

密级状态: 绝密() 秘密() 内部() 公开(√)

RKNN Compiler Support Operator List

(技术部,图形计算平台中心)

	当前版本:	V1.4.0
文件状态:	作 者:	NPU团队
[]正在修改	编辑:	刘雯君
[√]正式发布	审 核:	熊伟
	完成日期:	2022-09-02

瑞芯微电子股份有限公司 Rockchips Semiconductor Co., Ltd (版本所有,翻版必究)



更新记录

版本	修改人	修改日期	修改说明	核定人
v1.3.0	NPU 团队	2022-03-06	更新RK3588 OP列表,增加CPU OP List	熊伟
v1.3.1	NPU 团队	2022-03-26	增加首层输入说明列表	熊伟
v1.3.2	NPU 团队	2022-04-21	更新RV1103/1106 OP平台列表	熊伟
v1.4.0	NPU 团队 2022-09-02		1.新增RK3588多核协同运行支持情况 2.更新LSTM、transpose、softmax等OP支持情况 3.新增Conv-Add/Add-ReLu/Mul-ReLu Fuse OP支持情况	熊伟



目 录

第一章 RK356X NPU Operator List	3
第二章 RK3588 NPU Operator List	31
第三章 RK1103/1106 NPU Operator List	59
第四章 CPU Operator List	86
第五章 首层输入说明	91



第一章 RK356X NPU Operator List



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
			input tensor	batch/ 输入的batch		支持两个tensor的广播操作,以ONNX默认排列NCHW	
Add/Bias	支持	int8		channel/ 输入的channel	无限制	做 说 明 , 支 持 以 下 广 播 方 式 : 1.OP(A(N,C,H,W),B(N,C,H,W)),即两个维度相同的 tensor进行操作	per-layer/ per-channel
Aud/Blas	文 村	float16	[batch, channel, height, width]:tensor	height/ 输入的height		2. OP(A(N,C,H,W),B(C,1,1)), 即C维度做broadcasting 3. OP(A(N,C,H,W),B(scalar)), 即以单个标量做 broadcasting 说明: A或B都可以作为广播方。	
				width/ 输入的width		例子见 <u>注释(1)</u>	
	支持		input_tensor [batch, channel, height, width]:tensor	batch/ 输入的batch			
Sub		int l 6		channel/ 输入的channel	世 说 明 , 支 持 1.OP(A(N,C,H,W),B tensor进行操作 2. OP(A(N,C,H,W),B 3. OP(A(N,C,H,W),B) broadcasting		per-layer/ per-channel
				height/ 输入的height		3. OP(A(N,C,H,W),B(scalar)),即以单个标量做	
				width/ 输入的width			
			input_tensor [batch, channel, height, width]:tensor	batch/ 输入的batch		支持两个tensor的广播操作,以ONNX默认排列NCHW 做 说 明 , 支 持 以 下 广 播 方 式 : 1.OP(A(N,C,H,W),B(N,C,H,W)),即两个维度相同的 tensor进行操作 2.OP(A(N,C,H,W),B(C,1,1)),即C维度做broadcasting	per-layer/
		Int l 6		channel/ 输入的channel			
Mul/Scale	支持			height/ 输入的height	broadcasting 4.OP(A(N,C,H,W),B(H 目前仅支持RK356X	3.OP(A(N,C,H,W),B(scalar)),即以单个标量做broadcasting 4.OP(A(N,C,H,W),B(H,W)),即HW维度做broadcasting,目前仅支持RK356X FP16类型 说明: A或B都可以作为广播方。	per-channel
				width/ 输入的width			



	· · · · · · · · · · · · · · · · · · ·							
	operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				input_tensor [batch, channel, height, width]:tensor	batch/ 输入的batch		支持两个tensor的广播操作,以ONNX默认排列NCHW	
	Div	部分支持	float16		channel/ 输入的channel	无限制	做说明,支持以下广播方式: 1、OP(A(N,C,H,W),B(N,C,H,W)),即两个维度相同的 tensor进行操作 2、OP(A(N,C,H,W),B(C,1,1)),即C维度做broadcasting	per-layer/ per-channel
	OIV	即	moatro		height/ 输入的height			
					width/ 输入的width			
		支持	int8 int16 float16	input_tensor [batch, channel, height, width]:tensor	batch/ 输入的batch	无限制		
	Max				channel/ 输入的channel	[1,8192]		per-layer/ per-channel
					height/ 输入的height			
					width/ 输入的width	[1,8176]		
				input_tensor [batch, channel, height, width]:tensor	batch/ 输入的batch	无限制		
	Min	支持	int8 int16 float16		channel/ 输入的channel	做 i 1.OF tensc 2. OF 3. OF broa		per-layer/
					height/ 输入的height			per-channel
					width/ 输入的width	[1,8176]		



	海心WET								
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式		
				batch/ 输入的batch	1				
Global AveragePool 支持	本 林	int8 int16 float16 input_tensor [batch, channel, height, width]:tensor int8 int16 float16 input_tensor [batch, channel, height, width]:tensor height/ 输入的height width/ 输入的width (1,343)		per-layer					
	XIII								
				width/					
	batch/ 输入的batch								
GlobalMaxPool	支持	int8 int16	input tensor [batch, channel, height,	channel/ 输入的channel	[1,8192]		per-layer		
Globaliviani	文持 width]:tensor width]:tensor height/输入的height [1,343] width/输入的width		po. Myor						
				width/	[[1,573]				



如心队亡 1 以							
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch	1		
			input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	[1,8192]		
				height/ 输入的height	[1,8192]		
				width/ 输入的width	[1,0172]	,,0192]	
			auto_pad:string	auto_pad/ pad的方式	仅支持NOTSET		per-layer
			ceil_mode:int64	ceil_mode/ 使用ceil或floor的方 式计算输出的shape	不支持		
			count_include_pad:int64	count_include_pad/ 是否包含pad数值进 行计算	1		
		int8	kernel_shape [kernel_h,	kernel_h/ height方向的kernel大 小	Tilldel Approbability on the Logarith		
AveragePool	支持	int16 float16	kernel_w]:int64[]	kernel_w/ width方向的kernel大 小	- 无限制,NPU支持[1,7];其它由CPU支持。		
				pads_left/ left方向的pads大小			
			pads [pads_top, pads_left,	pads_right/ right方向的pads大小	[0,7]		
			pads_bottom, pads_right]:int64[]	pads_top/ top方向的pads大小	[0,7]		
				pads_bottom/ bottom方向的pads大 小			
		strides [strides_h, strides_w]:int64[]		stride_h/ height方向的strides大 小			
			strides [strides_h, strides_w]:int64[] str wi	stride_w/ width方向的strides大 小	[1,8]		



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch	1		
			input_tensor [batch, channel, height,	channel/ 输入的channel	[1,8192]		
			width]:tensor	height/ 输入的height	[1,0100]		
				width/ 输入的width	[1,8192]		
			auto_pad:string	auto_pad/ pad的方式	仅支持NOTSET		per-layer
			ceil_mode:int64	ceil_mode/ 使用ceil或floor的方 式计算输出的shape	不支持		
			dilations [dilations_h,	dilations_h/ height方向的dilations 大小			
			dilations_w]:int64[]	dilations_w/ widtht方向的dilations 大小			
MaxPool	支持	int8 int16 float16	kernel shape [kernel h,	kernel_h/ height方向的kernel大 小	无限制,NPU支持[1,7]; 其它由CPU支持。		
			kernel_w]:int64[]	kernel_w/ width方向的kernel大 小	7.ZPK中9, NFU又河[1,/]; 共と田CFU又河。		
				pads_left/ left方向的pads大小			
			pads [pads_top, pads_left,	pads_right/ right方向的pads大小	[0,7]		
			pads_bottom, pads_right]:int64[]	pads_top/ top方向的pads大小	[0,7]		
				pads_bottom/ bottom方向的pads大 小			
			storage_order: int64	storage_order/优先储存方式	0		
		stride_h/ height方向的strides大 小 strides_h, strides_w]:int64[] stride_w/ width方向的strides大 小					
			strides [strides_h, strides_w]:int64[]	width方向的strides大			



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
			epsilon:double	epsilon/ 除以标准差时加上防 止除0的实数	非0实数,参考值为1e-5		
		int8	momentum:double	momentum/ 训练时的滑动平均参 数	无限制		
Batch Normalization	支持	int16 float16		batch/ 输入的batch	1		per-layer/ per-channel
			input_tensor [batch, channel, height,	channel/ 输入的channel	[1,0102]		
			width]:tensor	height/ 输入的height	[1,8192]		
				width/ 输入的width	[1,8176]		
				batch/ 输入的batch	支持多batch		
			input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	[1,8192]		
				height/ 输入的height			
				width/ 输入的width			
				layernorm_weight [channel, height,	channel/ 输入的channel	等于input_channel	
			width]:tensor(const) layernorm_bias[channel,height,	height/ 输入的height	等于input_height		
Layer Normalization	部分支持	float16	width]:tensor(const)	width/ 输入的width	等于input_width		per-layer
			normalized_shape:int64[]	normalized_ shape /参与每一批归一化的 Feature的尺寸	NPU仅支持,包含除第0维(batch维)以外的其他所有维度, 度, 如input_shape[n,c,h,w], 仅支持normalized_shape[c,h,w], 如input_shape[n,c,h], 仅支持normalized_shape[c,h], 如input_shape[n,c], 仅支持normalized_shape[c], 其余情况会转到CPU执行。		
			elementwise_affine:int64	elementwise_affine/ 是否具有可学习数	0 或 1 (默认为 0)。 当为1时拥有LayerNorm.weight与LayerNorm.bias,仅支持 weight/bias的尺寸: elementwise_shape与normalized_shape一 致; 当为0时LayerNorm.weight为全1值, LayerNorm.bias为全0 值。		
			eps:double	eps/ 防止除法溢出的偏移 参数	无限制		



					和心灰电1灰仍有依公司		
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch			
CV: v/D · L L L		int8 int16		channel/ 输入的channel	了. [FE] And		
Clip/ReLU6	支持	float16	width]:tensor	height/ 输入的height	无限制		per-layer
				width/ 输入的width			
	支持	int8 int16	input_tensor [batch, channel, height, width]:tensor	batch/ 输入的batch			
Elu				channel/ 输入的channel	无限制		
		float16		height/ 输入的height			
				width/ 输入的width			
				batch/ 输入的batch			
Gelu	支持	int8 int16	input tensor [batch, channel, height,	channel/ 输入的channel	无限制		
	A.19	float16		height/ 输入的height			
				width/ 输入的width			



				设置项/			
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch			
Relu	支持	int8 int16		channel/ 输入的channel	无限制		per-layer
Reiu	XH	float16	width]:tensor	height/ 输入的height	√L ÞÞ\$ 仲均		per-layer
				width/ 输入的width			
	支持		input_tensor [batch, channel, height, width]:tensor	batch/ 输入的batch			
LeakyRelu		int8 int16 float16		channel/ 输入的channel	无限制		per-layer
LeakyReiu				height/ 输入的height			
				width/ 输入的width			
				batch/ 输入的batch	无限制		
				channel/ 输入的channel			per-layer/ per-channel
PRelu	支持	int8 int16 float16	input_tensor [batch, channel, height, width]:tensor	height/ 输入的height			
				width/ 输入的width			
				slope/ PReLU系数	仅支持单个标量或C维度系数		



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch	1		
			input_tensor [sequence, batch, input_size] :tensor	sequence/ 输入的sequence	无限制,建议4对齐		
				input_size/ 输入的input_size	无限制,建议8对齐		
			direction:string	direction/ 指定GRU的运算方向	forward: 指定GRU的运算方向为前向 reverse: 指定GRU的运算方向为反向 bidirectional: 指定GRU的运算方向为双向		per-layer
			batch_size:int64 (extern)	batch_size/ 指定GRU输入的 batchsize	1		
GRU	部分支持 GRU扩展以及变体 命名为exGRU算	float16	sequence_size :int64 (extern)	sequence_size/ 指定GRU输入的 seqsize	无限制,建议4对齐		
GRE	子,参数项中指明 (extern)的项为 exGRU独有的参数 项。	nout to	hidden_size:int64 (extern)	hidden_size/ GRU单元中的 hiddensize	无限制,建议8对齐		
		linear_before_ reset:int64 linear_before_ reset/ LBR变种的选择					
			input_layout:string (extern)	input_layout/指定与 对应输入shape含义一 致的layout	1、snc: 指定layout对应的输入shape为[seqs, batches, input_size] 2、(sn)c: 指定layout对应的输入shape为[seqs*batches, input_size,1,1] 要求填写指定的layout,同时要求填写该op实际对应的		
					batch_size、sequence_size、hidden_size。		
		output_layout	output_layout:string (extern)	output_layout/指定与 对应输出shape含义一	1、sbnc: 指定layout对应的输出shape为 [seqs,directions,batches, hidden_size] 2、(sn)c: 指定layout对应的输出shape为[seqs*batches, directions*input_size,1,1]		
				致的layout	要求填写指定的layout,同时要求填写该op实际对应的batch_size、sequence_size、hidden_size。directions>1时仅支持batches=1。		



