DRP-AI Translator i8 V1.04

Release Note

Introduction

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

Key Features and Enhancements

- Newly support RZ/V2N device
- Support ConvTranspose operator
- Support “Group Convolution”

Contents

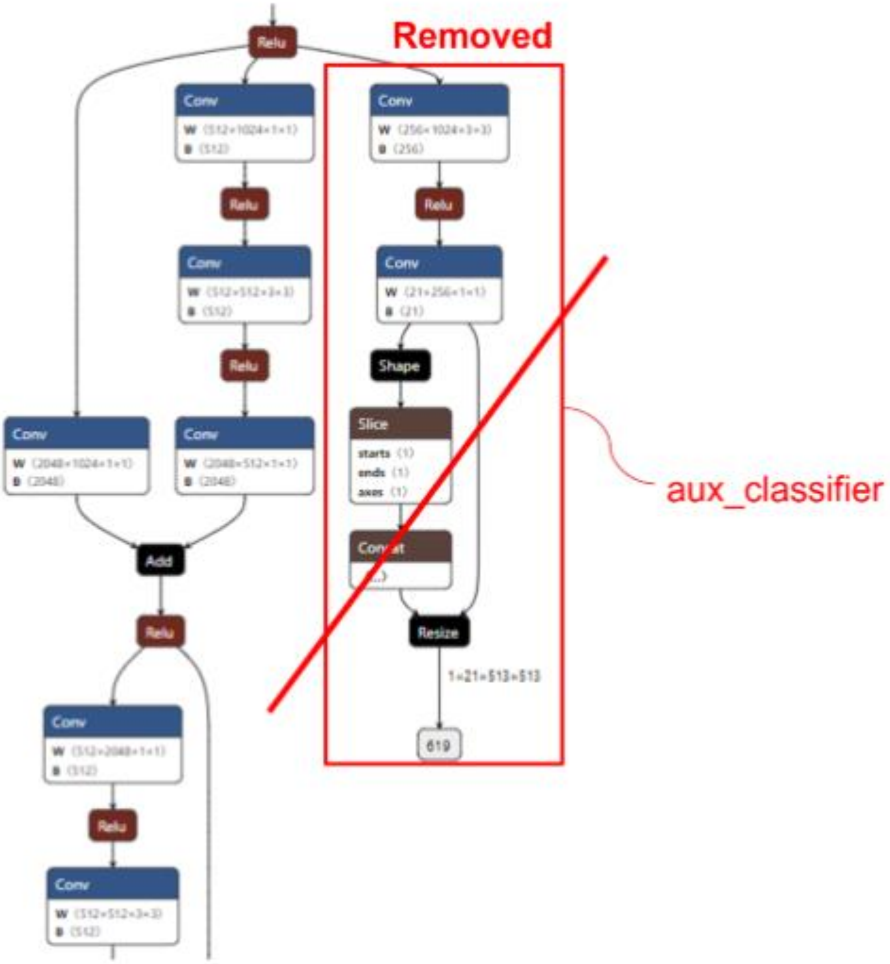
1. Improvement

1.1 RZ/V2N is newly supported

DRP-AI Translator i8 supports RZ/V2N. The usage of DRP-AI Translator i8 is compatible with V2H. A new “Translator_v2n.sh” has been added as a translation script for V2N. By using this script, the inference estimation time for V2N is generated.

