DRP-AI Translator i8 V1.11

Release Note

Introduction

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

Key Features and Enhancements

- Extension of supported attribute conditions for operators such as Convolution, MaxPool, Add, Mul, and Pad
- Optimization for activation function
- Enhancement of error analysis function for Pre & Post-definition YAML file

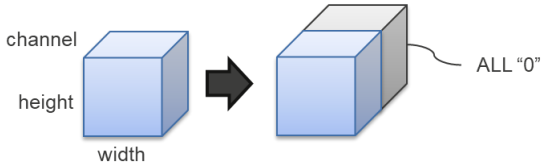
Contents

1. Improvement

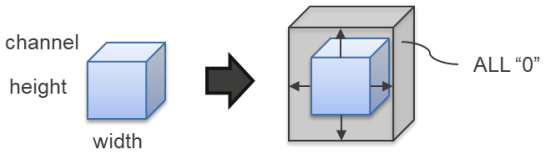
1.1 Operator / Attribute updates

- ***Convolution***
  - Newly support ker2x2, stride 1 or 2, pad [l,r,t,b] (pad size <= 1), group = 1, dilation = 1
  - Newly support ker3x3, stride 2, pad [l,r,t,b] (pad size <= 2), group = 1, dilation = 1
  - Newly support ker3x3, stride 2, pad [l,r,t,b]] (pad size <= 2), group = n, dilation = 1, * n = ich = och
- ***MaxPool***
  - Newly support ker1x1, stride 1 or 2, pad [l,r,t,b] (pad size <= 0)
- ***Add***
  - Expand input shape condition.
    - Input A: feature map with (1,ch,H,W)
    - Input B : feature map with [(1,ch,H,W) or (1,ch,1,1)] or parameter with [(1,ch,H,W) , (1,ch,1,1) or (1)]
    - A and B are commutative.
- ***Mul***
  - Expand input shape condition.
    - Input A: feature map with (1,ch,H,W)
    - Input B : feature map with [(1,ch,H,W) or (1,ch,1,1)] or parameter with [(1,ch,H,W) , (1,ch,1,1) or (1)]
    - A and B are commutative.
- ***HardSigmoid***
  - Newly support HardSigmoid operator
- ***Pad***
  - Case 1: Channel direction padding
    - mode: "constant“, constant_value: 0.0
    - Inputs: data: (1, ch, H, W), pads: [0, ch_begin, 0, 0, 0, ch_end, 0, 0]
    - Outputs: output: (1, ch+ch_begin+ch_end, H, W)
  - Case 2: Height & Width direction padding
    - mode: "constant“, constant_value: 0.0
    - Inputs: data: (1, ch, H, W), pads[0, 0, h_begin, w_begin, 0, 0, h_end, w_end]
    - Outputs: output: (1, ch, H+h_begin+h_end, W+w_begin+w_end)

