Huawei HiAI DDK Operator Specifications

1. Parameter Description

Parameter	Description
ni	Indicates the size of a batch.
ci	feature dimension, e.g. for RGB images ci=3 Indicates the number of channels.
hi/ho/	Indicates the height.
wi/wo	Indicates the width.
pad_h(pad_y)/ pad_w(pad_x)	Indicates the number of pixels to be padded. By default, the padding size is less than the kernel size.
stride_h(stride_y, shift_h)/ stride_w(stride_x, shift_w)	specifies the intervals at which to apply the filters to the input Indicates the step.
kernel_h(kernel_y)/ kernel_w(kernel_x)	specifies height and width of each filter Indicates the size of a convolution core.
window_h(window_y)/ window_w(window_x)	Indicates a window size.
zoom/shrink	Indicates the zoom-in or zoom-out multiple.

2. Operator Boundary

2.1 Caffe Operator Boundary

No.	Operator	Boundary
1	AvgPool/MaxPool	stride_h < = window_h; stride_w < = window_w Note: better to make window_h < =hi,window_w < =wi xo x window_h+xo < =2000 wi+xo < =2000 xo=(wi+pad_w-window_w)/stride_w+1; **pad != 0**: pad_w < kw pad_h < kh (pad_w+wi) x ci < 8000

2	interp	wi < =10000 ho < =10000 wo < =2500 hi < =12000 ci>=256: 4 x wo+wi+ho < 1000 Ci < 256: (4wo+wi+ho) x (ci+15) < =256000
3	ABSVAL	The result of ni x ci x hi x wi is affected by the memory size. The precision is limited. The value range is [-10, 10].
4	Convolution	kernel_h < =32 kernel_w < =32 ni < =512 ci < =3200 wi < =5400000 hi < =5400000 pad = 0: The result of ni x hi x wi x (ci+co) is affected by the memory size. pad != 0 : pad_w < kw pad_h < kh (pad_w+wi) x ci < 8000 The result of 2ni x (hi+pad_h) x (wi+pad_w) x (ci+co) is affected by the memory size. When groups exist: group>=2 (ci/group)%16=0 (co/group)%16=0 The result of ni x (hi+pad_h) x (wi+pad_w) x (ci+co) is affected by the memory size.
5	RELU1	The result of ni x ci x hi x wi is restricted by the memory size.

		The precision is limited. The value range is [-10, 10].
6	scale	The result of ni x ci x hi x wi is restricted by the memory size.
7	greater	The result of ni x ci x hi x wi is restricted by the memory size.
8	lessEqual	The result of ni x ci x hi x wi is restricted by the memory size.
9	RELU6	The result of ni x ci x hi x wi is restricted by the memory size. The precision is limited. The value range is [-10, 10].
10	exp	The result of ni x ci x hi x wi is restricted by the memory size.
11	l2_pooling	pool1_stride_h < =pool1_window_h < =hi pool1_stride_w < =pool1_window_w < =wi xo x window_h+xo < =2000 wi+xo < 2000 xo=(wi+pad_w-window_w)/stride_w+1; pad != 0: pad_w < kw pad_h < kh (pad_w+wi) x ci < 8000
12	concat_by_feature	ni < 128 The input c value for image stitching must be the same. The value is a multiple of 16. A maximum of eight tensors can participate in image stitching at a time. ni x hi x wi < 4096 x 4096
13	deconvolution	max(stride_h,stride_w) x no < =256 ci < 256 co < 256 Group parameters are not supported. pad != 0: pad_w < kw pad_h < kh (pad_w+wi) x ci < 8000
14	SIGMOID	The result of ni x ci x hi x wi is

		restricted by the memory size. The precision is limited. The value range is [-10, 10].
15	add	The result of ni x ci x hi x wi is restricted by the memory size.
16	mult	The result of ni x ci x hi x wi is restricted by the memory size.
17	RELU	The result of ni x ci x hi x wi is restricted by the memory size. The precision is limited. The value range is [-10, 10].
18	Sub	The result of ni x ci x hi x wi is restricted by the memory size.
19	floor	The result of ni x ci x hi x wi is restricted by the memory size.
20	select	The result of ni x ci x hi x wi is restricted by the memory size.
21	svdf	num_units < 256
22	cast	The result of ni x ci x hi x wi is restricted by the memory size. The following data types can be converted: CAST_FLOAT32_TO_UINT8 CAST_UINT8_TO_FLOAT32 CAST_INT8_TO_FLOAT16 CAST_FIX8_TO_FLOAT16 CAST_FLOAT16_TO_FIX8 CAST_FLOAT16_TO_FLOAT32 CAST_FLOAT32_TO_FLOAT16 CAST_INT16_TO_FLOAT16 CAST_FLOAT16_TO_INT16
23	conv_depthwise	kernel_h < =32 kernel_w < =32 ni < =512 wi < =5350000 hi < =5400000 ci x m < =640 ni x ci x hi x wi>=8 ci x m x kh x kw>=8 **pad = 0**:

