

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: | V1.5.2 |
|---------|-------|------------|
| 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: | 2023-08-04 |

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 | | 更新RK3588 OP列表,增加CPU OP List | 熊伟 |
| v1.3.1 | NPU 团队 | 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 | .0 NPU 团队 2022-09-02 2.更新LSTM、transpose、softmax | | 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 |



| 版本 | 修改人 | 修改日期 | 修改说明 | 核定人 |
|--------|-------|------------|-----------------------------------------------------------------------------------------------------------------------------|-----|
| v1.5.1 | NPU团队 | 2023-06-06 | 1. 新增部分CPU OP支持项 | |
| v1.5.2 | NPU团队 | 2023-08-04 | 新增RK3562 exSoftmaxMask OP支持项 新增where OP 支持项 新增MatMul GPU OP支持项 新增exGlu OP 支持 | |



目 录

| RKNN Compiler Support Operator List | |
|-------------------------------------|-----|
| 瑞芯微电子股份有限公司 | |
| 第一章 RK3566/3568 NPU Operator List | 5 |
| 第二章 RK3588 NPU Operator List | 36 |
| 第三章 RV1103/1106 NPU Operator List | 68 |
| 第四章 RK3562 NPU Operator List | 98 |
| 第五章 CPU Operator List | 131 |
| 第六章 GPU Operator List | 136 |



| · 선 1 · 상 | 工作 工工 4人 | ・ 4人 コーン公 ロロ | | 10 | _ |
|-------------------|-----------|-------------------|-----------------------------------------|------------------------------------|---|
| 卫七百 | 水見 兀り作前 。 | 八 46前 545 752 11日 | | 1 3' | ۶ |
| 77 U T | | | *************************************** | $\mathbf{I} \mathcal{J}^{\dagger}$ | L |



第一章 RK3566/3568 NPU Operator List



| | | | | | KKSS | | 7月12日 | |
|---|-----------|-------|---------|-----------------------------------------------------|------------------------|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------|
| | 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ONNX规范的四维tensor的所有广播操作,以 | |
| | Add/Bias | | | input_tensor | channel/ 输入的channel | | 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 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> | |
| | | | | input_tensor [batch, channel, height, width]:tensor | batch/ 输入的batch | | THE TALL MENTAL INC. IN CONTROL OF THE PROPERTY OF THE PROPERT | |
| | Sub | | | | channel/ 输入的channel | 无限制 | | per-layer/ per-channel |
| | | | | | height/ 输入的height | | | |
| Ę | | | | | width/ 输入的width | | | |
| | | | | | batch/ 输入的batch | | 12 ()P(A (N (' 日 W) R ((' 1 1)) 图 ('經 度 標 / Droadcasting | |
| | V 10 | | tloat16 | input_tensor [batch, channel, height, width]:tensor | channel/ 输入的channel | | | per-layer/ per-channel |
| | Mul/Scale | 支持 | | | height/ 输入的height | | | |
| | | | | | width/ 输入的width | | | |



| | | RKS300/3308 NI O Operator List 如心版电 1 放 协行 | | | | | | |
|--|----------|--------------------------------------------|-----------------|-----------------------------------------------------|------------------------|----------------|----------------------------------------------------------------------------------------------------------------------------------|---------------------------|
| | operator | 支持情况 | 输入数据类型 | 输入 | 设置项/ 输入参数含义 | 约束规格 | 广播支持(维度补齐) | 量化支持方式 |
| | | | | | batch/ 输入的batch | | 支持两个tensor的广播操作,以ONNX默认排列NCHW | |
| | Div | 部分支持 | float16 | input_tensor | channel/ 输入的channel | | 做说明,支持以下广播方式: I、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 | 即刀 又行 | Tioat 10 | [batch, channel, height, width]:tensor | height/ 输入的height | | 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] | 1 - 1 - 1 - 1 | per-layer/ per-channel |
| | vida | | | | height/ 输入的height | | | |
| | | | | | width/ 输入的width | [1,8176] | | |
| | | | | | batch/ 输入的batch | 无限制 | | |
| | Min | | | input_tensor [batch, channel, height, width]:tensor | channel/ 输入的channel | [1 9 102] | 1 = 11 11 11 | per-layer/ |
| | | XH | | | height/ 输入的height | 2. 3. b: | | per-channel |
| | | | | | width/ 输入的width | [1,8176] | | |



