

DRP-AI Translator i8 V1.03

Release Note

Introduction

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

Key Features and Enhancements

- Add new translation guide about depth estimation model
- Improve inference time of resize operator
- Support Split operator

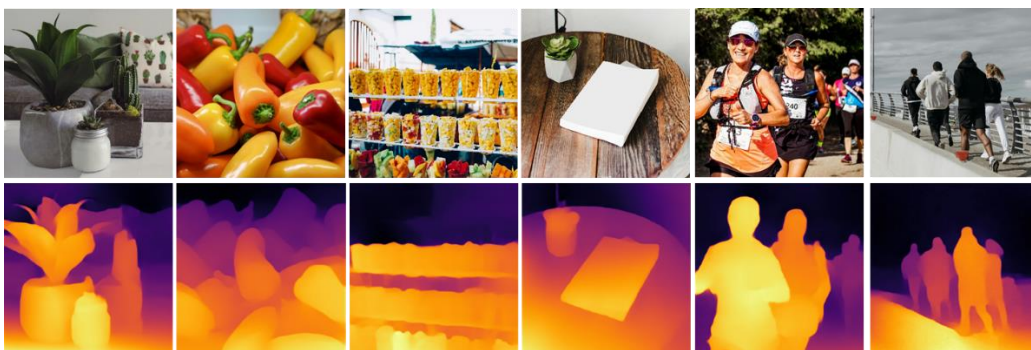
Contents

1.	Improvement.....	2
1.1	Add new translation guide about depth estimation model.....	2
1.2	Optimization: <i>resize</i> operator	2
1.3	Operator / Attribute updates	2
2.	Fixed Issues	4
2.1	Operator : <i>Convolution, Dilated Depthwise Convolution, Sparsed Convolution</i>	4
2.2	Graph: Shared input node for different <i>Concat/Add</i> operation.....	4
3.	Known Issues	5
3.1	A weight parameter is shared with multiple <i>Convolution</i>	5
3.2	<i>Error pattern</i> in specific combinations of heigh/width/input channel/output channel	5
4.	Getting Started Guide	6

1. Improvement

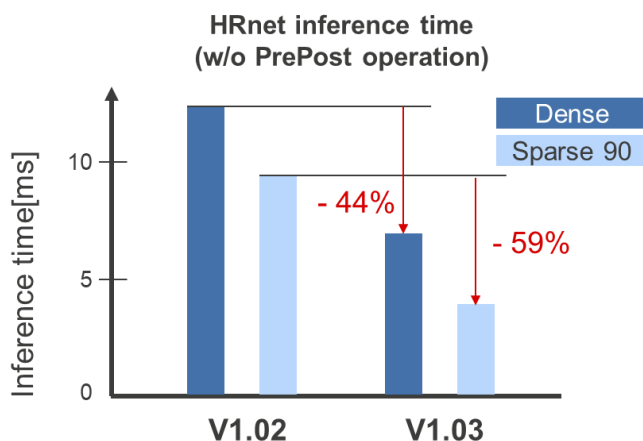
1.1 Add new translation guide about depth estimation model

Please see **GettingStarted/how-to/depth_estimation** folder in DRP-AI Translator i8 for the detail.
Following are sample images.



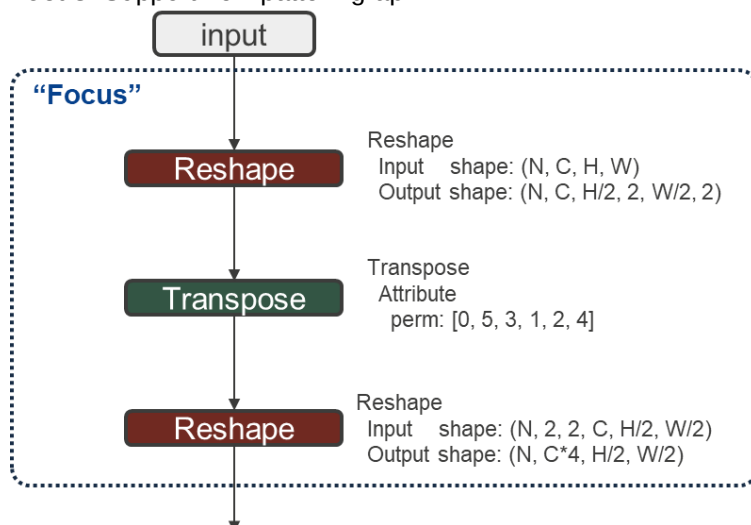
1.2 Optimization: *resize* operator

Improved inference time of **resize** operator. Especially **HRNet(Pose Estimation)** inference time has been drastically improved.