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch	batch>1时要求batch=4n,(n为正整数),建议n<=4。 注:LSTM单向:无限制,LSTM双向:不同时支持多batch。		
		input_tensor [sequence, batch, input_size]:tensor sequence/ 输入的sequence input_size/ 输入的input_size 和入的input_size 表限制,建议8对齐			无限制,建议4对齐		
			direction:string	direction/ 指定LSTM的运算方 向	forward: 指定LSTM的运算方向为前向 reverse: 指定LSTM的运算方向为反向 bidirectional: 指定LSTM的运算方向为双向		
			batch_size:int64 (extern)	batch_size/ 指定LSTM输入的 batchsize	大于1时仅支持4的倍数		
			sequence_size :int64 (extern)	sequence_size/ 指定LSTM输入的 seqsize	无限制,建议4对齐		
			hidden_size:int64 (extern)	hidden_size/ LSTM单元中的 hiddensize proj_size/ LSTM单元存在 0<=proj_size<=hiddensize			
	部分支持 LSTM 扩展以及变	int8	proj_size:int64 (extern)				
LSTM	体命名为exLSTM算子,参数项中指明(extern)的项为	项中指明 的项为 由有的参数 inpu	input_forget:int64	input_forget/ cifg变种的选择	1(T) or 0(F) 目前限定0, 即尚不支持		per-layer/ per-channel
	exLSTM独有的参数 项。		has_dropout:int64 (extern)	has_dropout/ caffe框架下的 indicator功能的选择	l(T) or 0(F) Caffe框架下,启用该功能要求输入indicator,工具端自动配置,无需手动配置。		
			has_projection:int64 (extern)	has_projection/ projection变种	l(T) or 0(F) 目前限定0, 即尚不支持		
			input_layout:string (extern)	input_layout/指定与 对应输入shape含义一 致的layout	1、snc: 指定layout对应的输入shape为[seqs, batches, input_size] 2、(sn)c: 指定layout对应的输入shape为[seqs*batches, input_size,1,1]		
				,	要求填写指定的layout,同时要求填写该op实际对应的 batch_size、sequence_size、hidden_size。		
		1、sbnc: 指定layout对应的输出shape为 [seqs,directions,batches, hidden_size] 2、(sn)c: 指定layout对应的输出shape为[seqs*batches, directions*input_size,1,1]					
				致的layout	要求填写指定的layout,同时要求填写该op实际对应的 batch_size、sequence_size、hidden_size。 directions>1时仅支持batches=1。		



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch			
			input_tensor [batch, channel, height,	channel/ 输入的channel	channel方向concat时,除了最后一个输入外,其他输入的channel大小需要对齐。对齐量: 8bit数据: 8对齐, 16bit数据: 4对齐 据: 4对齐 其他方向Concat无限制。		
Concat	部分支持	int8 int16 float16	width]:tensor	height/ 输入的height			per-layer
				width/ 输入的width			
			axis:int64	aixs/ 拼接的维度			
				batch/ 输入的batch			
Mish	支持	int8	input_tensor [batch, channel, height,	channel/ 输入的channel			
Mish ₹	XIT	float16	width]:tensor	height/ 输入的height	无限制		
				width/ 输入的width			



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
			b 年	batch/ 输入的batch	1		
		int8 int16	input_tensor [batch, channel, height,	channel/ 输入的channel	T. 1851 ded		
			width1:tensor	height/ 输入的height	无限制		
Pad	支持			width/ 输入的width	[1,8176]		
rad	又持	int64	pads:tensor	[n_begin,c_begin,h_begin,w_begin,n_end,c_end,h_end,w_end]/输入各轴上前后插入的pad大小	目前仅支持n_begin,c_begin,n_end,c_end为1 h_begin,w_begin,h_end,w_end无限制		
		float	constant_value:tensor	constant_value/ 填充入pad的值	无限制		
		string	mode:string	mode/pad模式	仅支持constant		
				batch/ 输入的batch	无限制		
			input_tensor [batch, channel, height,	channel/ 输入的channel			
	BadwaMaar 尚不支持 目前由	int8	width]:tensor	height/ 输入的height	[1,8192]		
ReduceMean CPU实现	CPU实现	int16 float16		width/ 输入的width			
			axes:int64[]	axes/ 指定reduce的轴	单轴:无限制,多轴:{2,3}		
			keepdims:int64[]	keepdims/ 是否需要保持维度不 变	0		



				设置项/			
operator	支持情况	输入数据类型	输入	输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch	无限制		
			input_tensor [batch, channel, height,	channel/ 输入的channel			
	尚不支持 目前由	int8	width]:tensor height/ 输入的height [1,8192]	per-layer/			
ReduceSum	CPU实现	int16 float16		width/ 输入的width			per-channel
			axes:int64[]	axes/ 指定reduce的轴	单轴:无限制,多轴:{2,3}		
			keepdims:int64[]	keepdims/ 是否需要保持维度不 变	0		
		batch/ 输入的batch channel/ 输入的channel input_tensor [batch, channel, height,			支持多batch		
	部分支持		width]:tensor	height/ 输入的height	[1,8192]		
Resize	目前NPU仅支持宽 高方向不超过8倍的 整倍数的最邻近插 值缩放,其余不支	int8 int16 float16		width/ 输入的width	1.[1,8176] 2.设放大倍数为s(s为正整数),width*s*(s-1)<=8192		per-layer
持部分的会Fallback 到CPU上实现。	持部分的会Fallback		mode:string	mode/resize采用的模 式	仅支持nearest		
			scales:int64[]	scales/尺寸放大倍数	仅支持1-8整数倍		
		roi:int64[]	roi/进行resize的输入 范围	仅支持全局([0,0,0,0,1,1,1,1])			



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch	无限制		
		int8		channel/ 输入的channel			
		int16 float16	width]:tensor	height/ 输入的height	[1,8192]		
				width/ 输入的width	[1,8176]		
		支持		batch_o/ 输出的batch_o	无限制		
Reshape	部分支持			channel_o/ 输出的channel	[1,8192]		
				height_o/ 输出的height			
		int64	shape (batch o, channel o, height o,	width_o/ 输出的width	[1,8176]		
		int64	width_o):tensor [[[[[[[[[[[[[[[[[[[n,c,h1,w1]- >[n,c,h2,w2]/ (h1*w1=h2*w2)	支持		
				[1,c,h,w]- >[c1,hw1,1,1]/ (c1=c/a, h*w=hw1/a,a 为整数)			
				[n,c,1,1]->[1,n1,h,w]/ (c=h*w/a, n1=n/a, a为 整数)	不支持		



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch	无限制		
	int8 尚不支持 int16 float16		input_tensor [batch, channel, height,	channel/ 输入的channel	[1,8192]		
			width]:tensor	height/ 输入的height	[[1,0192]		
Reverse		.8	width/ 输入的width	[1,8176]			
Sequence		batch_axis:int64	batch_axis/ 指定是否为batch维度	1			
			time_axis:int64	time_axis/ 指定是否为time维度	0		
			sequence_lens:int64[]	sequence_lens/ 指定序列翻转的数量	仅支持channel数		
				batch/ 输入的batch			
	gmoid 支持 int8 int16 float1	int8 input_tensor [int16 width]:tensor	input_tensor [batch, channel, height,	channel/ 输入的channel			
Sigmoid			width]:tensor	height/ 输入的height	无限制		
				width/ 输入的width			



和心厥电 7 放 仍 行 成 公 行							
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch			
Handelan aid	+++	int8 int16	input_tensor [batch, channel, height,	channel/ 输入的channel			
HardSigmoid	支持	float16	width]:tensor	height/ 输入的height	无限制		per-layer
			width/ 输入的width				
			batch/ 输入的batch				
Swish	支持	int8 持 int16	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制		
Swish	又行	float16	width]:tensor	height/ 输入的height			
				width/ 输入的width			
				batch/ 输入的batch			
HardSwish 支持	本 株	int8 int16	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制		per-layer
	1×14	float16	width]:tensor	height/ 输入的height	יער איב נהח		per-tayer
				width/ 输入的width			



	· · · · · · · · · · · · · · · · · · ·							
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式	
				batch/ 输入的batch				
	支持	int8 int16	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制		per-layer	
Softplus	XH	float16	width]:tensor	height/ 输入的height	ンL PR 申リ		per-iayer	
				width/ 输入的width				
				batch/ 输入的batch	无限制			
		input_tensor [batch, channel, height, width]:tensor float16 channel/ 输入的channel 理议4对齐 height/ 输入的height	[1,8192] 建议4对齐					
Softmax	尚不支持,目前由 CPU实现		width]:tensor	height/ 输入的height			per-layer	
				width/ 输入的width	•			
			axis:int64	axis/ 做softmax的轴	1,即channel方向			



						·····································	
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch			
	部分支持 in		input_tensor [batch, channel, height,	channel/ 输入的channel	T. B. Ad		
			width]:tensor	height/ 输入的height	无限制		
				width/ 输入的width			
Slice		int8 int16 float16	starts:int64[]	start/ 切分的起始位置	channel方向Slice时,channel_start要对齐。 对齐量: 8bit数据: 8对齐,16bit数据: 4对齐。 其他方向无限制。		per-layer
			ends:int64[]	ends/ 切分的终止位置	channel方向Slice时,channel_end要对齐。 对齐量: 8bit数据: 8对齐,16bit数据: 4对齐。 其他方向无限制。		
			axes:int64[]	axes/ 选取切分的轴	支持任意0~3轴, 支持同时多轴选择		
			steps:int64[]	steps/ 选取切分对应轴的步 长	1		
			1	batch/ 输入的batch			
			input_tensor [batch, channel, height,	channel/ 输入的channel			
			width]:tensor	height/ 输入的height	无限制		
Split	部分支持 intle	int8 int16		width/ 输入的width			per-layer
		float16	axis:int64	axis/ 切分的维度			
			split:int64[]	spilt/ 指定切分后维度的长 度	channel方向Split时,除了最后一个输出外,其他输出的channel需要对齐。对齐量: 8bit数据:8对齐,16bit数据:4对齐。 其他方向无限制。		



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch			
Toub	*	int8 int16	input_tensor [batch, channel, height,	channel/ 输入的channel	工門私		per-layer
	float16	width]:tensor	height/ 输入的height	无限制		per-layer	
			width/ 输入的width				
		batch/ 输入的batch [1,1024] channel/ 输入的channel input_tensor [batch, channel, height,			[1,1024]		
			[1,0102]				
			width]:tensor	height/ 输入的height	[1,8192]		
Transpose	部分支持	int8 int16 float16	int16	width/ 输入的width	[1,8176]		
			perm:int64[]	axis order/ 转置的轴顺序	仅支持 (1) perm=[3,1,2,0],in_shape=[n,c,1,1],且n,c要求8bit数据: 8 对齐, 16bit数据: 4对齐。 (2) perm=[3,1,2,0],in_shape=[1,c,1,w],且w,c要求8bit数据: 8 对齐, 16bit数据: 4对齐。 (3) perm=[2,1,0,3],in_shape=[n,c,1,1],且n,c要求8bit数据: 8 对齐, 16bit数据: 4对齐。 (4) perm=[2,1,0,3],in_shape=[1,c,h,1],且h,c要求8bit数据: 8 对齐, 16bit数据: 4对齐。		



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch	无限制		
			input_shape [batch, channel, height,	channel/ 输入的channel	无限制		
			width]:tensor	height/ 输入的height	无限制		
				width/ 输入的width	仅对首层输入width存在限制 详见 <u>首层输入说明</u>		
				num_output/ 输出的channel	无限制		
			kernel_shape [num_output,	num_input/ 输入的channel	之上 Mic (191)		
		num_input, kernel_h, kernel_w]:int64[]	num_input, kernel_h, kernel_w]:int64[]	kernel_h/ height方向的kernel大 小			
				kernel_w/ width方向的kernel大 小	[1,31]		
Convolution	支持	int8 int16 float16	ut16	stride_h/ height方向的strides大 小			per-layer/ per-channel
			strides [strides_h, strides_w]:int64[]	stride_w/ width方向的strides大 小	大 [1,7]		
				pads_left/ left方向的pads大小			
			pads [pads_top, pads_left,	pads_right/ right方向的pads大小	[0,15]		
			pads_bottom, pads_right]:int64[]	pads_top/ top方向的pads大小			
				pads_bottom/ bottom方向的pads大 小			
		group:int64	group:int64	group/ group的大小	无限制		
			attacione fattacione d	dilations_h/ height方向的dilations 大小			
		dilations [dilatio dilations_w]:int	dilations [dilations_h, dilations_w]:int64[]	dilations_w/ widtht方向的dilations 大小	[1,31]		



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch	无限制		
			input_shape [batch, channel, height,	channel/ 输入的channel	无限制		
			width]:tensor	height/ 输入的height	无限制		
				width/ 输入的width	仅对首层输入width存在限制 详见 <u>首层输入说明</u>		
				num_output/ 输出的channel	无限制		
				num_input/ 输入的channel	无限制		
			kernel_shape [num_output, num_input, kernel_h, kernel_w]:int64[]	kernel_h/ height方向的kernel大 小	mel大 [1,8] nel大 ides大		
		int8		kernel_w/ width方向的kernel大 小			per-layer/
Depthwise Convolution	支持	int16 float16	atridas (atridas la atridas policité (f.)	stride_h/ height方向的strides大 小			per-channel
		strides [strides_h, strides_w]:int64[] stride_w/ width方向的strides大 小	stride_w/ width方向的strides大				
				pads_left/ left方向的pads大小			
			pads [pads_top, pads_left,	pads_right/ right方向的pads大小			
			pads_bottom, pads_right]:int64[]	pads_top/ top方向的pads大小	[0,15]		
				pads_bottom/ bottom方向的pads大 小			
			dil he:	dilations_h/ height方向的dilations 大小	[1,31]		
			dilations_w]:int64[]	dilations_w/ widtht方向的dilations 大小	F-55-1		



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch	无限制		
			input shape [batch, channel, height,	channel/ 输入的channel	无限制		
			width]:tensor	height/ 输入的height	无限制		
				width/ 输入的width	仅对首层输入width存在限制 详见 <u>首层输入说明</u>		
			kernel_shape [num_output, num_input, kernel_h, kernel_w]:int64[]	num_output/ 输出的channel	无限制		
				num_input/ 输入的channel	无限制		
				kernel_h/ height方向的kernel大 小			per-layer/ per-channel
		int16 float16		kernel_w/ width方向的kernel大 小	[1,31]		
			strides [strides_h, strides_w]:int64[]	stride_h/ height方向的strides大 小	{2,4,8}		
ConvTranspose/ Deconvolution	支持			stride_w/ width方向的strides大 小			
			pads [pads_top, pads_left, pads_bottom, pads_right]:int64[]	pads_left/ left方向的pads大小	支持0-15 设置pad时注意: 不支持 kernel_h * dilations_h - dilations_h - pads_top < 0		
				pads_right/ right方向的pads大小			
				pads_top/ top方向的pads大小	不支持 kernel_w * dilations_w - dilations_w - pads_left < 0 不支持 stride_h * (height - 1) - pads_top + 1 < output_h 不支持 stride_w *(width - 1) - pads_left + 1 < output_w		
				pads_bottom/ bottom方向的pads大 小	The state of the s		
			group:int64	group/ group的大小	1 当且仅当num_input=num_output时,支持num_output		
			dilations [dilations_h, dilations_w]:int64[]	dilations_h/ height方向的dilations 大小	[1.31]		
				dilations_w/ widtht方向的dilations 大小	L. C.		



