Manage the Increased Scale and Complexity of Vision AI with Greater Power Efficiency of Mid-Range RZ/V2N MPU

Koichi Nose, **Masayuki Shimobeppu**, **Takao Toi**, Embedded Processor Product Management Department, Embedded Processing Marketing Division, Embedded Processing Product Group, Renesas Electronics

Kentaro Mikami, System Solution Department 1, Software Development Division, Software & Digitalization Group, Renesas Electronics

Overview

With the increased demand for systems performing advanced AI (Artificial Intelligence) within endpoint devices, particularly in applications requiring real-time performance such as robots, AI cameras, and drones, there are also expectations for increasingly sophisticated processing performance and power efficiency, to enable AI processing of more advanced and diverse tasks.

To meet these demands, Renesas offers the RZ/V series- a microprocessor (MPU) for Vision AI equipped with a highly efficient AI accelerator (DRP-AI). This white paper provides an overview of the latest products in the RZ/V series and introduces technologies for optimizing AI processing through hardware and software coordination.

Endpoint AI Trends and Challenges

Recent years have seen a rapid evolution of Endpoint AI, and it is increasingly being used in a variety of applications. Since AI processing in the cloud involves huge amounts of communication traffic and time, there is a growing shift from cloud to edge computing, especially in the field of Vision AI and for tasks requiring real-time performance.

Endpoint AI Trends

Improved real-time performance: Conventional cloud AI utilizes large amounts of data and abundant computing resources, but endpoint AI can process data locally to minimize communication latency and improve real-time performance. This enables real-time processing of sensor data and allows for immediate feedback (Figure 1).

Diversification of AI tasks: AI is now used in a wide range of predictive tasks, from recognizing images such as dogs, cats, and objects, to predicting 3D depth from still images. A wide range of applications have been developed in recent years that plan and make decisions with improved AI recognition accuracy, such as optimal robot operations and smart buildings with optimized energy consumption based on environment and human behavior.

Reduction of model weight: Advances are being made in lightweight technologies for AI model performance and improvements in memory access efficiency, making highly efficient AI processing possible on endpoint devices.

Development from CNN model to transformer model: A CNN (Convolutional Neural Network) is a model specialized for image recognition, while a transformer model uses a self-attention mechanism to efficiently handle long-distance dependencies. In addition to improving image recognition accuracy, this makes it possible to handle a wide range of tasks, such as time series information and natural language processing.

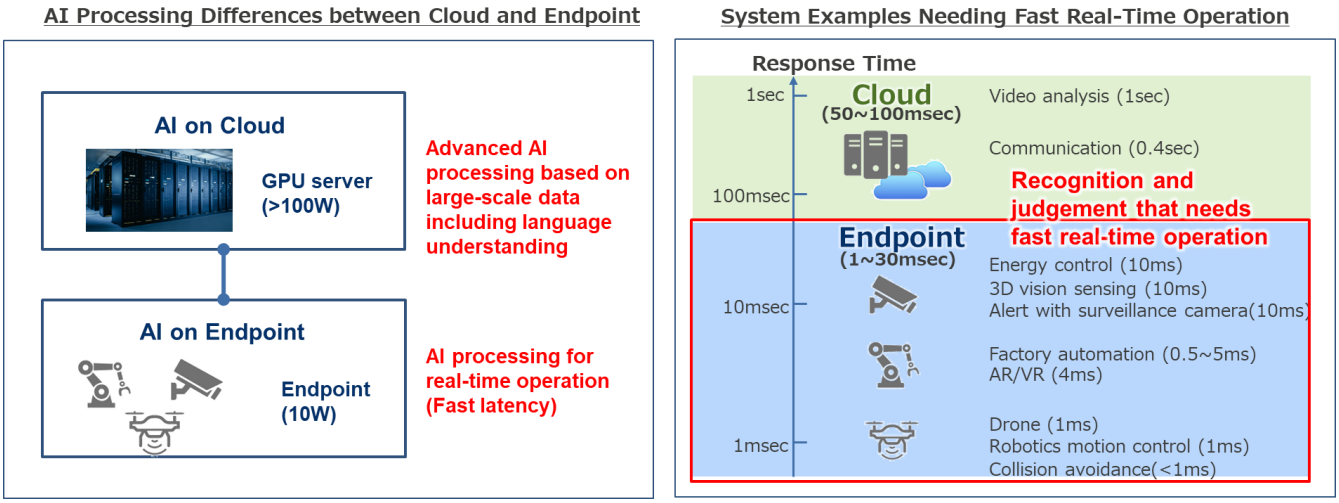

Figure 1: Positioning of Endpoint AI

Challenges of Conventional Endpoint AI Solutions

1) Large-scale AI models: Image AI originally required multiply-accumulate operations that were two to three orders of magnitude greater than conventional image processing algorithms, and so dedicated AI chips and AI accelerators with advanced AI processing capabilities were developed. However, AI models are increasing in scale each year in order to achieve higher recognition accuracy, resulting in an increasing amount of calculations required per image (Figure 2).

2) Increased complexity of AI models: AI models are increasing in complexity each year, especially with the appearance of advanced architectures such as transformer models. Increased computational resources are required for model training and inference due to this, which makes implementation difficult on endpoint devices. Furthermore, the increased memory usage of complex models is difficult to handle with the limited resources of endpoint devices.

3) Difficulty of AI tools: AI tools tend to be specialized and complex, therefore requiring advanced expertise to optimize endpoint devices. At the same time, users developing applications and systems for embedded devices have a great variety of AI expertise and needs, which means it is essential to prepare an AI tool environment to support a wide range of users. Selecting, optimizing and implementing AI models on devices also requires a lot of effort, while compatibility between different tools and lack of tools optimized for specific hardware can create other issues.

There are concerns that power consumption will continue to increase as models become larger and more complex. The RZ/V2N is a new MPU device that can solve these problems, make a significant contribution to the implementation of AI at endpoints and is equipped with DRP-AI3 enabling highly power efficient processing of lightweight models.

