

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:	V1.5.0
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:	2023-05-22

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支持列表	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
v1.4.1	NPU 团队	2022-12-05	1.新增Conv-Add-Relu Fuse OP支持情况; 2.新增输出接口的tensor和layout说行	熊伟
v1.4.1b20	NPU 团队/HPC团队	2023-01-12	1.更新RK3588首层输入宽的限制 2.更新RV1106 Conv-Add-Relu Fuse OP支持情况 3.更新RK3588/RV1106 Transpose限制	熊伟
v1.4.2	NPU团队	2023-02-13	1.新增RK3562 OP支持列表 2.修复部分描述错误	熊伟
v1.5.0	NPU团队	2023-05-22	1. 新增部分CPU OP支持项 2. 对所有平台新增add/mul更多广播支持项 3. 更新输入大分辨率规格支持	熊伟



目 录

第一章 RK3566/3568 NPU Operator List	3
第二章 RK3588 NPU Operator List	32
第三章 RV1103/1106 NPU Operator List	61
第四章 RK3562 NPU Operator List	89
第五章 CPU Operator List	118
第六章 模型输入输出说明	123



第一章 RK3566/3568 NPU Operator List



					00/3308 NPU Operator List			
	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
				input tensor	batch/ 输入的batch		支持ONNX规范的四维tensor的所有广播操作,以	
	Add/Bias		int8 float16		channel/ 输入的channel	无限制	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/
		X 19		[batch, channel, height, width]:tensor	height/ 输入的height	ען זיין אין	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
_					width/ 输入的width		说明: A或B都可以作为广播方。 例子见 <u>注释(1)</u>	
				input_tensor [batch, channel, height, width]:tensor	batch/ 输入的batch			
	Sub	支持			channel/ 输入的channel	- 无限制		per-layer/ per-channel
		2			height/ 输入的height			
-					width/ 输入的width			
				input_tensor [batch, channel, height, width]:tensor	batch/ 输入的batch			
					channel/ 输入的channel			per-layer/ per-channel
	Mul/Scale 支持	支持			height/ 输入的height	- 无限制 		
					width/ 输入的width			



	KK3500/5508 NI O Operator Elst 新心原在 1 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
					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		input_tensor	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/
	SIV.	邱	Hoatio	[batch, channel, height, width]:tensor	height/ 输入的height	ZL pic my	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 FP16类型说明: A 或B都可以作为广播方。	per-channel
					width/ 输入的width		例子见 <u>注释(1)</u>	
			int8 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		int8 float16		channel/ 输入的channel	[1,8192]		per-layer/ per-channel
		~."	noatro		height/ 输入的height			
					width/ 输入的width	[1,8176]		



INCOMODINE O OPERAN 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
				batch/ 输入的batch	1		
Global AveragePool		int8 float16	input tensor [batch, channel, height,	channel/ 输入的channel	[1,8192]		
	支持	width]:tensor height/ 输入的height width/ 输入的width/ 输入的width		per-layer			
				width/ 输入的width	[1,3-45] (***********************************		
				batch/ 输入的batch	1		
GlobalMaxPool		int8 float16	input tensor [batch, channel, height,	channel/ 输入的channel	[1,8192]		
S.Sounvari voi	支持 float16 width]:tensor height/输入的height width/ 输入的width/ 输入的width		per-layer				
				width/ 输入的width	[1,5-5] (1001A116又可范围)		



				设置项/	OU/3308 NI O Operator List	和心版电 1 放 1	
operator	支持情况	输入数据类型	输入	⁰	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	[1,8192]		
				width/ 输入的width	[1,6172]		
			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 count_include_pad:int64 count_include_pad/ 是否包含pad数值进 行计算 kernel_h/ height方向的kernel大 小 kernel_w]:int64[] kernel_w/ width方向的kernel大 小	不支持				
			count_include_pad:int64	是否包含pad数值进	1		
AveragePool	支持		kernel_w]:int64[]	width方向的kernel大	无帐制,NFU又转[1,/]; 共已田CFU又符。		per-layer
				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,/]		
				pads_bottom/ bottom方向的pads大 小			
			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]		



				00/3308 NI O Operator List	河心城屯 1 双切有限公司	
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
			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 0103]		
			width/ 输入的width	[1,0172]		
		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		
	kernel_shape [kernel_h, 小 工限制 NDII支持[17]。其它由CDII支持					
支持		st16 kernel_shape [kernel_h, kernel_w]:int64[]	height方向的kernel大小	无限制,NPU支持[1,7]; 其它由CPU支持。		per-layer
			kernel_w/ width方向的kernel大 小			
			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		
		strides [strides_h, strides_w]:int64[]	stride_h/ height方向的strides大 小			
			stride w/	[1,8]		
		int8	input_tensor [batch, channel, height, width]:tensor auto_pad:string ceil_mode:int64 dilations [dilations_h, dilations_w]:int64[] int8 float16 kernel_shape [kernel_h, kernel_w]:int64[] pads [pads_top, pads_left, pads_bottom, pads_right]:int64[] storage_order: int64	## A South A	Septimized Se	株式 株式 株式 株式 株式 株式 株式 株式



					00/3300 NI O Operator List	和心脉也 1 放下		
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	float16		batch/ 输入的batch	1		per-layer/ per-channel	
			input_tensor [batch, channel, height,	channel/ 输入的channel	无限制			
			width]:tensor	height/ 输入的height	ルドでは			
				width/ 输入的width	无限制			
				batch/ 输入的batch	支持多batch			
			input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	无限制			
				height/ 输入的height				
				width/ 输入的width				
			layernorm_weight [channel, height, width]:tensor(const)	channel/ 输入的channel	等于input_channel			
				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	
			nor	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, 应,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			



					00/3308 NI O 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
				batch/ 输入的batch			
Clip/ReLU6	本性	int8 float16	input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	工 (FH dad		per-layer
Спр/кедоб	支持	noatro		height/ 输入的height	无限制		per-layer
				width/ 输入的width			
		int8 float16	input_tensor [batch, channel, height, width]:tensor	batch/ 输入的batch			
Elu	支持			channel/ 输入的channel	无限制		
	XN			height/ 输入的height			
				width/ 输入的width			
				batch/ 输入的batch			
Gelu	支持	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel	工 (대 4대		
				height/ 输入的height	无限制		
				width/ 输入的width			



					INCOMOSTO IN O OPERATO EIST		710-13-18/-E 1 1/K I	
operato	or	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	Relu	支持	int8 float16	input tensor [batch, channel, height,	channel/ 输入的channel	无限制		per-layer
Refu		ΧĦ	Hoatio	width]:tensor	height/ 输入的height	入L pic 即J		per-rayer
					width/ 输入的width			
				input_tensor [batch, channel, height, width]:tensor	batch/ 输入的batch			
LeakyF	Relu		int8 float16		channel/ 输入的channel	无限制		per-layer
Deakyi	.coru	Χn			height/ 输入的height			
					width/ 输入的width			
					batch/ 输入的batch			
					channel/ 输入的channel	无限制		per-layer/ per-channel
PRelu		支持	int8 float16	input_tensor [batch, channel, height, width]:tensor	height/ 输入的height	ZG NC NO		
					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 λ	设置项/	约束规格	广播支持(维度补齐)	量化支持方式
	Z N IB VL	柳八双眉大王	THE	输入参数含义 batch/ 输入的batch	1	/ 1周人内《年及日月/	至此久的分為
			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		ner-laver
		E.持 Sequence_siz :int64 (exter sint) 的项为 独有的参数 linear_before reset:int64	batch_size:int64 (extern)	batch_size/ 指定GRU输入的 batchsize	1		per-layer
GRU	部分支持 GRU扩展以及变体 命名为exGRU算		sequence_size :int64 (extern)	sequence_size/ 指定GRU输入的 seqsize	无限制,建议4对齐		
	子,参数项中指明(extern)的项为 exGRU独有的参数 项。		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	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	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。		



					Constitution Else	710-G-10X-G-1 7X-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 无限制,建议4对齐 无限制,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/ LSTM单元中的 hiddensize proj_size/ LSTM单元存在 proj_size:int64 (extern) proj_size:int64 (extern) proj_size = hiddensize proj_size = hiddensize	无限制,建议8对齐				
	部分支持		1 ==	LSTM单元存在	. =		
LSTM	LSTM 扩展以及变 体命名为exLSTM算	int8 float16	proj_size 日即限定0,即同个文持projection功能	目前限定0,即尚不支持projection功能		per-layer/	
LSTM	子,参数项中指明 (extern)的项为	中指明 I项为	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-tayer/ 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	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	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]		
				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。		
			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。		



o	perator	支持情况	输入数据类型	输入	江里语/	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		
C	oncat	部分支持	int8 float16	width]:tensor	height/ 输入的height	据: 4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		
			int8 float16		batch/ 输入的batch	无限制		
	fish			input tensor [batch, channel, height,	channel/ 输入的channel			
	Mish	支持		width]:tensor	height/ 输入的height			
					width/ 输入的width			



