中端 MPU RZ/V2N，高能效、应对庞大结构及复杂的视觉 AI 应用

Koichi Nose, **Masayuki Shimobeppu**, **Takao Toi**, Embedded Processor Product Management Department, Embedded Processing Marketing Division, Embedded Processing Product Group, 瑞萨电子

Kentaro Mikami, System Solution Department 1, Software Development Division, Software & Digitalization Group, 瑞萨电子

概述

业内对在端侧设备中执行高级 AI（人工智能）指令的系统需求日益增加，尤以要求实时性能的机器人、AI 摄像头和无人机等为中心的应用需求不断增长。 此外，由于 AI 能够处理更加高级和多样化的任务，要求系统的处理性能和能效也进一步提升。

为了满足这一需求，瑞萨电子面向 视觉 AI 推出了搭载高效 AI 加速器 （DRP-AI）的 RZ/V 系列 MPU（微处理器）。 本白皮书将介绍 RZ/V 系列的最新产品概要、以及通过软硬件协同实现 AI 处理的优化技术。

端侧 AI 的趋势和挑战

端侧 AI 近年来发展迅猛，广泛运用于各种应用中。由于云端 AI 处理通信量庞大且耗时长，要求实时性的任务尤其是视觉 AI 领域，逐渐从云处理转移到端侧处理。

端侧 AI 的发展趋势

实时性能得到提升：传统的云端 AI 充分利用大量的数据和丰富的计算资源，而端侧 AI 可以在本地处理数据，能够最大限度地减少通信延迟并提升实时性。这样，能够实时处理传感器数据并提供即时反馈（图 1）。

AI 任务多样化：从狗猫识别到物体识别类图像的识别领域，再到依据静止图像预测 3D 深度，端侧 AI 支持各种预测任务。 近年来，不单识别 AI 的精度得到提升，用于规划和决策的 AI（例如，用于规划机器人最佳运动轨迹的 AI、以及基于环境和人类行为优化能源消耗的智能建筑）等广泛的应用开发也在不断推进。

模型轻量化：我们正在推进兼顾 AI 模型特性的轻量化技术和内存访问高效化技术，该技术可在端侧设备上实现高效 AI 处理。

从 CNN 模型到 Transformer 模型的发展：CNN（Convolutional Neural Networks，卷积神经网络）是一种专门用于图像识别的模型。Transformer 模型具备可高效处理长距离依赖关系的自注意力机制(Self-Attention)，除提升图像识别精度之外，它还能执行时间序列信息和自然语言处理等各种任务。

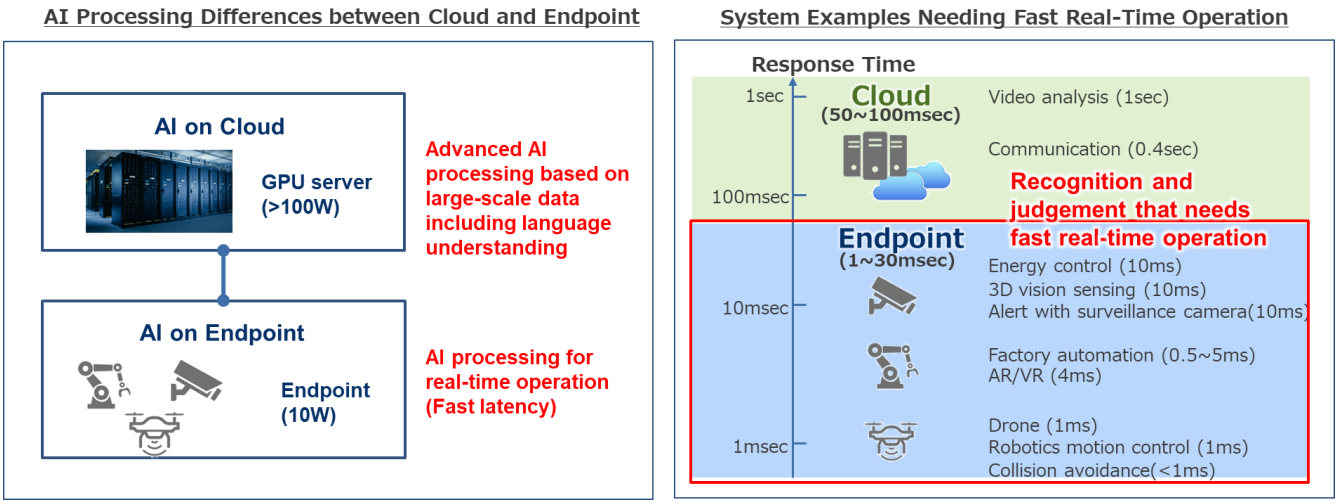

图 1：端侧 AI 的定位

传统端侧 AI 解决方案存在的问题

1）AI 模型规模庞大化：最初，图像 AI 需要比传统图像处理算法多出两到三位乘法累加运算，业内开发出了具有出色 AI 处理性能的专用 AI 芯片和 AI 加速器。然而，为了实现更高的识别精度，AI 模型越来越庞大，每张图像所需的计算量也在相应增加（图 2）。

2）AI 模型复杂化：AI 模型日益复杂化，尤其是伴随 Transformer 模型之类高级架构的出现。这样导致模型训练与推理所需的计算资源不断增加，从而使模型难以配置到端侧设备上。此外，复杂的模型还会导致内存使用量增加，而端侧设备有限的资源就无法支持这些模型。

3）AI 工具难度高：由于端侧设备的优化需要高水平的专业知识，故而 AI 工具往往是专业又复杂的。

另一方面，开发嵌入式设备系统与应用程序的用户对于 AI 的专业性与要求多种多样，开发能够满足广大用户需求的 AI 工具环境至关重要。 此外，AI 模型的选择、优化和设备配置需要大量的工作，不同工具之间的兼容性问题、针对特定硬件的优化工具不足也是待解决课题。

模型结构庞大化和复杂化导致人们担心功耗会进一步增加，而 RZ/V2N 系列搭载模型轻量化且功耗低的 DRP-AI3，有望成为解决难题的新产品，为端侧的 AI 配置做出重大贡献。

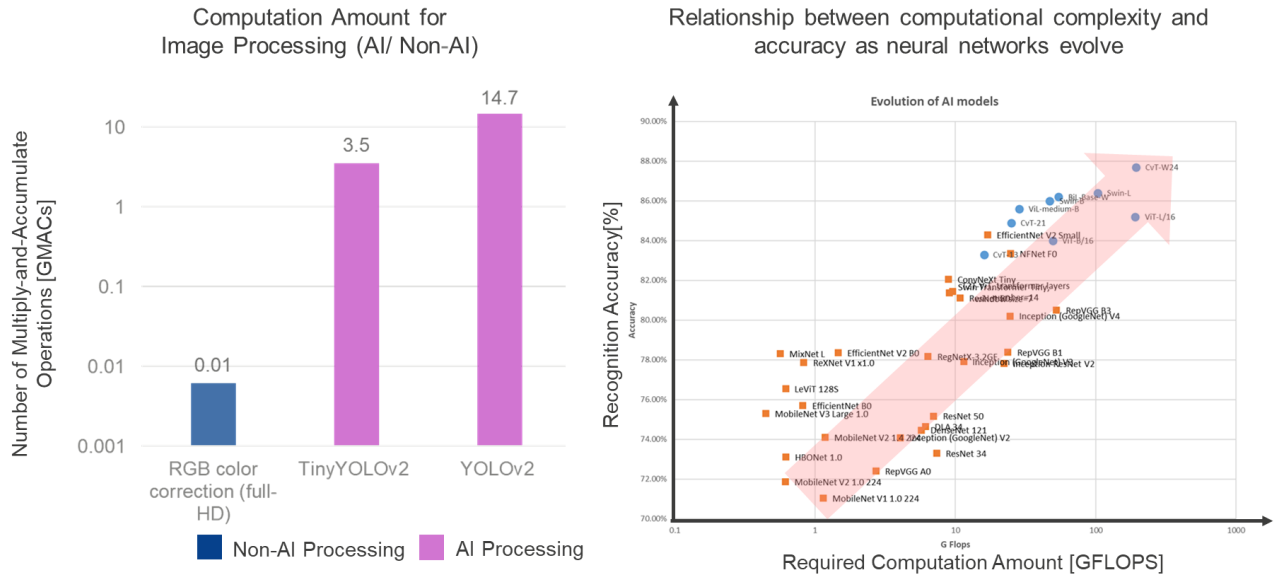

图 2：端侧 AI 解决方案面临的问题

适合 Vision AI 的 RZ/V 系列的特征

针对端侧视觉 AI 应用的 MPU 系列

RZ/V 系列是面向视觉 AI 的 MPU，采用瑞萨电子独创的高效 AI 加速器（DRP-AI）。使用瑞萨电子的动态可重构处理器（DRP: Dynamically Reconfigurable Processor），能够高效执行各种 AI 运算。除了最大算力为 80TOPS 的高端 AI MPU——RZ/V2H 之外，最佳性能可达 15TOPS 的 RZ/V2N 也将于 2025 年 3 月开始量产，后者具有 AI 性能可扩展性以支持各种端侧 AI 应用程序。