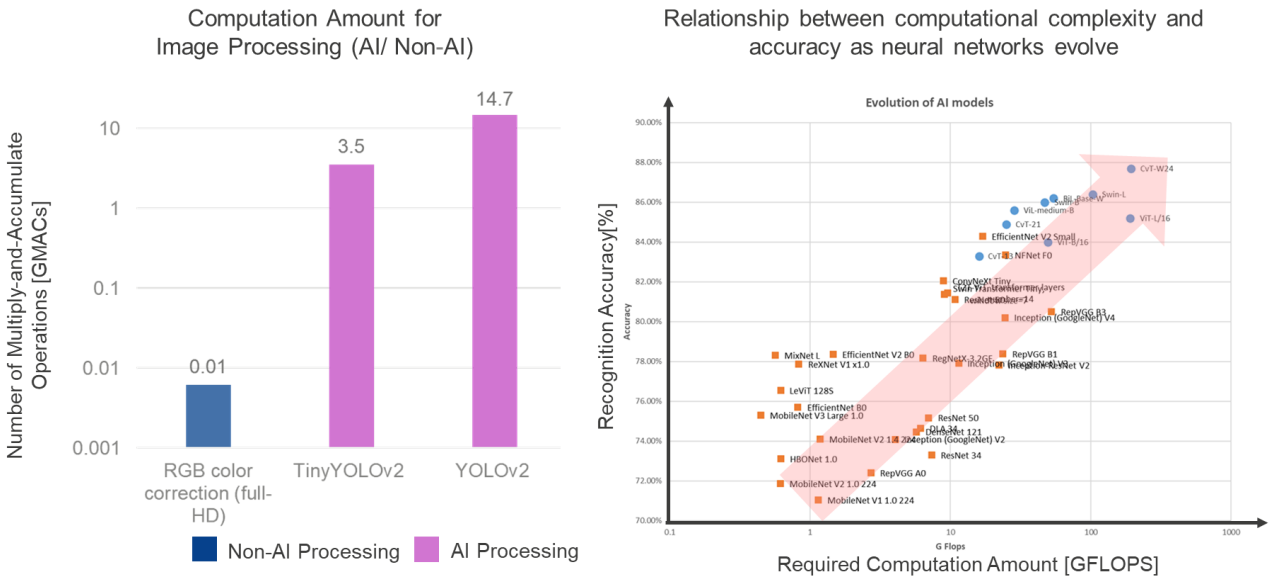

Figure 2: Challenges of Endpoint AI Solutions

Optimized for Vision AI – Features of the RZ/V Series

MPU Family for Endpoint Vision AI Applications

The RZ/V series is an MPU for Vision AI that uses Renesas's original highly efficient AI accelerator (DRP-AI). Varied AI calculations can be efficiently processed using Renesas' Dynamically Reconfigurable Processor (DRP). Mass production of the RZ/V2N with a maximum performance of 15 TOPS will begin from March 2025. The RZ/V2N expands the existing RZ/V portfolio that includes the recent RZ/V2H MPU, a high-end AI MPU with a maximum performance of 80 TOPS. The expanded portfolio provides AI performance scalability to support a wide variety of endpoint AI applications.

The RZ/V series is also designed to maximize the power efficiency of AI in embedded products, keeping the generation low. For example, the RZ/V2H without a fan maintains temperatures at approximately the same level as embedded GPUs using a large fan, making it suitable for use in embedded devices with strict restrictions on installation size and heat generation. The maximum power efficiency is approximately 10 TOPS/W, which is far superior to that of competing devices.

Target Applications

The RZ/V series is targeted primarily at the high-end (several tens of TOPS) to mid-range (1 TOPS to 10 TOPS) segments within the endpoint AI market (Figure 3). The high-end RZ/V2H is primarily intended for applications that require substantial real-time performance and advanced AI, such as collaborative robots, AGVs, and drones, while the mid-range RZ/V2N is primarily intended for applications such as DMS (Driver Monitoring System), mobile robots, and AI cameras. The cost-effective RZ/V2N is most suitable for mid-range AI market applications that require high performance AI at an affordable price.

Figure 3: Target Applications of the RZ/V Series

AI Accelerator Features (DRP-AI)

DRP-AI is an original Renesas AI accelerator built into the RZ/V series, which achieves both high performance and flexibility through the following technologies:

Flexible support for a wide range of AI processing: The AI multiply-accumulate unit (AI-MAC), optimized for image recognition AI (CNN) processing, works in coordination with the dynamically reconfigurable processor (DRP) (Figure 4), achieving both high performance and flexibility in large-scale and diversified Vision AI processing. Furthermore, performing image processing on the DRP prior to AI processing achieves an increase in speed for the whole system.

Support for reduced AI model weight: Renesas' proprietary lightweight technology achieves outstanding power efficiency (up to about 10 TOPS/W) with its high-speed processing of sparse models by reducing calculations that have minimal effect on AI model accuracy.

Hardware-software co-optimization: The dedicated DRP-AI tool from Renesas enables generation of execution data optimized for processing on the DRP-AI.

This white paper mainly introduces DRP-AI tool-related technologies. For more information on DRP-AI hardware technology, please refer to the DRP-AI white paper and the ISSCC2024 paper[1] in the reference and related information section.

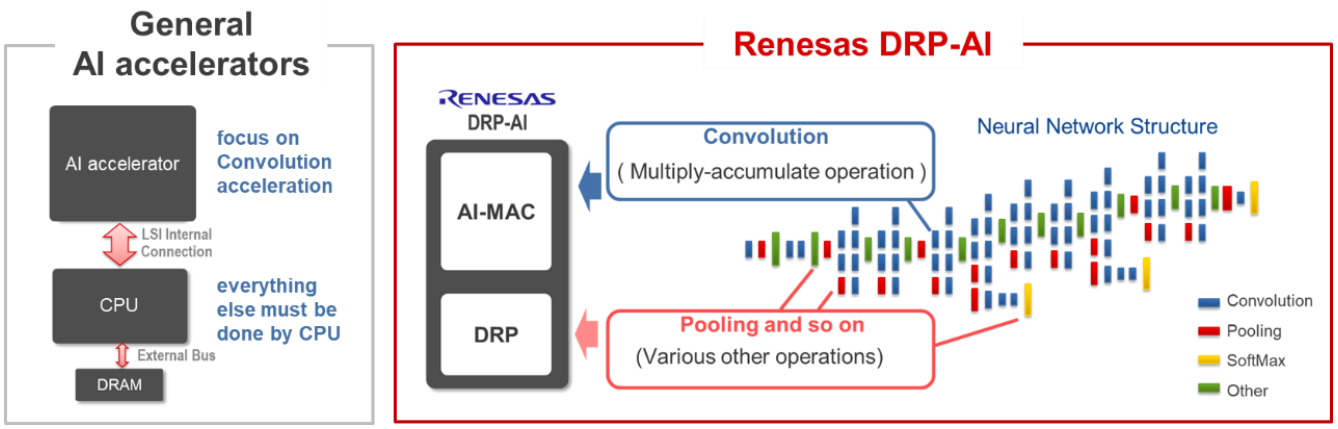

Figure 4: AI Accelerator Equipped with DRP (Dynamically Reconfigurable Processor) (DRP-AI)

DRP-AI Tool Configuration

Renesas DRP-AI tools for development of AI applications can be easily utilized by a wide range of users, from AI novices to AI experts. This section will introduce the overall structure and features of these development tools, and the second half will introduce techniques for reducing model weights, which is also a feature of these tools.

DRP-AI Tool Features

Renesas provides usable forms of the two AI tools shown in Figure 5 as a development environment for DRP-AI applications. This makes them easy to use for a wide range of users, from AI novices to experts.

Free AI application users: Renesas provides a variety of open-source AI applications using pre-trained models. Most applications include an executable file for the AI model and can run on the device as is (AS IS or Custom Application in Figure 6). Also, our Transfer learning tool (TLT) can be used to retrain the model with a dataset provided by the user. This allows users to customize the model according to their usage scenario. For more details, visit the AI application Github .

Users who use their own AI models: Implementation of user custom models is also supported (Custom Model in Figure 6). Using the AI compiler, (DRP-AI TVM) which is optimized for DRP-AI, enables the generation of highly efficient executable files that run on the RZ/V from the user's custom models. A DRP-AI Extension Pack tool is also provided for retraining that supports pruning to create reduced weight custom models.

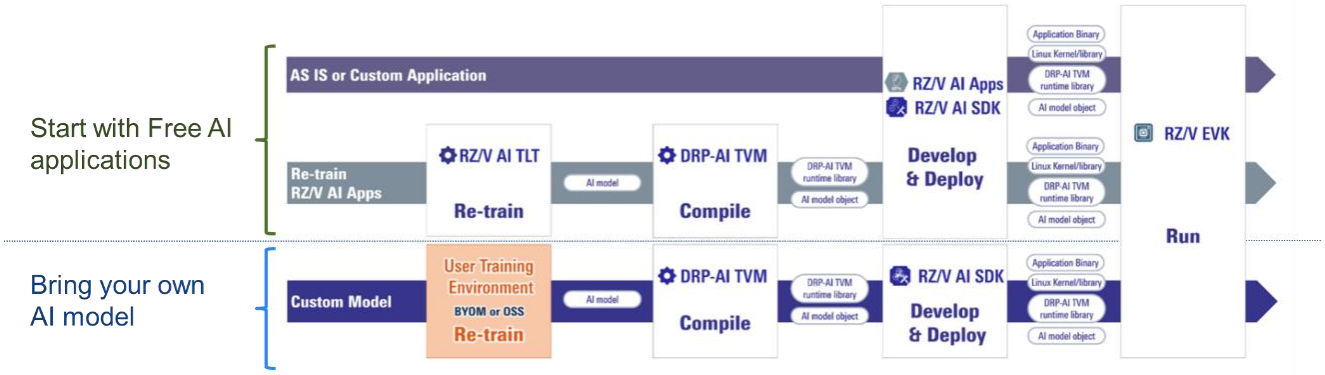

Figure 5: Configuration of DRP-AI Tools for Each User Application

Implementation flow for user-owned AI model (BYOM)

The model implementation flow for BYOM (Figure 6) has the following features:

Support for a wide range of Vision AI models: The device implementation and lightweight tools are designed to support a wide range of AI models and tasks to make extensive use of AI models that are evolving rapidly.

