DRP-AI Translator i8 V1.01

Release Note

Introduction

This release note describes the improvements of the DRP-AI Translator.

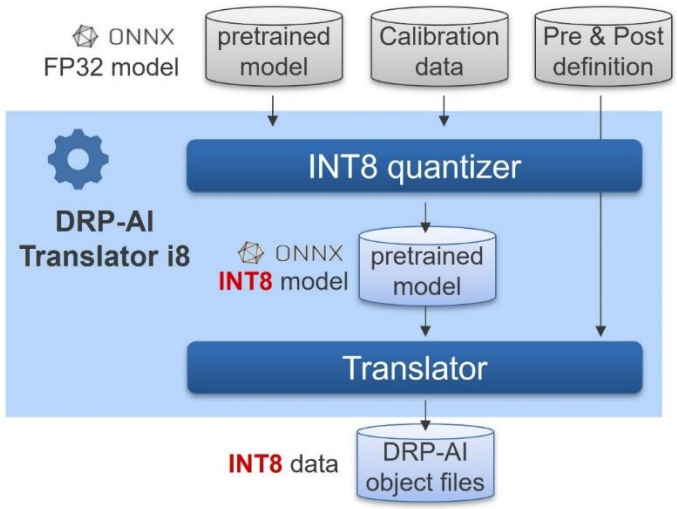
Key Features and Enhancements

- Supports Quantization function
- Supports sparse model translation
- Compatibility with conventional “DRP-AI Translator”
  - Input AI model is float ONNX format
  - Pre/Post Processing is supported and accelerated by DRP

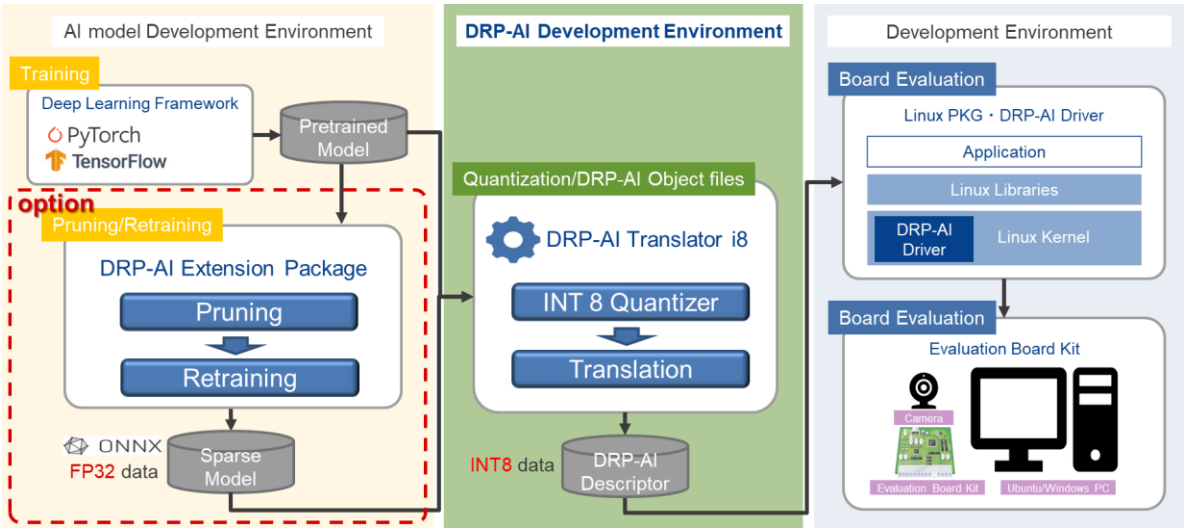
Contents

1. Overview

DRP-AI Translator i8 is a tool for translating the deep learning trained models(onnx) into executable files with DRP-AI. DRP-AI Translator i8 consists of 2 tools: INT8 Quantizer & Translator.
INT8 Quantizer is a tool to generate INT8 ONNX models. For conversion, calibration data and a pretrained float onnx model are required.
Translator is a tool that generates executable files from QDQ onnx model. For translation, rINT8 QDQ onnx and Pre&Post definition files are required.