					OU/3300 TH O Operator Eist	7周已版记1版	
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			
		float16	width]:tensor	height/ 输入的height	无限制		
Pad	+++			width/ 输入的width	[1,8176]		
rau	支持	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			
ReduceMean	尚不支持 目前由	int8	width]:tensor	height/ 输入的height	[1,8192]		
Reduceiviean	CPU实现	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		



KK3500/3506 N O Operator List 加心脉电力								
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_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	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	支持多batch		
			input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	[1,8192]			
	部分支持			height/ 输入的height				
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 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	
1	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])			



量化支持方式



an anatan	+ H	4A) #L III W. 701	<i>t</i> A)		LL LIT LIS		
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
			E s	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]		
Reverse	尚不支持	int8 float16		width/ 输入的width	[1,8176]		
Sequence	四个文訂		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	支持	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制		
Jighlord	大 打	noarro	width]:tensor	height/ 输入的height	עיף אַאַן אַען אַען אַען אַען אַען אַען אַע		
				width/ 输入的width			



					00/3308 NI O 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
				batch/ 输入的batch			
HardSigmoid	支持	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制		per-layer
	XII	Hoatro	width]:tensor	height/ 输入的height	入L pic 即J		per-rayer
				width/ 输入的width			
	支持		input_tensor [batch, channel, height, width]:tensor	batch/ 输入的batch			
Swish		int8 float16		channel/ 输入的channel	无限制		
	2.19			height/ 输入的height			
				width/ 输入的width			
				batch/ 输入的batch			
HardSwish	支持	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制		per-layer
		iloan o		height/ 输入的height	Zupa ng		
				width/ 输入的width			



						OUI 3300 THE OPERATOR EIST	Ald G. MY. C. J. W. D.	
opera	tor	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
	Softplus				batch/ 输入的batch			
Softn		支持	int8 input_tensor [batch, channel, height, width] tensor width] tensor [batch, channel, height, height, width] tensor [batch, channel, height, heig		per-layer			
Зопр	Ads	X1)	Hourty	width]:tensor	height/ 输入的height	Zu pre ipi		per layer
					width/ 输入的width			
				input_tensor [batch, channel, height, width]:tensor	batch/ 输入的batch	无限制		
					channel/ 输入的channel	[1,8192] 建议4对齐		
Softm	ax	尚不支持,目前由 CPU实现	float16		height/ 输入的height	1		per-layer
					width/ 输入的width			
					axis/ 做softmax的轴	1,即channel方向		



	RK3500/3506 NLC Operator List 如心脉电 1 放仍有限本							
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	7 02 44			
			width]:tensor	height/ 输入的height	无限制			
				width/ 输入的width				
Slice	部分支持	int8 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			
				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	int8 float16		width/ 输入的width			per-layer	
			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			
Tanh	支持	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制		per-layer
Tanh	又行	moatro	width]:tensor	height/ 输入的height	ZL PIC (PU)		per-rayer
				width/ 输入的width			
			input_tensor [batch, channel, height, width]:tensor	batch/ 输入的batch	[1,1024]		
		int8 float16		channel/ 输入的channel	-[1,8192]		
				height/ 输入的height			
Transpose	部分支持			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。		



					700/3300 141 C Operator Elst	7周-6-108-10-1 7次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>		
				num_output/ 输出的channel			
		kernel_shape [num_output, num_input, kernel_h, kernel_w]:int64[]					
				kernel_w/ width方向的kernel大			per-layer/ per-channel
Convolution	支持	int8 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]		
				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/ group的大小	无限制		
			h	dilations_h/ height方向的dilations 大小			
		dilations_w]:int64[]	dilations_w/ widtht方向的dilations	[1, 32]			
				大小			



operator	支持情况	输入数据类型	#\\ \)	设置项/	约束规格	广播支持(维度补齐)	量化支持方式
operator	义持简优	制入	制 人	输入参数含义	约 米, 放 恰	/	里化义 行力入
				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	[1,8]		per-layer/ per-channel
		int8		kernel_w/ width方向的kernel大 小			
Depthwise Convolution	支持	float16	strides [strides_h, strides_w]:int64[]	stride_h/ height方向的strides大 小	[1,7]		
				stride_w/ width方向的strides大 小			
				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			
		l a	1	dilations_h/ height方向的dilations 大小			
		dilations_w]:int64[] dil	dilations_w/ widtht方向的dilations 大小	[1, 32]			



	KK3500/3506 NI O Operator List 加心原屯 1 应							
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	(1.21)			
				kernel_w/ width方向的kernel大 小	[1,31]			
		int8	anida fanida la anida aniliar (fil	stride_h/ height方向的strides大 小				
ConvTranspose/ Deconvolution	支持	float16	strides [strides_h, strides_w]:int64[]	stride_w/ width方向的strides大 小	{2,4,8}		per-layer/ per-channel	
				pads_left/ left方向的pads大小	± ± 50.15			
			pads [pads_top, pads_left,	pads_right/ right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 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, pads_right]:int64[]	pads_top/ top方向的pads大小				
				pads_bottom/ bottom方向的pads大 小				
			group:int64	group/ group的大小	l 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 dilations_w/ widtht方向的dilations 大小	[1, 32]			



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/			量化支持方式 量化支持方式 per-layer/ per-channel
			input_tensor_2 [K,N]:tensor	输入数据的形状			
Comm	T to the CONTROL OF	int8	alpha:double	alpha/ 矩阵A*B乘法的scale			
Gemm	不支持 由CPU实现	moatro	beta:double	beta/ 输入C矩阵的scale	不支持		per-channel
			transA:int64	transA/ A矩阵是否转置			
			transB:int64	transB/ B矩阵是否转置			
				batch/ 输入的batch			
MatMul	不支持 由CPU实现	int8	C/	K/ 输入的K			per-layer/
	TI ZIN HOLOKA			C/ 输入的C	不支持		per-channel
				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
				batch/ 输入的batch			
		int8	input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel			
		float16		height/ 输入的height	无限制		
Expand	支持			width/ 输入的width			
	ZN	int64 sha		batch_o/ 输出的batch_o			
			shape (batch_o, channel_o, height_o,	channel_o/ 输出的channel	无限制		
			width_o):tensor	height_o/ 输出的height			
				width_o/ 输出的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
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						



					T	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 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
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 Co	onvolution				
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						



				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
Add+Relu	支持	同Add				
Mul+Relu	支持	同Mul				
Convolution + add + Relu	支持	同Convolution				
·						

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. 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



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 channel/ 输入的channel		支持ONNX规范的四维tensor的所有广播操作,以ONNX 默认排列NCHW做说明 , 支 持 以 下 广 播 方 式 :		
Add/Bias	支持	int8 float16	input_tensor [batch, channel, height, width]:tensor	和人的channel height/ 输入的height	无限 即	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。 例子见 <u>注释(1)</u>	per-layer/ per-channel	已支持
				width/ 输入的width				
				batch/ 输入的batch				
Sub	支持	int8 float16	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/	尚不支持
			[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。	per-channel	
				width/ 输入的width				
				batch/ 输入的batch				
Marl/Carla	+- 4-	int8	input_tensor	channel/ 输入的channel		支持ONNX规范的四维tensor的所有广播操作,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/	N T + H
Mul/Scale	支持	float16	[batch, channel, height, width]:tensor	height/ 输入的height		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	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				
Div		float16	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/	多核协协同运况 尚不支持 尚不支持
DIV	部分支持	Hoatro	[batch, channel, height, width]:tensor	height/ 输入的height	无限制	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FP16类型 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	无限制			
Max	支持	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel	[1,8192]	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/	尚不支持
			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。	per-channel	
				width/ 输入的width	[1,8176]			
				batch/ 输入的batch	无限制			
Min	支持	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel	[1,8192]	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	尚不支持
	Z19		width]:tensor	height/ 输入的height	[]			
				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		量化支持方式	多核协同运 行支持情况
				batch/ 输入的batch	1			
Global		int8 float16	input tensor [batch, channel, height,	channel/ 输入的channel	[1,8192]		per-layer	尚不支持
Global AveragePool 支持	文抒	noatro	input_tensor [batch, channel, height, width]:tensor	height/ 输入的height	[1 242] (4.4][(0 本株英国)		per-iayer	
				width/ 输入的width	[1,343](toolkit2支持范围)			
				batch/ 输入的batch	1			
GlobalMaxPool	支持	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel	[1,8192]			尚不支持
GIOGALIVIANI OUI	太打	HORELO	width]:tensor	height/ 输入的height	[1,343] (toolkit2支持范围)		per-layer	网小又特
				width/ 输入的width	[1,973] (1001K112X1978III)			