Multi-framework support: Supports a variety of AI frameworks, including PyTorch and TensorFlow.

Unified API across RZ/V products: Provides a unified API across RZ/V series product, ensuring consistency in user AI application development.

The specific process steps are as follows.

The DRP-AI Extension Pack retrains the model on an AI framework (Pytorch or TensorFlow) that has added DRP-AI support and creates a lightweight (pruned) model. It also supports retraining (Quantization Aware Training/QAT) to restore recognition accuracy after it is reduced due to model quantization (reduction from 32 bits to 8 bits). Pruning models and QAT support are optional for RZ/V2H and RZ/V2N.

Input the model and calibration data (a set of images to match the use case) into the DRP-AI TVM.

DRP-AI TVM estimates the optimal processing flow and memory access method for the DRP-AI hardware configuration and generates the executable file (runtime) data on the device.

Quantization of the model is automatically performed in the DRP-AI TVM using the calibration data (converting it from a 32-bit floating-point to an 8-bit integer). The lightweight model can also be simulated on the host PC and compared with the inference results of the original 32-bit model (interpreter mode).

AI inference can be performed on devices by deploying runtime data to the RZ/V2H or RZ/V2N.

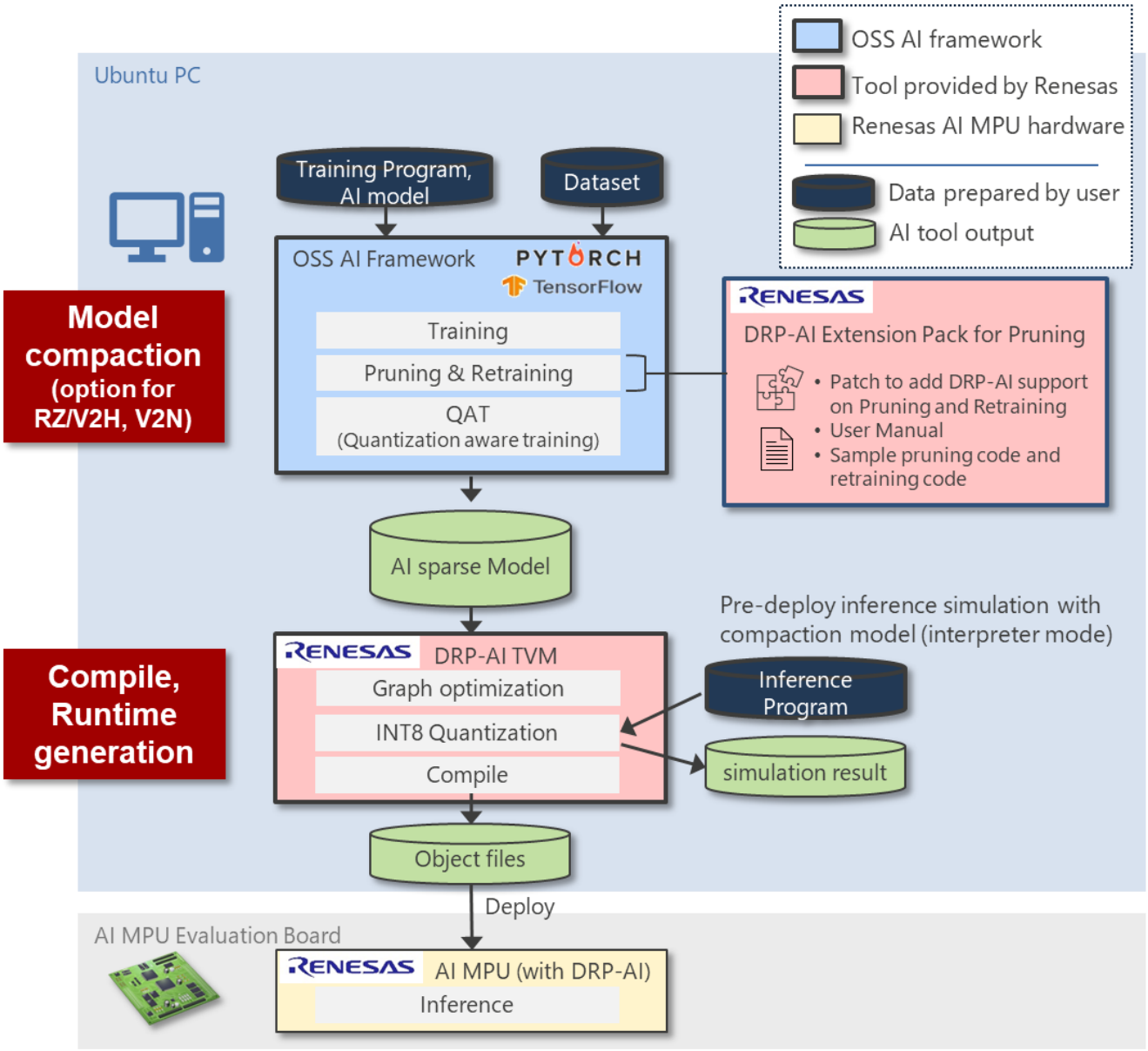

Figure 6: BYOM (Bring Your Own Model) Tool Flow

DRP-AI Extension Pack (Pruning Tool)

Reducing the weight of AI models with DRP-AI Extension Pack (pruning tool)

The latest DRP-AI (DRP-AI3) used on the RZ/V2H and RZ/V2N is equipped with a proprietary technology for high-speed, low-power processing of lightweight AI models using a technique called pruning, shown in Figure 7. Pruning is a technique for reducing the number of parameters by removing weights between nodes within a neural network. This reduces the hardware power consumption and increases the speed of the inference process.

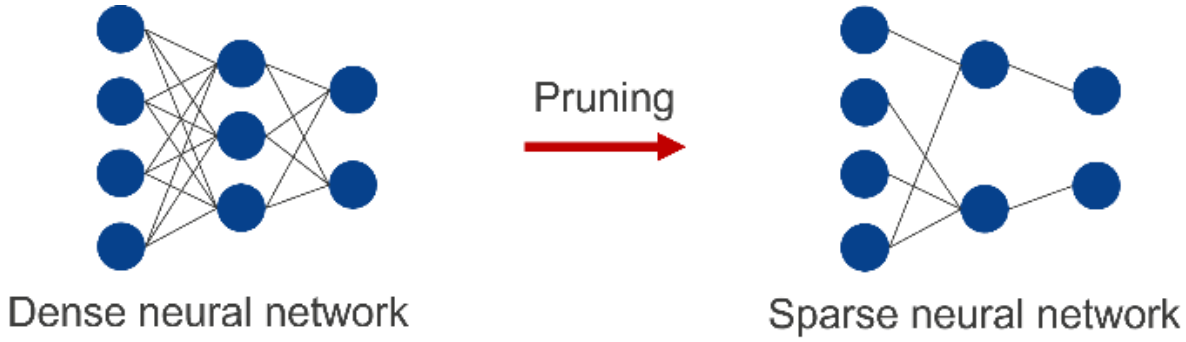

Figure 7: Reduction of Model Weight by Pruning

The DRP-AI Extension Pack provides optimized pruning functions for DRP-AI when combined with the user's PyTorch or TensorFlow training programs. The DRP-AI Extension Pack has the following features, which enable easy development even for users unfamiliar with lightweight technology, while supporting diverse user requirements.