					· 加心倾电 1 放忉有限公司		
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
			input_tensor_1 [M, K]:tensor	M,K,N/ 输入数据的形状			
			input_tensor_2 [K,N]:tensor				
Comm	不支持 由CPU实现	int8	alpha:double	alpha/ 矩阵A*B乘法的scale	不支持		per-layer/
Gemm	不义符 田CPU头巩	float16	beta:double	beta/ 输入C矩阵的scale			per-channel
			transA:int64	transA/ A矩阵是否转置			
			transB:int64	transB/ B矩阵是否转置			
			innut tensor 1 [batch K. Cl:tensor	batch/ 输入的batch	不支持		
MatMul	不支持 由CPU实现	int8	input_tensor_1 [batch, K, C]:tensor	K/ 输入的K			per-layer/ per-channel
	T Z II II II I Z Z Z	float16	input_tensor_2 [batch, C, H]:tensor	C/ 输入的C			
				H/ 输入的H			



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持 (维度补齐)	量化支持方式
Convolution + Relu	支持						
Convolution + Clip	支持						
Convolution + PRelu/LeakyRelu	支持						
Convolution + Add	支持						
Convolution + Mul	尚不支持						
Convolution + Sigmoid	尚不支持						
Convolution + Tanh	尚不支持	同Convolution					
Convolution + Softplus	尚不支持						
Convolution + HardSigmoid	尚不支持						
Convolution + HardSwish	尚不支持						
Convolution + Elu	支持						
Convolution + Swish	尚不支持						
Convolution + Mish	尚不支持						



							TKA 3
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
ConvTranspose + Relu	尚不支持						
ConvTranspose + Clip	尚不支持						
ConvTranspose + PRelu/LeakyRelu	尚不支持						
ConvTranspose + Add	尚不支持						
ConvTranspose + Mul	尚不支持						
ConvTranspose + Sigmoid	尚不支持						
ConvTranspose + Tanh	尚不支持	同ConvTranspo	se				
ConvTranspose + Softplus	尚不支持						
ConvTranspose + HardSigmoid	尚不支持						
ConvTranspose + HardSwish	尚不支持						
ConvTranspose +Elu	尚不支持						
ConvTranspose + Swish	尚不支持						
ConvTranspose + Mish	尚不支持						



operator	支持情况	输入数据类型 输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
Depthwise Convolution + Relu	支持					
Depthwise Convolution + Clip	支持					
Depthwise Convolution + PRelu/LeakyRelu	支持					
Depthwise Convolution + Add	尚不支持					
Depthwise Convolution + Mul	尚不支持					
Depthwise Convolution + Sigmoid	尚不支持					
Depthwise Convolution + Tanh	尚不支持	同Depthwise Convolution				
Depthwise Convolution + Softplus	尚不支持					
Depthwise Convolution + HardSigmoid	尚不支持					
Depthwise Convolution + HardSwish	尚不支持					
Depthwise Convolution + Elu	支持					
Depthwise Convolution + Swish	尚不支持					
Depthwise Convolution + Mish	尚不支持					



ol	perator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式		
A	.dd+Relu	支持	同Add	Add						
М	ſul+Relu	支持	同Mul							

注释:

- (1)广播支持举例:
- 1. OP(A(N,C,H,W),B(N,C,H,W)): OP(A(1,16,32,8),B(1,16,32,8))=C(1,16,32,8)
- 2. OP(A(N,C,H,W),B(C,1,1)): OP(A(1,16,32,8),B(16))=C(1,16,32,8)
- $3 \cdot OP(A(N,C,H,W),B(scalar)): OP(A(1,16,32,8),B(1))=C(1,16,32,8)$
- 4. OP(A(N,C,H,W),B(H,W)): OP(A(1,16,32,8),B(32x8))=C(1,16,32,8)

设计建议: 当除数是常量时,建议转换成除数倒数的乘法。乘法在运算效率显著大于除法。

(2) 约束规格中, [a,b]表示支持a-b; {a,b,c}表示支持a,b,c。



第二章 RK3588 NPU Operator List



	和心版电 7 放							
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式	多核协同运 行支持情况
				batch/ 输入的batch				
Add/Bias	支持	int8	input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	工作 仙		per-layer/	已支持
Aud Dias	ZN	float16		height/ 输入的height				
				width/ 输入的width				
			input_tensor [batch, channel, height, width]:tensor	batch/ 输入的batch	 无限制 		mon layran/	
Sub	支持			channel/ 输入的channel				尚不支持
				height/ 输入的height				PALLXIA
				width/ 输入的width				
				batch/ 输入的batch				
		int8	input tensor	channel/ 输入的channel		支持两个tensor的广播操作,以ONNX默认排列NCHW做说明,支持以下广播方式: 1.OP(A(N,C,H,W),B(N,C,H,W)),即两个维度相同的tensor进行操作 2.OP(A(N,C,H,W),B(C,1,1)),即C维度做broadcasting	per-layer/	
Mul/Scale	支持		[bateh, channel, height, width]:tensor	height/ 输入的height	- 无限制		per-channel	尚不支持
				width/ 输入的width				



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式	多核协同运 行支持情况
				batch/ 输入的batch				
			input tensor	channel/ 输入的channel		支持两个tensor的广播操作,以ONNX默认排列NCHW做说明,支持以下广播方式: 1、OP(A(N,C,H,W),B(N,C,H,W)),即两个维度相同的tensor进行操作 2、OP(A(N,C,H,W),B(C,1,1)),即C维度做broadcasting 3、OP((N,C,H,W),scalar),即以单个标量做broadcasting 4、OP(A(N,C,H,W),B(H,W)),即HW维度做broadcasting,目前仅支持RK356X FP16类型说明:A或B都可以作为广播方。例子见注释(1)	per-layer/ per-channel	
Div	部分支持	float16	[batch, channel, height, width]:tensor	height/ 输入的height	无限制			尚不支持
				width/ 输入的width				
			input_tensor [batch, channel, height, width]:tensor	batch/ 输入的batch	无限制	支持两个tensor的广播操作,以ONNX默认排列NCHW做说明,支持以下广播方式: 1、OP(A(N,C,H,W),B(N,C,H,W)),即两个维度相同的tensor进行操作 2.OP(A(N,C,H,W),B(C,1,1)),即 C 维 度 做 broadcasting 3.OP(A(N,C,H,W),B(scalar)),即以单个标量做broadcasting 说明: A或B都可以作为广播方。	per-layer/ per-channel	
Max	支持	int8		channel/ 输入的channel	[1,8192]			尚不支持
	Z.N	float16		height/ 输入的height	[-,]			
				width/ 输入的width	[1,8176]			
				batch/ 输入的batch	无限制			
Min	支持	int8	input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	[1,8192]		per-layer/	尚不支持
	227	float16		height/ 输入的height			per-channel	
				width/ 输入的width	[1,8176]			



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式	多核协同运 行支持情况
			input_tensor [batch, channel, height, width]:tensor	batch/ 输入的batch	[1,8192]			
Global AveragePool	支持	int8 int16		channel/ 输入的channel			per-layer	尚不支持
	ΧN	float16		height/ 输入的height	- [1,343]			间小又行
				width/ 输入的width				
				batch/ 输入的batch	1	per-layer		
GlobalMaxPool	支持	int8 int16	input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	[1,8192]		ner-laver	尚不支持
Giorgina del	XIT	float16		height/ 输入的height	-[1,343]		ры шуы	阿尔文社
				width/ 输入的width				



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式	多核协同运 行支持情况
				batch/ 输入的batch	1			
			input_tensor [batch, channel, height,	channel/ 输入的channel	[1,8192]			
				height/ 输入的height	EL 0102]	per		
				width/ 输入的width	-[1,8192]		per-layer j	
			auto_pad:string	auto_pad/ pad的方式	仅支持NOTSET			尚不支持
			ceil_mode:int64	ceil_mode/ 使用ceil或floor的方式 计算输出的shape	不支持			
			count_include_pad:int64	count_include_pad/ 是否包含pad数值进行 计算	1			
AveragePool	支持	int8 int16	t8 t16 kernel_shape [kernel_h, oat16 kernel_w]:int64[]	kernel_h/ height方向的kernel大 小	─无限制,NPU支持[1,7];其它由CPU支持。			
		float16		kernel_w/ width方向的kernel大 小				
			I	pads_left/ left方向的pads大小				
			pads [pads_top, pads_left,	pads_right/ right方向的pads大小	50.73			
			pads_bottom, pads_right]:int64[]	pads_top/ top方向的pads大小	[0,7]			
				pads_bottom/ bottom方向的pads大小				
				stride_h/ height方向的strides大 小				
			strides [strides_h, strides_w]:int64[] stri wid	stride_w/ width方向的strides大 小	[1,8]			



期心版电 1 放货有限公司								
支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式	多核协同运 行支持情况	
			batch/ 输入的batch	1				
		input tensor [batch, channel, height,	channel/ 输入的channel	[1,8192]		量化支持方式 多核协同i 行支持情i		
			height/ 输入的height	[1 0102]				
			width/ 输入的width	[[1,0192]				
		auto_pad:string	auto_pad/ pad的方式	仅支持NOTSET				
		ceil_mode:int64	ceil_mode/ 使用ceil或floor的方式 计算输出的shape	不支持		per-layer 尚不支持		
		dilations [dilations_h,	dilations_h/ height方向的dilations 大小	1				
		dilations_w]:int64[]	dilations_w/ widtht方向的dilations 大小					
支持	int8 int16 float16	kernel shane [kernel h.	kernel_h/ height方向的kernel大 小			per-layer	尚不支持	
	lioutiv	kernel_w]:int64[]	kernel_w/ width方向的kernel大 小	无限制,NPU支持[1,7];其它由CPU支持。				
			pads_left/ left方向的pads大小					
		pads [pads_top, pads_left,	pads_right/ right方向的pads大小	10 71				
		pads_bottom, pads_right]:int64[]	pads_top/ top方向的pads大小	[65,1]				
			pads_bottom/ bottom方向的pads大小					
		storage_order: int64	storage_order/优先储 存方式	0				
	S	strides [strides_h, strides_w]:int64[]	stride_h/ height方向的strides大 小 stride_w/ width方向的strides大	[1,8]				
		int8	input_tensor [batch, channel, height, width]:tensor auto_pad:string ceil_mode:int64 dilations [dilations_h, dilations_w]:int64[] int8 int16 float16 kernel_shape [kernel_h, kernel_w]:int64[] pads [pads_top, pads_left, pads_bottom, pads_right]:int64[] storage_order: int64	input_tensor [batch, channel, height, width]:tensor input_tensor [batch, channel, height, width]:tensor auto_pad:string auto_pad/pad的方式 ceil_mode:int64 ceil_mode:int64 ceil_mode/使用cilgdfoor的方式; 计算输出的shape dilations_b/ height/pi向的dilations_b/ height/pi向的dilations_t/, dilations_w/ width/pi向的dilations_t/, dilations_w/ width/pi向的dilations_t/, height/pi向的dilations_t/, hernel_b/ height/pi向的dilations_t/, hernel_b/ height/pi向的dilations_t/, hernel_b/ height/pi向的pads_top/ telf/piophpads_top/ top/ top/ piophpads_top/ top/ top/ top/ piophpads_top/ top/ top/ piophpads_top/ top/ piophpads_top/ top/ top/ piophpads_top/ top/ top/ piophpads_top/ top/ top/ piophpads_top/ top/ piophpads_top/ top/ top/ piophpads_top/ top/ top/ top/ piophpads_top/ top/ top/ top/ piophpads_top/ top/ top/ top/ piophpads_top/ top/ top/ piophpads_top/ top/ top/ top/ piophpads_top/ top/ top/ top/ top/ top/ top/ top/	## Action of Batch, channed, height, width jtensor [batch, channed, height, width jtensor [batch, channed, height, width jtensor [batch, channed, height, width	### Part Part	Table	



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式	多核协同运 行支持情况
			epsilon:double	epsilon/ 除以标准差时加上防 止除0的实数	非0实数,参考值为1e-5			
			momentum:double	momentum/ 训练时的滑动平均参 数	无限制			
Batch Normalization	支持	int8 int16 float16		batch/ 输入的batch	1		per-layer/ per-channel	尚不支持
			input tensor [batch, channel, height,	channel/ 输入的channel	H 0102			
			width]:tensor	height/ 输入的height	[1,8192]			
				width/ 输入的width	[1,8176]			
				batch/ 输入的batch	支持多batch			
			input_tensor [batch, channel, height,	channel/ 输入的channel				
			width]:tensor	height/ 输入的height	[1,8192]			
				width/ 输入的width				
			layernorm_weight [channel, height,	channel/ 输入的channel	等于input_channel			
Layer	部分支持	float16	width]:tensor(const) layernorm bias [channel, height,	height/ 输入的height	等于input_height		per-layer	尚不支持
Normalization	即刀叉打	ilout 10	width]:tensor(const)	width/ 输入的width	等于input_width		per layer	阿小文母
			normalized_shape:int64[]	normalized_ shape /参与每一批归一化的 Feature的尺寸	NPU仅支持,包含除第0维(batch维)以外的其他所有维度,如input_shape[n,c,h,w], 仅支持normalized_shape[c,h,w], 如input_shape[n,c,h], 仅支持normalized_shape[c,h], 如input_shape[n,c], 仅支持normalized_shape[c], 其余情况会转到CPU执行。			
			elementwise_affine:int64	elementwise_affine/ 是否具有可学习数	0 或 1(默认为 0)。 当为1时拥有LayerNorm.weight与LayerNorm.bias,仅支持weight/bias的 尺寸:elementwise_shape与normalized_shape一致:当为0时 LayerNorm.weight为全1值,LayerNorm.bias为全0值。			
			eps:double	eps/ 防止除法溢出的偏移 参数	无限制			