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			113213111392
			input_tensor [batch, channel, height,	channel/ 输入的channel	[1,8192]		per-layer ្រិ	
			width]:tensor	height/ 输入的height	[1,8192]			
		int8 float16		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			
			count_include_pad:int64	count_include_pad/ 是否包含pad数值进行 计算	1			
AveragePool	支持		int8 float16 kernel_shape [kernel_h, kernel_w]:int64[]	kernel_h/ height方向的kernel大 小	一无限制,NPU支持[1,7];其它由CPU支持。			尚不支持
				kernel_w/ width方向的kernel大 小				
				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	[6,7]			
				pads_bottom/ bottom方向的pads大小				
			stride heigh 小	stride_h/ height方向的strides大 小				
		Sti	strides [strides_h, strides_w]:int64[]	stride_w/ width方向的strides大 小	[1,8]			



					14455500 141 & Operator Elst				
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]				
			width]:tensor	height/ 输入的height	[1,8192]				
				width/ 输入的width	[1,0172]				
			auto_pad:string	auto_pad/ pad的方式	仅支持NOTSET				
		int8 float16	ceil_mode:int64 dilations [dilations_h, dilations_w]:int64[]	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 大小	1		per-layer		
				dilations_w/ widtht方向的dilations 大小					
MaxPool	支持			kernel_h/ height方向的kernel大 小	- 无限制,NPU支持[1,7];其它由CPU支持。			尚不支持	
				kernel_w/ width方向的kernel大 小	Дикир, M 0 X 19 [1,7]; ЖЕШСІ 0 X 19 .				
				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				
		strides [strides h, strides w]:int64	strides [strides_h, strides_w]:int64[]	stride_h/ height方向的strides大 小	- [1,8]				
			stride_w/ width方向的strides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	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		per-layer/ per-channel		
Batch Normalization	支持	int8 float16		batch/ 输入的batch	1			尚不支持	
			input tensor [batch, channel, height,	channel/ 输入的channel	- mr.dad				
			width]:tensor	height/ 输入的height	无限制				
				width/ 输入的width	无限制				
			batch/ 输入的batc	batch/ 输入的batch	支持多batch				
			input_tensor [batch, channel, height,	channel/ 输入的channel	无限制				
			width]:tensor	height/ 输入的height					
				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	即从文刊	riout 10	width]:tensor(const)	width/ 输入的width	等于input_width		per layer	阿小文羽	
			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_aff eps:double	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				
Clip/ReLU6	-t-let	int8 float16	input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel				
Clip/ReLU6	支持	iloat16		height/ 输入的height	无限制		per-layer	已支持
				width/ 输入的width				
				batch/ 输入的batch				
Elu		int8 float16	input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	T III del			N T + 4
Eiu	支持			height/ 输入的height	无限制			尚不支持
				width/ 输入的width				
				batch/ 输入的batch				
Gelu	支持	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制			尚不支持
Gelu	文 打	inoutiv	width]:tensor	height/ 输入的height				HAN, N. M.
			wic 输,	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	支持	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制			已支持
Keiu	又行	noatto	width]:tensor	height/ 输入的height	プロドマ 市リ		per-layer	C 又 持
				width/ 输入的width				
				batch/ 输入的batch				
LeakyRelu	支持	int8 float16	input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	无限制		per-layer	已支持
LeakyReiu	文科			height/ 输入的height				CXH
				width/ 输入的width				
				batch/ 输入的batch				
				channel/ 输入的channel	无限制			
PRelu 3	支持	int8 float16	input_tensor [batch, channel, height, width]:tensor	height/ 输入的height	ZUNRI		per-layer/ per-channel	已支持
			新 —	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				
		direct	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	1				
	部分支持 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 参数项。	指定 sequence_size 指定 seqsi: int64 (extern) 指定 seqsi: int64 (extern) hidde fix fix float 16 hidden_size:int64 (extern) float	float16 hidden_		sequence_size/ 指定GRU输入的 seqsize	无限制,建议4对齐		per-layer	
GRU				hidden_size:int64 (extern)	hidden_size/ GRU单元中的 hiddensize	无限制,建议8对齐			尚不支持
					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)e: 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/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]					
			Output_tayout.suring (extern) A 应相Ushape 5 文 致的layout 要 seq	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												
		寺 - 展以及 5 第子, - 申指明 -) か独有 - 頭。	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	po	per-channel										
LSTM	LSTM 扩展以及 变体命名为 exLSTM算子,		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			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]										
	的参数项。		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。												
						l			1	7	<u>-</u> !	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、sne: 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(ex	output_layout/指定与 output_layout/指定与 对应输出shape含义一 致的layout 要求 seque	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。		1										



ope	rator	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	channel方向concat时,除了最后一个输入外,其他输入的channel大小一需要对齐。对齐量:8bit数据:16对齐,16bit数据:8对齐。 其他方向Concat无限制。			
		部分支持		input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel				
Con	Concat $\frac{q}{r}$		int8 float16		height/ 输入的height			per-layer	已支持
					width/ 输入的width				
					aixs/ 拼接的维度	无限制			
					batch/ 输入的batch				
Mis	Mish 支持	支持	int8 float16	input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	无限制			尚不支持
					height/ 输入的height				
				wic 输,	width/ 输入的width				



			ACCOUNT C 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	多核协同运 行支持情况	
				batch/ 输入的batch	1				
		int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel	一 无限制				
		поатто	width]:tensor	height/ 输入的height					
Pad	支持			width/ 输入的width	[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						
		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					
ReduceMean	尚不支持 目前 由CPU实现	int8 float16	width]:tensor	height/ 输入的height	[1,8192]			尚不支持	
ReduceMean É	由CPU实现	iioatio		width/ 输入的width				网尔文哲	
			axes:int64[]	axes/ 指定reduce的轴	单轴:无限制, 多轴:{2、3}				
	keepdims/ keepdims:int64[] 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	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				
ReduceSum	尚不支持 目前	int8 float16	width]:tensor	height/ 输入的height	[1,8192]		per-layer/	尚不支持
ReduceSum	由CPU实现	noacro		width/ 输入的width			per-channel	同小义材
			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			
		ir		batch/ 输入的batch	支持多batch			
			input_tensor [batch, channel, height,	channel/ 输入的channel	- [1,8192]			
			width]:tensor	height/ 输入的height	[1,0172]			
	如八士林			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			
D .:	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。	PU仅支持 r向不超过 整倍数的 近插值缩 余不支持 分的会 ck到CPU 实现。	mode:string	mode/resize采用的模式	仅支持nearest			N. T. Lebe
Kesize			scales:int64[]	scales/尺寸放大倍数	仅支持1-8整数倍		per-layer	尚不支持
			roi:int64[]	roi/进行resize的输入范围	仅支持全局([0,0,0,0,1,1,1,1])			



				设置项/			1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 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 float16	input_tensor [batch, channel, height,	channel/ 输入的channel				
			width j: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		î	
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			
			为!	[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			



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, width]:tensor	channel/ 输入的channel								
Reverse Sequence	尚不支持			height/ 输入的height	- [1,8192]							
				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							
				t:	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								
a	1.15	int8	input tensor [batch, channel, height,	channel/ 输入的channel				Mer to the				
Sigmoid	支持	float16	width J.tensor heigh 输入	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				
H 161	_t_t	int8	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	尚不支持
				width/ 输入的width				
			batch/ 输入的batch					
Swish	支持	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel	- 无限制			尚不支持
SWISH	XH	Hoatro	width]:tensor	height/ 输入的height	入上 PRC 中 切			同小文符
				width/ 输入的width				
				batch/ 输入的batch				
HardSwish	int8 HardSwish 支持 float16	int8	input_tensor [batch, channel, height,	channel/ 输入的channel	- 无限制		per-layer	尚不支持
a ratus wish		induit)	width]:tensor	height/ 输入的height) ZL pre ipg		per rayer	HAVI-X14
			w ¥	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				
Softplus	支持	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制		per-layer	尚不支持
Sortplas	ZN	nout	width]:tensor	height/ 输入的height	ACPICIPAL TO A STATE OF THE STA		per layer	1971-X19
			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					
			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	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				
			width]:tensor	height/ 输入的height	无限制			
				width/ 输入的width				
Slice	int8 部分支持 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	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[]		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	无 陳 即			
Split		int8 float16		width/ 输入的width			per-layer	尚不支持
			axis:int64	axis/ 切分的维度	无限制			
		sp	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				
Tanh	士柱	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel	工門也		per lover	尚不支持
Taiiii	支持	noatro	width]:tensor	height/ 输入的height	的h/height	per-rayer	同个文符	
				width/ 输入的width				
				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	[1,0122]			
Transpose	部分支持	int8 float16		width/ 输入的width	[1,8176]			尚不支持
Truispose	IIP/J X IV	Touri	perm:int64[]	axis order/ 转置的轴顺序	RK3588支持所有RK3566/3568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<512: 16bit 数据时,h1*w1<8176,w1*c1<512: 16bit 数据时,h1*w1<8176,w1*c1<512: 16bit 数据时,h1*w1<8176。(2)perm=[0,3,1,2]时,h1*w1<8176。(3)perm=[0,3,2,1]时,h1*w1<1024。			INTERNATION OF THE PROPERTY OF