1) Users can automatically generate lightweight models that are appropriate for their hardware, simply by setting the pruning rate.

2) If a learning program is available, it can be applied to any CNN model.

AI Model Compression (Pruning) Flow using DRP-AI Extension Pack

Figure 8 shows the pruning flow of a user-owned AI model (BYOM) using the DRP-AI Extension Pack.

Preparation of training program and dataset: When training the AI model, it is necessary to prepare the training program, and the dataset used for training (Pytorch and TensorFlow are supported frameworks).

Application of patches: Partially rewrite learning programs and apply DRP-AI extension pack patch.

Performing pruning and retraining: Performance of retraining is possible on servers or PCs with Pytorch or TensorFlow installed.

Analysis of pruning results: The pruning model recognition accuracy is analyzed by the board implementation evaluation or in interpreter mode (described later) to determine the optimal pruning rate following compilation and quantization by DRP-AI TVM.

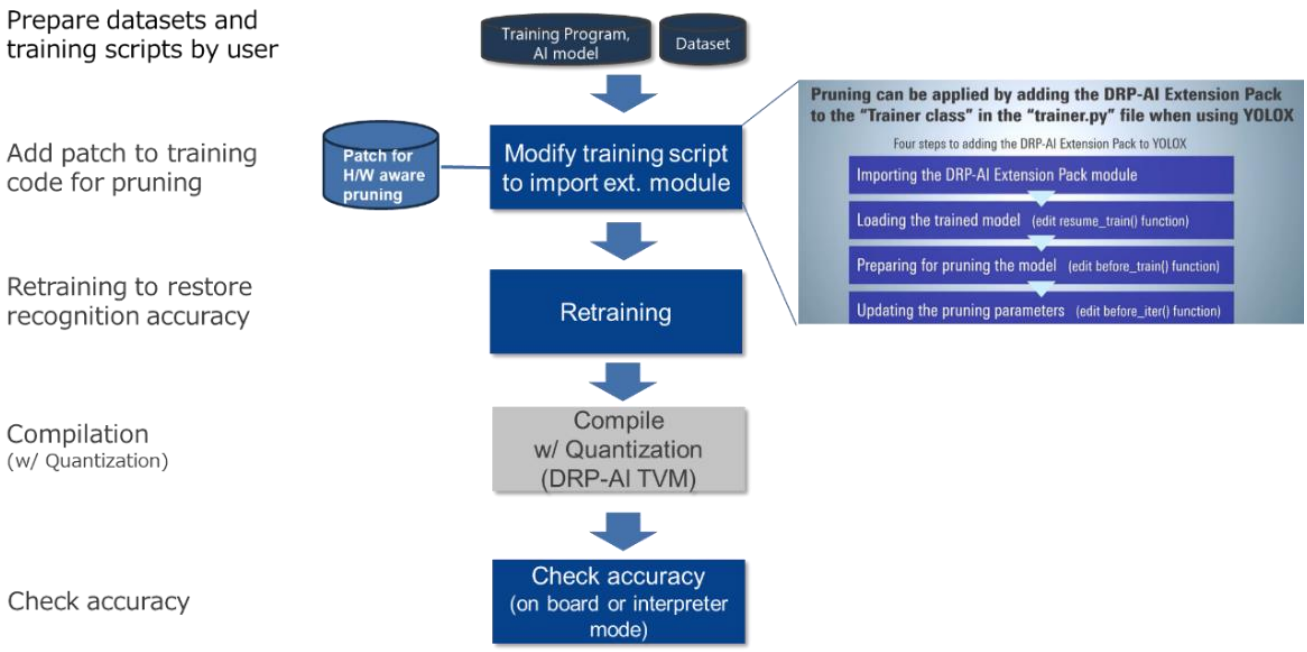

Figure 8: Model Compression Flow with DRP-AI Extension Pack

Videos showing how to modify the learning program, web guides detailing the flow, and sample scripts, etc. are provided on the DRP-AI TVM github.

Highly-flexible pruning settings that take recognition accuracy into consideration

To maximize the acceleration effect of DRP-AI it is important to select pruning nodes that are suitable for the hardware configuration of DRP-AI. The selection rules are incorporated in a patch in the DRP-AI Extension Pack, so it is possible for users to create relevant pruned models without regard for the selection rules. To use this patch, you need to prepare a training script and modify some of its code. At the same time, it is applicable to a wide range of tasks since it has the advantage of having no restrictions on the CNN models that can be pruned.

The node reduction rate (pruning rate) can be set to any value (70%, 80%, 90%, etc.), allowing fine adjustment of the tradeoff between recognition accuracy and speed. Furthermore, layers that may affect recognition accuracy if pruned are automatically excluded from the layers to be pruned. The deterioration of recognition accuracy is therefore suppressed even when a high pruning rate is set. At present the pruning rate is uniform except for the excluded layer, but Renesas plans to make improvements to allow more flexibility in the future.

Countermeasures for decreased recognition accuracy when pruning

Scalable CNN models have become the mainstream in recent years, and sizes such as S, M, and L may be selected according to the required recognition accuracy. In general, larger model sizes tend to result in less degradation of recognition accuracy during pruning compared to the pre-pruning base model (Figure 9). Therefore, when recognition accuracy drops more than necessary in a specific AI model, in addition to setting the pruning rate to a lower level, selecting a large model and setting the pruning rate to be higher is also effective.

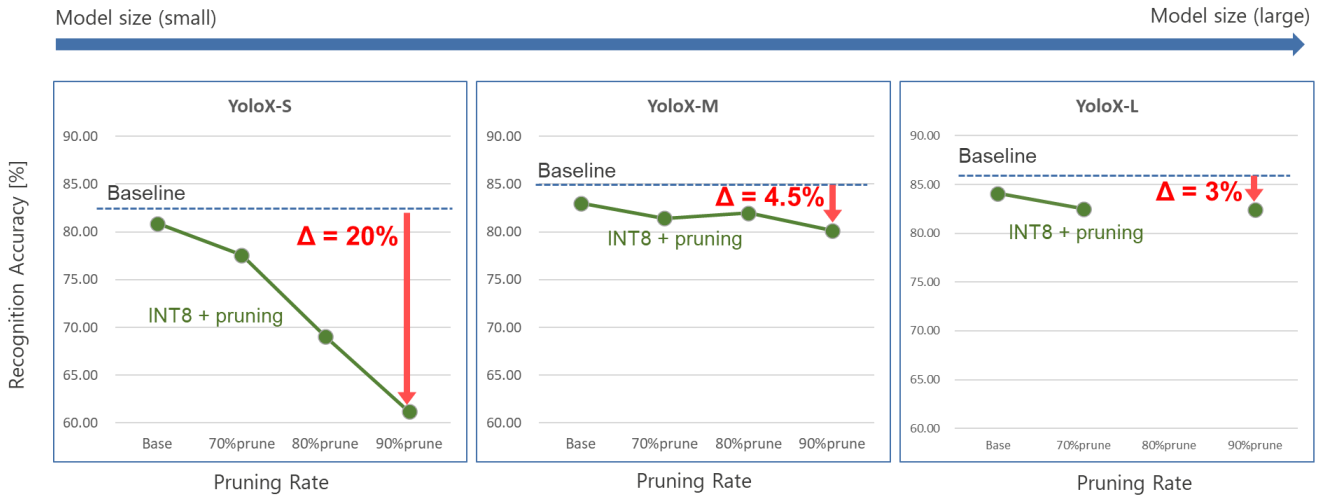

Figure 9: Relationship Between Model Size, Pruning Rate and Recognition Accuracy

A method to estimate inference speed before pruning

To balance the improvement of inference speed through pruning and the maintenance of recognition accuracy, it is necessary to change the pruning rate, perform re-learning, and repeat and adjust the loop to check the inference speed and recognition accuracy. To adjust with fewer loops, it is effective to estimate the inference speed when pruning and determine the target pruning rate in advance before performing time-consuming relearning. DRP-AI TVM provides a flow to estimate the inference time before and after pruning before performing the time-consuming relearning. To be specific, a provisional pruning model is created after setting the pruning rate, and the inference speed is measured on the actual machine. This can help determine whether to adopt a pruning model. For more information, read the "DRP-AI Extension Pack (Pruning Tool) Sparse Model Processing Speed Check Guide".