此外，RZ/V 系列旨在最大限度地发挥嵌入式产品的 AI 能效，因此能够降低发热量。例如，即使不使用风扇，RZ/V2H 也可将温度控制在与使用大型风扇的嵌入式 GPU 几乎相同的水平，因此适用于对安装尺寸和发热有严格限制要求的嵌入式设备。该系列最大能效约为 10TOPS/W，远优于其他公司。

目标应用

RZ/V 系列主要面向端侧 AI 市场中的高端（10TOPS 以上）市场到中端（1TOPS-10TOPS）市场(图 3)。面向高端市场的 RZ/V2H 主要用于需要高实时性能和高级 AI 的应用领域，例如协作机器人、AGV 和无人机；面向中端市场的 RZ/V2N 主要用于 DMS（驾驶员监控系统）、移动机器人和 AI 摄像头等应用领域。RZ/V2N 性价比高，非常适合 AI 性能要求高且要求合理价格的中端 AI 市场应用。

图 3： RZ/V 系列的目标应用

AI 加速器 (DRP-AI)的特征

DRP-AI 是瑞萨电子在上述 RZ/V 系列中搭载的独创 AI 加速器，采用下列技术实现了高性能和灵活性。

灵活支持广泛的 AI 处理：通过适用于图像识别 AI（CNN） 处理的优化 AI 乘法累加单元（AI-MAC）与动态重建处理器（DRP)之间的协同动作(图 4)，在大规模化与多样化的视觉 AI 处理中兼具高性能和灵活性。此外，DRP 可在 AI 处理之前执行图像处理，由此能够提升整个系统的速度。

支持 AI 模型轻量化：采用瑞萨电子独创的轻量化技术高速处理剪枝模型，能够减少计算量且不影响识别精度，可实现出色的能效(最大约 10TOPS/W)。

软硬件协同优化：瑞萨电子的专用 DRP-AI 工具可以生成优化 DRP-AI 处理的执行数据。

在本白皮书中，我们将主要介绍与 DRP-AI 工具相关的技术。 有关 DRP-AI 硬件技术的更多信息，请参阅 DRP-AI 白皮书和 ISSCC2024 论文[1]。

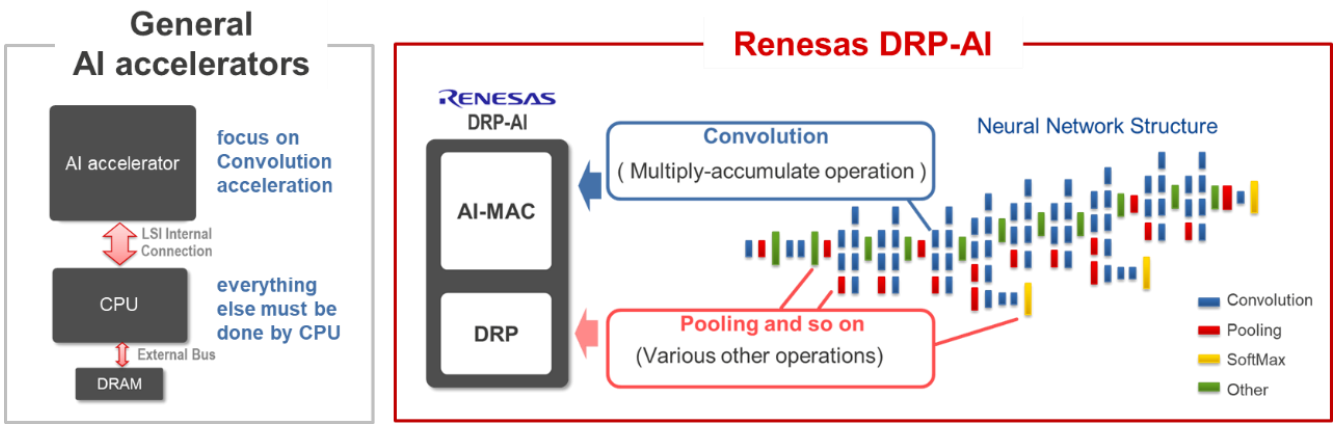

图 4： 搭载 DRP(Dynamically Reconfigurable Processor)的 AI 加速器(DRP-AI)

DRP-AI 工具配置

为了支持客户的 AI 应用程序开发，瑞萨电子面向 DRP-AI 提供 AI 开发工具，从 AI 初学者到专家的广大用户均可使用这些工具。本章介绍这些开发工具的整体结构和特征。在后半部分，我们将介绍本工具的特征——模型轻量化技术等。

DRP-AI 工具的特征

瑞萨电子提供图 5 所示的两个 AI 工具作为 DRP-AI 应用的开发环境。这样，从 AI 初学者到专家的广大用户都可以轻松使用产品。

使用瑞萨免费 AI 应用程序的用户：我们提供多种开源 AI 应用程序，这些应用采用预训练的模型。大多数应用程序都包含 AI 模型可执行文件，该可执行文件能够直接在设备上运行（图 6 中的 AS IS 或 Custom Application）。并且，我们还提供模型迁移学习工具（TLT），使用客户的数据重新训练模型，从而用户可以根据使用场自定义模型（图 6 的 Re-train RZ/V AI Apps）。有关详细信息，请参阅 AI 应用程序的 Github 。

自带 AI 模型的用户：我们还支持用户配置自定义模型（图 6 中的 Custom Model）。使用针对 DRP-AI 优化的 AI 编译器（DRP-AI TVM），可以生成可执行文件在 RZ/V 上高效运行用户的自定义模型。 此外，我们还提供支持自定义模型轻量化的工具（DRP-AI Extension Pack）供用户进行重新训练模型。

