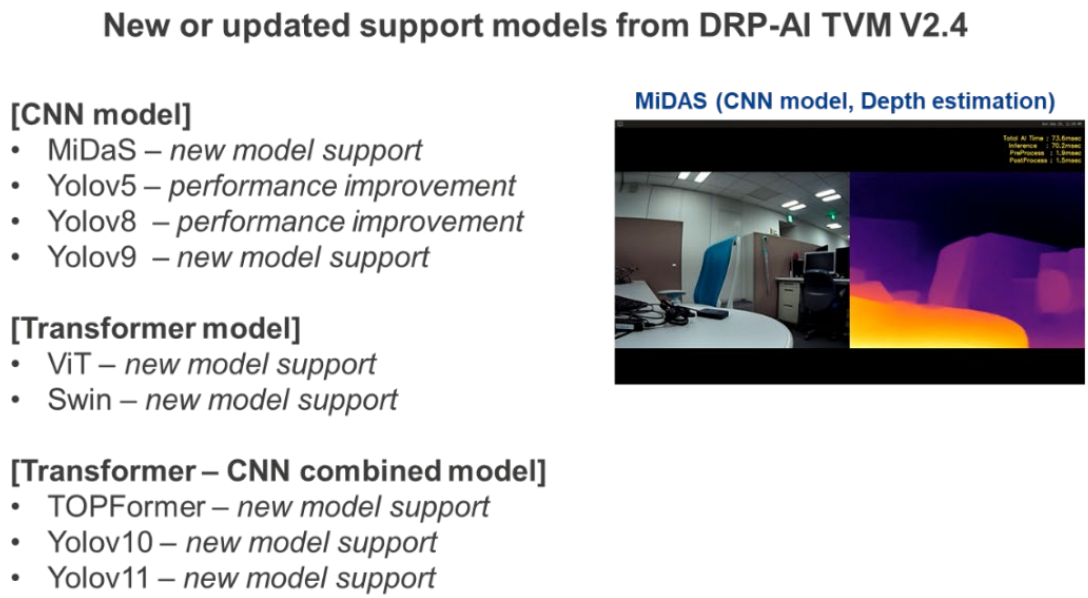


図 15: 2025 年 3 月の DRP-AI TVM (Version2.4)で新規対応・アップデートした主な AI モデル

トランスフォーマ実装実例

Topformer は、Token Pyramid Transformer の略であり、モバイルデバイス向けのセマンティックセグメンテーションに特化したアーキテクチャです[9]。ここで「トークン」とは、画像やデータを細かい部分に分けたものを指します。Topformer は、さまざまな大きさのトークンを入力として受け取り、それぞれの大きさに応じた意味を持つ特徴を生成します。これにより、トークン同士の関係を強化し、データの表現力を向上させます。本モデルでは、CNN とトランスフォーマのレイヤを組み合わせることで、CNN やトランスフォーマだけのモデルに比べて、精度と速度のバランスが高くなっています。

今回、DRP-AI TVM V2.4 にて Topformer モデルをコンパイルし、RZ/V2H での実装に成功しました。モデル部分の処理時間(inference time)は 100msec 以下であり、高精度かつリアルタイムなセグメンテーション処理も可能となってきております。

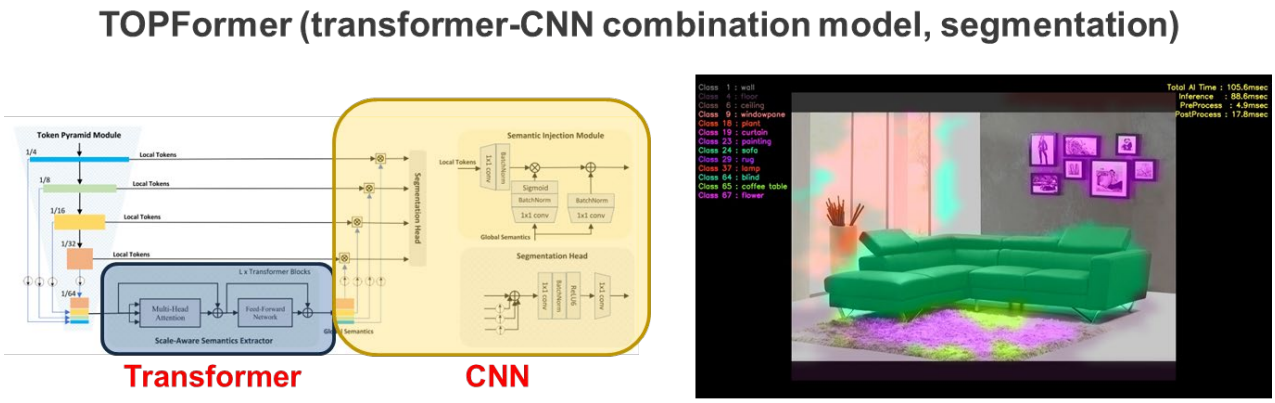


図 16: CNN-トランスフォーマ混載モデルの例(Topformer)