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式	多核协同运 行支持情况
				batch/ 输入的batch			per-layer 已支持	
Cli- ID -LLIG	-1-4-1	int8 int16	input_tensor [batch, channel, height,	channel/ 输入的channel	THE Art			
Clip/ReLU6	支持	float16	width]:tensor	height/ 输入的height	无限制		per-iayer	亡 文符
			width/ 输入的width					
				batch/ 输入的batch				
Elu	支持	int8 int16	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制			从不士柱
Eiu	大 科	float16	width]:tensor	height/ 输入的height	人 PR 申 リ			尚不支持
				width/ 输入的width				
				batch/ 输入的batch				
Gelu	支持	int8 int16	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制			尚不支持
	~13	float16	width]:tensor	height/ 输入的height				1871 VN
				width/ 输入的width				



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持 (维度补齐)	量化支持方式	多核协同运 行支持情况
				batch/ 输入的batch				
Relu	-1-4-	int8 int16	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制			7 + b
Keiu	支持	float16	width]:tensor	height/ 输入的height		per-layer	已支持	
			width/ 输入的width					
				batch/ 输入的batch				
Looky Doly	-t-4t	int8 int16	input_tensor [batch, channel, height,	channel/ 输入的channel	工程在		per-layer 已支持	77 + kt
LeakyRelu	支持	float16	width]:tensor	height/ 输入的height	无限制			亡文 捋
				width/ 输入的width				
				batch/ 输入的batch				
				channel/ 输入的channel	无限制			
PRelu	支持	int8 int16 float16	input_tensor [batch, channel, height, width]:tensor	height/ 输入的height	Ne pre pre		per-layer/ per-channel	已支持
				width/ 输入的width				
			slope/ PReLU系数 仅支持单个标量或C维度系数					



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持 (维度补齐)	量化支持方式	多核协同运 行支持情况
				batch/ 输入的batch	1			
			input_tensor [sequence, batch, input_size] :tensor	sequence/ 输入的sequence	无限制,建议4对齐			
				input_size/ 输入的input_size	无限制,建议8对齐			
			direction:string	direction/ 指定GRU的运算方向	forward: 指定GRU的运算方向为前向 reverse: 指定GRU的运算方向为反向 bidirectional: 指定GRU的运算方向为双向			
			batch_size:int64 (extern)	batch_size/ 指定GRU输入的 batchsize	I			
	部分支持 GRU 扩展以及 变体命名为		sequence_size :int64 (extern)	sequence_size/ 指定GRU输入的 seqsize	无限制,建议4对齐			
GRU	exGRU算子,参 数项中指明 (extern)的项 为exGRU独有的	float16	hidden_size:int64 (extern)	hidden_size/ GRU单元中的 hiddensize	无限制,建议8对齐		per-layer	尚不支持
	参数项。		linear_before_ reset:int64	linear_before_ reset/ LBR变种的选择	1(T) or 0(F)		年化文行力人 行支持信	
			input_layout:string (extern)	input_layout/指定与对应输入shape含义一致的layout	1、snc: 指定layout对应的输入shape为[seqs, batches, input_size] 2、(sn)c: 指定layout对应的输入shape为[seqs*batches, input_size,1,1] 要求填写指定的layout,同时要求填写该op实际对应的batch_size、sequence_size、hidden_size。			
		outp	output_layout:string (extern)	output_layout/指定与 对应输出shape含义一 致的layout	1、sbnc: 指定layout对应的输出shape为[seqs,directions,batches, hidden_size] 2、(sn)c: 指定layout对应的输出shape为[seqs*batches, directions*input_size,1,1] 要求填写指定的layout,同时要求填写该op实际对应的batch_size、sequence_size、hidden_size。 directions>1时仅支持batches=1。			



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式	多核协同运 行支持情况
				batch/ 输入的batch	batch>1时要求batch=4n,(n为正整数),建议n<=4。 注:LSTM单向:无限制,LSTM双向:不同时支持多batch。			
			input_tensor [sequence, batch, input_size]:tensor	sequence/ 输入的sequence	无限制,建议4对齐			
				input_size/ 输入的input_size	无限制,建议8对齐			
			direction:string	direction/ 指定LSTM的运算方向	forward: 指定LSTM的运算方向为前向 reverse: 指定LSTM的运算方向为反向 bidirectional: 指定LSTM的运算方向为双向			
			batch_size:int64 (extern)	batch_size/ 指定LSTM输入的 batchsize	大于1时仅支持4的倍数			
			sequence_size :int64 (extern)	sequence_size/ 指定LSTM输入的 seqsize	无限制,建议4对齐		per-layer/	
	部分支持		hidden_size:int64 (extern)	hidden_size/ LSTM单元中的 hiddensize	无限制,建议8对齐		per-channel	
LSTM	LSTM 扩展以及 变体命名为 exLSTM算子,	int8 int16	proj_size:int64 (extern)	proj_size/ LSTM单元存在 projection时的proj_size	0<=proj_size<=hiddensize 目前限定0,即尚不支持projection功能			尚不支持
	参数项中指明 (extern)的项 为exLSTM独有 的参数项。	float16	input_forget:int64	input_forget/ cifg变种的选择	1(T) or 0(F) 目前限定0, 即尚不支持			
	的多数项。		has_dropout:int64 (extern)	has_dropout/ caffe框架下的indicator 功能的选择	I(T) or 0(F) Caffe框架下,启用该功能要求输入indicator,工具端自动配置,无需手动配置。			
			has_projection:int64 (extern)	has_projection/ projection变种	1(T) or 0(F) 目前限定0, 即尚不支持			
			input_layout:string (extern)	input layout/指定与对应输入shape含义一致的layout	1、snc: 指定layout对应的输入shape为[seqs, batches, input_size] 2、(sn)c: 指定layout对应的输入shape为[seqs*batches, input_size,1,1] 要求填写指定的layout,同时要求填写该op实际对应的batch_size、sequence_size、hidden_size。]		
			out	output_layout:string (extern)	output_layout/指定与 对应输出shape含义一 致的layout	1、sbnc: 指定layout对应的输出shape为[seqs,directions,batches, hidden_size] 2、(sn)c: 指定layout对应的输出shape为[seqs*batches, directions*input_size,1,1]		
			Same	要求填写指定的layout,同时要求填写该op实际对应的batch_size、sequence_size、hidden_size。directions>1时仅支持batches=1。				



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式	多核协同运 行支持情况
				batch/ 输入的batch				
			input_tensor [batch, channel, height,	channel/ 输入的channel	channel方向concatl时,除了最后一个输入外,其他输入的channel大小			
Concat 部分	部分支持	int8 int16 float16	width]:tensor	mput_tensor [batch, channel, height, width]:tensor 需要对齐。对齐量: 8bit数据: 16对齐,16bit数据: 8对齐。 其他方向Concat无限制。 height/ 输入的height			已支持	
				width/ 输入的width				
			axis:int64	aixs/ 拼接的维度	无限制			
				batch/ 输入的batch				
Mish	支持	int8	input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	- 无限制		尚不支	尚不支持
		float16	width]:tensor	height/ 输入的height				
				width/ 输入的width				



	+447	₩ HI 水平	tA.)	设置项/ 输入参数含义	从立场的		目 // 士杜士	
operator	支持情况	输入数据类型	输入	输入参数含义	约束规格	广播支持(维度补齐)	重化文符力式	行支持情况
				batch/ 输入的batch	1			
		int8 int16	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制		量化支持方式 多核	
		float16	loat16 widing:tensor	height/ 输入的height	Zupkung			
Pad 支持	支持			相D人口J WIGHT	[1,8176]			尚不支持
		int64	pads:tensor	[n_begin,c_begin,h_beg in,w_begin,n_end,c_en d,h_end,w_end]/ 输入各轴上前后插入 的pad大小	目前仅支持n_begin,c_begin,n_end,c_end为1 h_begin,w_begin,h_end,w_end无限制			
		float	constant_value:tensor	constant_value/ 填充入pad的值	无限制			
		string mode:string mode/pad模式 无限制						
				batch/ 输入的batch	无限制			
			input_tensor [batch, channel, height,	channel/ 输入的channel				
D. dura-Mari	尚不支持 目前	int8	width]:tensor	height/ 输入的height	[1,8192]		尚不支	ルエナ
ReduceMean	由CPU实现	int16 float16		width/ 输入的width				同个文符
			axes:int64[]	axes/ 指定reduce的轴	单轴:无限制,多轴:{2、3}			
			keepdims:int64[]	keepdims/ 是否需要保持维度不 变	0			



	ALCINC TIND THAT							
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持 (维度补齐)	量化支持方式	多核协同运 行支持情况
				batch/ 输入的batch	无限制			
			input_tensor [batch, channel, height,	channel/ 输入的channel				
D. I. C	尚不支持 目前	int8	width]:tensor	height/ 输入的height	[1,8192]		per-layer/	WITHE
ReduceSum	th CDI ISSUE	int16 float16		width/ 输入的width			per-channel	尚不支持
			axes:int64[]	axes/ 指定reduce的轴	单轴:无限制, 多轴:{2、3}			
			keepdims:int64[]	keepdims/ 是否需要保持维度不 变	0			
				batch/ 输入的batch	支持多batch			
			input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	<u>[1,8192]</u>			
				height/ 输入的height				
	如公支柱			width/ 输入的width	1.[1,8176] 2.设放大倍数为s(s为正整数),width*s*(s-1)<=8192			
	部分支持 目前NPU仅支持 宽高方向不超过 8倍的整倍数的 最邻近插值缩 放,其余不支持 部分的会 Fallback到CPU 上实现。	int8	mode:string	mode/resize采用的模式	仅支持nearest			
Resize		float16	scales:int64[]	scales/尺寸放大倍数	仅支持1-8整数倍		per-layer	尚不支持
		ck到CPU 实现。	roi:int64[]	roi/进行resize的输入范围	仅支持全局([0,0,0,0,1,1,1,1])			



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式	多核协同运 行支持情况
				batch/ 输入的batch	无限制			
		int8	input_tensor [batch, channel, height,	channel/ 输入的channel				
		int16 float16	width]:tensor	height/ 输入的height	[1,8192]			
				width/ 输入的width	[1,8176]			
				batch_o/ 输出的batch_o	无限制			
Reshape	部分支持			channel_o/ 输出的channel	[1,8192]			尚不支持
				height_o/ 输出的height				
		int64	shape (batch_o, channel_o, height_o, width_o):tensor	width_o/ 输出的width	[1,8176]			
				[n,c,h1,w1]- >[n,c,h2,w2]/ (h1*w1=h2*w2)	支持		尚不支持	
				[1,c,h,w]- >[c1,hw1,1,1]/ (c1=c/a, h*w=hw1/a, a 为整数)	当c,c1,hw,hw1均(i8 16对齐,fp16 8对齐)时支持			
				[n,c,1,1]->[1,n1,h,w]/ (c=h*w/a, n1=n/a, a为 整数)	当n,c,n1,hw均(i8 16对齐,fp16 8对齐)时支持			



	· 加心似电子 放彻有限公司							
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式	多核协同运 行支持情况
				batch/ 输入的batch	无限制			
			input_tensor [batch, channel, height,	channel/ 输入的channel	EL 81923			
			width]:tensor	height/ 输入的height	[1,8192]			
Reverse Sequence 尚不支持	尚不支持	int8 int16 float16		width/ 输入的width	[1,8176]			尚不支持
			batch_axis:int64	batch_axis/ 指定是否为batch维度	1			
			time_axis:int64	time_axis/ 指定是否为time维度	0			
			sequence_lens:int64[]	sequence_lens/ 指定序列翻转的数量	仅支持channel数			
				batch/ 输入的batch				
Sigmoid	士性	int8 int16	input_tensor [batch, channel, height,	channel/ 输入的channel	工門伍山			以 不士柱
Sigmoid	支持	float16	width]:tensor	height/ 输入的height	- 无限制			尚不支持
				width/ 输入的width				



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式	多核协同运 行支持情况
				batch/ 输入的batch				
11. 10		int8	input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel				North Lab
HardSigmoid	支持	int16 float16	width]:tensor	height/ 输入的height	无限制		per-layer	尚不支持
				width/ 输入的width				
				batch/ 输入的batch			尚不支持	
	-to tet	int8	channel/ 输入的chan input_tensor [batch, channel, height,	channel/ 输入的channel	THE MALE			
Swish	支持	int16 float16	width]:tensor	height/ 输入的height	无限制			向 个 文 符
				width/ 输入的width				
				batch/ 输入的batch				
HardSwish	支持	int8 int16	input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	无限制		per-layer	尚不支持
Hatuswish	X14	float16	width]:tensor	height/ 输入的height	No bis that		per-layer	同小又行
				width/ 输入的width				



				设置项/		対明心以生		
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式	多核协同运 行支持情况
				batch/ 输入的batch				
S. A. L.	++++	int8 int16	input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	T Wilded			WITH
Softplus 支	支持	float16		height/ 输入的height	无限制		per-layer	尚不支持
				width/ 输入的width				
				batch/ 输入的batch	无限制			
			input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	[1,8192] 建议8对齐			
Softmax	尚不支持,目前 由CPU实现			height/ 输入的height	1		per-layer	尚不支持
				width/ 输入的width			per-layer 尚不支	
			axis:int64	axis/ 做softmax的轴	1,即channel方向			