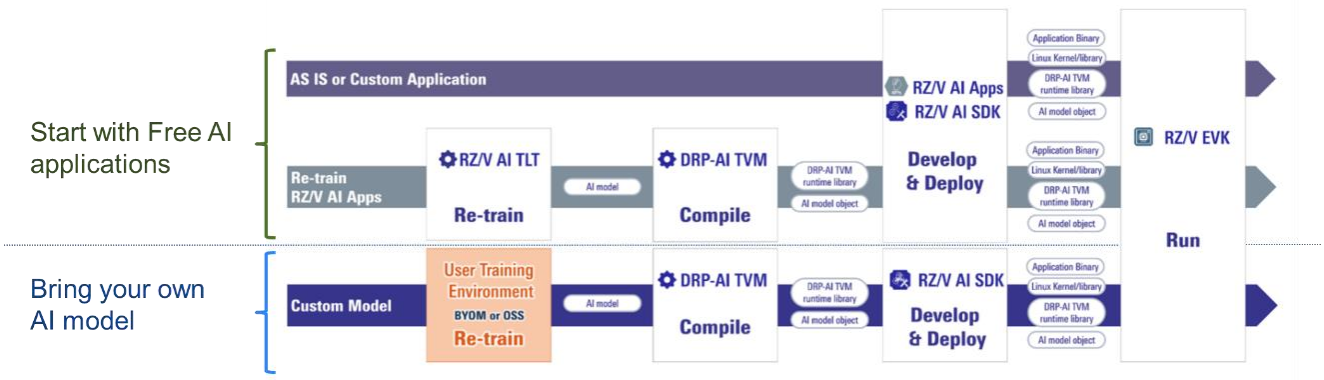

图 5：不同用户用途的 DRP-AI 工具配置

用户自带 AI 模型（BYOM）的安装流程

BYOM 模型安装流程（图 6）具有以下特点：

兼容各种视觉 AI 模型：设备安装工具与轻量化工具采用支持各种 AI 模型和任务的结构，确保用户能够广泛运用日新月异的 AI 模型。

多框架支持：支持 PyTorch 和 TensorFlow 等多种 AI 框架

RZ/V 产品之间统一 API：提供覆盖 RZ/V 系列产品的统一 API，以确保用户 AI 应用开发的一致性。

具体处理步骤如下。

在使用 DRP-AI Extension Pack 添加 DRP-AI 支持包的 AI 框架（Pytorch 或者 Tensorflow）上进行模型压缩，创建轻量化（剪枝）模型。此外，它还支持重新学习以恢复因模型量化而降低的识别精度（从 32 位减少到 8 位）(QAT: Quantization Aware Training）。（剪枝模型与 QAT 支持对于 RZ/V2H 和 V2N 是可选的）

在 DRP-AI TVM 中，输入模型和校准数据（用与训练相类似的图像数据集）。

DRP-AI TVM 生成 DRP-AI 硬件配置的最佳处理流程和内存访问方法等，并生成可在设备上执行模型数据。

在 DRP-AI TVM 内部自动量化（从 32 位浮点型转换为 8 位整数型）使用校准数据的模型。轻量化的模型通过在主机 PC 上进行模拟，可与原来的 32 位模型推理结果进行比较确认（解释器模式）。

将模型数据安装（部署）到 RZ/V2H 上，可以在设备上执行 AI 推理。

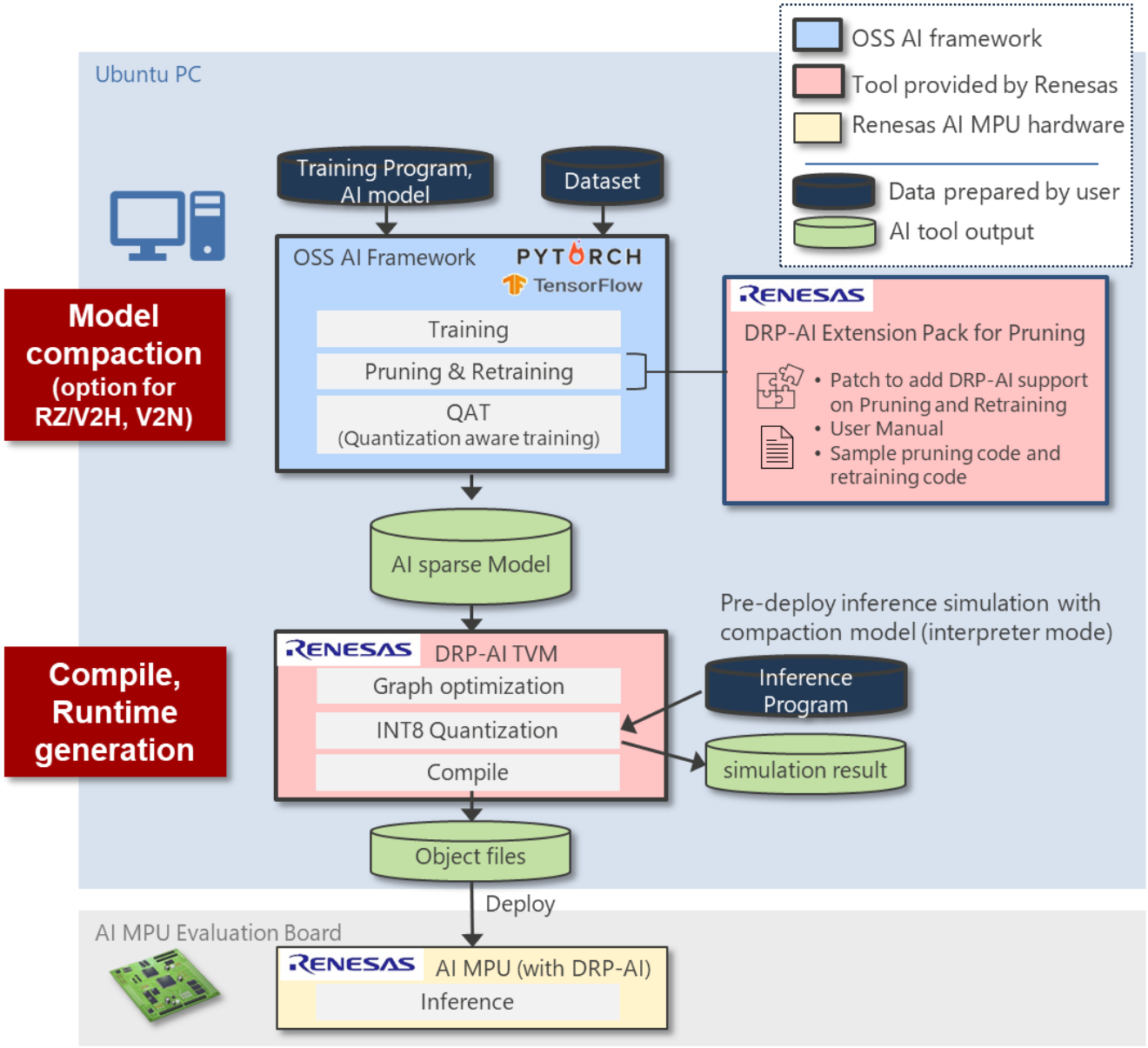

图 6：BYOM (Bring Your Own Model)工具流程

DRP-AI Extension Pack（剪枝工具）

通过 DRP-AI Extension Pack（剪枝工具）使 AI 模型轻量化

RZ/V2H 和 RZ/V2N 中搭载的最新 DRP-AI（DRP-AI3），采用独创技术，用于在高速和低功耗下对通过图 7 所示的“剪枝”技术实现 AI 模型轻量化处理。剪枝是一种通过删除神经网络中节点之间的权重来减少参数数量的技术，能够降低硬件功耗并加快推理过程。

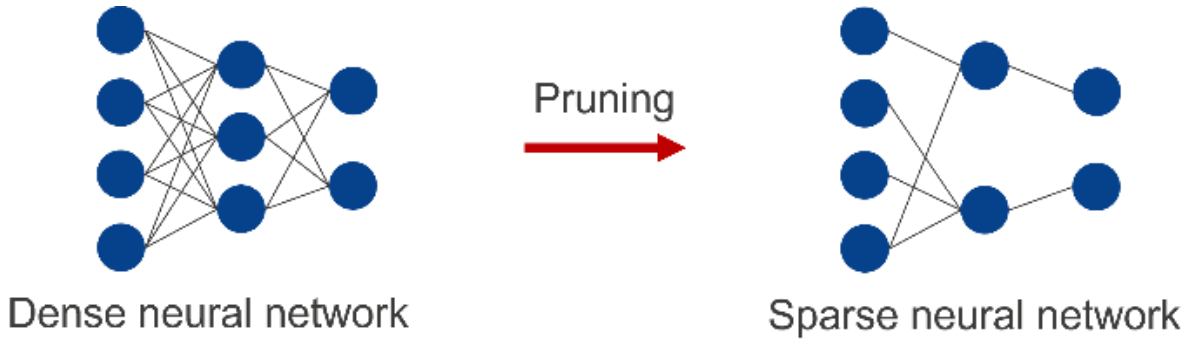

图 7：通过剪枝实现模型轻量化

DRP-AI Extension Pack 与用户的 PyTorch 或 TensorFlow 训练程序结合起来，可提供优化 DRP-AI 的剪枝功能。DRP-AI Extension Pack 具备下列特征，不但能够满足多样化的用户需求，即使不熟悉轻量化技术的用户也可轻松参与开发工作。