1.2 Optimization of the included DLV3 model

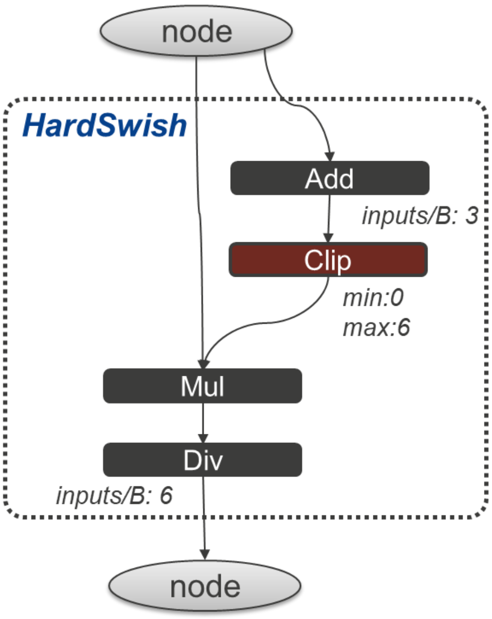
In the DLV3 onnx model included in DRP-AI Translator i8, graph that is not necessary for inference was cut. The inference time was also improved.



1.3 Operator / Attribute updates

- ***Convolution***:
  - Newly support ker6x6, stride 2, pad [l,r,t,b], pad size <=3
  - Newly support “GroupConvolution”. See User’s Manual for the details
- ***ConvTranspose***:
  - Newly support “ConvTranspose”.
  - Supported kernels are 2x2, 3x3 and 4x4 with stride 2
  - See User’s Manual for the details.
- ***InstanceNormalization***:

- ***HardSwish***:
  - Support another graph structure.



- ***Sigmoid(PostProcessing)***:
  - Sigmoid operation can be defined at postprocessing.

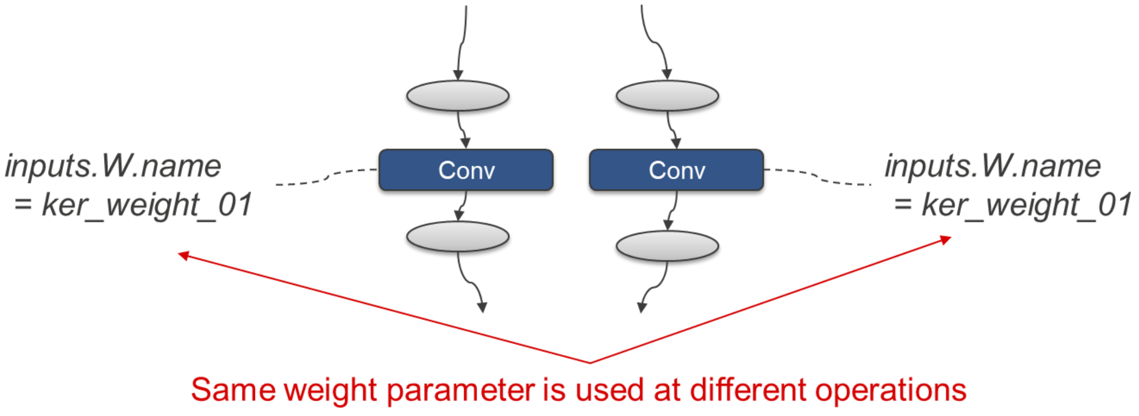
2. Fixed Issues

2.1 Operator : *Convolution, Dilated Convolution, MatMul, MaxPool and Concat*

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 : *Weight parameter shared with multiple Convolution*

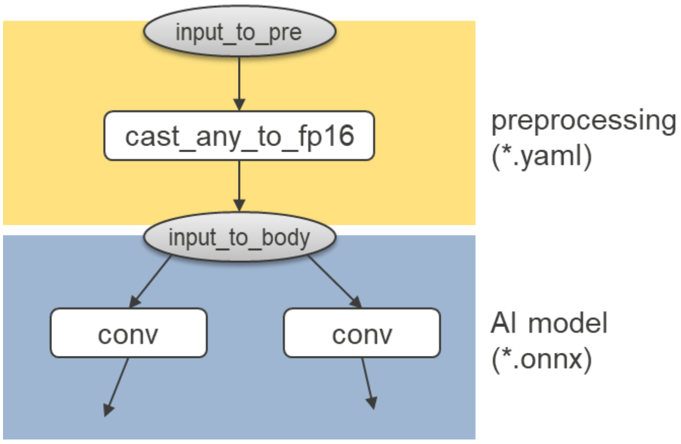
Fixed the issue where the subgraph below may cause translation error.