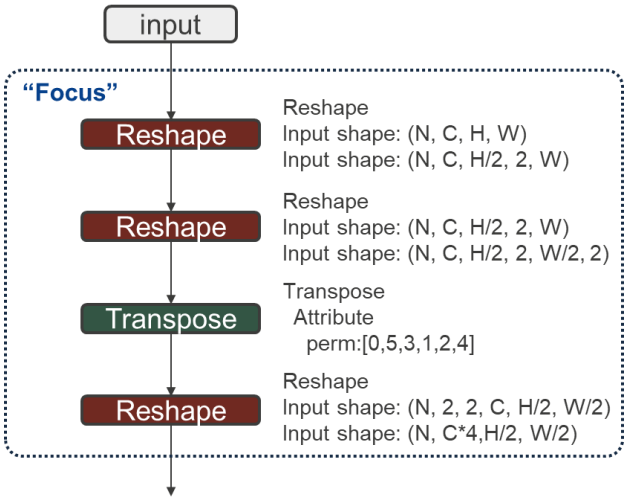
Case1: Channel direction



Case2:Height & Width direction



- ***Focus***
  - Support new pattern graph

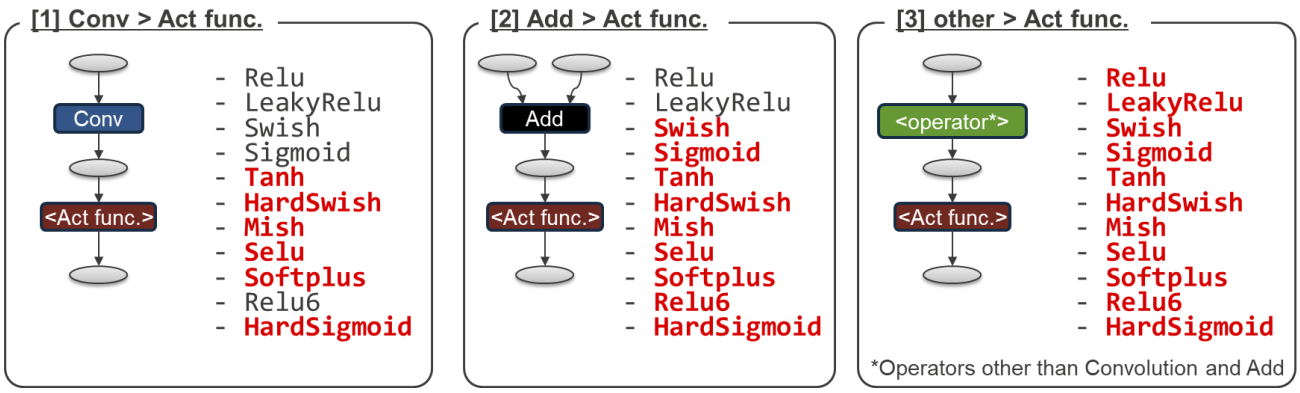


- ***Focus***
  - Support new pattern graph

1.2 Optimization for activation function(Act func.)

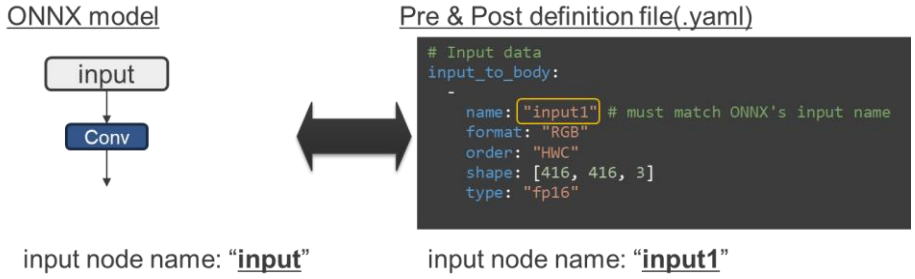
The processing of activation functions has been optimized, resulting in faster inference times. The optimized activation functions for the graph structure shown below are indicated respectively.

Highlighted in red red: Improvements in Translator i8 V1.11



1.3 Enhancement of error analysis function for Pre & Post-definition YAML file

The functionality for analyzing syntax checks of Pre & Post-definition YAML files has been enhanced. During translation, the analysis results are displayed as error messages or warning messages. Please check the message and, if necessary, consider modifying the Pre & Post-definition YAML file.



Error message example

[ERROR] ONNX model does not contain input_to_body 'input1'

2. Fixed Issues

2.1 Operator : Gemm, Add and Softmax

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 Sparse mode error occurred in models with shared bias parameters

The problem occurred when converting a model in sparse mode where bias parameters were shared across multiple Convolution operations.

3. Known Issues

3.1 Error pattern in specific combinations of height/width/input channel/output channel

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

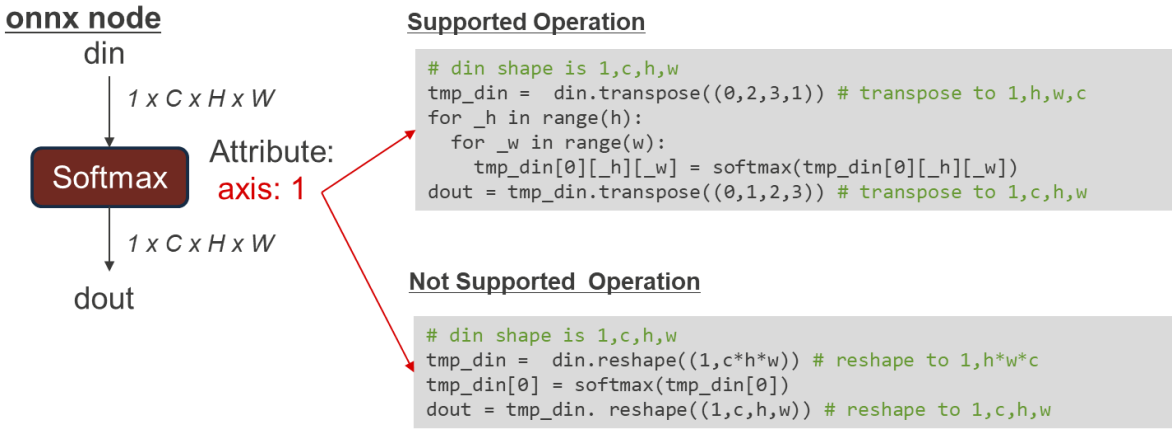
1. operator: *Convolution* or *MaxPool* or *AveragePool*

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

3.2 Softmax attribute

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



4. Getting Started Guide

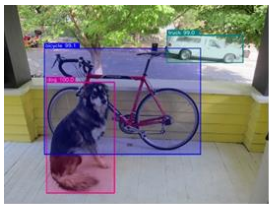
After installing DRP-AI Translator i8, sample pruned/dense onnx models and the **Getting Started** guide are extracted along with the INT8 Quantizer & Translator. **Getting Started** helps you learn how to use DRP-AI Translator i8. If you use Translator i8 for the first time, please refer to Getting_Started/README.md. 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 (Single)
	MMPose YoloX-Pose (Multi)
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

Sematic Segmentation