1.3 Operator / Attribute updates

- **Convolution**: Newly support ker16x16, stride 16, pad [0,0,0,0] attribute
- **Focus**: Support new pattern graph



- ***Slice***: Support new use case: “*Split*” operator equivalent processing
- ***Split***: Newly support *Split* operator
- ***Softmax***(PostProcessing): Support 3D shape input(HWC format)

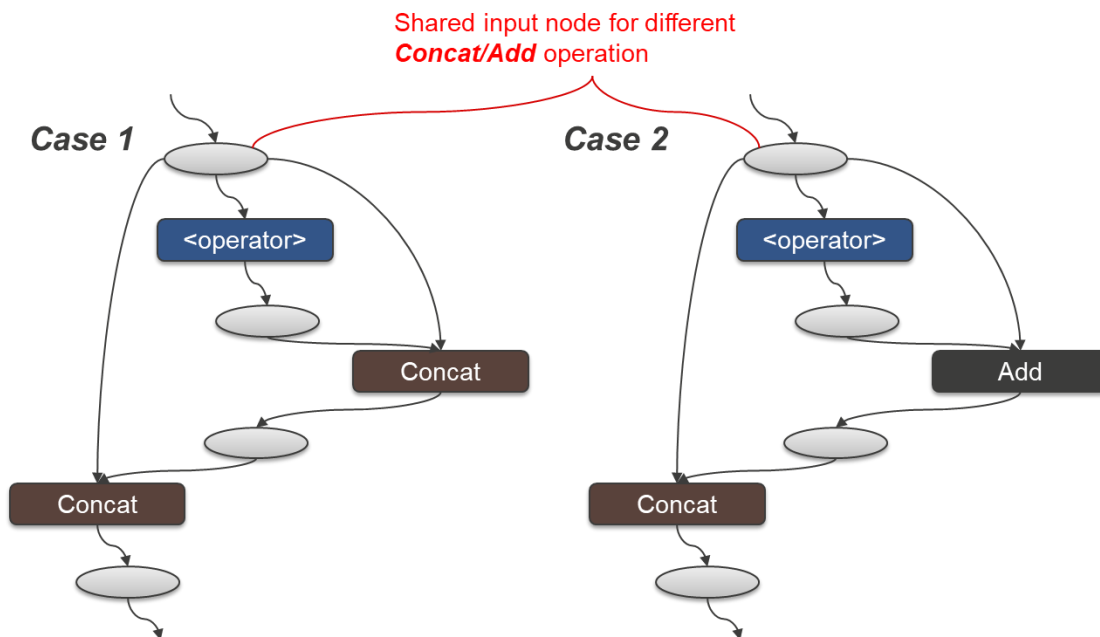
2. Fixed Issues

2.1 Operator : *Convolution, Dilated Depthwise Convolution, Sparsed Convolution*

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

2.2 Graph: Shared input node for different *Concat/Add* operation

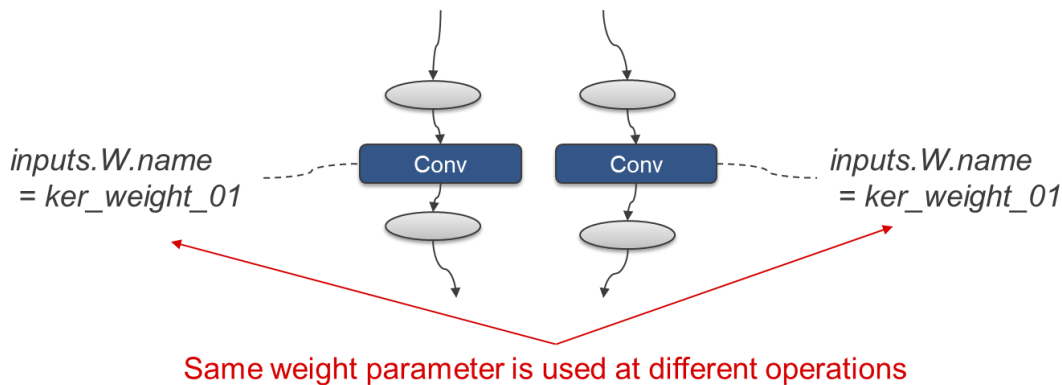
Fixed the issue that there may be errors in the inference results after *conversion*.



3. Known Issues

3.1 A weight parameter is shared with multiple *Convolution*

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



3.2 Error pattern in specific combinations of heigh/width/input channel/output channel

If the following conditions 1, 2, and 3 are true, there may be an error in the inference results.

1. operator: *Convolution* or *MaxPool* or *AveragePool*
2. $(Ker \% 2 == 0)$ or $(Ker \% 2 != 0 \ \& \ pad != (ker - 1) / 2)$
3. Feature map size is large
e.g. $ih = iw = 80, ich = 512, och = 512$

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

```

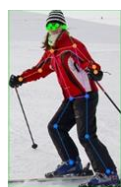
The Getting Started guide describes how to translate the following AI models.

Category	AI model
Object Detection	Lightnet YOLOv2
	Megvii-BaseDetection YOLOX
Semantic Segmentation	torchvision DeepLabv3
Classification	torchvision ResNet50
Human Pose Estimation	MMPose HRNet
Depth Estimation	PyTorch Hub MiDas (*1)

*1: Sample pruning model is not included in DRP-AI Translator i8. Please follow the guide to download the model.



Object Detection



Pose Estimation



Semantic Segmentation