2.3 Preprocessing : *cast_any_to_fp16*

Fixed the issue related to cast_any_to_fp16 operation in preprocessing. If the following two conditions were satisfied, a conversion error occurred.

1. There is no normalize operation after cast_any_to_fp16
2. cast_any_to_fp16 output is branched in onnx model



3. Known Issues

3.1 Error pattern condition

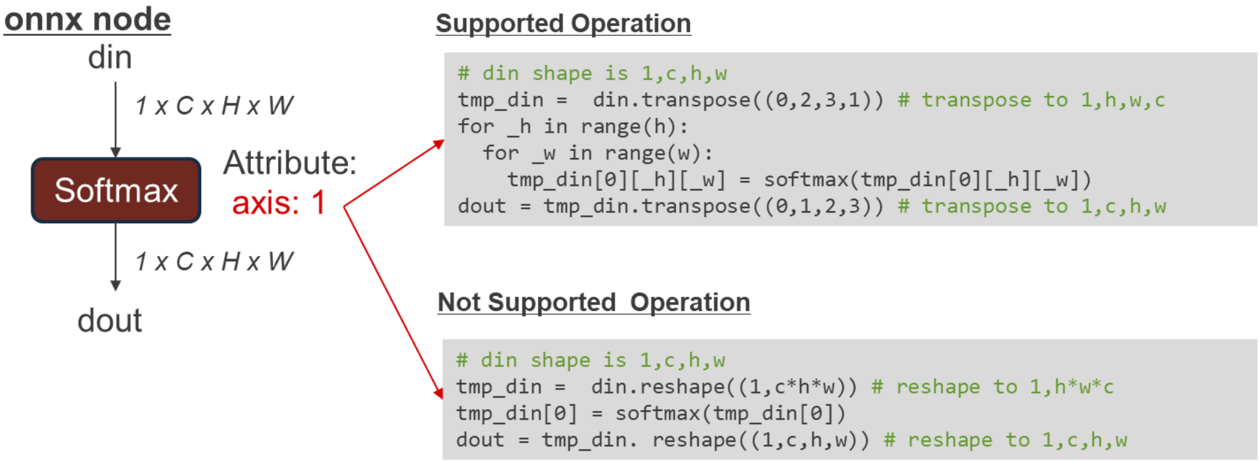
If the following conditions 1, 2, and 3 are true, there may be an error in the inference results.

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

3.2 softmax attributes

Due to a change in the interpretation of axis attribute, old opset softmax operation is not supported. There are error in output value.



Supported Operation

```
# din shape is 1,c,h,w
tmp_din =  din.transpose((0,2,3,1)) # transpose to 1,h,w,c
for _h in range(h):
  for _w in range(w):
    tmp_din[0][_h][_w] = softmax(tmp_din[0][_h][_w])
dout = tmp_din.transpose((0,1,2,3)) # transpose to 1,c,h,w
```

Not Supported Operation

```
# din shape is 1,c,h,w
tmp_din =  din.reshape((1,c*h*w)) # reshape to 1,h*w*c
tmp_din[0] = softmax(tmp_din[0])
dout = tmp_din. reshape((1,c,h,w)) # reshape to 1,c,h,w
```

4. Getting Started Guide

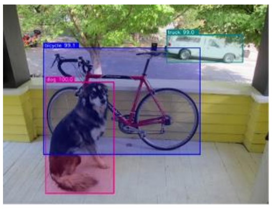
After installing DRP-AI Translator i8, sample pruned onnx models and the **Getting Started** guide are extracted along with the INT8 Quantizer & Translator. **Getting Started** helps you learn how to use DRP-AI Translator i8. If you use Translator i8 for the first time, please refer to *Getting_Started/README.md*.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



Object Detection

Pose Estimation

Semantic Segmentation