	和心族电1放闪行!							
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式	多核协同运 行支持情况
				batch/ 输入的batch				
			input_tensor [batch, channel, height,	channel/ 输入的channel	T m ted			
			width]:tensor	height/ 输入的height	无限制			
				width/ 输入的width			per-layer 尚不立	
Slice	部分支持	int8 int16 float16	starts:int64[]	start/ 切分的起始位置	channel方向Slice时,channel_start要对齐。 对齐量: 8bit数据: 16对齐,16bit数据: 8对齐。 其他方向无限制。			尚不支持
			ends:int64[]	ends/ 切分的终止位置	channel方向Slice时,channel_end要对齐。 对齐量: 8bit数据: 16对齐,16bit数据: 8对齐。 其他方向无限制。			
			axes:int64[]	axes/ 选取切分的轴	支持任意0~3轴, 支持同时多轴选择			
			steps:int64[]	steps/ 选取切分对应轴的步 长	1			
				batch/ 输入的batch				
			input_tensor [batch, channel, height,	channel/ 输入的channel	无限制			
			width]:tensor	height/ 输入的height	Zuprip			
Split	部分支持	int8 int16 float16		width/ 输入的width			per-layer	尚不支持
			axis:int64	axis/ 切分的维度	无限制			
			split:int64[]	spilt/ 指定切分后维度的长 度	channel方向Split时,除了最后一个输出外,其他输出的channel需要对齐。对齐量: 8bit数据: 16对齐, 16bit数据: 8对齐。 其他方向无限制。			



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式	多核协同运 行支持情况
				batch/ 输入的batch				
Tank		int8		channel/ 输入的channel	俞入的channel			ルナセ
Tanh	支持	float16	width]:tensor	height/ 输入的height	- 无限制		per-layer	同个文符
				width/ 输入的width				尚不支持
			batch/ 输入的batch	无限制				
			input_tensor [batch, channel, height,	channel/ 输入的channel	[1,8192]			
			width]:tensor	height/ 输入的height	[1,0172]			
Transpasa	部分支持	int8 int16		width/ 输入的width	[1,8176]			出不去柱
Transpose	中乃又付	float16	perm:int64[]	axis order/ 转置的轴顺序	RK3588支持所有RK356X上支持的transpose操作,在该基础上支持: n轴不参与转置时允许c、h、w三轴如下四种转置。限制与说明如下: 1、假设in_shape[n1,c1,h1,w1],out_shape[n2,c2,h2,w2] 2、四种转换分别为(1)perm=[0,2,3,1], NCHW->NHWC。(2)perm=[0,2,1,3], NCHW->NHCW。(3)perm=[0,3,1,2], NCHW->NWCH。(4)perm=[0,3,2,1], NCHW->NWHC。 3、以上四种转置无对齐要求。但在满足对齐要求时效率更高。对齐要求为:第1点中参数的c1、c2均要满足8bit数据: 16对齐,16bit数据: 8对齐。 4、NPU限制项:(1)perm=[0,2,3,1]时,8bit数据时,h1*w1<8176,w1*c1<1023。(2)perm=[0,3,1,2]时,h1*w1<8176。(3)perm=[0,3,2,1]时,h1*w1<8*8192,h2*w2<8*8192, w1<1024。			间小又行



		州心城屯 1 双切有帐公司						
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持 (维度补齐)	量化支持方式	多核协同运 行支持情况
				batch/ 输入的batch	无限制			
			input_shape [batch, channel, height,	channel/ 输入的channel	无限制			
			width]:tensor	height/ 输入的height	无限制			
				width/ 输入的width	仅对首层输入width存在限制 详见 <u>首层输入说明</u>			
				num_output/ 输出的channel				
			kernel_shape [num_output,	num_input/ 输入的channel	无限制			
			num_input, kernel_h, kernel_w]:int64[]	kernel_h/ height方向的kernel大 小				
				kernel_w/ width方向的kernel大 小	-[1,31]			
Convolution	支持	int8 int16 float16	strides [strides_h, strides_w]:int64[]	stride h/ height方向的strides大小 stride_w/ width方向的strides大小	[1,7]		per-layer/ per-channel	已支持
				pads_left/ left方向的pads大小				
			pads [pads_top, pads_left,	pads_right/ right方向的pads大小	[0,15]			
			pads_bottom, pads_right]:int64[]	pads_top/ top方向的pads大小	[0,13]			
				pads_bottom/ bottom方向的pads大小				
			group:int64	group/ group的大小	无限制			
		dilations [dilations h, 大バ	dilations_h/ height方向的dilations 大小	[1,31]				
			dilations_w]:int64[]	dilations_w/ widtht方向的dilations 大小	[reser]			



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式	多核协同运 行支持情况	
				batch/ 输入的batch	无限制			17 大小田が	
			input_shape [batch, channel, height,	channel/ 输入的channel	无限制		per-layer/		
			width]:tensor	height/ 输入的height	无限制				
				width/ 输入的width	仅对首层输入width存在限制 详见 <u>首层输入说明</u>				
				num_output/ 输出的channel	无限制		per-layer/		
			kernel_shape [num_output,	num_input/ 输入的channel	无限制				
			num_input, kernel_h, kernel_h/ kernel_w]:int64[] kernel_t/ height方向的kernel大			per-layer/			
Depthwise	支持	int8		kernel_w/ width方向的kernel大 小	[1,8]			3.**	
Convolution	又行	float16		stride_h/ height方向的strides大 小	(1.21)			二 文符	
			strides [strides_h, strides_w]:int64[]	stride_w/ width方向的strides大 小	[1,31]				
				pads_left/ left方向的pads大小					
			pads [pads_top, pads_left,	pads_right/ right方向的pads大小	[0,15]				
			pads_bottom, pads_right]:int64[]	pads_top/ top方向的pads大小	[[0,12]				
				pads_bottom/ bottom方向的pads大小					
			dilations [dilations_h,	dilations_h/ height方向的dilations 大小	[1,31]				
			dilations_w]:int64[]	dilations_w/ widtht方向的dilations 大小	[1,41]				



			刘中以代 J						
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持 (维度补齐)	量化支持方式	多核协同运 行支持情况	
				batch/ 输入的batch	无限制				
			input_shape [batch, channel, height,	channel/ 输入的channel	无限制		量化支持方式 多核协同设行支持情况 的不支持 尚不支持		
			width]:tensor	height/ 输入的height	无限制				
				width/ 输入的width	仅对首层输入width存在限制 详见 <u>首层输入说明</u>				
				num_output/ 输出的channel	无限制				
			kernel_shape [num_output, num_input, kernel_h,	num_input/ 输入的channel	无限制				
			kernel_w]:int64[]	kernel_h/ height方向的kernel大 小	F1 212				
				kernel_w/ width方向的kernel大 小	[1,31]				
		int8		stride_h/ height方向的strides大 小	(0.40)		per-layer/ per-channel 尚不支柱		
ConvTranspose/ Deconvolution	支持	int16 float16	strides [strides_h, strides_w]:int64[]	stride_w/ width方向的strides大 小	{2,4,8}		per-layer/ per-channel	尚不支持	
				pads_left/ left方向的pads大小					
			pads [pads_top, pads_left,	pads_right/ right方向的pads大小	支持0-15 设置pad时注意:				
			pads_bottom, pads_right]:int64[]	pads_top/ top方向的pads大小	不支持 kernel_h * dilations_h - dilations_h - pads_top < 0 不支持 kernel_w * dilations_w - dilations_w - pads_left < 0 不支持 stride_h *(height - 1) - pads_top + 1 < output_h 不支持 stride_w *(width - 1) - pads_left + 1 < output_w				
				pads_bottom/ bottom方向的pads大小					
			group:int64	group/ group的大小	支持1 当且仅当num_input=num_output时,支持num_output				
			dilations [dilations_h,	dilations_h/ height方向的dilations 大小	[1,31]				
			dilations_w]:int64[]	dilations_w/ widtht方向的dilations 大小	[[t]				



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持 (维度补齐)	量化支持方式	多核协同运 行支持情况		
			input_tensor_1 [M, K]:tensor	M,K,N/	the law of the second of the law					
			input_tensor_2 [K,N]:tensor	输入数据的形状	转为Matmul实现,约束同Matmul					
Gemm	尚不支持 目前	int8	alpha:double	alpha/ 矩阵A*B乘法的scale	无限制		per-layer/	尚不支持		
Genin	由CPU实现	int16	beta:double	beta/ 输入C矩阵的scale	无限制		per-channel	问小又行		
			transA:int64	transA/ A矩阵是否转置	仅静态tensor支持转置					
			transB:int64	transB/ B矩阵是否转置	大肝恋telisol 文行 7 直					
	部分支持 目前该支持仅针 对双feature输入		input_tensor_1 [batch, K, C]:tensor K/ input_tensor_2 [batch, C, H]:tensor H/	batch/ 输入的batch	双feature时: batch、H无限制 K支持[8,8192],对齐要求为8bit数据: 16对齐,16bit数据:8对齐 C支持[32,19384],对齐要求:32对齐					
MatMul		int8		K/ 输入的K	K*C <=65532 C*H <=65532 K*H <=65532		per-layer/	ルエナは		
Mauviui	未来将支持输入 为 feature+constant	float16		C/ 输入的C	feature+constant时: 若input_tensor_!为feature,则转为batch个feature[K,C,1,1] + weight[H,C,1,1]的conv;		per-channel	尚不支持		
				H/ 输入的H	若input_tensor_2为feature,则转为batch个feature[1,C,H,1] + weight[K,C,1,1]的conv; C对齐要求:32对齐 其他约束和conv相同					
			input_tensor_1 [batch, channel, K,	batch/ 输入的batch	双feature时: batch无限制 channl、K支持[8,8192],对齐要求为8bit数据: 16对齐,16bit数据;					
	部分支持		N]:tensor	channel/ 输入的channel	8对齐 N支持[32,19384],对齐要求: 32对齐 K*N <=65532					
MatMul (4d)	目前该支持仅针 对双feature输入 未来将支持输入	int16 int16 float16 input_tensor_2 [batch, channel, N, M]:tensor	入 into K/ 输入的	K/ 输入的K	K*M <=65532 M*N <=65532 feature+constantlh]:		per-layer/ per-channel			
	为 feature+constant		N/ 输入的M	若input_tensor_1为feature,则转为batch*channel个feature[K,N,1,1] + weight[M,N,1,1]的conv;若input_tensor_2为feature,则转为batch*channel个feature[1,N,M,1] +						
				M/ 输入的M	weight[K,N,1,1]的conv; N对齐要求: 32对齐 其他约束和conv相同					



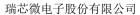
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式	多核协同运 行支持情况
Convolution + Relu	支持							已支持
Convolution + Clip	支持							已支持
Convolution + PRelu/LeakyR elu	支持							已支持
Convolution + Add	支持							已支持
Convolution + Mul	尚不支持							尚不支持
Convolution + Sigmoid	尚不支持							尚不支持
Convolution + Tanh	尚不支持	同Convolution						尚不支持
Convolution + Softplus	尚不支持							尚不支持
Convolution + HardSigmoid	尚不支持							尚不支持
Convolution + HardSwish	尚不支持							尚不支持
Convolution + Elu	支持							尚不支持
Convolution + Swish	尚不支持							尚不支持
Convolution + Mish	尚不支持							尚不支持



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持 (维度补齐)	量化支持方式	多核协同运 行支持情况
ConvTranspose + Relu	尚不支持							尚不支持
ConvTranspose + Clip	尚不支持							尚不支持
ConvTranspose + PRelu/LeakyRelu	尚不支持							尚不支持
ConvTranspose + Add	尚不支持							尚不支持
ConvTranspose + Mul	尚不支持							尚不支持
ConvTranspose + Sigmoid	尚不支持							尚不支持
ConvTranspose + Tanh	尚不支持	同ConvTranspose	е					尚不支持
ConvTranspose + Softplus	尚不支持							尚不支持
ConvTranspose + HardSigmoid	尚不支持							尚不支持
ConvTranspose + HardSwish	尚不支持							尚不支持
ConvTranspose + Elu	尚不支持							尚不支持
ConvTranspose + Swish	尚不支持							尚不支持
ConvTranspose + Mish	尚不支持							尚不支持



		· · · · · · · · · · · · · · · · · · ·						
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式	多核协同运 行支持情况
Depthwise Convolution + Relu	支持							已支持
Depthwise Convolution + Clip	支持							已支持
Depthwise Convolution + PRelu/LeakyR elu	支持							已支持
Depthwise Convolution + Add	尚不支持							已支持
Depthwise Convolution + Mul	尚不支持							尚不支持
Depthwise Convolution + Sigmoid	尚不支持							尚不支持
Depthwise Convolution + Tanh	尚不支持	同Depthwise Con	volution					尚不支持
Depthwise Convolution + Softplus	尚不支持							尚不支持
Depthwise Convolution + HardSigmoid	尚不支持							尚不支持
Depthwise Convolution + HardSwish	尚不支持							尚不支持
Depthwise Convolution + Elu	支持							尚不支持
Depthwise Convolution + Swish	尚不支持							尚不支持
Depthwise Convolution + Mish	尚不支持							尚不支持





operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持 (维度补齐)	量化支持方式	多核协同运 行支持情况
Add+Relu	支持	同Add						
Mul+Relu	支持	同Mul						
2、OP(A(N,C,H, 3、OP(A(N,C,H, 4、OP(A(N,C,H, 设计建议: 当除	W),B(N,C,H,W)): OP(A,W),B(C,1,1)): OP(A,W),B(scalar)): OP(W),B(H,W)): OP(A,W),B(H,W)): OP(A,W)	A(1,16,32 8),B(16) A(1,16,32,8),B(1)) A(1,16,32,8),B(32x8	=C(1,16,32,8) 8))=C(1,16,32,8) 的乘法。乘法在运算效率显著大于除	法。				



