DRP-AI Translator i8 V1.02

Release Note

Introduction

This release note describes the improvements of the DRP-AI Translator i8.

Key Features and Enhancements

- Simulation api for Pre & Post processing by DRP
- Update Execution time summary sheet
- Sparse effect estimation option

Contents

1. Improvement

1.1 Simulation api for Pre & Post processing by DRP

Support python api which run the simulation of Pre & Post processing by DRP. The guide is described at appendix of User's Manual. See Appendix B in User's Manual.

1.2 Update Execution time summary sheet

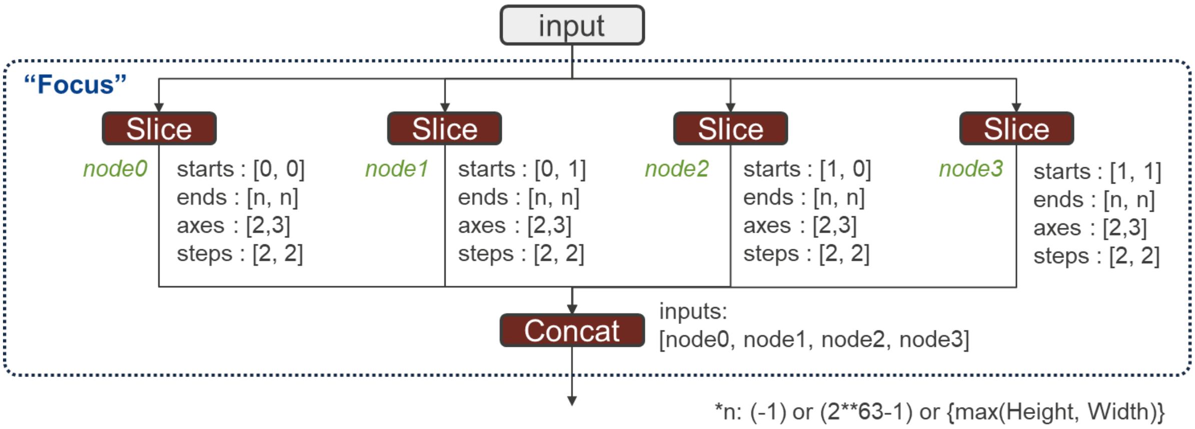
Update the excel sheet format. Information about data type and module (AI-MAC or DRP) are added to the summary sheet.

1.3 Sparse effect estimation option

This option is intended for users who want to know the inference time improvement when using sparse mode with different sparse ratio without retraining.
[Note] Please note that this option should only be used to estimate the sparse model inference time. The object files generated by this option CAN NOT inference correctly.

1.4 Support additional "Focus" graph structure

Following graph (called "Focus" layer) is newly supported.



1.5 Overflow detection

If there is a node which is overflow from a FP16 range in Quantized onnx, a warning message will be shown.

1.6 Optimized inference speed

To improve inference speed, a part of graph/operator is optimized in translation process.

2. Fixed Issues

2.1 Operator : *Convolution*, *Add, Concat* and *Resize*(not pre-processing)

Fixed the issues where DRP-AI object file was not generated correctly when certain combinations of heigh/width/input channel/output channel are used.