1）用户只需设置剪枝率，即可自动生成适合硬件的轻量化模型。

2）如果有训练程序，可适用于任意的 CNN 模型。

使用 DRP-AI Extension Pack 的 AI 模型压缩（剪枝）流程

图 8 表示使用 DRP-AI Extension Pack 的用户自带 AI 模型(BYOM)的剪枝流程。

准备训练程序和数据集：在训练 AI 模型时，需要准备训练程序和用于训练的数据集。（可支持的框架包括 Pytorch 和 Tensorflow）

补丁应用：重写部分训练程序，运用 DRP-AI Extension Pack 补丁。

剪枝与重新训练的执行：可以在安装 Pytorch 或 Tensorflow 的服务器或 PC 上执行重新训练。

剪枝结果分析：经过 DRP-AI TVM 编译和量化后，在开发板上实施评估或者在下述解释器模式下分析剪枝模型的识别精度，以确定最佳剪枝率。

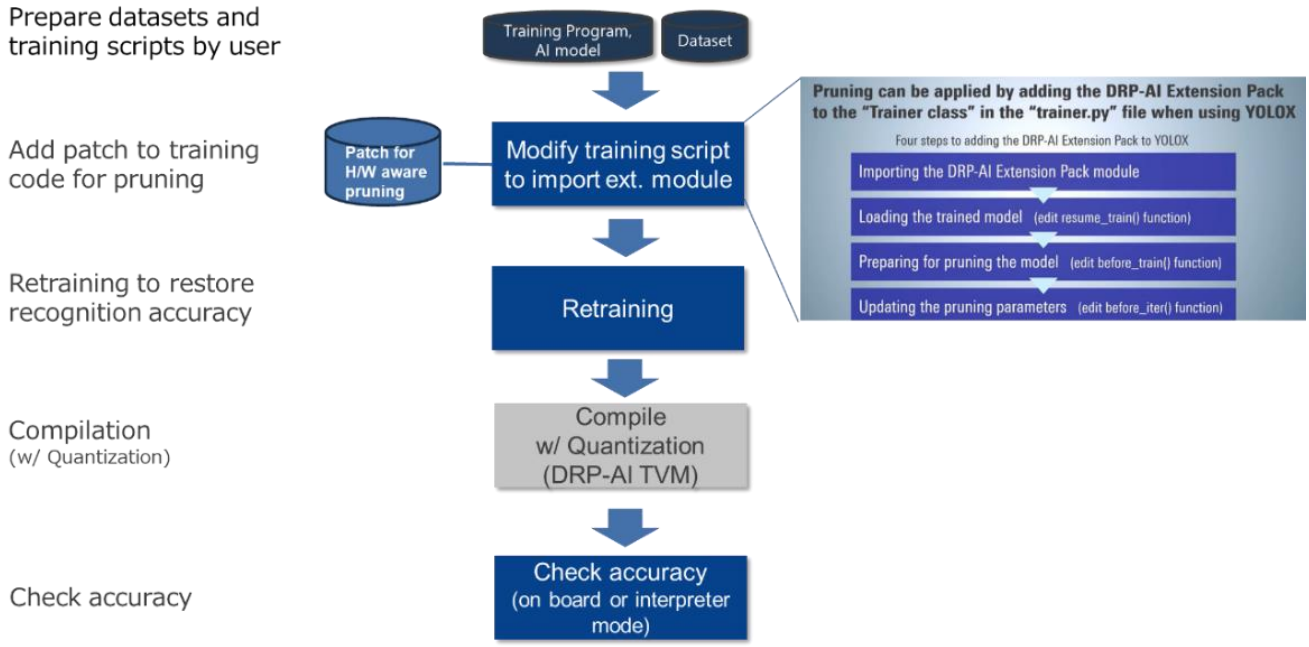

图 8： 使用 DRP-AI Extension Pack 的模型压缩流程

DRP-AI TVM 的 Github 提供了展示训练流程修改方法的视频、详细介绍流程的 Web 指南和示例脚本等。

考虑识别精度，随意设置修枝参数

为了通过剪枝最大限度地发挥 DRP-AI 的加速效果，选择适合 DRP-AI 硬件配置的剪枝节点非常重要。这个选择规则内置于 DRP-AI Extension Pack 的补丁之中，用户可在不了解选择规则的情况下创建适当的剪枝模型。若要使用该补丁，用户需要准备训练脚本并修改其部分代码。 另一方面，它的特点是对能够修剪的 CNN 模型没有限制，可以广泛用于各种任务。

由于节点缩减率（剪枝率）可以设置为任意值（70%、80%、90%等），因此能够极为细致地调整识别精度和速度之间的权衡关系。 此外，它还会从修剪目标层自动排除修剪后影响识别精度的层。这样，在设置较高的剪枝率之际也能抑制识别精度的恶化。当前，除了排除层之外，它的剪枝率设置是一致的，但我们也计划在未来进行改进以增加自由度。

防止剪枝时识别精度下降的措施

近年来，CNN 模型以可扩展模型为主流，用户可根据所需识别精度选择 S、M 或 L 等大小。一般来说，与剪枝前的基本模型相比，模型尺寸越大，剪枝过程中识别精度的降低程度就越小（图 9）。因此，在特定的 AI 模型中，当剪枝时识别精度降低程度超过必要水平之际，作为对策，除了 1）中 将剪枝率设为较低值之外，采用 2）中选择更大尺寸的模型并将剪枝率设置得更高也是一种行之有效的方法。

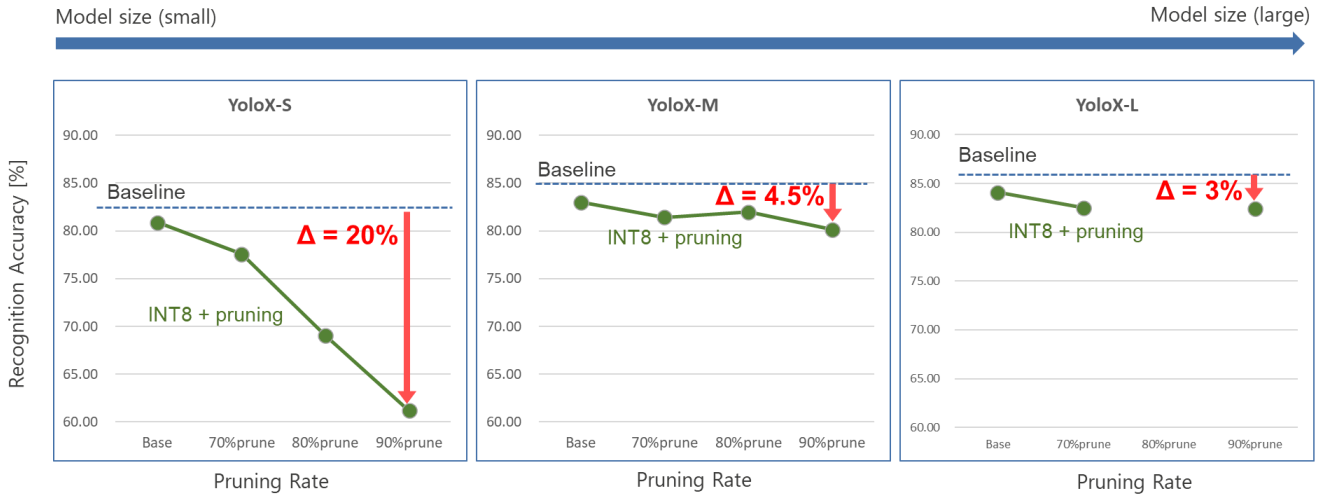

图 9： 模型大小与剪枝率和识别精度之间的关系

剪枝时推理速度的预算方法

为了在利用剪枝提升推理速度和保持识别精度之间取得平衡，需要改变剪枝率和进行重新训练，并反复调整循环次数以确认推理速度和识别精度。为了在更少的循环次数下进行调整，在执行耗时的重新训练之前，计算剪枝时的推理速度并提前确定目标剪枝率行之有效。DRP-AI TVM 提供了一个流程，用于在执行耗时的重新训练之前计算剪枝前后的推理时间。 具体来说，该方法就是在设置剪枝率后创建一个临时剪枝模型，然后在实机上测量推理速度。这样，就有助于用户部署剪枝模型。有关详细信息，请参阅《DRP-AI Extension Pack (Pruning Tool) Sparse Model Processing Speed Check Guide 》。