					100 300 10 O Operator Enst		11 100 11 100 2		
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, kernel_h,	num_input/ 输入的channel	无限制				
		r l		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 float16		stride_h/ height方向的strides大 小			per-layer/	已支持	
			strides [strides_h, strides_w]:int64[]	stride_w/ width方向的strides大 小	per-layer/ per-channel	per-channel			
				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]				
	dil		pads_bottom/ bottom方向的pads大小						
		group:int64	group/ group的大小	无限制					
		dilations [dilations_h, 大/dilations_w]:int64[] dilations_wice	dilations_h/ height方向的dilations 大小	H 20					
			dilations_w/ widtht方向的dilations 大小	[1, 32]					



					KK3388 NI O 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	多核协同运 行支持情况
				batch/ 输入的batch	无限制			
	n 支持 int8 float16	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	无限制			
			num_input, kernel_h, kernel_w]:int64[]	kernel_h/ height方向的kernel大 小	EL 01			
Depthwise	本体			kernel_w/ width方向的kernel大 小		per-layer/	已支持	
Convolution	XIV			stride_h/ height方向的strides大 小			per-chamier	LXN
			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	[0,13]			
				pads_bottom/ bottom方向的pads大小				
		dilations [dilations_h,	dilations_h/ height方向的dilations 大小	rt 207				
			dilations_w]:int64[]	dilations_w/ widtht方向的dilations 大小	[1, 32]			



					1005500 141 C Operator Elst					
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	无限制					
			kernel_shape [num_output,	num_input/ 输入的channel	无限制					
			1		num_input, kernel_h, kernel_w]:int64[]	kernel_h/ height方向的kernel大 小	nel大 [1,31]			
				kernel_w/ width方向的kernel大 小	[[1,31]	per-layer/ per-channel				
				stride_h/ height方向的strides大 小						
ConvTranspose/ Deconvolution	支持	int8 float16	strides [strides_h, strides_w]:int64[]	stride_w/ width方向的strides大 小	{2,4,8}			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					
			dil he	dilations_h/ height方向的dilations 大小	H 221					
		dilations_w]:int64[] dil	dilations_w/ widtht方向的dilations 大小	[1, 32]						



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/	转为Matmul实现,约束同Matmul			
			input_tensor_2 [K,N]:tensor	输入数据的形状	投 / ywaimui 失兆,约来问Maimui			
Carren	尚不支持 目前	int8	alpha:double	alpha/ 矩阵A*B乘法的scale	无限制		per-layer/	WITHE
Gemm	由CPU实现		beta:double	beta/ 输入C矩阵的scale	无限制		per-channel	尚不支持
			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	汉府念tensor又存转直			
				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			
W 24 1	部分支持 目前该支持仅针 对双feature输入	int8 float16	input_tensor_1 [batch, K, C]:tensor	K/ 输入的K	- C X 付 2 2,19384 J 、		per-layer/	Mer lak
MatMul	未来将支持输入 为 feature+constant	float16		C/ 输入的C	feature+constant时: 若input_tensor_1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
	reature (constant		input_tensor_2 [batch, C, H]:tensor	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)	MatMul 対双feature输入 4d)	int8 float16	input_tensor_2 [batch, channel, N,	K/ 输入的K	K*M <=65532 M*N <=65532 feature+constant ^[h] :		per-layer/ per-channel	尚不支持
		M]:tensor N/	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			



					300 TH C Operator Elst		
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			
		int8 float16					
		noatio	width]:tensor	height/ 输入的height	无限制		
Expand				width/ 输入的width			
		S		batch_o/ 输出的batch_o			
			shape (batch_o, channel_o, height_o, width_o):tensor	channel_o/ 输出的channel	无限制		
			widin_o):tensor	height_o/ 输出的height			
				width_o/ 输出的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
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



			RK3566 Ni U Operator List 如心原它1成仍有核本可					
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	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						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	ivolution					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



					RR5500 141 O Operator Elst		和心脉电 7 放历 6 成2	7 -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					
Add+Relu	支持	同Add	dd										
Mul+Relu	支持	同Mul											
Convolution+ Add+Relu	支持	同Convolution											
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)): OW),B(C,1,1)): OP(AW),B(scalar)): OP(AW),B(H,W)): OP(AW)OP(AW),B(H,W)): OP(AW)OP(AW),B(H,W)): OP(AW)OP(AW),B(H,W)): OP(AW)OP(AW),B(H,W)	A(1,16,32 8),B(16)) A(1,16,32,8),B(1))= (1,16,32,8),B(32x8 转换成除数倒数自	=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 RV1103/1106 NPU Operator List



	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
		2000		input_tensor [batch, channel, height, width]:tensor	输入参数含义 batch/ 输入的batch	-		
	Add/Bias	支持	int8 float16		channel/ 輸入的channel height/ 输入的height	- 无限制		per-layer/ per-channel
					width/ 输入的width			
					batch/ 输入的batch			
	Sub		int8 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 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
					height/ 输入的height width/			
					输入的width			
					batch/ 输入的batch		士柱ONNV相类的III徐	
	Mul/Scale	支持	int8 float16	input_tensor	channel/ 输入的channel	无限制	支持ONNX规范的四维tensor的所有广播操作,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。 例子见 <u>注释(1)</u>	per-layer/
	wan seale	又行	noatro		height/ 输入的height	June 192		per-channel
					width/ 输入的width			



				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		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	
Div	部分支持	float16	input_tensor	channel/ 输入的channel	无限制		per-layer/
Div	祁	Hoatro	[batch, channel, height, width]:tensor	height/ 输入的height	入L PRE 即1	 OP((N,C,H,W),scalar),即以单个标量做broadcasting 说明: A或良都可以作为广播方。 例子见<u>注释(1)</u> 	per-channel
				width/ 输入的width			
	暂不支持	int8 float16	input_tensor [batch, channel, height, width]:tensor	batch/ 输入的batch	1		
Max				channel/ 输入的channel	[1,8192]		per-layer/
				height/ 输入的height			per-channel
				width/ 输入的width	[1,8176]		
				batch/ 输入的batch	1		
Min	暂不支持	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel	[1,8192]	2T 399.4/F	per-layer/
Min	1 1 2.79	Iloat16		height/ 输入的height	E / · · · J		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
				batch/ 输入的batch	1		
Global AveragePool	支持	int8 input_tensor [batch, channel, height, width]:tensor		per-layer			
	X.N	int8		height/ 输入的height	[1,343] (toolkit2支持范围)	per kayer	
				width/ 输入的width	[1,343] (tootkite) Zinyasiin		
				batch/ 输入的batch	1		
GlobalMaxPool	古株	ints	input_tensor [batch, channel, height,	channel/ 输入的channel	[1,8192]		per-layer
GlobalMaxPool	, ДИ	int8 input_tensor [batch, channel, height, width]:tensor height/输入的height width/ 输入的width/ 输入的width/		per rayer			
				width/ 输入的width	[1,545](TOO1KIT2文/行犯問)		