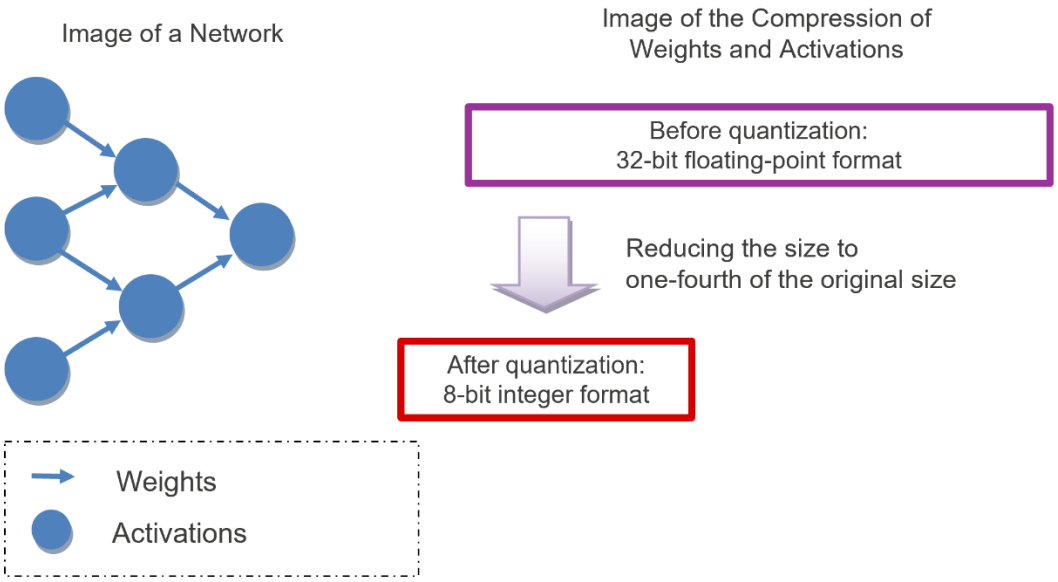
Below is a RZ/V2H DRP-AI implementation flow.



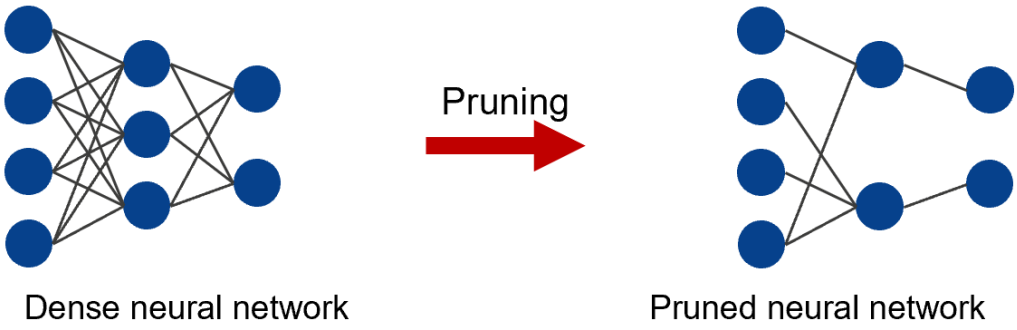
If you want to generate Sparse Model, please download "DRP-AI Extension pack" from Renesas web site.
“AI model development Environment” and “Development Environment” are out of scope of DRP-AI Translator i8.

2. What is Quantization and Pruning?

Quantization is the process of reducing the sizes of models by representing parameters of networks such as weights with a lower bit width.



Nodes are interconnected in a neural network as shown in the figure below.



Methods of reducing the number of parameters by removing weights between nodes or removing nodes are referred to as “**pruning**”. A neural network to which pruning has not been applied is generally referred to as a dense neural network. Applying pruning to a neural network lead to a slight deterioration in the accuracy of the model but can reduce the power required by hardware and accelerate the inference process.

If you want to generate Sparse model, please download "DRP-AI Extension pack" from Renesas web site. Pruning procedure is out of scope of DRP-AI Translator i8.

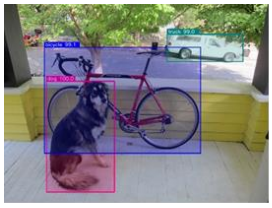
3. Getting Started Guide

After installing DRP-AI Translator i8, sample pruned onnx models and the **Getting Started** guide are extracted along with the INT8 Quantizer & Translator. **Getting Started** helps you learn how to use DRP-AI Translator i8. If you use Translator i8 for the first time, please refer to *Getting_Started/README.md*. Below is a directory structure.

```
DRP-AI_Translator_i8(install directory)
├── Getting_Started  … Guide for DRP-AI Translator i8
|       ├── README.md … Overview of Getting Started
├── onnx_models      … Sample pruned onnx models
├── drpAI_Quantizer  … Root directory of INT8 Quantizer
└── translator       … Root directory of Translator
```

The Getting Started guide describes how to translate the following AI models.

Category	AI model
Object Detection	Lightnet YOLOv2
	Megvii-BaseDetection YOLOX
Semantic Segmentation	torchvision DeepLabv3
Classification	torchvision ResNet50
Human Pose Estimation	MMPose HRNet



Object Detection

Pose Estimation

Sematic Segmentation