AI 编译器（DRP-AI TVM）

DRP-AI TVM 的特征

DRP-AI TVM 是基于 Apache TVM 的 DRP-AI 优化编译器，Apache TVM 被广泛用作开源 AI 编译器（图 10）。DRP-AI TVM 具有以下特征，实现了最新型号应用的灵活性、最佳处理性能和产品之间的可扩展性。

支持异构配置：为了支持多样化的 AI 模型，可将 DRP-AI 不支持的层结构（算子）分配给 CPU，并在 DRP-AI 和 CPU 之间执行协同动作。将各层分配给 DRP-AI 和 CPU 都是在 DRP-AI TVM 输入模型结构，识别之后自动执行操作。

支持多种 AI 框架：DRP-AI TVM 支持主流 AI 框架，包括 ONNX、PyTorch 和 TensorFlow 等。

DRP-AI 和 CPU 协同环境：为了在 AI 加速器（DRP-AI）和 CPU 之间实现高效协同动作，它具备规划数据传输与运算的有效调度（比如，CPU 处理的 Linux 虚拟内存空间和 DRP-AI 处理的物理内存空间两者的内

存管理、权重数据的再利用、通过突增传输使外部 DRAM 流量最小化等）、并自动生成集成 DRP-AI 和 CPU 协同环境 runtime 的功能。

DRP-AI TVM 支持和编译轻量化模型： AI 模型通过 DRP-AI TVM 量化和编译转换成可在 DRP-AI 执行的模型。具体来说，就是进行从一般的 AI 模型数据类型即 32 位浮点形式（FP32）到 16 位(FP16)的量化（RZ/V2L, V2M, V2MA）或者到 INT8 的量化（RZ/V2H, V2N）。在 DRP-AI Extension Pack 中输入轻量化的剪枝模型后，系统将自动进行优化和编译以运用 DRP-AI 的剪枝功能。

RZ/V 系列之间的兼容性：DRP-AI TVM 能在多个产品中使用相同的编译器和 API，因此在 DRP-AI TVM 上运行的 AI 应用程序可以跨产品通用（可能需要更改某些配置文件）。

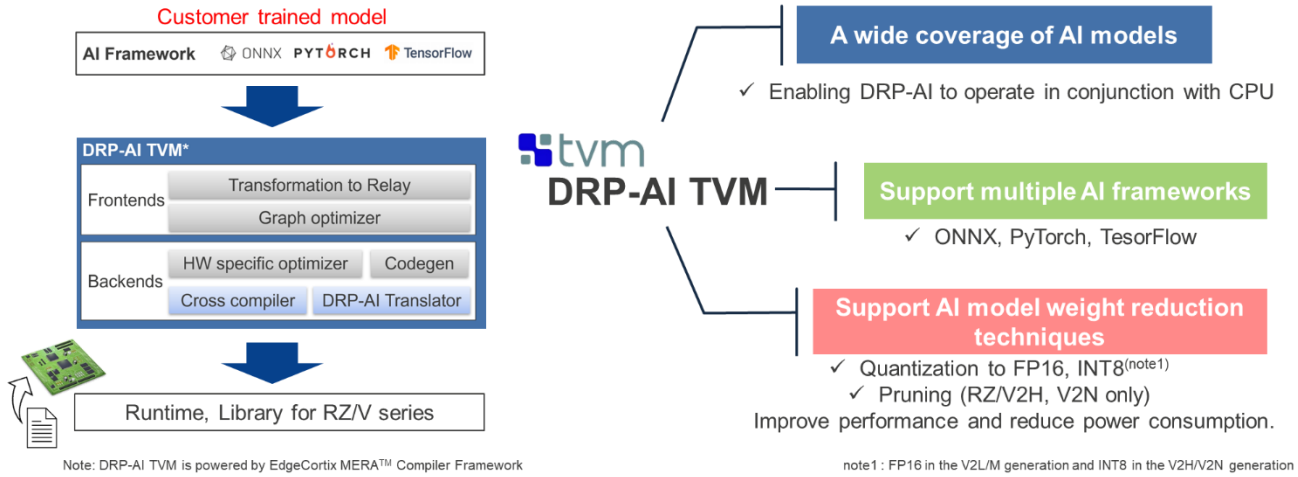

图 10： 用于 DRP-AI 的 AI 编译器 (DRP-AI TVM)

DRP-AI 的量化

在 DRP-AI 的 AI 模型和应用配置流程中，无需重新训练的量化手法 PTQ（Post-Training Quantization）在 DRP-AI TVM 内部执行。另一种手法就是 QAT（Quantization-Aware Training），通过重新训练来提高量化模型的识别精度。它是一种可选流程，在输入 DRP-AI TVM 之前由用户执行操作。关于 QAT 的具体流程，请参考 DRP-AI TVM 的 GitHub 章节

兼容 DRP-AI 剪枝模型

当用户使用经过 DRP-AI Extension Pack 轻量化的剪枝模型时，系统将自动应用 DRP-AI 剪枝功能。编译器自动分析剪枝后的权重数据，并生成可最大限度发挥 DRP-AI 剪枝功能的代码。使用剪枝后的模型，能够大幅减少每张图像的 AI 处理量和计算量、以及内存和 DRP-AI 之间的权重数据传输量，从而缩短推理时间并实现高能推效理。

使用 DRP 管线处理加快预处理速度

在使用 DRP-AI 加快 CNN 推理之际，AI 处理的图像调整的前处理时间相对较长。在这种架构中，可以使用 DRP 代替嵌入式 CPU 来提高整体性能。在使用 DRP-AI 测试芯片的示例中，对象识别应用程序(YOLOv2)的总体推理时间（包括预处理）比嵌入式 CPU 快 6.5 倍[1]。 此外，我们正在逐步扩充针对 DRP 优化的图像处理库。通过配置这些，与 CPU 动作相比，可将速度提高大约一个数量级。

RZ/V2H 和 RZ/V2N 的 CNN 模型推理性能评估

CNN 模型的推理性能

图 11 表示搭载 DRP-AI 的几代产品之间典型 CNN 模型的性能比较结果。搭载 DRP-AI3 的 RZ/V2N（最大性能 4TOPS（未剪枝）、15TOPS（已剪枝））与上一代搭载 DRP-AI 的 RZ/V2M（最大性能 1TOPS））相比，未剪枝模型的性能提升约 4 倍，剪枝率为 90%的模型的性能提升达到 10 倍以上。AI 模型运行时的性能提升，除了 AI-MAC 数量增强带来的峰值性能强化以外，通过量化实现轻量化、包括内置内存在内的内存管理等技术也做出了贡献。

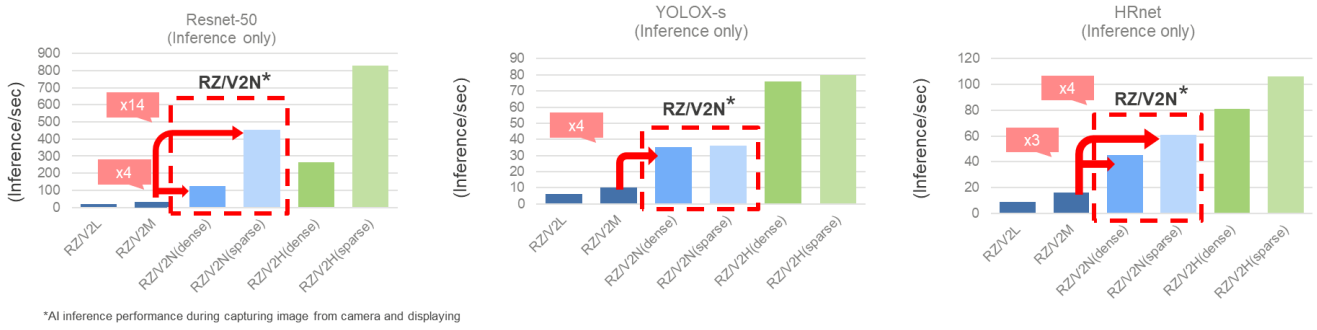

图 11： RZ/V 系列之间的 CNN 推理速度比较

通过 DRP-AI3 提升速度的效果变动因素

通过增强最大性能和剪枝技术提升速度的效果因模型而异，其理由如下。

首先，最大处理性能（峰值性能）通过增加 AI-MAC 内部积和运算器的搭载数量等技术来实现。但是，即使增加积和运算器的数量，由于受到所用 AI 模型的大小与内存带宽限制，有时运算器也无法一直得到充分利用。此外，由于 RZ/V2N 对内存带宽的限制比 RZ/V2H 更加严格，在同时运行 AI 以外的高负载应用程序时，RZ/V2N 的 AI 处理性能波动可能会更大。因此，选择设备不仅要考虑 AI 处理性能，考虑 AI 以外的处理负载也非常重要。