	KVIIO/IIIO NI O Operator Est.							
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				
				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	ad/ 值进行			
AveragePool	支持	int8	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	- 无限制, NPU支持[1,8]; 其它由CPU支持。			
				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	[0.7]			
			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[]	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		
				channel/ 输入的channel	[1,8192]		
			width]:tensor	height/ 输入的height	[1,9102]		
				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 [dilations_h,	dilations_h/ height方向的dilations大 小			
			dilations_w]:int64[]	dilations_w/ widtht方向的dilations大 小	ns大		
MaxPool	支持	int8	kernel_shape [kernel_h, 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。		per-layer
				kernel_w/ width方向的kernel大小	7517477		
				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	[0.7]		
			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		
		stri	otnidos Istnidos la strida- valvinté 453	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
			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	-t-kt	int8 float16		batch/ 输入的batch	1		per-layer/
Normalization	支持	noatro	input tensor [batch, channel, height,	channel/ 输入的channel	T. 1711 (fed		per-channel
			width]:tensor	height/ 输入的height	- 无限制		
				width/ 输入的width	无限制		
				batch/ 输入的batch	支持多batch		
			input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	无限制		
				height/ 输入的height			
				width/ 输入的width			
			layernorm weight [channel, height,	channel/ 输入的channel	等于input_channel		
Layer	North Life	g . 16	width]:tensor(const) layernorm bias [channel, height,	height/ 输入的height	等于input_height		
Normalization	尚不支持	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,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。	pias	
			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
				batch/ 输入的batch	1		
Clip/ReLU6	支持	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel			per-layer
	又行	noatro	width]:tensor	height/ 输入的height	无限制		per-iayer
				width/ 输入的width			
	支持	int8 float16	input_tensor [batch, channel, height, width]:tensor	batch/ 输入的batch			
Elu				channel/ 输入的channel	无限制		
Biu				height/ 输入的height			
				width/ 输入的width			
				batch/ 输入的batch			
Gelu	支持	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制		
Seid	AN		width]:tensor	height/ 输入的height			
				width/ 输入的width			



					TITOTALE OPERATOR BISE		
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		
p.,	de let	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel			
Relu	支持	Hoatro	width]:tensor	height/ 输入的height	无限制		per-layer
				width/ 输入的width			
				batch/ 输入的batch	1		
LeakyRelu	支持	int8 float16	input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	无限制		per-layer
				height/ 输入的height			
				width/ 输入的width			
				batch/ 输入的batch	1		
			input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	[1,8192]		
PRelu	支持	int8 float16		height/ 输入的height			per-layer/ per-channel
				width/ 输入的width	[1,8176]		
				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			
			<pre>input_tensor [sequence, batch, input_size] :tensor</pre>	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			
			batch_size:int64 (extern)	batch_size/ 指定GRU输入的 batchsize	1			
	尚不支持 GRU 扩展以及变体命 名为exGRU算子,参 数项中指明(extern)		sequence_size :int64 (extern)	sequence_size/ 指定GRU输入的seqsize	无限制,建议4对齐		per-layer	
GRU		子,参 extern) 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	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	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
	部分支持 LSTM 扩展以及变体		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	_	
LSTM	命名为exLSTM算子, 参数项中指明 (extern)的项为		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	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	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/指 应输出shape含 的layout	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			
				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:16对齐,16bit数据:8对齐。		
Concat		部分支持	int8 float16	width]:tensor	height/ 输入的height	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	无限制		
Mish		支持	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel			
	Mish	~19	float16	width]:tensor	height/ 输入的height			
				y y	width/ 输入的width			



operato	or	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 float16	input_tensor [batch, channel, height,	channel/ 输入的channel	工匠机		
			noatro	width]:tensor	height/ 输入的height	一 无限制		
					width/ 输入的width	[1,8176]		
Pad		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		
			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]		
Reduce	eMean	尚不支持	int8 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	
ReduceSum				batch/ 输入的batch	无限制			
				input_tensor [batch, channel, height,	channel/ 输入的channel			
	尚不支持	int8 float16	width]:tensor	height/ 输入的height	[1,8192]		per-layer/	
reduces	PATEN			width/ 输入的width			per-channel	
			axes:int64[]	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	约束规格: (1) height * width * type_bytes <= 130816; (2) input_tensor非四维时,shape无限制			
		int8 float16	input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel				
		Trout o	(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; 约束规格:			
		int64 w	width_o):tensor (输出shape指定维度为4维时看作nchw)	height_o/ 输出的height	(1) height_o * width_o * type_bytes <= 65535; (2) Align(height o * width o, alignment) <= 8192;			
			w	width_o/ 输出的width	(3)输出shape非四维时,shape无限制			



_ <u></u>					TITOTALE OPERATOR BISE		111111111111111111111111111111111111111						
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								
		input_tensor [batch, channel, height,	input tensor [batch, channel, height,	channel/ 输入的channel	1								
	部分支持 目前NPU仅支持宽高		width]:tensor	height/ 输入的height	[1,8192]								
Resize	方向不超过8倍的整倍 数的最邻近插值缩	int8 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								
		int8 float16						channel/ 输入的channel	[1,8192]				
								width]:tensor	height/ 输入的height	[1,0172]			
Reverse Sequence	尚不支持				width/ 输入的width	[1,8176]							
					-				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									
S:i.l	+-44:	int8	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制								
Sigmoid	支持	float16	width]:tensor	height/ 输入的height									
				width/ 输入的width									



					71100 141 C Operator Eist		
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				batch/ 输入的batch			
	支持	int8 float16	input tensor [batch, channel, height,	channel/ 输入的channel	无限制		per-layer
matusiginoid	XN	Hoatro	width]:tensor	height/ 输入的height	A. Pic (P)		per-rayer
				width/ 输入的width			
		batch/ 输入的batch					
Swish	支持	int8 float16	input tensor [batch, channel, height,	channel/ 输入的channel	无限制		
SWISH Z.	A19		width]:tensor	height/ 输入的height			
				width/ 输入的width			



						7 TTOO TATE OF PERMICE EAST		
operator	支持情况	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			
HardSwish	+++		int8 float16		channel/ 输入的channel	无限制		per-layer
HardSwish	支持		HOAT16		height/ 输入的height			per-layer
					width/ 输入的width			
					batch/ 输入的batch			
Softplus	+++		int8 float16	input tensor [batch, channel, height,	channel/ 输入的channel			mon lovon
Sortplus	支持	width]:tensor		per-layer				
					width/ 输入的width			
		batch/ 输入的batch 无限制	无限制					
				input_tensor [batch, channel, height,	channel/ 输入的channel	硬件支持[1,4096] 建议8对齐		
Softmax	部分支持	持	float16	width]:tensor	height/ 输入的height			per-layer
					width/ 输入的width	axis=1, 无限制 axis=3/-1,width[1, 8192], height无限制		
			ε		axis/ 做softmax的轴	1,即channel方向		



					71100 111 C Operator Bist		7 11 10 4 1 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			
			input_tensor [batch, channel, height,	channel/ 输入的channel	7 10 4-1		
			width]:tensor	height/ 输入的height	无限制		
				width/ 输入的width			
Slice	部分支持	int8 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/ 对齐量:	8bit数据: 16对齐, 16bit数据: 8对齐。					
			axes:int64[]	axes/ 选取切分的轴		_	
			steps:int64[]		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	入L PKC 即9		
Split	部分支持	int8 float16		width/ 输入的width			per-layer
			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
Tanh				batch/ 输入的batch			
	支持	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制		per-layer
1 aiii	文 村	noatro	width]:tensor	height/ 输入的height	ZUNEJ		per-rayer
		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]		
		width]:tensor			[1,0192]		
Transpose	部分支持	int8 float16	perm:int64[]	axis order/ 转置的轴顺序	RV1106、RV1103支持所有RK3566/3568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均要满足移bt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<8176。(3) perm=[0,3,2,1]时,h1*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		
				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	-[1,31]		
			kernel_w/ width方向的kernel大小				
Convolution	支持	int8	strides [strides_h, strides_w]:int64[]	stride_h/ height方向的strides大 小	[1,7]		per-layer/ per-channel
			saides [suides_ii, suides_w].into-[]	stride_w/ width方向的strides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_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		
		dilatio height dilations [dilations h, dilations_w]:int64[] dilations_w]		-[1, 32]			
			dilations [dilations_h, dilations_w]:int64[] dilat widt	dilations_w/ widtht方向的dilations大 小			



				71100 NI O Operator Eist	和心族七 1 及[
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	pht方向的kernel大小 el w/		
. Le let	int8		kernel_w/ width方向的kernel大小			per-layer/ per-channel
支持	stride_h/ height方向的strides大 strides [strides_h, strides_w]:int64[] stride_w/ width方向的strides大小		per-channel			
		stride w/				
			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	10.17		
		pads_right]:int64[]	pads_top/ top方向的pads大小	[[0,15]		
			pads_bottom/ bottom方向的pads大小			
		dilations [dilations h,	dilations_h/ height方向的dilations大 小			
		dilations_w]:int64[] dil	dilations_w/ widtht方向的dilations大 小			
	支持情况	支持	input_shape [batch, channel, height, width]:tensor kernel_shape [num_output, num_input, kernel_h, kernel_w]:int64[] int8	input shape [batch, channel, height, width]:tensor input shape [batch, channel, height, 物入的channel] height/物入的height width/物入的width num_output/物记的channel kernel_h, kernel_w]:int64[] int8 i	A	