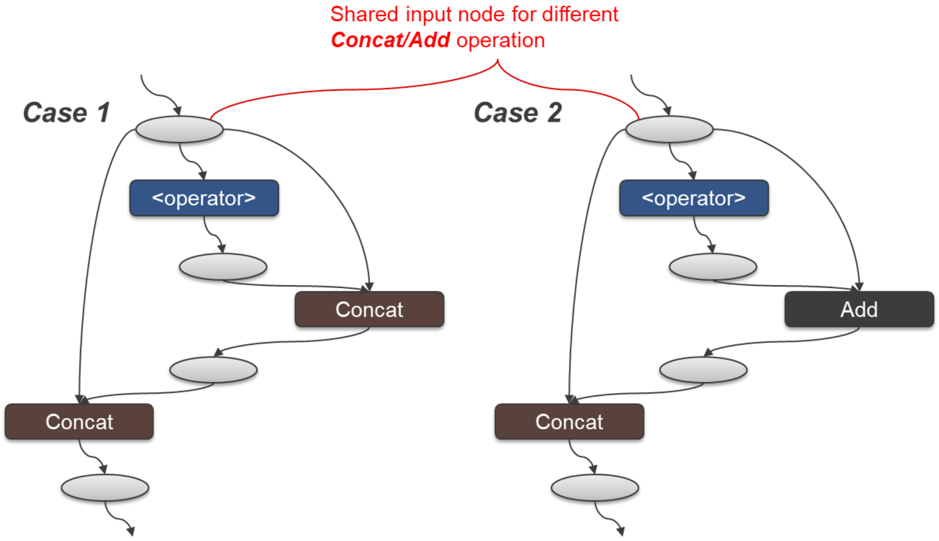
2.2 Graph: *Add > BatchNorm > ReLu*

Fixed the issue to support ***Add*** > ***BatchNorm*** > ***ReLu*** graph, correctly.

3. Known Issues

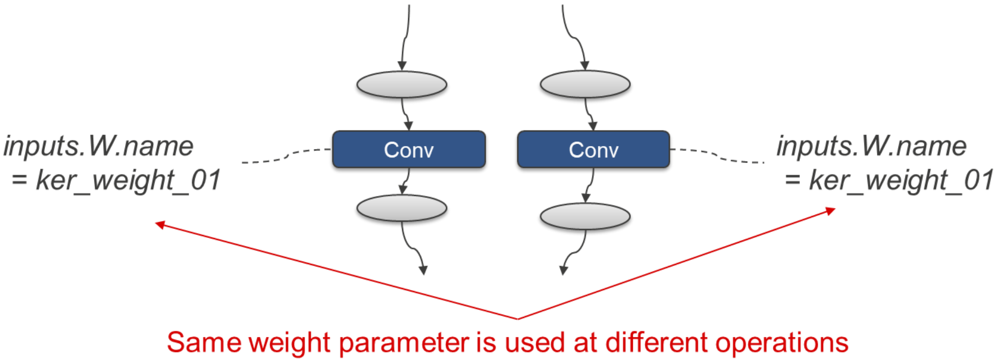
3.1 *Shared input node to different Concat/Add operatin*

There may be errors in the inference results after conversion. An example of the structure is as follows.



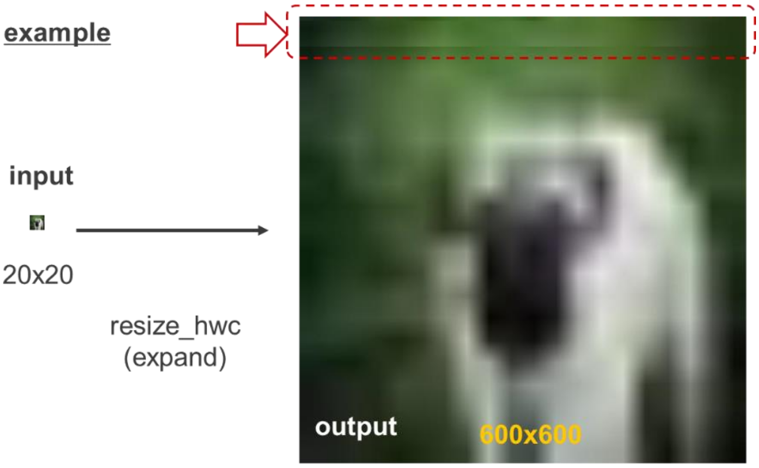
3.2 A weight parameter is shared with multiple Convolution

Translator does not support below graph structure which has Convolutions sharing a same weight parameters.



3.3 resize_hwc pre-processing by DRP

In the case of expanding an image with DRP preprocessing, there may be an error at the top edge of the image.



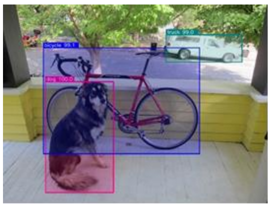
4. Getting Started Guide

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