接下来展示的是剪枝效果的模型依赖性。DRP-AI 的剪枝模型高速化技术，通过剪枝减少神经网络的节点数（即运算次数）来削减每张图像的推理时间和功耗。通过剪枝可以减少求和运算的次数，因此相应地

提高了功耗的能效。另一方面，推理时间的效果根据所用 AI 模型不同而变化。其理由是，虽然通过剪枝可以减少积和运算时间与节点信息（重量）的通信量（图 12（A）），从而能够缩短每个图像的推理时间，但是由于各层的运算结果（特征图）即使剪枝也不会被削减（图 12(B)），故而高剪枝会导致特征图的通信量增加，而内存通信量（带宽）限制可能导致速度达到极限。但是，与使用 GPU 时剪枝速度提升 10%左右[2]相比，瑞萨的剪枝性能提高了 20%-300%左右，因此在提升性能方面也是卓有成效的。

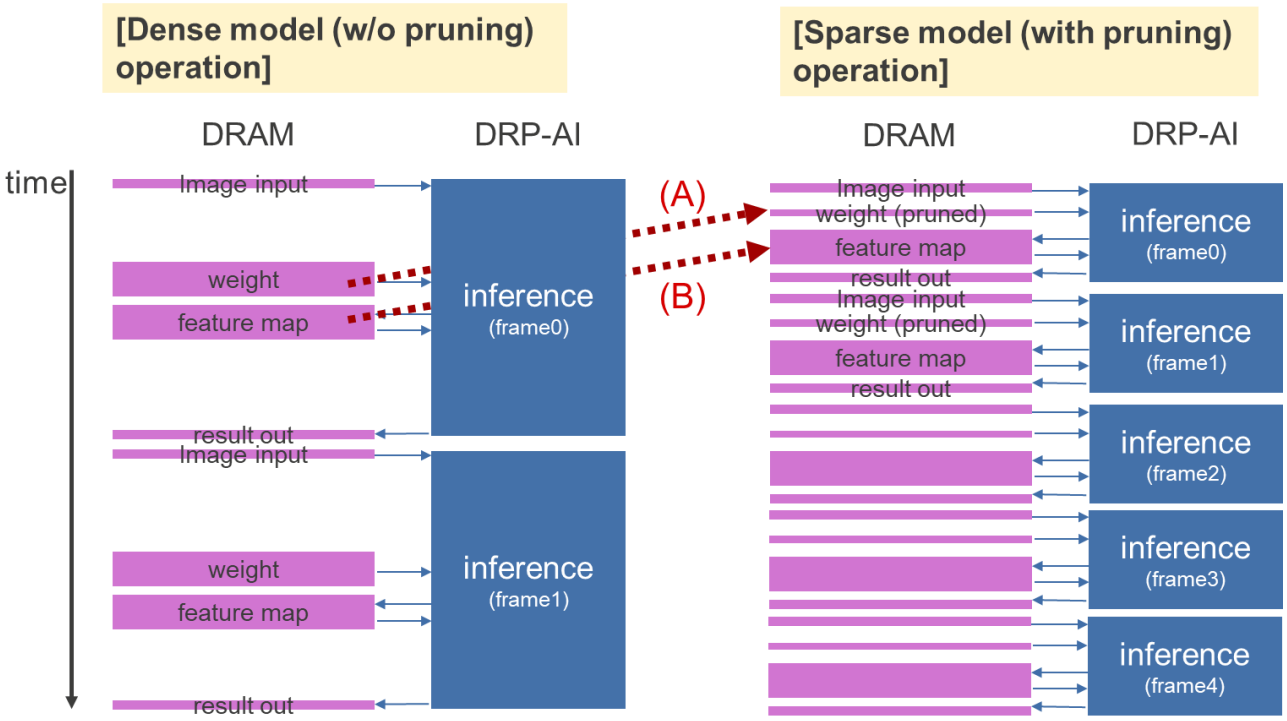

图 12： 预剪枝模型(Dense)和已剪枝模型(Sparse)之间的内存访问比较（图像）

Transformer 模型的端侧部署

近年来，由于 Transformer 模型的高性能和多用途性，其越来越多地部署在端侧设备中。尤其是除了 ChatGPT 等对话模型以外，ViT[3]等 Vision Transformer 模型、SayCan[4]/RT-2[5]等机器人行为生成模型也引起了人们的关注。为了应对这种新趋势，我们为 Transformer 改进了高灵活性的 DRP-AI 工具，确保可在 RZ/V2H 和 RZ/V2N 中使用 Transformer 模型进行推理。

另一方面，Transformer 模型存在诸如频繁使用与 CNN 模型不同的运算类型、所需内存大小和计算量大于 CNN 模型等问题。 除了进一步改进 DRP-AI 工具之外，瑞萨计划继续推动下一代 DRP-AI 的开发，从而提高 Transformer 模型的运算效率和增强剪枝等模型小型化技术。

Transformer 模型的嵌入式配置趋势和挑战

扩大 Transformer 模型应用

由于近年的技术进步，视觉 AI 领域已经从当前的关键因素即 CNN（Convolutional Neural Network）发展为 Transformer 模型。在这个背景后有几个重要因素（图 13）。

首先，Transformer 模型比 CNN 模型具有更高的识别精度。 这是因为 Transformer 使用自注意力机制（Self-Attention）来有效地处理图像中长距离信息的依赖关系。这样，不仅可处理图像识别任务，还可处理时间序列预测和自然语言处理等各种任务。此外，Transformer 模型也适合多模态数据处理。例如，它在组合多种数据格式的任务中表现出色，譬如同时处理图像和文本、或者集成音频和视频等。此外，随着进一步开发 Transformer 模型轻量化技术，低功耗实时处理 Transformer 模型也得以实现。

最近，融合 CNN 和 Self-Attention 的模型（例如 YOLOv10 和 Topformer 等）广受瞩目，因为其比传统的图像 AI（CNN 模型）更能维持精度与性能之间的平衡。这些模型将 CNN 的特征提取能力与 Self-Attention 的长距离依赖关系把握能力相结合，实现了更高的精度与效率。

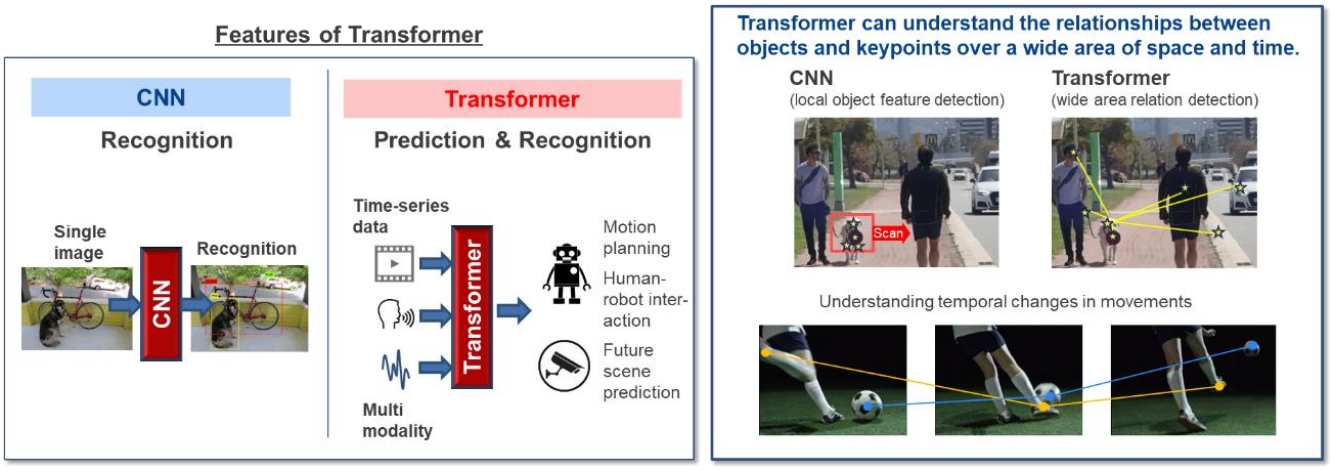

图 13： Transformer 模型的特征

Transformer 端侧配置面临的挑战

然而，Transformer 模型存在频繁使用与 CNN 模型种类不同的运算等问题，并且其模型规模也大于 CNN（图 14）。因此，为了在端侧配置 Transformer 模型，必须解决这些问题。