第三章 RK1103/1106 NPU Operator List



							圳心城电 1 放历有 2							
opera	ator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式						
					batch/ 输入的batch									
Add	/Rias	支持	int8 int16	input_tensor	channel/ 输入的channel	- 无限制	支持两个tensor的广播操作,以ONNX默认排列NCHW做说明,支持以下广播方式: $1.OP(A(N,C,H,W),B(N,C,H,W))$,即两个维度相同的tensor进行操作	per-layer/						
rad	Dias	XN	float16	[batch, channel, height, width]:tensor	height/ 输入的height	ZGRC192	2.OP(A(N,C,H,W),B(C,1,1)), 即 C 维 度 做 broadcasting 3.OP(A(N,C,H,W),B(scalar)), 即以单个标量做broadcasting 说明: A或B都可以作为广播方。例子见 <u>注释(1)</u>	per-channel						
					width/ 输入的width									
					batch/ 输入的batch									
Sub		支持 int8 int16 int16 float16 input_tensor [batch, channel, height, width]:tensor binding float f	int8 int16 float16 input_tensor [batch, channel, height, width]:tensor [batch, channel, height/ 输入的height			无限制	1.OP(A(N,C,H,W),B(N,C,H,W)), 即两个维度相同的tensor进	per-layer/						
				2.OP(A(N,C,H,W),B(C,1,1)), 即 C 维 度 做 broadcasting 3.OP(A(N,C,H,W),B(scalar)), 即以单个标量做broadcasting	per-channel									
											batch/ 输入的batch			
	Mul/Scale 支持		int8	input tensor	channel/ 输入的channel		支持两个tensor的广播操作,以ONNX默认排列NCHW做说明,支持以下广播方式: 1.OP(A(N,C,H,W),B(N,C,H,W)),即两个维度相同的tensor进行操作 2.OP(A(N,C,H,W),B(C,1,1)),即C维度做broadcasting	per-layer/						
Mul/		文持	int16	nt16 input_tensor [batch, channel, height, width]:tensor h	height/ 输入的height	3.OF 4.OF 仅支 说明	3.OP(A(N,C,H,W),B(scalar)),即以单个标量做broadcasting 4.OP(A(N,C,H,W),B(H,W)),即HW维度做broadcasting,目前 仅支持RK356X FP16类型 说明: A或B都可以作为广播方。 例子见注释(1)	per-channel						
				w #										



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch			
Div	部分支持	float16	input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	无限制	支持两个tensor的广播操作,以ONNX默认排列NCHW做说明,支持以下广播方式: 1、OP(A(N,C,H,W),B(N,C,H,W)),即两个维度相同的tensor进行操作 2、OP(A(N,C,H,W),B(C,1,1)),即C维度做broadcasting	per-layer/ per-channel
	IP // X3T	inoarro		height/ 输入的height	Данстру	3、OP((N,C,H,W),scalar), 即以单个标量做broadcasting 4、OP(A(N,C,H,W),B(H,W)), 即HW维度做broadcasting, 目 前仅支持RK356X FP16类型 说明: A或B都可以作为广播方。 例子见注释 (1)	
				width/ 输入的width		VI J JULIET VI J	
		int8 int16 float16	input_tensor [batch, channel, height, width]:tensor	batch/ 输入的batch	1	支持两个tensor的广播操作,以ONNX默认排列NCHW做说明,支持以下广播方式: 1、OP(A(N,C,H,W),B(N,C,H,W)),即两个维度相同的tensor进行操作 2.OP(A(N,C,H,W),B(C,1,1)),即 C 维 度 做 broadcasting 3.OP(A(N,C,H,W),B(scalar)),即以单个标量做broadcasting 说明: A或B都可以作为广播方。	per-layer/ per-channel
Max	支持			channel/ 输入的channel	[1,8192]		
				height/ 输入的height			
				width/ 输入的width	[1,8176]		
				batch/ 输入的batch	1		
Min	支持	int8	input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	明 , 支 持 以 下 厂 播 方 式 : 1.OP(A(N,C,H,W),B(N,C,H,W)), 即两个维度相同的tensor设	1.OP(A(N,C,H,W),B(N,C,H,W)), 即两个维度相同的tensor进	per-layer/ per-channel
		float16		height/ 输入的height		2.OP(A(N,C,H,W),B(C,1,1)), 即 C 维 度 做 broadcasting 3.OP(A(N,C,H,W),B(scalar)), 即以单个标量做broadcasting	
				width/ 输入的width	[1,8176]		



何心脉电 J 放 D 行 附 A 引							
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch	1		
Global AveragePool	支持	int8 int16	input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	[1,8192]	per-layer	
	X i i	intro		height/ 输入的height	[1,343]	pe	рстаусі
			width/ 输入的width				
				batch/ 输入的batch	1		
GlobalMaxPool	支持	int8 int16	<pre>input_tensor [batch, channel, height,</pre>	channel/ 输入的channel	[1,8192]		per-laver
GlobalMaxPool	, X, IT		width]:tensor	height/ 输入的height	-[1,343]		per-layer
				width/ 输入的width			



	师心城电 J 放仍有限公司							
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式	
				batch/ 输入的batch	1			
			input_tensor [batch, channel, height,	channel/ 输入的channel	[1,8192]			
			width]:tensor	height/ 输入的height	[1 0102]			
				width/ 输入的width	[1,8192]			
			auto_pad:string	auto_pad/ pad的方式	仅支持NOTSET			
			ceil_mode:int64	计算输出的shape	不支持			
			count_include_pad:int64	count_include_pad/ 是否包含pad数值进行 计算	性に対して			
AveragePool	支持	int8 int16	kernel_shape [kernel_h,	kernel_h/ height方向的kernel大小			per-layer	
			kernel_w]:int64[]	kernel_w/ width方向的kernel大小				
				pads_left/ left方向的pads大小				
			pads [pads_top, pads_left, pads_bottom,	pads_right/ right方向的pads大小	10.71			
			pads_right]:int64[]	pads_top/ top方向的pads大小	[0,7]			
				pads_bottom/ bottom方向的pads大小				
			ctridae [ctridae h. ctridae wikint641]	stride_h/ height方向的strides大 小	[1 0]			
			strides [strides_h, strides_w]:int64[]	stride_w/ width方向的strides大小	[1,8]			



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch	1		
			input_tensor [batch, channel, height,	channel/ 输入的channel	[1,8192]		
			width]:tensor	height/ 输入的height	EL 01021		
				width/ 输入的width	[1,8192]		
			auto_pad:string	auto_pad/ pad的方式	仅支持NOTSET		
			ceil_mode:int64	ceil_mode/ 使用ceil或floor的方式 计算输出的shape	不支持		
		dilations_h/ height方向的dilations大 dilations [dilations_h,					
			dilations_w]:int64[]	dilations_w/ widtht方向的dilations大 小			per-layer
MaxPool		int8 int16 kernel_shape [kernel_h, kernel_w]:int64[] kernel_w] int64[] kernel_m/ kernel_m/ kernel_w/ width方向的kernel大小 width方向的kernel大小	kernel_shape [kernel_h,	kernel_h/ height方向的kernel大小	无限制,NPU支持[1,7]; 其它由CPU支持。		
			- 元曜制,NPU又存[1,/]; 共七田CPU又存。	Siken in existing it Maliter exists			
				pads_left/ left方向的pads大小			
			pads [pads_top, pads_left, pads_bottom,	pads_right/ right方向的pads大小	50.71		
			pads_right]:int64[]	pads_top/ top方向的pads大小	[0,7]		
				pads_bottom/ bottom方向的pads大小			
			storage_order: int64	storage_order/优先储存 方式	0		
			strides [strides_h, strides_w]:int64[]	stride_h/ height方向的strides大 小			
				stride_w/ width方向的strides大小	-[1,8]		



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持 (维度补齐)	量化支持方式
			epsilon:double	epsilon/ 除以标准差时加上防止 除0的实数	非0实数,参考值为1e-5		
Batch Normalization 支持			momentum:double	momentum/ 训练时的滑动平均参数	无限制		
	支持	int8 int16		batch/ 输入的batch	1		per-layer/
	文刊	float16 channel/ input tensor floatch channel height 输入的channel		per-channel			
			width]:tensor	height/ 输入的height	[1,8192]		
				width/ 输入的width	[1,8176]		
				batch/ 输入的batch	支持多batch		
			input_tensor [batch, channel, height,	channel/ 输入的channel			
			width]:tensor	height/ 输入的height [1,8192]			
				width/ 输入的width			
			layernorm_weight [channel, height,	channel/ 输入的channel	等于input_channel		
Layer	尚不支持	float16	width]:tensor(const) layernorm_bias [channel, height,	height/ 输入的height	等于input_height		per-layer
Normalization	同个文持	noatro	width]:tensor(const)	width/ 输入的width	等于input_width		per-tayer
			normalized_shape:int64[]	normalized_ shape /参与每一批归一化的 Feature的尺寸	NPU仅支持,包含除第0维(batch维)以外的其他所有维度,如input_shape[n,c,h,w], 仅支持normalized_shape[c,h,w], 如input_shape[n,c,h], 仅支持normalized_shape[c,h], 如input_shape[n,c], 仅支持normalized_shape[c], 其余情况会转到CPU执行。		
				elementwise_affine/ 是否具有可学习数	0 或 1(默认为 0)。 当为1时拥有LayerNorm.weight与LayerNorm.bias,仅支持weight/bias 的尺寸: elementwise_shape与normalized_shape一致; 当为0时 LayerNorm.weight为全1值,LayerNorm.bias为全0值。		
			eps:double	eps/ 防止除法溢出的偏移参 数	无限制		



「中国							
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch	1		
Clip/ReLU6 支	支持	int8 int16	input_tensor [batch, channel, height,	channel/ 输入的channel	[1,8192]		per-layer
	· 文付	float16	width]:tensor	height/ 输入的height	[1,0172]		per-rayer
				width/ 输入的width	[1,8176]		
	支持	int8 int16 float16	input_tensor [batch, channel, height, width]:tensor	batch/ 输入的batch			
Elu				channel/ 输入的channel	无限制		
Liu				height/ 输入的height			
				width/ 输入的width			
				batch/ 输入的batch			
Gelu	支持	muo	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制		
		float16	width]:tensor	height/ 输入的height			
				width/ 输入的width			



	- 加心似电 J 放 切 有 N 公 可						
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch	1		
Relu	+44	int8 int16	input_tensor [batch, channel, height,	channel/ 输入的channel	[1,8192]		per-layer
	支持	float16	width]:tensor	height/ 输入的height	[1,0172]		per-rayer
				width/ 输入的width	[1,8176]		
LeakyRelu 支持		int8 int16 int16 float16 input_tensor [batch, channel, height, width]:tensor heigh	batch/ 输入的batch	1			
	支持		input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	-[1,8192]		per-layer
				height/ 输入的height			
				width/ 输入的width	[1,8176]		
				batch/ 输入的batch	1		
				channel/ 输入的channel	[1,8192]		per-layer/ per-channel
PRelu		int8 int16 float16	width]:tensor	height/ 输入的height	[]		
				width/ 输入的width	[1,8176]		
				slope/ PReLU系数	仅支持单个标量或C维度系数		



支持情况	输入数据类型		An market			
	棚八数16天型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
			batch/ 输入的batch	1		
		input_tensor [sequence, batch, input_size] :tensor	sequence/ 输入的sequence	无限制,建议8对齐		
			input_size/ 输入的input_size	无限制,建议8对齐		
		direction:string	direction/ 指定GRU的运算方向	forward:指定GRU的运算方向为前向 reverse:指定GRU的运算方向为反向 bidirectional:指定GRU的运算方向为双向		per-layer
		batch_size:int64 (extern)	batch_size/ 指定GRU输入的 batchsize	1		
尚不支持		sequence_size :int64 (extern)	sequence_size/ 指定GRU输入的seqsize	无限制,建议4对齐		
	RU算子,参 明(extern) GRU独有的	hidden_size:int64 (extern)	hidden_size/ GRU单元中的 hiddensize	无限制,建议8对齐		
的项为exGRU独有的 参数项。		linear_before_ reset:int64	linear_before_ reset/ LBR变种的选择	1(T) or 0(F)		
		input_layout:string (extern)	input_layout/指定与对应输入shape含义一致的layout	1、snc: 指定layout对应的输入shape为[seqs, batches, input_size] 2、(sn)c: 指定layout对应的输入shape为[seqs*batches, input_size,1,1] 要求填写指定的layout,同时要求填写该op实际对应的batch_size、sequence_size、hidden_size。		
		output_layout:string (extern)	output_layout/指定与对应输出shape含义一致的layout	1、sbnc: 指定layout对应的输出shape为[seqs,directions,batches, hidden_size] 2、(sn)c: 指定layout对应的输出shape为[seqs*batches, directions*input_size,1,1] 要求填写指定的layout,同时要求填写该op实际对应的batch_size、		
	GRU 扩展以及变体命 名为exGRU算子,参 数项中指明(extern) 的项为exGRU独有的	GRU 扩展以及变体命 名为exGRU算子,参 数项中指明(extern) 的项为exGRU独有的	itensor direction:string batch_size:int64 (extern) sequence_size :int64 (extern) hidden_size:int64 (extern) hidden_size:int64 (extern) linear_before_ reset:int64 input_layout:string (extern)	input_tensor [sequence, batch, input_size] input_size] itensor direction/ firection/ f	input_tensor [sequence, batch, input_size] tensor Sequence	