		The result of ni x hi x wi x (ci+co) is affected by the memory size. **pad != 0:** pad_w < kw pad_h < kh (pad_w+wi) x ci < 8000 The result of 2ni x (hi+pad_h) x (wi+pad_w) x (ci+co) is affected by the memory size.
24	LRN	ci < =256
25	Elu	The result of ni x ci x hi x wi is restricted by the memory size. The precision is limited. The value range is [-10, 10].
26	Exp	The result of ni x ci x hi x wi is restricted by the memory size. The precision is limited. The value range is [-10, 10].
27	Log	The result of ni x ci x hi x wi is restricted by the memory size.
28	Lstm_unit	The result of ni x ci x hi x wi is restricted by the memory size.
29	normalize	kernel_h < =32 kernel_w < =32 ni < =512 ci < =3200 wi < =5400000 hi < =5400000 pad = 0: The result of ni x hi x wi x (ci+co) is affected by the memory size. pad != 0: pad_w < kw pad_h < kh (pad_w+wi) x ci < 8000 The result of 2ni x (hi+pad_h) x (wi+pad_w) x (ci+co) is affected by the memory size.
30	power	The result of ni x ci x hi x wi is restricted by the memory size. y=power (x) is implemented by e^ (y x lnx). The precision is affected by exp

		and ln.
31	greptensor	no+shift_n < =ni ho+shift_h < =hi wo+shift_w < =wi wo x fo < =12000

32	InnerProduct	if (ni>1&&(hi>1||wi>1)
		Same as convolution
		else;
		Restricted by the memory size

33	batchnorm	The result of ni x ci x hi x wi is restricted by the memory size.
34	Eltwise	The result of ni x ci x hi x wi is restricted by the memory size.
35	Sqrt	The result of ni x ci x hi x wi is restricted by the memory size.
36	tanh	The result of ni x ci x hi x wi is restricted by the memory size.
37	Slice	The result of ni x ci x hi x wi is restricted by the memory size.
38	Silence	The result of ni x ci x hi x wi is restricted by the memory size.

2.2 Tensorflow Operator Boundary

No.	Python API	C++ API	Boundary
1	tf.nn.avg_pool	AvgPool	stride_h < = window_h; stride_w < = window_w Note: better to make window_h < =hi,window_w < =wi xo x window_h+xo < =2000 wi+xo < =2000 xo=(wi+pad_w-window_w)/stride_w+1; **pad != 0**: pad_w < kw pad_h < kh (pad_w+wi) x ci < 8000
2	tf.nn.max_pool	MaxPool	stride_h < = window_h; stride_w < =

			window_w Note: better to make window_h < =hi,window_w < =wi xo x window_h+xo < =2000 wi+xo < =2000 xo=(wi+pad_w-window_w)/stride_w+1; **pad != 0**: pad_w < kw pad_h < kh (pad_w+wi) x ci < 8000
3	tf.image.resize_images (ResizeMethod.BILINEAR)	ResizeBilinear	wi < =10000 ho < =10000 wo < =2500 hi < =12000 ci>=256: 4 x wo+wi+ho < 1000 Ci < 256: (4wo+wi+ho) x (ci+15) < =256000
4	tf.image.resize_images (ResizeMethod.NEAREST_NEIGHBOR)	ResizeNearestNeighbor	wi < =10000 ho < =10000 wo < =2500 hi < =12000 ci>=256: 4 x wo+wi+ho < 1000 Ci < 256: (4wo+wi+ho) x (ci+15) < =256000
5	tf.abs	Abs	The result of ni x ci x hi x wi is restricted by the memory size. The precision is limited. The value range is [-10, 10].
6	tf.nn.conv2d	Conv2D	kernel_h < =32 kernel_w < =32 ni < =512 ci < =3200 wi < =5400000 hi < =5400000 **pad = 0**: The result of ni x hi x wi x (ci+co) is affected by the memory size. **pad != 0**: pad_w < kw

			pad_h < kh (pad_w+wi) x ci < 8000 The result of 2ni x (hi+pad_h) x (wi+pad_w) x (ci+co) is affected by the memory size.
7	tf.greater	Greater	The result of ni x ci x hi x wi is restricted by the memory size.
8	tf.less_equal	LessEqual	The result of ni x ci x hi x wi is restricted by the memory size.
9	tf.nn.relu	Relu	The result of ni x ci x hi x wi is restricted by the memory size. The precision is limited. The value range is [-10, 10].
10	tf.nn.relu6	Relu6	The result of ni x ci x hi x wi is restricted by the memory size. The precision is limited. The value range is [-10, 10].
11	tf.contrib.keras.layers.LeakyReLU	/	The result of ni x ci x hi x wi is restricted by the memory size. The precision is limited. The value range is [-10, 10].
12	tf.exp	Exp	The result of ni x ci x hi x wi is restricted by the memory size.
13	tf.concat	Concat	1. Input restrictions: ni < 128, ni x hi x wi < 4096 x 4096 2. values[i].dim(axis)==16x (integer multiple of 16)
14	tf.nn.conv2d_tra	Conv2DBackpropInput	max(stride_h,stride_w) x num_output <=256
15	tf.sigmoid	Sigmoid	The result of ni x ci x hi x wi is restricted by the memory size. The precision is limited. The value range is [-10, 10].
16	tf.add	Add	The result of ni x ci x hi x wi is restricted by the memory size.
17	tf.add_n	AddN	The result of ni x ci x hi x wi is restricted by the memory size.
18	tf.multiply	Multiply	The result of ni x ci x hi x wi is restricted by the memory size.