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, 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	无限制		
		kernel_h, kernel_w]:int64[] kernel_h/ height方向的kernel大小					
ConvTranspose/ Deconvolution	支持		[2.8]		per-layer/ per-channel		
				stride_w/ width方向的strides大小			
			tetous				
			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: - 不支持 kernel_h * dilations_h - dilations_h - pads_top < 0		
			pads_right]:int64[]	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	7 × 5 state_s (state s) pass_text s compa_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_h/ height方向的dilations大 小	[1, 32]		
			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
			input_tensor_1 [M, K]:tensor	M,K,N/	than one was series		
			input_tensor_2 [K,N]:tensor	输入数据的形状	转为Matmul实现,约束同Matmul		
	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/
Gemm	实现		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			
	batch/ 输入的batch 双feature时: batch、H无限制 K之持[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:			batch/ 输入的batch			
Madel			per-layer/				
	实现		input tensor 2 [batch, C, H]:tensor	C/ 输入的C	若input_tensor_1为feature,则转为batch个feature[K,C,1,1] + weight[H,C,1,1]的conv; 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		per-channel
			input_tensor_2 [batch, C, H]:tensor 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
				batch/ 输入的batch			
		int8	input_tensor [batch, channel, height,	channel/ 输入的channel			
		width]:tensor height/ 输入的height width/ 输入的width/					
Expand	支持			width/ 输入的width			
	~.,	int64		batch_o/ 输出的batch_o	无限制		
			shape (batch o, channel o, height o,	channel_o/ 输出的channel			
			width_o):tensor	height_o/ 输出的height			
				width_o/ 输出的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
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					
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	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	volution				
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						



KVIIO5/IIO0 IV O Operator Est.							7 17 100
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					
Convolution+Add+Relu	支持	同Convolution					

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、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。



第四章 RK3562 NPU Operator List



						302 NI O Operator List	圳心城屯;双边	
	pperator	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/Bias	支持	int8 float16	input_tensor	channel/ 输入的channel	无限制	支持ONNX规范的四维tensor的所有广播操作,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/
				[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 4.OP(A(N,C,H,W),B(H,W)),即HW维度做broadcasting 说明:A或B都可以作为广播方。例子见注释(1)	per-channel
-					width/ 输入的width		列丁	
					batch/ 输入的batch			
	Sub	支持	int8 float16	input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	一 无限制		per-layer/
					height/ 输入的height			per-channel
-					width/ 输入的width			
					batch/ 输入的batch			
	Mul/Scale		int8	input tensor	channel/ 输入的channel		支持ONNX规范的四维tensor的所有广播操作,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/
		支持	float16	height/ width/	height/ 输入的height	无限制	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 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		P34 Zellari SAZ	



					502 NI O 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
				batch/ 输入的batch			
Div	部分支持	float16	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 2、OP(A(N,C,H,W),B(C,1,1)),即C维度做broadcasting	per-layer/
	四八 义打	Hoatro	[batch, channel, height, width]:tensor	height/ 输入的height	ALIVE (PI)	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FP16类型 说明: A或B都可以作为广播方。 例子见注释(1)	per-channel
				width/ 输入的width			
	暂不支持			batch/ 输入的batch	1		
Max			innut tensor hatch channel height	channel/ 输入的channel	- [1,8192]	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/
	1 1 2.39			height/ 输入的height			per-channel
				width/ 输入的width	[1,8176]		
				batch/ 输入的batch	1		
Min	暂不支持	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel	[1,8192]	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 : 1.0P(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/
		N文字 noat16 width]:tensor h	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。	per-channel	
				width/ 输入的width	[1,8176]		



						1302 NI O Operator Eist		
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		支持	int8	input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	[1,8192]		per-layer
AveragePoo	ol	X 17	into		height/ 输入的height			per-rayer
					width/ 输入的width	[1,343] (toolkit2支持范围)		
		batch/ 输入的batch						
GlobalMaxI	Pool	支持	int8 float16	input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel [1,8192]		per-layer	
GlobalMaxPool	12.19	float16		height/ 输入的height	[1,343] (toolkit2支持范围)		ры шуы	
				width/ 输入的width	[1,545](UUIKI12又可记四)			



					5502 IVI O Operator List	河心吸电 1 双[
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			
				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	1		
AveragePool	支持	int8	kernel_shape [kernel_h, kernel_w]:int64[]	kernel_h/ height方向的kernel大小	、 - 无限制,NPU支持[1,8]: 其它由CPU支持。		per-layer
				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	-[0,7]		
			pads_right]:int64[]	pads_top/ top方向的pads大小	[0,7]		
				pads_bottom/ bottom方向的pads大小			
		strides	stridas lotridas la otridas valiatéAE	stride_h/ height方向的strides大 小	[1,8]		
			strides [strides_h, strides_w]:int64[]	stride_w/ width方向的strides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
				输入参数含义 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]		
			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_w]:int64[] dilations_w/ width方向的dilations大 小					
MaxPool	支持		kernel_shape [kernel_h, kernel_w]:int64[]		· 天限制,NPU支持[17],其它由CPU支持。		per-layer
				kernel_w/ width方向的kernel大小	ZERGRAS MIOZNIIS, ALIBOTOZNI		
				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	[0,7]		
			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		
			anida facida la anida antino 473	stride_h/ height方向的strides大 小			
		S	strides [strides_h, strides_w]:int64[]	stride_w/ width方向的strides大小	[1,8]		



					202 11 C Operator Elst		
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		
			momentum:double	momentum/ 训练时的滑动平均参数	无限制		
Batch	支持	int8 float16		batch/ 输入的batch	1		per-layer/
Normalization	又14	noatro	input tensor [batch, channel, height,	channel/ 输入的channel	7 112 4-1		per-channel
			width]:tensor	height/ 输入的height	- 无限制		
	width/ 输入的width 无限制						
		batch/ 输入的batch channel/ 输入的channel height/ 输入的height width/ 输入的width layernorm weight [channel, height,] hatch/ 输入的channel batch/ 输入的channel channel/ 输入的width/ 输入的width should specified and specif			支持多batch		
			input tensor [batch, channel, height,				
				height/ 输入的height	无限制		
Layer	+- 4+	G-416	width]:tensor(const) layernorm bias [channel, height,	height/ 输入的height	等于input_height		
Normalization	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, 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		



	2447	松) 松田米町	44.)	设置项/ 输入参数含义	Marte In Ma	- IN C 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2 1/2	
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	1		
Clip/ReLU6	支持	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel			per-layer
Спр/кедоо	XH	Hoatro	width]:tensor	height/ 输入的height	无限制		per-rayer
				width/ 输入的width			
				batch/ 输入的batch			
Elu	支持	int8 float16	input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	- 无限制		
				height/ 输入的height			
				width/ 输入的width			
				batch/ 输入的batch			
Gelu	支持	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制		
				height/ 输入的height			
				width/ 输入的width			



			3 13 17 17 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
				batch/ 输入的batch	1		
Relu	支持	int8 float16	input tensor [batch, channel, height,	channel/ 输入的channel			per-layer
Refu	文 村	noatro	width]:tensor	height/ 输入的height	无限制		per-layer
				width/ 输入的width			
				batch/ 输入的batch	1		
LeakyRelu	支持	int8 float16	input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	无限制		per-layer
LeakyKeiu	X17			height/ 输入的height			per-layer
				width/ 输入的width			
				batch/ 输入的batch	1		
				channel/ 输入的channel			
PRelu	支持	int8 float16	input_tensor [batch, channel, height, width]:tensor	height/ 输入的height	无限制		per-layer/ per-channel
				width/ 输入的width			
				slope/ PReLU系数	仅支持单个标量或C维度系数		



		RRESOLITIO Operator Est							
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				
			<pre>input_tensor [sequence, batch, input_size] :tensor</pre>	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				
		扩展以及变体命 exGRU算子,参 中指明(extern) 为exGRU独有的 项。	sequence_size :int64 (extern)	sequence_size/ 指定GRU输入的seqsize	无限制,建议4对齐				
GRU			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,l,l] 要求填写指定的layout, 同时要求填写该op实际对应的batch size、				
					要求填与指定的layout,回时要求填与该op头标对应的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	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
	部分支持 LSTM 扩展以及变体		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		
LSTM	参数项中指明 (extern)的项为		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 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]		
				Hayout	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	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]		
				ή/11	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。		



						302 IVI O Operator List	和心脉毛 1 从 1	
	pperator	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:16对齐,16bit数据:8对齐。		
	Concat	部分支持	int8 float16	width]:tensor	height/ 输入的height	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		
			int8 float16	input_tensor [batch, channel, height, width]:tensor	batch/ 输入的batch	无限制		
	Mish				channel/ 输入的channel			
	IVIISII				height/ 输入的height			
					width/ 输入的width			