	和心族电子及仍有限公司						
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch	batch>1时要求batch=4n,(n为正整数),建议n<=4。 注:LSTM单向: 无限制,LSTM双向: 不同时支持多batch。		
			input_tensor [sequence, batch, input_size]:tensor	sequence/ 输入的sequence	无限制,建议4对齐		
				input_size/ 输入的input_size	无限制,建议8对齐		
			direction:string	direction/ 指定LSTM的运算方向	forward: 指定LSTM的运算方向为前向 reverse: 指定LSTM的运算方向为反向 bidirectional: 指定LSTM的运算方向为双向		
			batch_size:int64 (extern)	batch_size/ 指定LSTM输入的 batchsize	大于1时仅支持4的倍数		
			sequence_size :int64 (extern)	sequence_size/ 指定LSTM输入的 seqsize	无限制,建议4对齐		per-layer/
			hidden_size:int64 (extern)	hidden_size/ LSTM单元中的 hiddensize	无限制,建议8对齐		per-channel
LSTM	部分支持 LSTM 扩展以及变体	into	proj_size:int64 (extern)	proj_size/ 0<=proj_size<=hiddensize			
LSTM	命名为exLSTM算子, 参数项中指明 (extern)的项为	int8 int16	input_forget:int64	input_forget/ cifg变种的选择	1(T) or 0(F) 目前限定0, 即尚不支持		
	exLSTM独有的参数项 。		has_dropout:int64 (extern)	has_dropout/ caffe框架下的indicator 功能的选择 1(T) or 0(F) Caffe框架下,启用该功能要求输入indicator,工具端自动配置,无 需手动配置。			
			has_projection:int64 (extern)	has_projection/ projection变种	1(T) or 0(F) 目前限定0,即尚不支持		
			input_layout:string (extern)	input_layout/指定与对应输入shape含义一致	1、snc: 指定layout对应的输入shape为[seqs, batches, input_size] 2、(sn)c: 指定layout对应的输入shape为[seqs*batches, input_size,1,1]		
				的layout	要求填写指定的layout,同时要求填写该op实际对应的batch_size、sequence_size、hidden_size。		
			output_layout:string(extern)	output_layout/指定与对应输出shape含义一致	1、sbnc: 指定layout对应的输出shape为[seqs,directions,batches, hidden_size] 2、(sn)c: 指定layout对应的输出shape为[seqs*batches, directions*input_size,1,1]		
				的layout	要求填写指定的layout,同时要求填写该op实际对应的batch_size、sequence_size、hidden_size。directions>1时仅支持batches=1。		



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch			
Concat			input_tensor [batch, channel, height,	channel/ 输入的channel height channel方向concat时,除了最后一个输入外,其他输入的channel大			
	部分支持	int8 int16 float16	width]:tensor	height/ 输入的height	— 小s需要对齐。对齐量: 8bit数据: 16对齐,16bit数据: 8对齐。 其他方向Concat无限制。		per-layer
				width/ 输入的width			
			axis:int64	aixs/ 拼接的维度	无限制		
				batch/ 输入的batch			
Mish	支持	int8 int16	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制		
Mish	12.19	float16	width]:tensor	height/ 输入的height			
				width/ 输入的width			



	(电子									
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式			
				batch/ 输入的batch	1					
		int8	input_tensor [batch, channel, height,	channel/ 输入的channel	The state of the s					
		int16 float16	width]:tensor	height/ 输入的height	无限制					
Pad 支				width/ 输入的width	[1,8176]					
	支持	int64	pads:tensor	[n_begin,c_begin,h_begi n,w_begin,n_end,c_end, h_end,w_end]/ 输入各轴上前后插入的 pad大小	目前仅支持n_begin,c_begin,n_end,c_end为1 h_begin,w_begin,h_end,w_end无限制					
		float	constant_value:tensor	constant_value/ 填充入pad的值	无限制					
		string	mode:string	mode/pad模式	仅支持constant					
				batch/ 输入的batch	无限制					
			input_tensor [batch, channel, height,	channel/ 输入的channel						
			width]:tensor	height/ 输入的height	[1,8192]					
ReduceMean	尚不支持	int8 int16 float16		width/ 输入的width						
			axes:int64[]	axes/ 指定reduce的轴	单轴:无限制, 多轴:{2,3}					
			keepdims:int64[]	keepdims/ 是否需要保持维度不变	0					



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch	无限制		
				channel/ 输入的channel			
ReduceSum	尚不支持	int8 int16	width]:tensor	height/ 输入的height	[1,8192]		per-layer/
Reducesum	问小义行	float16		width/ 输入的width			per-channel
			axes:int64[]	axes/ 指定reduce的轴	单轴:无限制, 多轴:{2,3}		
			keepdims:int64[]	keepdims/ 是否需要保持维度不变	0		
		int8 int16	input_tensor [batch, channel, height,	batch/ 输入的batch	约束规格: (1) height * width * type_bytes <= 130816; (2) input_tensor非四维时,shape无限制		
				channel/ 输入的channel			
		float16	(input_tensor的维度为4维时看作nchw)	height/ 输入的height			
				width/ 输入的width			
Reshape	部分支持			batch_o/ 输出的batch_o			
			shape (batch_o, channel_o, height_o,	channel_o/ 输出的channel	计算量: alignment=16/type_bytes;		
		int64	width_o):tensor (输出shape指定维度为4维时看作nchw)	height_o/ 输出的height	约束规格: (1) height_o * width_o * type_bytes <= 65535; (2) Align(height_o * width_o, alignment) <= 8192; (3) 输出shane 非贝维斯,shane 无限制		
		v	width_o/ 输出的width	- (3) 输出shape非四维时,shape无限制			



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式																																		
				batch/ 输入的batch	支持多batch																																				
				channel/ 输入的channel	[1,8192]																																				
部分支持		width]:tensor	height/ 输入的height	[1,0172]																																					
Resize	数的最邻近插值缩	int8 int16 float16		width/ 输入的width	1.[1,8176] 2.设放大倍数为s(s为正整数),width*s*(s-1)<=8192		per-layer																																		
	放,其余不支持部分 的会Fallback到CPU上 实现。		mode:string	mode/resize采用的模式	仅支持nearest																																				
			scales:int64[]	scales/尺寸放大倍数	仅支持1-8整数倍																																				
			roi:int64[]	roi/进行resize的输入范 围	仅支持全局([0,0,0,0,1,1,1,1])																																				
				batch/ 输入的batch	无限制																																				
			input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	[1 8102]																																				
				height/ 输入的height	[1,8192]																																				
Reverse Sequence	尚不支持			width/ 输入的width	[1,8176]																																				
		noutro	batch_axis:int64	batch_axis/ 指定是否为batch维度	1																																				
																						-	-	-			-				t	-	-	-	-		time_axis:int64	time_axis/ 指定是否为time维度	0		
			sequence_lens:int64[]	sequence_lens/ 指定序列翻转的数量	仅支持channel数																																				
				batch/ 输入的batch																																					
G: 11		int8	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制																																				
Sigmoid	支持	int16 float16	width]:tensor	height/ 输入的height																																					
				width/ 输入的width																																					



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式	
HardSigmoid 支持			₽ \$	batch/ 输入的batch				
	支线	int8 int16	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制		per-layer	
	Z.N	float16	width]:tensor	height/ 输入的height	ZUN IPJ		por anyon	
				width/ 输入的width				
				batch/ 输入的batch				
Swish	channel/ 输入的channel 支持 int16 input_tensor [batch, channel, height, width]tensor							
Swish 支持		float16	width]:tensor	height/ 输入的height				
				width/ 输入的width				



operator	支持情况	输入数据类型	输入	设置项/	约束规格	广播支持(维度补齐)	量化支持方式
	22.7.11.72	1147 799497 (-1.1		输入参数含义 batch/ 输入的batch		, man descript	1,000,000
H 10 : 1	HardSwish 支持	int8	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制		
HardSwisn		int16 float16	width]:tensor	height/ 输入的height			per-layer
			width/ 输入的width				
				batch/ 输入的batch			
Softplus	支持	int8 int16	input_tensor [batch, channel, height,	channel/ 输入的channel			per-layer
Sortplus	又付	float16	width]:tensor	height/ 输入的height	无限制		per-tayer
				width/ 输入的width			
				batch/ 输入的batch	无限制		
			input_tensor [batch, channel, height,	channel/ 输入的channel	硬件支持[1,8192] 建议8对齐		
Softmax	Softmax 部分支持	float16	width]:tensor	height/ 输入的height	1		per-layer
	2		width/ 输入的width				
			axis/ 做softmax的轴	1,即channel方向			



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch			
			input tensor [batch, channel, height,	channel/ 输入的channel	THE AND		
			width]:tensor	height/ 输入的height	无限制		
				width/ 输入的width			
Slice	部分支持	int8 int16 float16	starts:int64[]	start/ 切分的起始位置	channel方向Slice时,channel_start要对齐。 对齐量: 8bit数据: 16对齐,16bit数据: 8对齐。 其他方向无限制。		per-layer
			ends:int64[]	ends/ 切分的终止位置	channel方向Slice时,channel_end要对齐。 对齐量: 8bit数据: 16对齐,16bit数据: 8对齐。 其他方向无限制。		
			axes:int64[]	axes/ 选取切分的轴	支持任意0~3轴, 支持同时多轴选择		
			steps:int64[]	steps/ 选取切分对应轴的步长	1		
				batch/ 输入的batch			
			input tensor [batch, channel, height,	channel/ 输入的channel	THE AND		
			width]:tensor	height/ 输入的height	无限制		
Split	部分支持	int8 int16		width/ 输入的width			per-layer
		float16	axis:int64	axis/ 切分的维度	无限制		
			split:int64[]	spilt/ 指定切分后维度的长度	channel方向Split时,除了最后一个输出外,其他输出的channel需要对齐。 对齐量: 8bit数据: 16对齐,16bit数据: 8对齐。 其他方向无限制。		



				3九里诺/			
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch			
Tanh 支持	int8	input tensor [batch, channel, height,	channel/ 输入的channel	无限制		per-layer	
	float16	width]:tensor	height/ 输入的height	ZUNEJ		per-tayer	
				width/ 输入的width			
		batch/ 输入的batch channel/ 输入的channel width]:tensor tinput_tensor [batch, channel, height, width]:tensor theight/ 输入的height width/ 输入的width [1,8176]			无限制		
			input tensor [batch, channel, height,		[1 0102]		
			[1,0172]				
Transpose	部分支持	int8 int16 float16	perm:int64[]	axis order/ 转置的轴顺序	RV1106、RV1103支持所有RK356X上支持的transpose操作, 在该基础上支持: n轴不参与转置时允许c、h、w三轴如下四种转 置。限制与说明如下: 1、假设in_shape[n1,c1,h1,w1],out_shape[n2,c2,h2,w2] 2、四种转换分别为(1) perm=[0,2,3,1], NCHW->NHWC。(2) perm=[0,2,1,3], NCHW->NHCW。(3)perm=[0,3,1,2], NCHW->NWCH。(4)perm=[0,3,2,1], NCHW->NWHC。 3、以上四种转置无对齐要求。但在满足对齐要求时效率更高。对 齐要求为: 第1点中参数的c1、c2均要满足8bit数据: 16对齐, 16bit数据: 8对齐。 4、NPU限制项:(1)perm=[0,2,3,1]时,8bit数据时, h1*w1<8176,w1*c1<512: 16bit 数据时,h1*w1<8176,w1*c1<1023。(2)perm=[0,3,1,2]时,h1*w1<8176。(3) perm=[0,3,2,1]时,h1*w1<8*8192,h2*w2<8*8192, w1<1024。		



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch	无限制		
		input_shape [batch, channel, height,	channel/ 输入的channel	无限制			
		width]:	width]:tensor	height/ 输入的height	无限制		
			width/ 输入的width	仅对首层输入width存在限制 详见 <u>首层输入说明</u>			
				num_output/ 输出的channel	无限制		
			kernel_shape [num_output, num_input,	num_input/ 输入的channel	无限制		
			kernel_h, kernel_w]:int64[]	kernel_h/ height方向的kernel大小	[1,31]		
				kernel_w/ width方向的kernel大小	[1,51]		
Convolution	支持	int8 int16	strides [strides_h, strides_w]:int64[]	stride_h/ height方向的strides大 小	[1,7]		per-layer/ per-channel
				stride_w/ width方向的strides大小	11,7]		
			1	pads_left/ left方向的pads大小			
			pads [pads_top, pads_left, pads_bottom,	pads_right/ right方向的pads大小	[0,15]		
			pads_right]:int64[]	pads_top/ top方向的pads大小	[[0,12]		
				pads_bottom/ bottom方向的pads大小			
		group:int64	group/ group的大小	无限制			
			dila hei	dilations_h/ height方向的dilations大 小	[2,31]		
		dilations_w]:int64[]	dilations_w/ widtht方向的dilations大 小				



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch	无限制		
				channel/ 输入的channel	无限制		
			width]:tensor	height/ 输入的height	无限制		
				width/ 输入的width	仅对首层输入width存在限制 详见 <u>首层输入说明</u>		
				num_output/ 输出的channel	无限制		
			kernel_shape [num_output, num_input,	num_input/ 输入的channel	无限制		
			kernel_h, kernel_w]:int64[]	kernel_h/ height方向的kernel大小	[1,8]		
Denthuis Constitution	-t- let	int8		kernel_w/ width方向的kernel大小			per-layer/ per-channel
Depthwise Convolution	文抒	int16	strides [strides_h, strides_w]:int64[]	stride_h/ height方向的strides大 小		per-channel	
			strides [strides_ii, strides_w].iiii04[]	stride_w/ width方向的strides大小	[1,7] des大小		
				pads_left/ left方向的pads大小			
			pads [pads top, pads left, pads bottom,	pads_right/ right方向的pads大小	[0,15]		
			pads [pads_top, pads_left, pads_bottom, pads_right]:int64[]	pads_top/ top方向的pads大小	[0,13]		
				pads_bottom/ bottom方向的pads大小			
		dilations_h/ height方向的dilations大 dilations [dilations_h,	[1,31]				
			dilations_w]:int64[]	dilations_w/ widtht方向的dilations大 小			



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
				batch/ 输入的batch	无限制		
		width]:tensor	channel/ 输入的channel	无限制			
			height/ 输入的height	无限制			
			width/ 输入的width	仅对首层输入width存在限制 详见 <u>首层输入说明</u>			
				num_output/ 输出的channel	无限制		
			kernel_shape [num_output, num_input,	num_input/ 输入的channel	/ 		
			kernel_h, kernel_w]:int64[]	kernel_h/ height方向的kernel大小			
				kernel_w/ width方向的kernel大小	大小 [1,31]		
ConvTranspose/ Deconvolution	支持	int8 int16	strides [strides_h, strides_w]:int64[]	stride_h/ height方向的strides大 小	[2.8]		per-layer/ per-channel
				stride_w/ width方向的strides大小	[2,8]		
			1	pads_left/ left方向的pads大小			
			pads [pads_top, pads_left, pads_bottom,	pads_right/ right方向的pads大小	支持0-15 设置pad时注意:		
			pads_right]:int64[]	pads_top/ top方向的pads大小	不支持 kernel_h * dilations_h - dilations_h - pads_top < 0 不支持 kernel_w * dilations_w - dilations_w - pads_left < 0 不支持 stride_h *(height - 1) - pads_top + 1 < output_h 不支持 stride_w *(width - 1) - pads_left + 1 < output_w		
				pads_bottom/ bottom方向的pads大小	- T. Z Suride_w (Widdin - 1) - paus_leit + 1 < output_w		
			group:int64	group/ group的大小	1,当且仅当num_input=num_output时,支持num_output		
			dila hei	dilations_h/ height方向的dilations大 小	[1,31]		
		dilations_w]:int64[]	dilations_w/ widtht方向的dilations大 小				