计算复杂性增大：与 CNN[6]不同，CNN 超过 90%的运算由卷积处理(Convolution)组成，而 Transformer 不使用权重数据，多用特征图之间的积和运算以及 Softmax、GELU、LayerNorm 等 Transformer 特有的复杂运算，结构繁琐[7]。因此，后者存在 AI 编译器不被完全支持且无法配置到设备中的情形、以及由于运算成为性能瓶颈而导致 AI 推论速度低下之类的问题。

模型大小和计算复量增加：Transformer 模型与传统 CNN 模型相比，能够提升识别精度。另一方面，它的模型大小和计算量往往要增加数倍，这是导致其难以配置到内存与计算资源有限的端侧设备上。

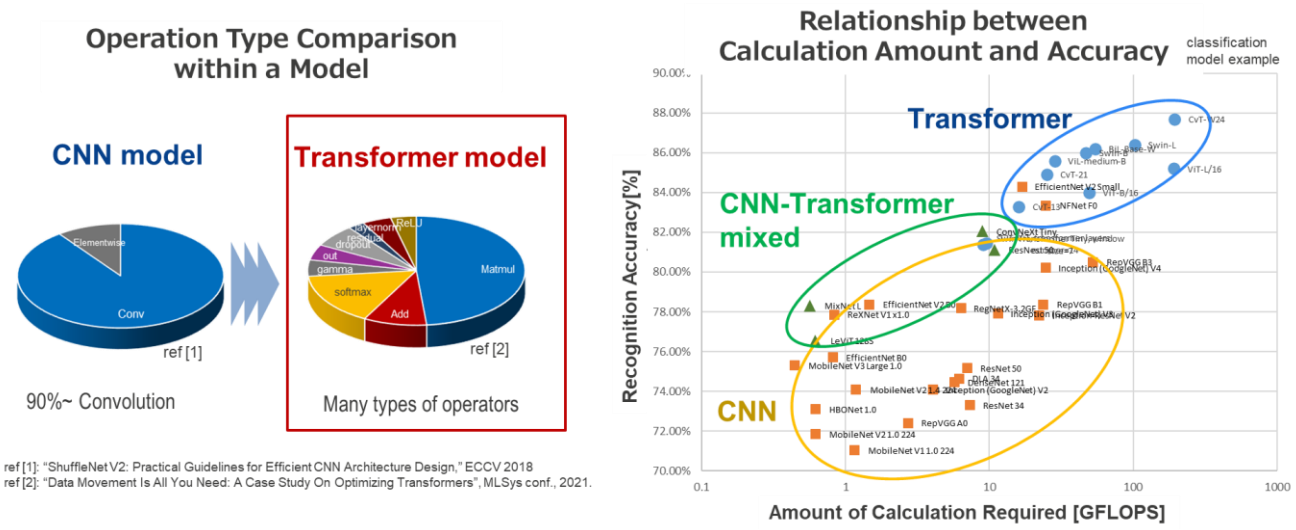

图 14：　Transformer 模型的端侧 AI 配置问题

DRP-AI3 支持的 Transformer 模型

DRP-AI3 的 Transformer 模型优化技术

RZ/V2H 和 RZ/V2N 中配置的 DRP-AI3，采用的是对 CNN 进行优化的架构。瑞萨电子正在推动 AI 编译器和其他环境，以便这些技术也可支持 Transformer 模型。具体来说，我们正在不断推动下列技术的研发。

加速多样化的 Activation：继续开发高速库，利用动态重建处理器 （DRP） 实现 Transformer 模型的多样化大规模运算处理。

优化矩阵运算：在 DRP-AI TVM 中将占 Transformer 大部分运算的矩阵运算转换为 CNN 特定运算（运算符），通过 DRP-AI 实现高速处理。

运用剪枝技术与量化技术：DRP-AI 的量化与剪枝工具，不单可用于 CNN，还可用于 Transformer。因此，它与 CNN 一样可用于剪枝模型高速化。还有，它还支持量化感知训练（QAT）。

支持 AI 编译器（DRP-AI TVM）Transformer

瑞萨电子加强了对 DRP-AI TVM 的 Transformer 的支持力度，从 2024 年 10 月发布的 Version 2.4 开始，部分 Transformer 模型和运算符模型可在 RZ/V2H 和 RZ/ V2N 上运行。截止 2025 年 3 月，除了 ViT 和 Swin[8]等 Vision transformer 之外，它还支持融合 Transformer 模型和 CNN 模型的模型（图 15）。DRP-AI3 原本是针对 CNN 的架构，而 DRP-AI 工具也处于持续进步的阶段，因此存在一些局限性。

・不支持某些运算符，因此某些模型无法转换。

・有时会出现 INT8 量化导致精度大幅下降的情形。

・不支持 DRP-AI 的运算符由 CPU 处理，与 CPU 相比速度提升最多可达 10 倍左右。

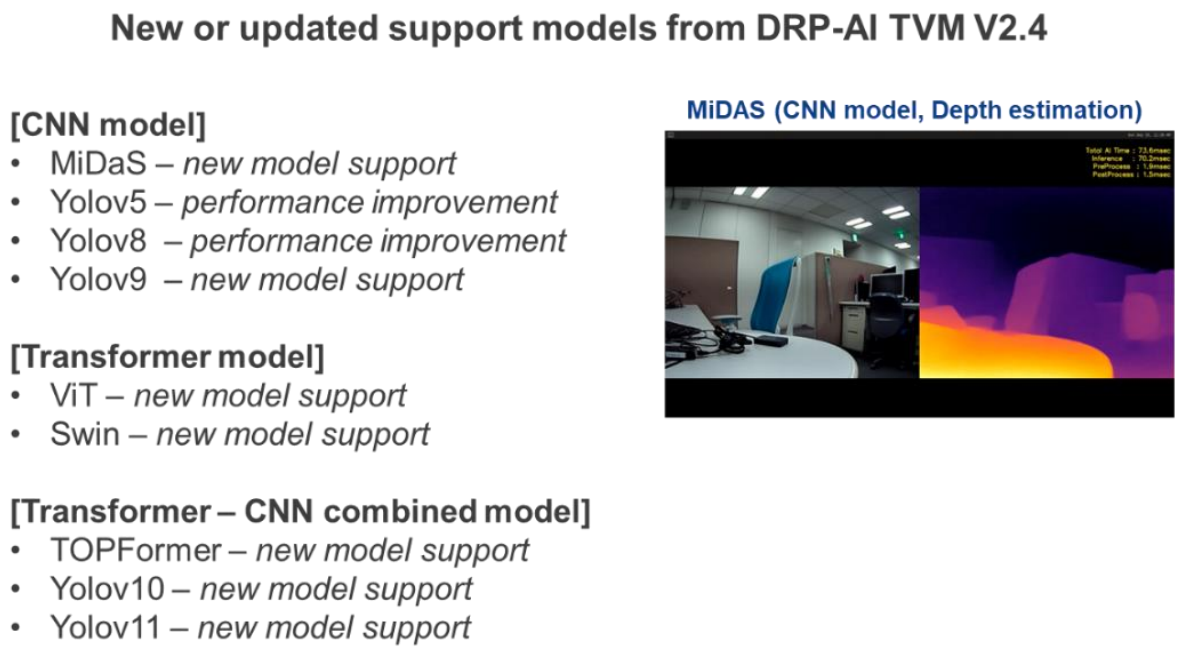

图 15： 2025 年 3 月的 DRP-AI TVM (Version2.4)新增支持和更新的主要 AI 模型

Transformer 配置示例

Topformer 是 Token Pyramid Transformer 的缩写，是一种专门用于移动设备语义分割的架构[9]。在这里，“Token”是指被分成较小部分的图像或数据。Topformer 将各种大小的 Token 作为输入，并根据其大小生成带有意义的特征。这样一来，能够加强 Token 之间的关系，并提升数据的表现力。在该模型之中，与单独使用 CNN 和 Topformer 的模型相比，将 CNN 和 Topformer 的层组合起来使得精度和速度之间达到了更高的平衡。