19	tf.subtract	Subtract	The result of ni x ci x hi x wi is restricted by the memory size.
20	tf.matmul	MatMul	The result of ni x ci x hi x wi is restricted by the memory size.
21	tf.nn.bias_add	BiasAdd	The result of ni x ci x hi x wi is restricted by the memory size.
s22	tf.nn.fused_batch norm	FusedBatchNorm	The result of ni x ci x hi x wi is restricted by the memory size.
23	tf.nn.lrn	LRN	The result of ni x ci x hi x wi is restricted by the memory size.
24	tf.where	Select	The result of ni x ci x hi x wi is restricted by the memory size.
25	tf.summary.merge	Merge	Not restricted
26	tf.nn.elu	Elu	The result of ni x ci x hi x wi is restricted by the memory size. The precision is limited. The value range is [-10, 10].
27	tf.rsqrt	Rsqrt	The result of ni x ci x hi x wi is restricted by the memory size. The precision is limited. The value range is [-10, 10].
28	tf.exp	Exp	The result of ni x ci x hi x wi is restricted by the memory size.
29	tf.log	Log	The result of ni x ci x hi x wi is restricted by the memory size.
30	tf.tanh	Tanh	The result of ni x ci x hi x wi is restricted by the memory size.
31	tf.slice	Slice	1. Input restrictions: ni < 128, ni x hi x wi < 4096 x 4096 2. Parameter restrictions: begin[i]==16x, size[i]==16x
32	tf.contib.layers.fl atten	Flatten	If IPU operators exist before and after the flatten API, the operators that either directly or indirectly depend on the output of the flatten API are considered as activation operators (such as relu, sigmoid, and tanh) or MatMul operators.
33	tf.split	Split	Input restrictions: ni < 128, ni x hi x wi < 4096 x 4096 Parameter restrictions:

			2.1. Currently, num_or_size_splits can be only an integer, instead of an array. 2.2. value.dim(axis) ==num_or_size_splits x 16x
34	tf.nn.depthwise_conv2d	DepthwiseConv2dNative	kernel_h < =32 kernel_w < =32 ni < =512 wi < =5350000 hi < =5400000 ci x m < =640 ni x ci x hi x wi>=8 ci x m x kh x kw>=8 **pad = 0**: The result of ni x hi x wi x (ci+co) is affected by the memory size. **pad != 0**: pad_w < kw pad_h < kh (pad_w+wi) x ci < 8000 The result of 2ni x (hi+pad_h) x (wi+pad_w) x (ci+co) is affected by the memory size.
35	tf.cast	Cast	The result of ni x ci x hi x wi is restricted by the memory size. The following data types can be converted: CAST_FLOAT32_TO_UINT8 CAST_UINT8_TO_FLOAT32 CAST_INT8_TO_FLOAT16 CAST_FIX8_TO_FLOAT16 CAST_FLOAT16_TO_FIX8 CAST_FLOAT16_TO_FLOAT32 CAST_FLOAT32_TO_FLOAT16 CAST_INT16_TO_FLOAT16 CAST_FLOAT16_TO_INT16
36	tf.floor	Floor	The result of ni x ci x hi x wi is restricted by the memory size.
37	tf.contrib.keras.backend.switch	Switch	Not restricted
38	tf.identity	Identity	Not restricted
39	tf.nn.softplus	Softplus	The result of ni x ci x hi x wi is restricted

			by the memory size.
40	tf.nn.softsign	Softsign	The result of ni x ci x hi x wi is restricted by the memory size.
41	tf.pad	Pad, PadV2	The result of ni x ci x hi x wi is restricted by the memory size.
42	tf.contrib.rnn.LSTMCell	/	Parameter restrictions: num_units==16x && num_proj==16x
43	tf.contrib.rnn.GRUCell	/	Parameter restriction: num_units==16x
44	tf.contrib.rnn.GRUBlockCell	GRUBlcokCell	The result of ni x ci x hi x wi is restricted by the memory size.
45	tf.contrib.rnn.LSTMBlockCell	LSTMBlockCell	The result of ni x ci x hi x wi is restricted by the memory size.