		RK3502 IV C Operator List								
•	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		
					batch/ 输入的batch	1				
			int8		channel/ 输入的channel					
			float16	width]:tensor	height/ 输入的height	无限制				
					width/ 输入的width	[1,8176]				
1	² ad	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				
			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					
					height/ 输入的height	[1,8192]				
]	ReduceMean	尚不支持	int8 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				



			TREESON THE CONTINUE OF THE PROPERTY OF THE PR								
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_tensor [batch, channel, height,	channel/ 输入的channel							
ReduceSum	WITH	int8 float16	width]:tensor	height/ 输入的height	[1,8192]		per-layer/ per-channel				
ReduceSum	尚不支持	noatro		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						
		int8 float16	input tensor [batch, channel, height, width]:tensor	batch/ 输入的batch	约束规格: (1) height * width * type_bytes <= 8192*8192*16; (2) input_tensor非四维时,shape无限制						
				channel/ 输入的channel							
				height/ 输入的height							
				width/ 输入的width							
Reshape	支持			batch_o/ 输出的batch_o							
		int64	shape (batch_o, channel_o, height_o, width_o):tensor (输出shape指定维度为4维时看作nchw)	channel_o/ 输出的channel	计算量: alignment=16/type_bytes;						
				height_o/ 输出的height	约束规格: (1) height_o * width_o * type_bytes <= INT32_MAX; (2) Align(height_o * width_o, alignment) <= 8192*8192; (3) 输出shape非四维时,shape无限制						
				width_o/ 输出的width							



					302 111 C Operator Elst			
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 0102]			
	部分支持 目前NPU仅支持宽高		width]:tensor	height/ 输入的height	[1,8192]			
Resize	方向不超过8倍的整倍 数的最邻近插值缩	int8 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	支持nearest/linear			
			scales:int64[]	scales/尺寸放大倍数	H scale * W scale <= 64			
			roi:int64[]	roi/进行resize的输入范围	仅支持全局([0,0,0,0,1,1,1,1])			
		int8 float16		batch/ 输入的batch	无限制			
			input_tensor [batch, channel, height, width]:tensor	channel/ 输入的channel	[1,0103]			
				height/ 输入的height	[1,8192]			
Reverse Sequence	尚不支持			width/ 输入的width	[1,8176]			
			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				
C:i.l	北 杜	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制			
Sigmoid	支持		width]:tensor	height/ 输入的height	7.53,7702			
				width/ 输入的width				



	Table 2.1.4 o b p.mor. Elec.							
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			
HardSigmoid	moid	支持	int8 float16	<pre>input_tensor [batch, channel, height, width]:tensor</pre>	channel/ 输入的channel	无限制		per-layer
Transis.		文持			height/ 输入的height			per myer
					width/ 输入的width			
		支持	int8 float16	input_tensor [batch, channel, height, width]:tensor	batch/ 输入的batch	无限制		
Swish					channel/ 输入的channel			
					height/ 输入的height			
					width/ 输入的width			





	INCOME TIXTHEA								
(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	
					batch/ 输入的batch				
	IardSwish	支持	int8 float16		channel/ 输入的channel	无限制		per-layer	
	iaiuswisii	又付	lioatio	width]:tensor	height/ 输入的height			per-layer	
					width/ 输入的width				
					batch/ 输入的batch				
	oftplus		int8 float16	input tensor [batch, channel, height,	channel/ 输入的channel			per-layer	
,	ortpius	又行	Hoatio	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			
					channel/ 输入的channel	硬件支持[1,8192]			
:	oftmax	支持	int8 float16	width]:tensor	height/ 输入的height	axis=1, 无限制		per-layer	
					width/ 输入的width	axis=3/-1,width[1, 8192], height无限制			
				axis:int64 a a	axis/ 做softmax的轴	1,3,即channel和width方向			



	KK3502 IV O Operator List 如心原心 1 成仍有限。							
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	工門也			
			width]:tensor	height/ 输入的height	无限制			
				width/ 输入的width				
Slice	部分支持	int8 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			
			input_tensor [batch, channel, height,	batch/ 输入的batch				
				channel/ 输入的channel				
			width]:tensor	height/ 输入的height	Zupkup			
Split	部分支持	int8 float16		width/ 输入的width			per-layer	
			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
				batch/ 输入的batch			
Tanh	+	int8 float16	input_tensor [batch, channel, height,	channel/ 输入的channel	无限制		per-layer
rann	支持	поатть	width]:tensor	height/ 输入的height	ZUPANI		per-layer
				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]		
				height/ 输入的height			
				width/ 输入的width	[1,8176]		
Transpose	支持	int8 float16	perm:int64[]	axis order/ 转置的轴顺序	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<2048*2048, w1*c1<2048*512; 16bit数据 据 时,h1*w1<2048*2048, w1*c1<2048*256。 (2) perm=[0,3,1,2] 时,h1*w1<2048*2048。 (3) perm=[0,3,2,1]时,h1*w1<2048*2048, d1*w2<2048*2048。		



					202 TO O Operator Elst		
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	kernel_h, kernel_w]:int64[]	kernel_h/ height方向的kernel大小	[121]		per-layer/ per-channel
				kernel_w/ width方向的kernel大小	[1,31]]	
Convolution	支持	int8 float16	strides [strides_h, strides_w]:int64[]	stride_h/ height方向的strides大 小	-[1,7]		
				stride_w/ width方向的strides大小			
				pads_left/ left方向的pads大小			
				pads_right/ right方向的pads大小	1		
			pads_right]:int64[]	pads_top/ top方向的pads大小	[0,15]		
				pads_bottom/ bottom方向的pads大小			
			group:int64	group/ group的大小	无限制		
			dilations [dilations_h, dilations_w]:int64[] di w	dilations_h/ height方向的dilations大 小			
				dilations_w/ widtht方向的dilations大 小	[1, 32]		



					502 147 G Operator Elst			
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_shape [batch, channel, height,	batch/ 输入的batch	无限制			
				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	无限制			
		kernel_h, kernel_w]:int64[] kernel_h/ height方向的kernel大小 kernel_w/ width方向的kernel大小 width方向的kernel大小						
Depthwise Convolution	士体			per-layer/ per-channel				
Deputwise Convolution	XN		strides [strides_h, strides_w]:int64[]	height方向的strides大			per-channel	
				stride_w/ width方向的strides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_right]:int64[]	pads_top/ top方向的pads大小	[0,12]			
			dilations [dilations h,	pads_bottom/ bottom方向的pads大小				
				dilations_h/ height方向的dilations大 小 dilations_w/	[1, 32]			
			unauons_wj.mo+[j		widtht方向的dilations大小			



					T O Operator Elst	7 N. G. W. G. J. W. L.	
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 kernel_w width方 int8 float16 strides [strides_h, strides_w]:int64[] kernel_h height方 小 stride h height方	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]			
ConvTranspose/ Deconvolution	支持		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			
				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: - 不支持 kernel_h * dilations_h - dilations_h - pads_top < 0		
			pads_right]:int64[]	pads_top/ top方向的pads大小	不支持 kernel_w * dilations_w - dilations_w - pads_left < 0 不支持 ktride_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	1,当且仅当num_input=num_output时,支持num_output		
			dilations [dilations_h, dilations_w]:int64[] di w	dilations_h/ height方向的dilations大 小			
				dilations_w/ widtht方向的dilations大 小	[1, 32]		



					T South to operator hist		TIME								
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										
	尚不支持 目前由CPU int8 实现 float16		input_tensor_2 [K,N]:tensor	输入数据的形状	マクJwiaunui 矢苑, ジ水下可wiaunui										
Gamm		int8	alpha:double	alpha/ 矩阵A*B乘法的scale	无限制		per-layer/								
Gemm		float16	beta:double	beta/ 输入C矩阵的scale	无限制		per-channel								
		transA:into4	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	(Afrecumsof 又)可有直	
MatMul(4d)	部分支持 目前该支持仅针对双 feature输入 未来将支持输入为 feature+constant	int8 float16	input_tensor_1 [batch, channel, K, N]:tensor input_tensor_2 [batch, channel, N, M]:tensor	batch/ 输入的batch channel/ 输入的channel K/ 输入的K N/ 输入的M	双feature时:batch无限制channl、K支持[1,8192] feature+constant时: 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]+weight[K,N,1,1]的conv;N对齐要求:32对齐其他约束和conv相同		per-layer/ per-channel								