AI Compiler (DRP-AI TVM)

DRP-AI TVM Features

DRP-AI TVM is a compiler optimized for DRP-AI based on Apache TVM, which is widely used as an open-source AI compiler (Figure 10). DRP-AI TVM has the following features, including flexibility for the latest models, optimal processing performance, and scalability between products.

Support for heterogeneous configurations: Layer structures (operators) not supported by DRP-AI can be assigned to the CPU and operated in coordination between DRP-AI and the CPU to support diversifying AI models. The allocation of each layer to DRP-AI and the CPU is automatically done by the DRP-AI TVM understanding the input model structure.

Support for multiple AI frameworks: DRP-AI TVM supports major AI frameworks, including ONNX, PyTorch, and TensorFlow.

Generation of DRP-CPU integration runtime: This function has the ability to plan efficient scheduling of data transfers and calculations (such as memory management of both the Linux virtual memory space handled by the CPU and the physical memory space handled by DRP-AI, reuse of weight data, and minimization of external DRAM traffic through burst transfers) to efficiently coordinate between the AI accelerator (DRP-AI) and the CPU and automatically generate an integrated runtime.

Lightweight model support and compilation for DRP-AI: The input AI model is quantized and compiled for DRP-AI in the DRP-AI TVM. Specifically, it performs quantization from 32-bit floating-point format (FP32), a data type for common AI models. to 16-bit (FP16) (RZ/V2L, V2M, V2MA) or INT8 (RZ/V2H, V2N). When a lightweight pruned model made with the DRP-AI Extension Pack is input, optimization and compilation are performed so that the pruning function of DRP-AI is automatically applied.

Compatibility between RZ/V series: The DRP-AI TVM can use the same compiler and API across multiple products, so AI applications that run on the DRP-AI TVM can be used across products (some configuration file changes may be required).

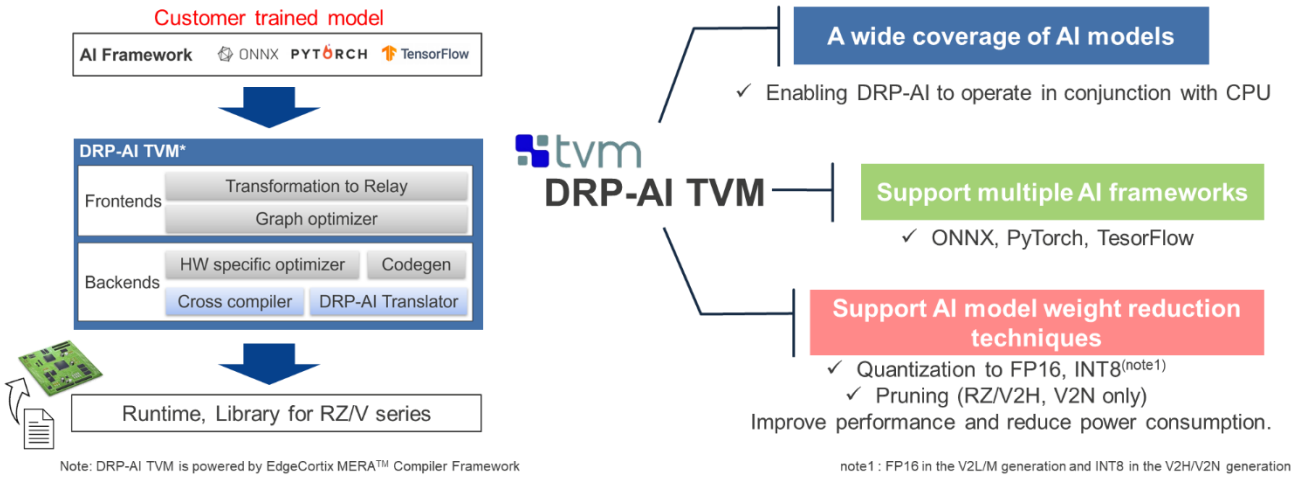

Figure 10: AI Compiler for DRP-AI (DRP-AI TVM)

Quantization for DRP-AI

PTQ (Post-Training Quantization), a quantization method that does not require retraining, is executed in the DRP-AI TVM in the DRP-AI AI model application implementation flow. Another method is QAT (Quantization-Aware Training), which improves the recognition accuracy of quantized models by retraining, and this is an optional flow that the user executes before inputting to DRP-AI TVM. Please refer to the DRP-AI TVM GitHub for the specific flow of QAT.

Compatible with pruning model for DRP-AI

The DRP-AI pruning function is automatically applied when using a lightweight pruned model with the DRP-AI Extension Pack. The compiler automatically analyzes the pruned weight data and generates code that maximizes the DRP-AI pruning function effectiveness. Using a pruned model allows for significantly reduced AI processing per image as well as reduced weight data transferred between computational memory and DRP-AI. This enables shorter inference times and a more power-efficient inference.

Increased pre-processing speed with DRP pipeline process

The pre-processing time for image adjustment for AI processing is relatively large when accelerating CNN inference using DRP-AI. In this architecture, the overall performance can be improved by using the DRP instead of the embedded CPU. In the DRP-AI test chip example, the overall inference time, including preprocessing, for an object recognition application (YOLOv2) is 6.5 times faster than when processed by an embedded CPU[1]. Furthermore, the image processing library optimized for the DRP is gradually being

expanded. By implementing these, it is possible to improve the speed by about one order of magnitude compared to CPU operation.

CNN model inference performance evaluation for RZ/V2H, RZ/V2N

CNN model inference performance

Figure 11 shows a comparison of the performance between generations of DRP-equipped products for a representative CNN model. Compared to the RZ/V2M equipped with the previous generation DRP-AI (maximum performance of 1 TOPS), the RZ/V2N equipped with DRP-AI3 (maximum performance of 4 TOPS without pruning or 15 TOPS with pruning) achieves an improvement in performance of approximately 4 times in the non-pruning model and up to 10 times or more in the model with a pruning rate of 90%. The improvement in performance when executing AI models is also due to technologies such as weight reduction through quantization and memory management, including built-in memory, as well as improved peak performance due to the increase in the number of AI-MACs.