本次，我们使用 DRP-AI TVM V2.4 编译了 Topformer 模型，并成功将其配置在 RZ/V2H 上。模型部分的处理时间(inference time)小于 100msec，并且可实现高精度和实时化的分割处理。

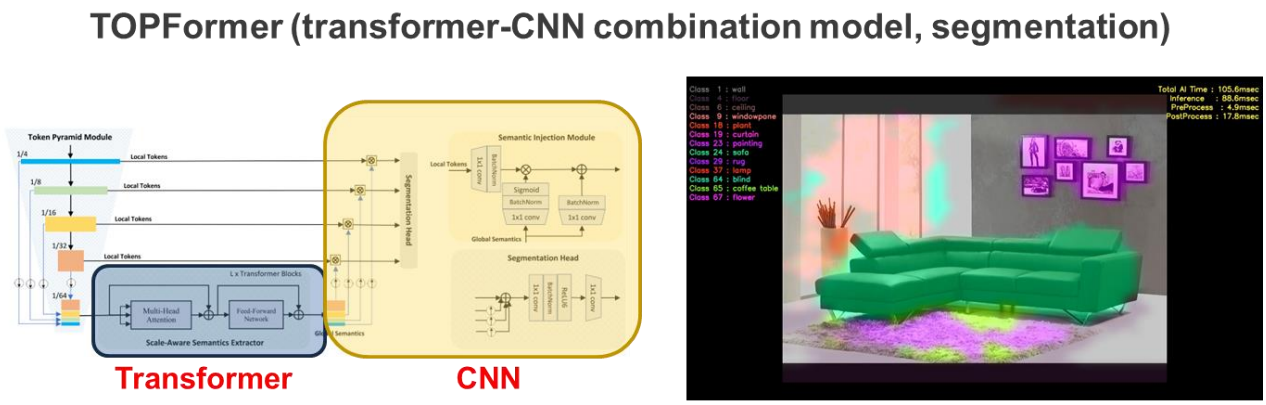

图 16： CNN-Topformer 混合模型示例(Topformer)

今后预定：开发新一代 AI 加速器（DRP-AI4）

新一代 AI 加速器(DRP-AI4)旨在进一步提升当前的 DRP-AI3，以实现更高性能和更高效率的 AI 处理。特别是，我们正在进一步强化对 DRP-AI 架构灵活性与轻量化模型的支持，加速研发 AI 加速器以确保 Topformer 和 CNN 均可高速且低功耗地实施处理。

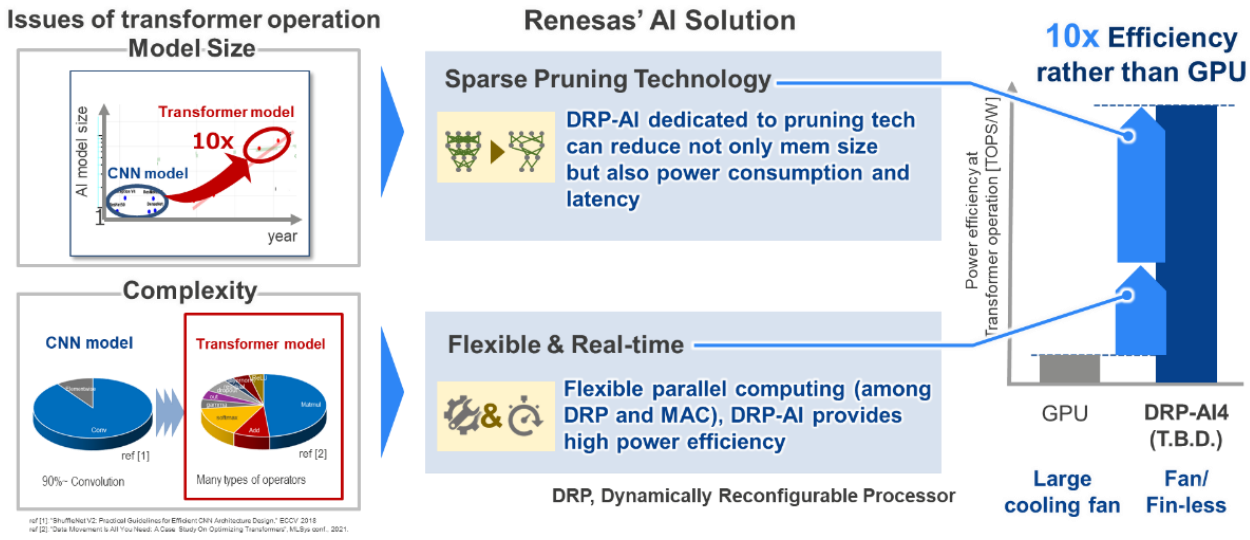

图 17： 面向新一代 DRP-AI (DRP-AI4)的措施

结论

瑞萨电子不断捕捉最新的技术趋势，研发产品并将其推向市场。我们的目标是确保客户可以放心使用我们的产品。今后，我们将继续提供创新的解决方案，助力客户的企业取得成功。

参考

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.

相关信息

RZ/V2N: 四核视觉 AI MPU、15TOPS 、双摄像头、高能效

RZ/V2H: 四核视觉 AI MPU，采用 DRP-AI3 加速器和高性能实时处理器

Embedded AI-Accelerator DRP-AI 白皮书

Next Generation Highly Power-Efficient AI Accelerator (DRP-AI3) 白皮书

重要通知和免责声明

瑞萨电子株式会社及其关联公司（以下简称“瑞萨”）的技术规范和可靠性数据（包括数据手册）、设计资源（包括参考设计）、应用或其他设计建议、Web 工具、安全信息以及其他资源“按原样”提供，不保证无瑕疵。瑞萨不做任何明示或暗示保证，包括但不限于产品适销性、特定用途适合性或不侵犯第三方知识产权的保证。

这些资源的适用对象为使用瑞萨产品熟练进行设计的开发人员。以下事宜请自行负责：(1)为您的应用选择合适的产品，(2)设计、验证和测试您的应用，(3)确保您的应用符合适用标准以及安全性等所有其他要求。这些资源如有更改，恕不另行通知。瑞萨仅授权您将这些资源用于开发采用瑞萨产品的应用。严禁复制这些资源或用于其他用途。我们未授予任何其他瑞萨知识产权或任何第三方知识产权的许可。

瑞萨对因使用这些资源而产生的任何索赔、损害、成本、损失或负债概不负责，且瑞萨及其代表的全部损失须由您赔偿。瑞萨的产品仅遵守瑞萨的销售通用条款和条件，或书面签订的其他适用条款。使用瑞萨的任何资源不会扩大或更改这些产品的任何适用保修或保修免责声明。

(Rev.1.0 Mar 2020)

公司总部

135-0061, 日本东京江东区

豊洲 3-2-24, TOYOSU FORESIA

https://www.renesas.com

联系信息

有关产品、技术的更多信息，文档的最新版本，或

离您最近的销售办公室，请访问：

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

商标

瑞萨电子的名称和徽标是瑞萨电子公司的商标。 所有商

标和注册商标均为其各自合法所有者的财产。

Doc Number: R01WP0027CC0100