	海霉·斯·							
operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式	
			input_tensor_1 [M, K]:tensor	M,K,N/	转为Matmul实现,约束同Matmul			
		input_tensor_2 [K,N]:tensor	输入数据的形状	ヤグjviaunui天死,をj水円viaunui				
Gemm	尚不支持 目前由CPU	int8	alpha:double	alpha/ 矩阵A*B乘法的scale	无限制		per-layer/	
Comm	实现	int16	beta:double	beta/ 输入C矩阵的scale	无限制		per-channel	
			transA:int64	transA/ A矩阵是否转置	- 仅静态tensor支持转置			
			transB:int64	transB/ B矩阵是否转置				
		工士林 日茶中CDU 1.00		batch/ 输入的batch				
Mari	尚不支持 目前由CPU		input_tensor_1 [batch, K, C]:tensor	imput_tensor_1 [batch, K, C]:tensor		双feature时: batch、H无限制 K支持[8,8192], 对齐要求为8bit数据: 16对齐, 16bit数据: 8对齐 C支持[32,19384], 对齐要求为32对齐 feature+constant时:		per-layer/
	int16	input_tensor_2 [batch, C, H]:tensor		若input_tensor_l为feature,则转为batch个feature[K,C,1,1] + weight[H,C,1,1]的conv; 若input_tensor为feature,则转为batch个feature[1,C,H,1] + weight[K,C,1,1]的conv; C对齐要求:32对齐 其他约束和conv相同		per-channel		
		input_tensor_2 [batch,		H/ 输入的H				



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
Convolution + Relu	支持						
Convolution + Clip	支持						
Convolution + PRelu/LeakyRelu	支持						
Convolution + Add	支持						
Convolution + Mul	尚不支持						
Convolution + Sigmoid	尚不支持						
Convolution + Tanh	尚不支持	同Convolution					
Convolution + Softplus	尚不支持						
Convolution + HardSigmoid	尚不支持						
Convolution + HardSwish	尚不支持						
Convolution + Elu	支持						
Convolution + Swish	尚不支持						
Convolution + Mish	尚不支持						



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持 (维度补齐)	量化支持方式
ConvTranspose + Relu	尚不支持						
ConvTranspose + Clip	尚不支持						
ConvTranspose + PRelu/LeakyRelu	尚不支持						
ConvTranspose + Add	尚不支持						
ConvTranspose + Mul	尚不支持						
ConvTranspose + Sigmoid	尚不支持						
ConvTranspose + Tanh	尚不支持	同ConvTranspose	e				
ConvTranspose + Softplus	尚不支持						
ConvTranspose + HardSigmoid	尚不支持						
ConvTranspose + HardSwish	尚不支持						
ConvTranspose + Elu	尚不支持						
ConvTranspose + Swish	尚不支持						
ConvTranspose + Mish	尚不支持						



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
		- 個八 数 加 天 生	1897	输入参数含义	ETA MIT	/ 加文时(非汉州)/	重化X 打刀式
Depthwise Convolution + Relu	支持						
Depthwise Convolution + Clip	支持						
Depthwise Convolution + PRelu/LeakyRelu	支持						
Depthwise Convolution + Add	支持						
Depthwise Convolution + Mul	尚不支持						
Depthwise Convolution + Sigmoid	尚不支持						
Depthwise Convolution + Tanh	尚不支持	同Depthwise Con	evolution				
Depthwise Convolution + Softplus	尚不支持						
Depthwise Convolution + HardSigmoid	尚不支持						
Depthwise Convolution + HardSwish	尚不支持						
Depthwise Convolution + Elu	支持						
Depthwise Convolution + Swish	尚不支持						
Depthwise Convolution + Mish	尚不支持						



operator	支持情况	输入数据类型	输入	设置项/ 输入参数含义	约束规格	广播支持(维度补齐)	量化支持方式
Add+Relu	支持	同Add					
Mul+Relu	支持	同Mul					

注释: (1) 广播支持举例:

- 1、OP(A(N,C,H,W),B(N,C,H,W)): OP(A(1,16,32,8),B(1,16,32,8))=C(1,16,32,8)
- 2、OP(A(N,C,H,W),B(C,1,1)): OP(A(1,16,32 8),B(16))=C(1,16,32,8)
- 3. OP(A(N,C,H,W),B(scalar)): OP(A(1,16,32,8),B(1))=C(1,16,32,8)
- (4、0P(A(N,C,H,W),B(H,W)); OP(A(1,16,32,8),B(32,8))=C(1,16,32,8) 设计建议: 当除数是常量时,建议转换成除数倒数的乘法。乘法在运算效率显著大于除法。

(2) 约束规格中, [a,b]表示支持a-b; {a,b,c}表示支持a,b,c。



第四章 CPU Operator List



Operator	描述	规格约束	说明
Add	加法操作	无限制	
AveragePool	平均池化	无限制	
ArgMin	取最小值的index	无限制	
ArgMax	取最大值的index	无限制	
BatchNormalization	批量归一化	无限制	
Cast	数据类型转换	SRC 支 持: FP32/BOOL/INT8/FP16/INT64 DST支持: FP32/INT8/FP16	
Clip	数据截断激活层	无限制	
Concat	合并操作	axis仅支持{0,1,2,3}	
Convolution	卷积操作	无限制	
ConvTranspose/Deconvolution	转置卷积	无限制	
DataConvert	数据类型转换	仅支持 bool/int8/float 类型转换	
DepthToSpace	通道方向空间方向转换	无限制	
Div	除法操作	无限制	
Equal	等于	无限制	
Ехр	指数函数	无限制	
Flatten	拉平操作	无限制	



Operator	描述	规格约束	说明
Gather	聚集操作	无限制	
Greater	大于	无限制	
GreaterOrEqual	大等于	无限制	
HardSwish (extern)	激活函数	无限制	ONNX扩展算子
InstanceNormalization	单例归一化	无限制	
LayerNorm (extern)	层归一化	无限制	ONNX扩展算子
Less	小于	无限制	
LessOrEqual	小等于	无限制	
LogSoftmax	激活函数	batchsize 仅支持 1	
LpNormalization	Lp归一化	无限制	
LRN (extern)	局部响应归一化	无限制	ONNX扩展算子
MatMul	多维矩阵相乘	无限制	
Max	取最大值	无限制	
MaxPool	最大池化	无限制	
MaxRoiPool	区域最大池化	无限制	
MaxUnpool	反向最大池化	无限制	
Mish(extern)	激活函数	无限制	ONNX扩展算子



			机心队电 7 及历有限公司
Operator	描述	规格约束	说明
Min	取最小值	无限制	
Mul	乘法	无限制	
Pad	填充	无限制	
Pow	指数计算	无限制	
Proposal (extern)	区域提议网络	batchsize 仅支持 1	ONNX扩展算子
ReduceMax	沿指定维度计算Max	输出维度不能超过4维	
ReduceMean	沿指定维度计算Mean	输出维度不能超过4维	
ReduceSum	沿指定维度计算Sum	输出维度不能超过4维	
ReduceMin	沿指定维度计算Min	输出维度不能超过4维	
Reorg	数据重排	无限制	
Reshape	数据形状改变	无限制	
Resize	数据宽高方向缩放	支持插值方式 bilinear; nearest2d	
ReverseSequence	序列翻转	无限制	
RoiAlign	区域对齐池化	仅支持Avg Pool Mode,batchsize 仅支持 1	
Slice	切片操作	batchsize 仅支持 1	
Softmax	激活函数	batchsize 仅支持 1	与ONNX OPSET 11规范一致



Operator	描述	规格约束	说明
Softmax (extern)	激活函数	batchsize 仅支持 1	ONNX扩展算子,与ONNX OPSET 13规 范一致
SpaceToDetph	空间方向向通道方向转换	无限制	
Split	拆分数据	无限制	
Sqrt	求平方根	无限制	
Squeeze	压缩数据维度	无限制	
Sub	减法	无限制	
Tile	扩充拷贝数据	batchsize 仅支持 1,不支持broadcasting	
Transpose	转置计算	无限制	
Upsample	上采样	支持插值方式 bilinear; nearest2d	



第五章首层输入说明



MADE AND ADDRESS TO					,,,,,,,,,												
	模型首层精度		首层设置输入数据类型	mean/scale/quant 后端实现设备	输入宽(width)对齐要求 单位:元素个数		输入宽(width)大小限制										
	类型	输入维度			当输入通道 (channel) 为1,3,4	当输入通道 (channel) 非1,3,4	当输入通道(channel)为1,3,4(声明见 <u>注释9</u>)	当输入通道 (channel) 非1,3,4									
			uint8	NPU	0		各卷积类型的width/kernel_h/kernel_w需要满足以下两式: 1. width * dilation kernel h < 1024*N										
	int8		int8	NI O		1	2. widh <= 2047 其中N必须为1到7的整数,超出范围的卷积不受支持,各卷积类型N的计算方式如下:	无限制									
	into		float16	CPU			Convolution: N = 8 - CEIL((dilation_kernel_h * dilation_kernel_w) / 128) Depthwise Convolution: N = 8 - CEIL((dilation_kernel_h * dilation_kernel_w) / 4096)) LIPIC IPI									
		-4维度	其他类型(* <u>注释</u> 8)	Cro			ConvTranspose/Deconvolution: N = 8 - CEIL((dilation_kernel_n * dilation_kernel_w) / 4096)										
RK356X		マジエ/文	uint8		4		各卷积类型的width/kernel_h/kernel_w需要满足以下两式: 1. width * dilation kernel h < 1024*N										
	float16		int8	CPU			2. width <= 2047	无限制									
			float16				Convolution: N = 8 - CEIL((dilation_kernel_h * dilation_kernel_w) / 128) Depthwise Convolution: N = 8 - CEIL((dilation_kernel_h * dilation_kernel_w) / 4096)										
			其他类型(* <u>注释</u> 8)				ConvTranspose/Deconvolution: N = 8 - CEIL((dilation_kernel_h * dilation_kernel_w) / 128)										
	无限制	非4维	无限制	CPU	1	1	无限制	无限制									
		-4维度	uint8	NPU	16		各卷积类型的width/kernel_h/kernel_w需要满足以下两式: 1. width * dilation kernel h <= 2048 * N										
	int8		int8	1110			1	2. width <= 2047	无限制								
			float16	-CPU				Convolution: N = 12 - MAX(CEIL((dilation kernel h * dilation kernel w) / 128), 3) Depthwise Convolution: N = 12 - MAX(CEIL((dilation kernel h * dilation kernel w) / 2048), 3)	JUPK IP3								
			其他类型(* <u>注释</u> 8)				ConvTranspose/Deconvolution: N = 12 - MAX(CEIL((dilation_kernel_h * dilation_kernel_w) / 128), 3)										
RK3588		174-1/2	uint8	-CPU		1	各卷积类型的width/kernel_h/kernel_w需要满足以下两式: 1. width * dilation kernel h <= 1024 * N										
	float16		int8		8		2. width <= 2047 其中N必须为1到7的整数,超出范围的卷积不受支持,各卷积类型N的计算方式如下: Convolution: N = 12 - MAX(CEIL((dilation_kernel_h * dilation_kernel_w) / 128), 3)		无限制								
			float16) Physikii								
													其他类型(* <u>注释</u> 8)				ConvTranspose/Deconvolution: N = 12 - MAX(CEIL((dilation_kernel_ir dilation_kernel_w) / 2048), 3)
	无限制	非4维	无限制	CPU	1	1	无限制	无限制									
RV1103/ RV1106	. 10		uint8	NINU	16			各卷积类型的width/kernel_h/kernel_w需要满足以下两式: 1. width * dilation_kernel_h <= 2048 * N 2. width <= 4096	7 170 4.1								
	int8	4维	int8	NPU	16	1	其中N必须为1到7的整数,超出范围的卷积不受支持,各卷积类型N的计算方式如下: Convolution: N = 8 - MAX(CEIL((dilation_kernel_h * dilation_kernel_w) / 128), 2) Depthwise Convolution: N = 8 - CEIL((dilation_kernel_h * dilation_kernel_w) / 4096) ConvTranspose/Deconvolution: N = 8 - MAX(CEIL((dilation_kernel_h * dilation_kernel_w) / 128), 2)	无限制									



注释:

- 1、该对齐约束仅针对零拷贝API,普通API无此对齐约束
- 2、输入宽的对齐要求可从零拷贝API中的w stride属性查询到,注意: w stride不支持更改
- 3、仅对输入宽(width)在不同的通道(channel)条件下有对齐要求,其他无约束
- 4、若输入不需要mean和scale,需要将mean和scale配置为0和1
- 5、若通道(channel) > 4,则mean/scale将统一使用第一个数值,即mean[0]和scale[0]
- 6、若首层为浮点类型则没有quant操作
- 7、CPU的mean/scale/quant操作仅针对RK356X和RK3588,RV1106/RV1103 不支持
- 8、输入对齐要求可能变动
- 9、声明:

|CEIL(x)将x向上取整 (示例: CEIL(0.4)=1)

MAX(x, y)将获取x、y中的较大值(示例: MAX(2,3) = 3)

dilation kernel h = kernel h * dilations h - dilations h + 1

dilation_kernel_w = kernel_w * dilations_w - dilations_w + 1

- 10、4维/非4维不是表示原始模型的输入维度,是rknn模型通过rknn_query接口查询的输入的维度信息
- 11、详细的用法请参考《Rockchip_RKNPU_User_Guide_RKNN_API》