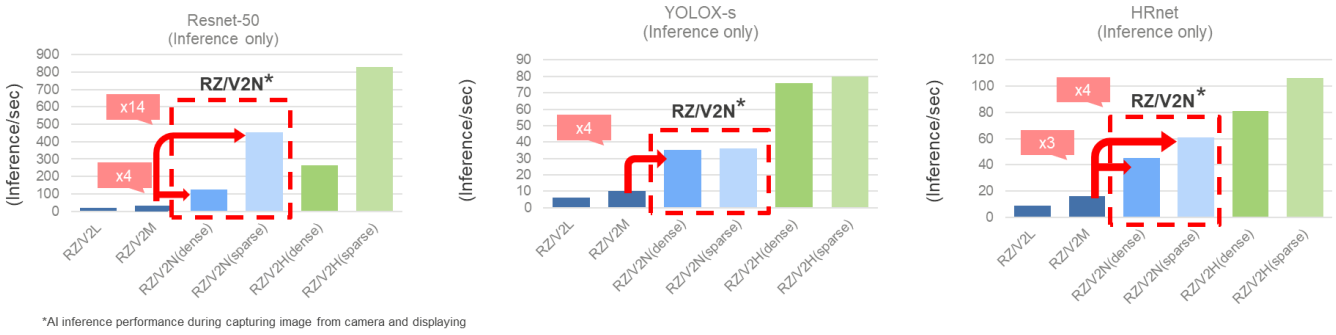

Figure 11: Comparison of CNN Inference Speeds Between RZ/V Series MPUs

Variables affecting the speed improvement for DRP-AI3

Here are some reasons why enhanced maximum performance and speed gains with pruning technology vary depending on the model.

The maximum processing performance (peak performance) is achieved by increasing the number of multiply-accumulate units in the AI-MAC. However, the performance improvement may not necessarily be proportional to the peak performance. Even if the number of multiply-accumulate operators is increased, the operators may not be used all the time due to the size of the AI model used and memory bandwidth restrictions. Furthermore, the RZ/V2N may have larger fluctuations in AI processing performance when running high-load applications other than AI simultaneously, since the memory bandwidth restrictions of the RZ/V2N are stricter than those of the RZ/V2H. It is therefore important to select a device that considers not only AI processing but also non-AI processing loads.

Next, we will show the model dependency of the pruning effect. The DRP-AI pruning model acceleration technology is a function that uses pruning to reduce the number of nodes (i.e., the number of operations) of the neural network, thereby reducing the inference time and power consumption per image. When it comes to power consumption, pruning reduces the number of multiply-accumulate calculations, thereby improving power efficiency. At the same time, the effect on inference time varies depending on the AI model used. While pruning can reduce the multiply-accumulate operation time and the amount of communication of node information (weight) (Figure 12 (A)), to shorten the inference time per image, the calculation result for each layer (feature map) is not reduced even by pruning (Figure. 12(B)). Therefore, as pruning speeds up

the amount of communication of the feature map increases and the speed tends to plateau due to limitations in memory communication volume (bandwidth). Renesas's pruning method can achieve performance improvements of around 20% to 300%, compared to a speed improvement of around 10% achieved by pruning on a GPU[2], which means it is an effective technology in terms of performance as well.

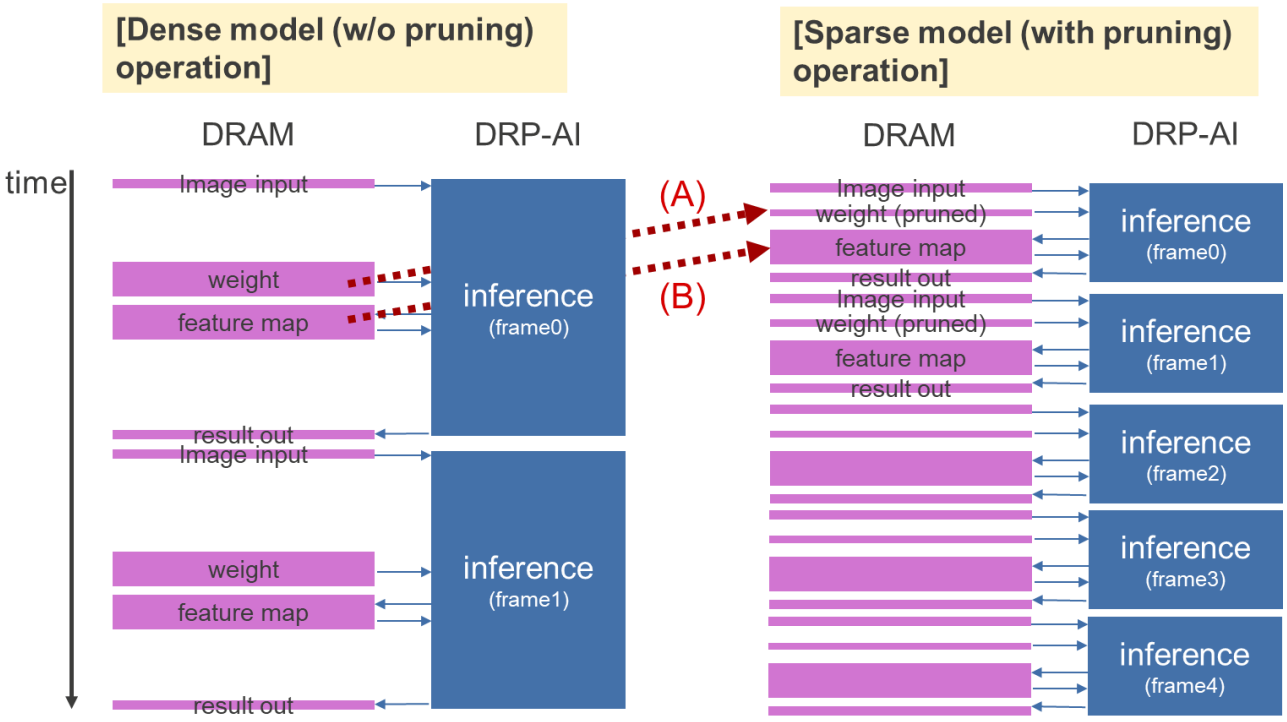

Figure 12: Comparison of Memory Access Between the Pre-Pruning Model (Dense) and the Post-Pruning Model (Sparse)

Introduction of Endpoints for Transformer Model

Transformer models have been increasingly adopted in endpoint devices in recent years due to their high performance and versatility. Vision Transformer models such as ViT3 and robot behavior generation models such as SayCan4/RT-25 have attracted attention as well as conversational models such as ChatGPT. To address this new trend, we have improved the highly flexible DRP-AI tool for transformers, which enables transformer models on the RZ/V2H and RZ/V2N.

At the same time, transformers have issues such as heavy use of different types of operations than those of CNN models. The memory size and computation requirements are also greater. Renesas plans to promote development of next-generation DRP-AI to improve transformer model computational efficiency and enhance model miniaturization techniques such as pruning, as well as improving the DRP-AI tool.

Trends and Challenges for Embedded Implementation of Transformer Models

Expanding the utilization of transformer models

Technological advancements in recent years have caused Vision AI to evolve from its current key element, the Convolutional Neural Network (CNN), to the transformer model. There are several key factors behind this evolution (Figure 13).

A transformer model has higher recognition accuracy than a CNN, because the transformer uses self-attention mechanisms to efficiently process dependencies of information at large distances in an image. This makes it possible to handle a wide range of tasks, including not only image recognition but also time series prediction and natural language processing. A transformer model is also well suited for multimodal data processing. For example, it performs well in tasks that combine multiple data formats, such as simultaneous processing of images and text and integration of audio and video. Furthermore, advances in the development of technology to reduce the weight of transformer models have enabled the processing of advanced transformer models in real time, with low power consumption.

More recently, it has been shown that models combining CNN and Self-Attention (e.g., YOLO v 10 and Topformer) are more effective than traditional image AI (CNN models) and are attracting attention as models that can achieve a good balance of accuracy and performance. These models combine the feature extraction capability of a CNN with Self-Attention's ability to comprehend long-range dependencies, in order to achieve greater accuracy and efficiency.