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			
		int8	input_tensor [batch, channel, height,	channel/ 输入的channel			
		float16	width]:tensor	height/ 输入的height	无限制		
Expand	支持			width/ 输入的width			
	ZN			batch_o/ 输出的batch_o			
		int64	shape (batch o, channel o, height o,	channel_o/ 输出的channel	无限制		
				height_o/ 输出的height			
				width_o/ 输出的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
Convolution + Relu	支持						
Convolution + Clip	支持						
Convolution + PRelu/LeakyRelu	支持						
Convolution + Add	支持						
Convolution + Mul	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	尚不支持						
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
Convolution + Relu	支持						
Convolution + Clip	支持						
Convolution + PRelu/LeakyRelu	支持						
Convolution + Add	支持						
Convolution + Mul	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	尚不支持						
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					
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
operator	又持有优	制八 奴旂尖空	測 人	输入参数含义	约米 规恰)	里化又行刀八
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 Con	volution				
Depthwise Convolution + Softplus	尚不支持						
Depthwise Convolution + HardSigmoid	尚不支持						
Depthwise Convolution + HardSwish	尚不支持						
Depthwise Convolution + Elu	尚不支持						
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	ful					
Convolution+Add+Relu	支持	同Convolution						

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、 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。



第五章 CPU Operator List



瑞芯微电子股份有限公司 描述 规格约束 说明 Operator 加法操作 无限制 Add AveragePool 无限制 平均池化 取最小值的index 无限制 ArgMin 取最大值的index 无限制 ArgMax BatchNormalization 无限制 批量归一化 SRC 支 持: float32/bool/int8/float16/int32/int64 数据类型转换 Cast DST支持: float32/int8/int32/float16 Clip 数据截断激活层 无限制 axis仅支持{0,1,2,3} Concat 合并操作 Convolution 卷积操作 无限制 ConvTranspose/Deconvolution 转置卷积 无限制 无限制 Cos 余弦函数 DataConvert 仅支持 bool/int8/float类型转换 数据类型转换 DepthToSpace 通道方向空间方向转换 无限制 除法操作 无限制 Div Equal 等于 无限制 Exp

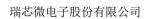
无限制

指数函数



瑞芯微电子股份有限公司

Operator	描述	规格约束	说明
Flatten	拉平操作	无限制	
Gather	聚集操作	无限制	
Greater	大于	无限制	
GreaterOrEqual	大等于	无限制	
GRU	门控循环单元	无限制	
GRU (extern)	门控循环单元	无限制	ONNX扩展算子
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	





Operator	描述	规格约束	说明
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
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	
RMSNorm (extern)	均方根归一化	无限制	ONNX扩展算子



瑞芯微电子股份有限公司

Operator	描述	规格约束	说明
RoiAlign	区域对齐池化	仅支持Avg Pool Mode,batchsize 仅支持 1	
ScatterND	N维索引取数	无限制	
Sin	正弦函数	无限制	
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
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	
Tanh	双曲正切函数	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	



第六章 模型输入输出说明



1、模型输入说明

芯片平台 模型首层精度 类型	構刑	臣	首层设置输入数	mean/scale/quant	输入宽(width)对齐要求 单位:元素个数		输入宽(width)大小限制		
		据类型	后端实现设备	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	NIDI	- 8	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 2. width <= 4096 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 - CEIL((dilation_kernel_h * dilation_kernel_w) / 128)	无限制	
	int8	→ 4维度	int8	NPU					
	ilito		float16	CPU					
			其他类型(* <u>注释</u> 8)	CPU			$\label{eq:convolution: N = 8 - CEIL((dilation_kernel_h * dilation_kernel_w) / 4096)} \\ ConvTranspose/Deconvolution: N = 8 - CEIL((dilation_kernel_h * dilation_kernel_w) / 128) \\$		
RK3566/3568		4年及	uint8	CPU	4	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 2. width <= 4096 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 - CEIL((dilation_kernel_h * dilation_kernel_w) / 128) Depthwise Convolution: N = 8 - CEIL((dilation_kernel_h * dilation_kernel_w) / 4096) ConvTranspose/Deconvolution: N = 8 - CEIL((dilation_kernel_h * dilation_kernel_w) / 128)	无限制	
	float16		int8						
	noatro		float16						
			其他类型(* <u>注释</u> 8)						
	无限制	非4维	无限制	CPU	1	1	无限制	无限制	
		4维度	uint8	NPU		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 <= 2048 * N 2. width <= 8192 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) Depthwise Convolution: N = 12 - MAX(CEIL((dilation_kernel_h * dilation_kernel_w) / 2048), 3) ConvTranspose/Deconvolution: N = 12 - MAX(CEIL((dilation_kernel_h * dilation_kernel_w) / 128), 3)		
	int8		int8	1110	-16			无限制	
			float16	-CPU	10				
			其他类型(* <u>注释</u> 8)						
RK3588		14/2	uint8	-CPU	8	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				2. width <= 8192 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) Depthwise Convolution: N = 12 - MAX(CEIL((dilation_kernel_h * dilation_kernel_w) / 2048), 3)	无限制	
			float16						
			其他类型(* <u>注释</u> 8)				ConvTranspose/Deconvolution: N = 12 - MAX(CEIL((dilation_kernel_h * dilation_kernel_w) / 128), 3)		
	无限制	非4维	无限制	CPU	1	1	无限制	无限制	
RV1103/ RV1106 in	int8	4维 	uint8	NPU	16	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 <= 2048 * N 2. width <= 4096 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	
			int8				兵中N必須为1到7的整数,超出犯菌的を核不文文材,合金核关型N的订算力式如下: 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)	√∟	

Rockchip 瑞芯微电子 模型输入输出说明 瑞芯微电子股份有限公司

int8		uint8 int8	NPU	16	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 <= 2048 * N 2. width <= 4096	T. VH dail	
	-4维度	float16 其他类型(* <u>注释</u> 8)	CPU			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 	
RK3562	float16	144	uint8 int8 float16 其他类型(* <u>注释</u>	CPU	8	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 <= 2048 * N 2. width <= 4096 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
	无限制	非4维	无限制	CPU	1	1	无限制	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、RV1106/RV1103不支持CPU的mean/scale/quant操作
- 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、详细的用法请参考《Rockchip_RKNPU_User_Guide_RKNN_API》



2、模型输出说明

芯片平台	模型输出精度 类型 (* <u>注释</u> 2)	输出维度	设置输出Layout	Channel对齐要求	H*W对齐要求
	int8	.4维度	NCHW	无	无
			NHWC	8对齐(* <u>注释</u> 1)	无
			NC1HWC2	最后一层卷积类算子,16对齐,最后一层非 卷积类算子8对齐	H*W要4对齐
			UNDEFINE	无	无
RK3566/3568			NCHW	无	无
			NHWC	4对齐(* <u>注释</u> 1)	
	float16		NC1HWC2	最后一层卷积类算子,8对齐,最后一层非卷 积类算子4对齐	H*W要4对齐
			UNDEFINE	无	无
	无限制	非4维	UNDEFINE	无	无
	int8	4维度	NCHW	无	无
			NHWC	16对齐(* <u>注释</u> 1)	
			NC1HWC2	最后一层卷积类算子,32对齐,最后一层非 卷积类算子16对齐	H*W要4对齐
			UNDEFINE	无	无
RK3588	float16		NCHW	无	无
			NHWC	8对齐(* <u>注释</u> 1)	无
			NC1HWC2	最后一层卷积类算子,16对齐,最后一层非 卷积类算子8对齐	H*W要4对齐
			UNDEFINE	无	无
	无限制	非4维	UNDEFINE	无	无

	确心似电子			快至:	制入制出 况 明
RV1103/			NC1HWC2	最后一层卷积类算子,32对齐,最后一层非 卷积类算子16对齐	H*W要4对齐
K V 1100	RV1106 int8	4维	NHWC	无	
			NCHW	无	无
		nt8	NHWC	无	
	int8		NC1HWC2	最后一层卷积类算子,32对齐,最后一层非 卷积类算子16对齐	H*W要4对齐
		4维度	UNDEFINE	无	无
RK3562			NCHW	无	无
	0 16		NHWC	无	无
float	float16	10at16	NC1HWC2	最后一层卷积类算子,16对齐,最后一层非 卷积类算子8对齐	H*W要4对齐
			UNDEFINE	无	无
	无限制	非4维	UNDEFINE	无	无

注释:

- 1、如果输出tensor类型是NHWC的,输出转换是NPU实现的输出,则有对齐要求,cpu实现的没有对齐要求; 2、输出精度类型int8/float16表示模型最后一层原始输出的数据类型;
- 3、NCHW输出,如果是NPU实现采用零拷贝接口则输出内存开辟的size以query出来的size为准;
- 4、NC1HWC2输出,输出内存开辟的size以query出来的size为准;