今後の予定: 次世代 AI アクセラレータ（DRP-AI4）開発

次世代の AI アクセラレータ(DRP-AI4)は、現行の DRP-AI3 をさらに進化させ、より高性能で効率的な AI 処理を実現することを目指しています。特に DRP-AI のアーキテクチャの柔軟性や軽量化モデル対応をさらに強化し、トランスフォーマ・CNN ともに高速・低電力に処理可能な AI アクセラレータを目指した開発を進めています。

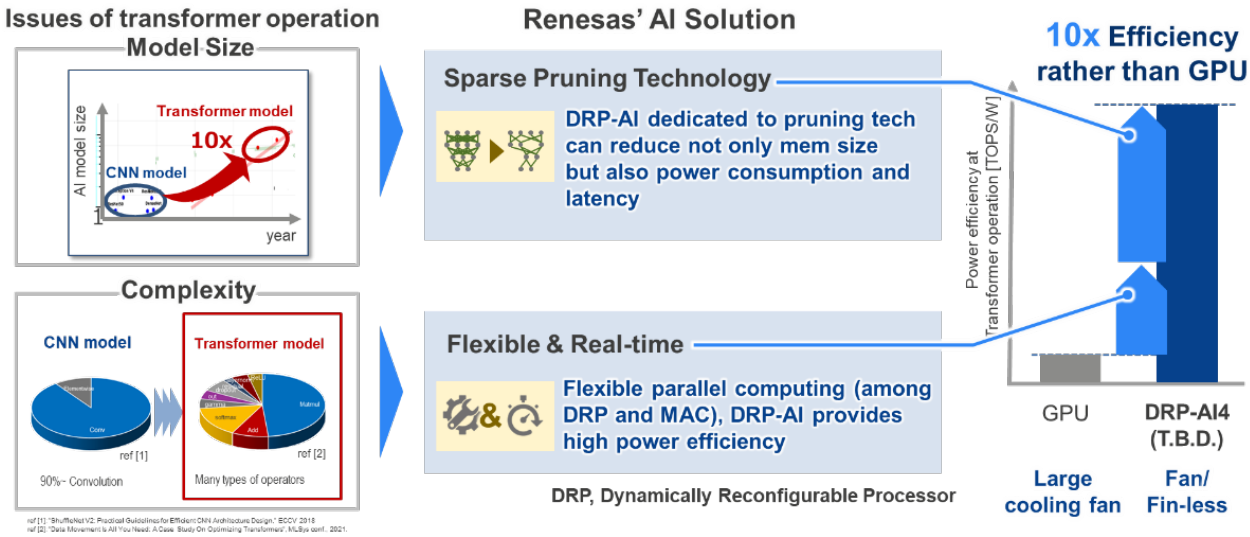


図 17: 次世代 DRP-AI (DRP-AI4)に向けた取り組み

結論

ルネサスは、常に最新の技術トレンドを捉えた技術開発と製品開発を行い、市場に提供しています。私たちの目標は、お客様が安心して当社の製品を使用できるようにすることです。これからも、革新的なソリューションを提供し続け、皆様のビジネスの成功をサポートしてまいります。

参考文献

1. K. Nose, et. al., “A 23.9TOPS/W @ 0.8V, 130TOPS AI Accelerator with 16× Performance-Accelerable Pruning in 14nm Heterogeneous Embedded MPU for Real-Time Robot Applications,” ISSCC2024.
2. Accelerating Inference with Sparsity Using the NVIDIA Ampere Architecture and NVIDIA TensorRT
3. D. Alexey et. al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". arXiv:2010.11929.
4. SayCan:Grounding Language in Robotic Affordances
5. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control, arXiv 2307.15818, 2023
6. N. Ma, et. al., “ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design,” Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 116-131.

7. A. Ivanov, et. al., “Data Movement is All You Need: A Case Study on Optimizing Transformers”. Proceedings of Machine Learning and Systems 3, pp. 711–732, 2021.
8. L. Ze, et. al. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows". arXiv:2103.14030
9. W. Zhang, et. al. “TopFormer: Token Pyramid Transformer for Mobile Semantic Segmentation”, CVPR 2022, pp. 12083-12093.

関連情報

- RZ/V2N: 高い AI 性能と電力効率を両立する 2 カメラ入力対応、15TOPS クアッドコアビジョン AI MPU
- RZ/V2H: 高効率 AI 推論と高速リアルタイム制御を実現できる 4 コアビジョン AI プロセッサ
- Embedded AI-Accelerator DRP-AI ホワイトペーパー
- Next Generation Highly Power-Efficient AI Accelerator (DRP-AI3) ホワイトペーパー

本社所在地

〒 135-0061 東京都江東区豊洲 3-2-24
（豊洲フォレシア）
https://www.renesas.com

お問合せ窓口

弊社の製品や技術、ドキュメントの最新情報、最寄りの営業お問合せ窓口に関する情報などは、弊社ウェブサイトをご覧ください。
http://www.renesas.com/contact/

Doc Number: R01WP0027JJ0100