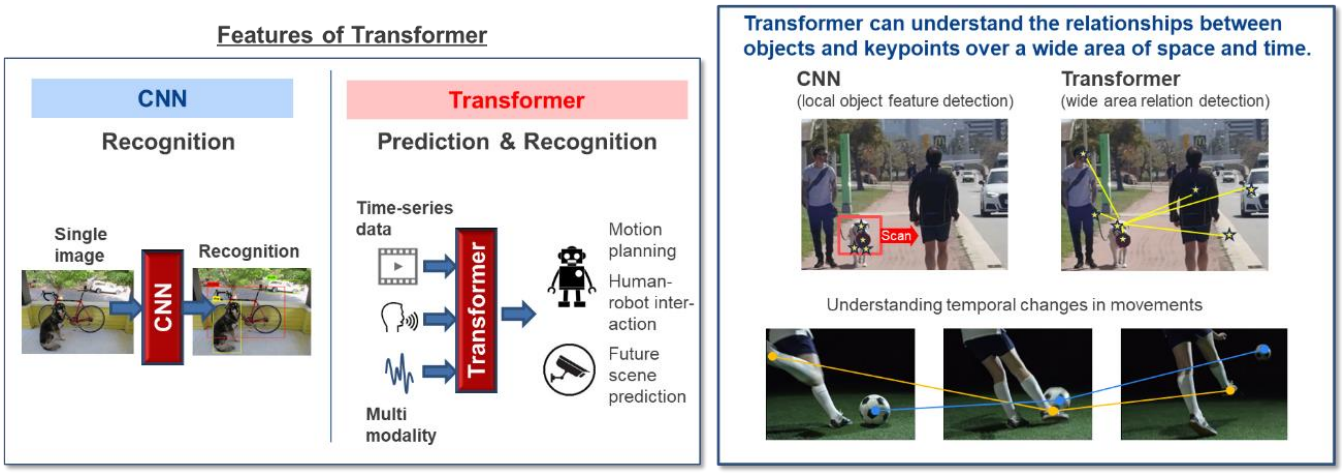

Figure 13: Features of Transformer Models

Challenges for Implementing Transformer Endpoints

However, transformers have issues such as the heavy use of different types of calculations and a larger model size when compared to CNN models. (Figure 14). It is necessary to solve these challenges to implement transformer models at endpoints.

Increased computational complexity: Unlike CNNs[6], where more than 90% of the computations are convolutional, transformers have a complex configuration that makes extensive use of highly complex operations specific to transformers such as multiply-accumulate operations between feature maps that do not use weight data, Softmax, GELU, and LayerNorm[7]. This can result in issues such as the AI compiler not being fully supported and impossible to implement on the device, and these operations can become performance bottlenecks and slow down the AI inference speed.

Increased model size and computational complexity: Transformer models can achieve higher recognition accuracy than conventional CNN models. On the other hand, model sizes and the number of calculations tend to be several times larger, which is a barrier to implementation on endpoint devices with limited memory and computational resources.

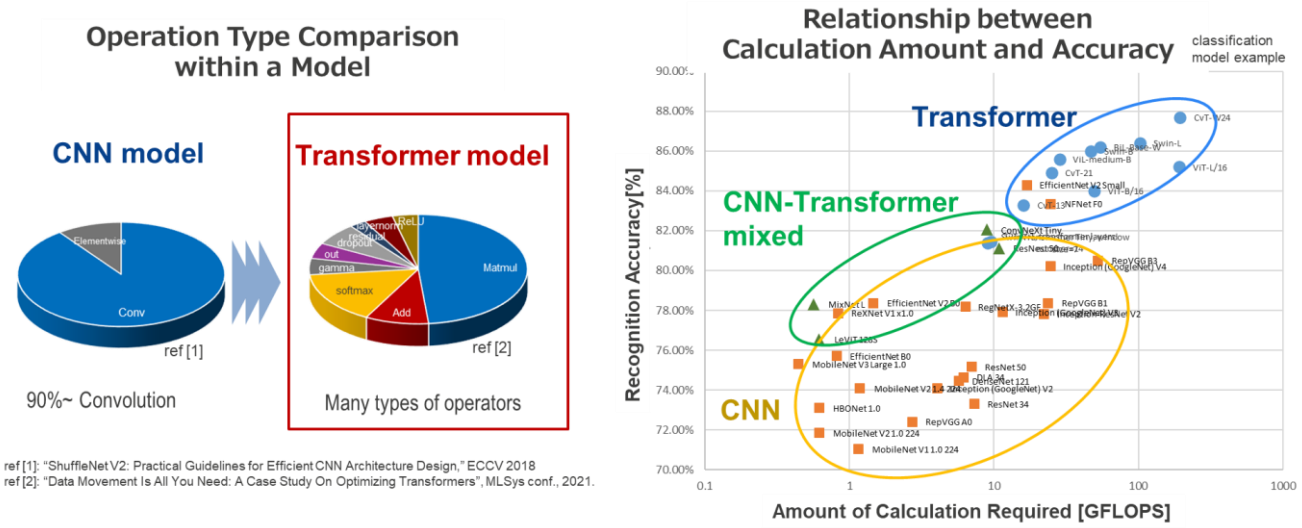

Figure 14: Challenges for Implementing Endpoint AI for Transformer Models

Transformer Model Support with DRP-AI3

Transformer Optimization Technology with DRP-AI3

The DRP-AI3 implemented in RZ/V2H and RZ/V2N has an architecture optimized for CNN. Renesas is evolving its AI compiler and other environments to make these technologies compatible with transformers. To be specific, we are promoting the development of the following technologies on a continuous basis.

Accelerating Diverse Activations: Continue to develop high-speed libraries that utilize Dynamically Reconfigurable Processor (DRPs) for the diverse and large-scale computational processes of transformers.

Optimization of matrix operations: Enable high-speed processing by DRP-AI through conversion of matrix operations, which account for most of the transformer's operations within the DRP-AI TVM to operations (operators) for CNN.

Application of pruning and quantization technology: DRP quantization and pruning tools can be applied to transformers as well as CNNs. As a result, it can also be used to increase the speed of pruning models, as with CNNs. Quantization Aware Training (QAT) will also be supported.

Transformer support for AI compiler (DRP-AI TVM)

Renesas has enhanced transformer support for DRP-AI TVM, and some transformer models and operators can be executed on RZ/V2H and RZ/V2N from version 2.4 released in October 2024. As of March 2025, models that combine transformer models and CNN models are supported (Figure 15) in addition to vision transformers such as ViT and Swin[8]. Furthermore, there are several restrictions, since DRP-AI3 is originally an architecture for CNN, and the DRP-AI tool is still in the process of evolving.

- It is not possible to convert some models because several operators are not supported.

- INT8 quantization may lead to a large decrease in accuracy.

- Operators not compatible with DRP-AI are processed by the CPU, so the speed improvement is up to about 10 times that of the CPU.

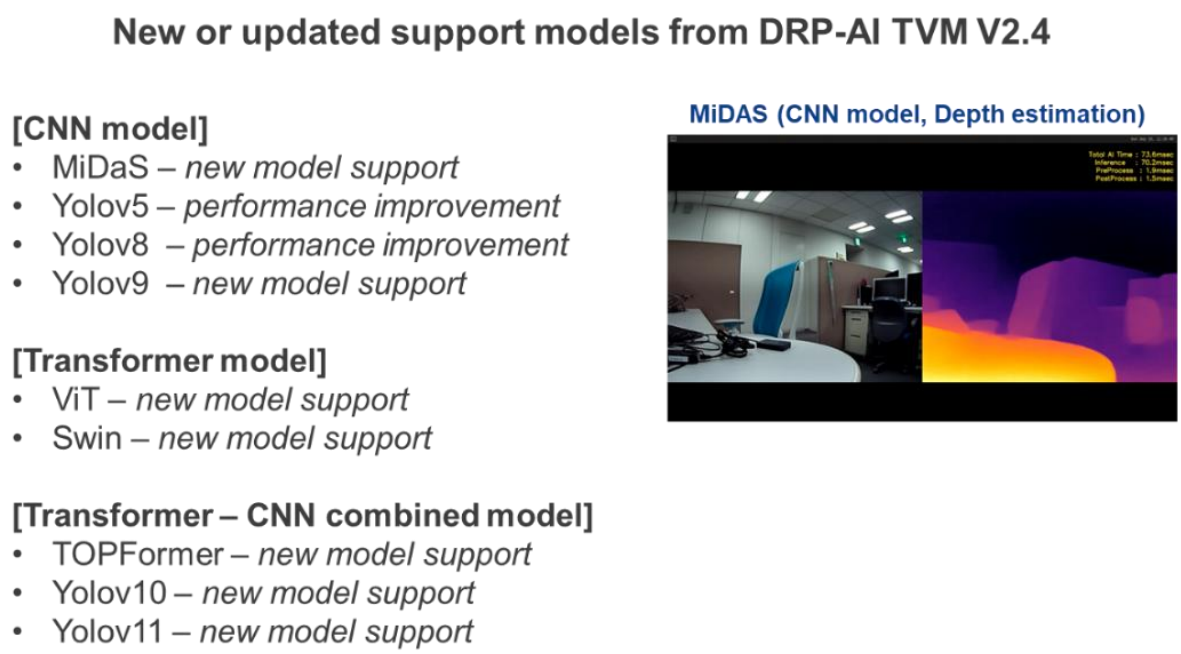

Figure 15: Main AI Models Newly Supported and Updated in DRP-AI TVM (Version 2.4)

Transformer Implementation Example

Topformer is an abbreviation for Token Pyramid Transformer, and is an architecture specialized for semantic segmentation for mobile devices 9. Here, a "token" refers to an image or data that is divided into smaller parts. Topformer takes tokens of various sizes as input and generates features with meanings according to their size. This enhances the relationship between tokens and improves data expressiveness. By combining CNN and transformer layers, this model provides a greater balance between accuracy and speed than models with only CNN or transformers.

This time, we compiled the Topformer model using DRP-AI TVM V2.4 and successfully implemented it on the RZ/V2H. The inference time of the model part is less than 100 msec, enabling performance of highly accurate real-time segmentation processing.

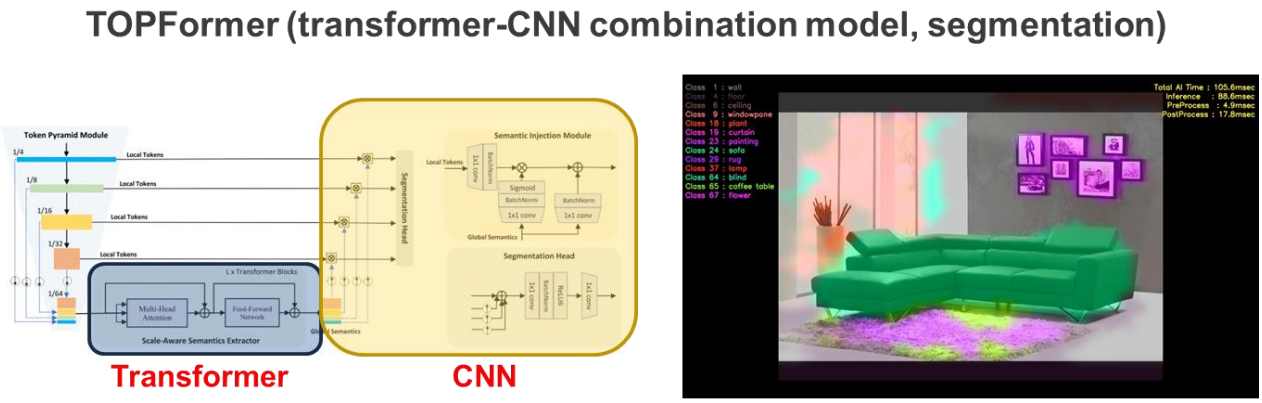

Figure 16: CNN-Transformer Mixed Model Example (Topformer)

What's next: Development of Next-generation AI accelerators (DRP-AI4)

The next-generation AI accelerator (DRP-AI4) is an evolution of the current DRP-AI3 and aims to achieve greater performance and more efficient AI processing. Specifically, we are further enhancing the flexibility of the DRP-AI architecture and its support for lightweight models, and developing an AI accelerator that can process both transformers and CNNs at high speed and with lower power consumption.

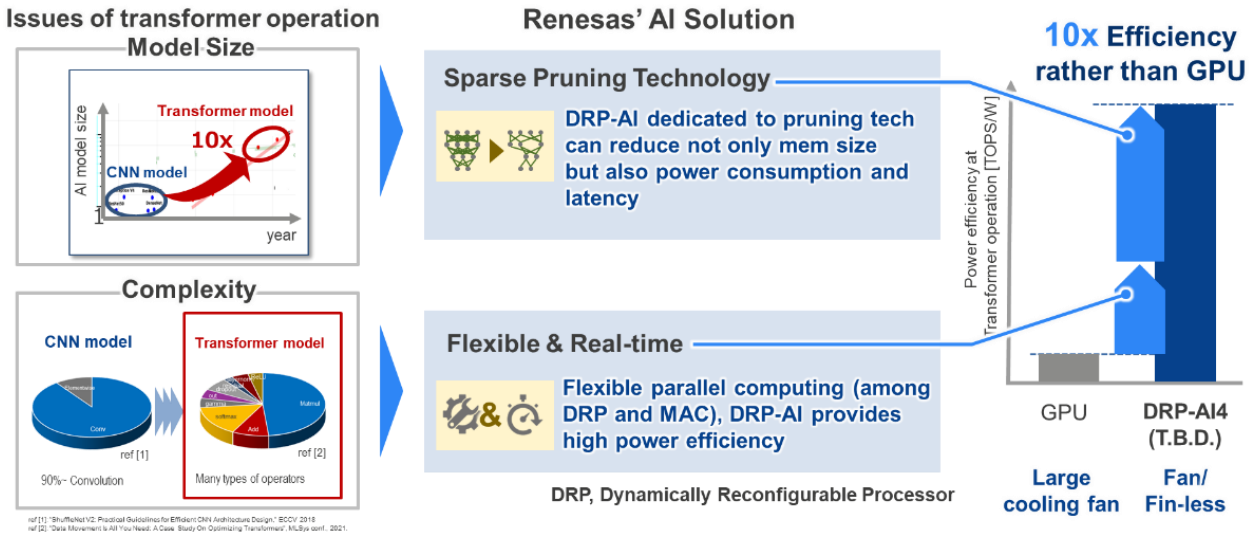

Figure 17: Initiatives for next-generation DRP-AI (DRP-AI4)

Conclusion

Renesas consistently develops technologies and products that capture the latest technological trends and bring them to market. Our goal is to ensure that customers can feel comfortable using our products. We will continue to offer innovative solutions and support the success of your business going forward.

Reference

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.

Related Information

RZ/V2N: 15TOPS Quad-Core Vision AI MPU with 2-Camera Connection and Excellent Power Efficiency

RZ/V2H: Quad-core Vision AI MPU with DRP-AI3 Accelerator and High-Performance Real-time Processor

Embedded AI-Accelerator DRP-AI White Paper

Next Generation Highly Power-Efficient AI Accelerator (DRP-AI3) White Paper

(Rev.1.0 Mar 2020)

Corporate Headquarters

TOYOSU FORESIA, 3-2-24 Toyosu, Koto-ku, Tokyo 135-0061, Japan

https://www.renesas.com

Contact Information

For further information on a product, technology, the most up-to-date version of a document, or your nearest sales office, please visit:

https://www.renesas.com/contact-us

Doc Number: R01WP0027EU0100