| | | RK3500/3508 NPU Operator List | | | | | | |
|---|-----------------|-------------------------------|-----------------|---------------------------------------|------------------------|---------------------------|------------|-----------|
| C | 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 | | per-layer |
| C | ilobal | 支持 | int8 float16 | input_tensor [batch, channel, height, | channel/ 输入的channel | [1,8192] | | |
| F | veragePool | XII | illoario | width]:tensor | height/ 输入的height | -[1,343] (toolkit2支持范围) | | |
| | | | | | width/ 输入的width | [1,543] (t001K1t2XN46III) | | |
| | | | int8 float16 | | batch/ 输入的batch | 1 | | |
| | JobalMaxPool | | | input_tensor [batch, channel, height, | channel/ 输入的channel | [1,8192] | | per-layer |
| | GiobaliviaxPool | X ft | | width]:tensor | height/ 输入的height | [1,343](toolkit2支持范围) | | per layer |
| | | | | | width/ 输入的width | | | |



| | KK3500/3506 NI C Operator List 加心版电 | | | | | | |
|-------------|-------------------------------------|-----------------|----------------------------------------------------------------|---------------------------------------------|----------------------------|------------|-----------|
| operator | 支持情况 | 输入数据类型 | 输入 | 设置项/ 输入参数含义 | 约束规格 | 广播支持(维度补齐) | 量化支持方式 |
| | | | | batch/ 输入的batch | 1 | | |
| | | | input_tensor [batch, channel, height, | channel/ 输入的channel | [1,8192] | | |
| | | | width]:tensor | height/ 输入的height | [1,8192] | | |
| | | | | 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 | | |
| | 支持 | int8 float16 | count_include_pad:int64 | count_include_pad/ 是否包含pad数值进 行计算 | 1 | | per-layer |
| | | | kernel_shape [kernel_h, kernel_w]:int64[] kernel_w/ | height方向的kernel大 | - 无限制,NPU支持[1,7];其它由CPU支持。 | | |
| AveragePool | | | | width方向的kernel大 | | | |
| | | | pads [pads_top, pads_left, pads_bottom, pads_right]:int64[] | pads_left/ left方向的pads大小 | [0,7] | | |
| | | | | pads_right/ right方向的pads大小 | | | |
| | | | | pads_top/ top方向的pads大小 | | | |
| | | | | 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] | | |



| | | | | | 3 13 10 4 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 | | |
| | | | 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 | [,,,,,,] | | |
| | | | 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 | | |
| | 支持 | int8 float16 | dilations [dilations_h, | dilations_h/ height方向的dilations 大小 | | | |
| | | | | dilations_w/ widtht方向的dilations 大小 | | | |
| MaxPool | | | | kernel_h/ height方向的kernel大 小 | 无限制,NPU支持[1,7];其它由CPU支持。 | p | 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[] height | 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 |
|------------------------|------|---------|-----------------------------------------------------|--------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|---------------------------|
| | | | 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 | プルドで 中リ | | |
| | | | | 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 | | |
| | | | width]:tensor(const) layernorm_bias[channel,height, | height/ 输入的height | 等于input_height | | |
| Layer Normalization | 部分支持 | float16 | width]:tensor(const) | width/ 输入的width | 等于input_width | | per-layer |
| | | | normalized_shape:int64[] | normalized_ shape /参与每一批归一化的 Feature的尺寸 | NPU仅支持,包含除第0维(batch维)以外的其他所有维度, 度, 如input_shape[n,c,h,w], 仅支持normalized_shape[c,h,w], 如input_shape[n,c,h], 仅支持normalized_shape[c,h], 如input_shape[n,c], 仅支持normalized_shape[c], 其余情况会转到CPU执行。 | | |
| | | | elementwise_affine:int64 | elementwise_affine/ 是否具有可学习数 | 0 或 1(默认为 0)。 当为1时拥有LayerNorm.weight与LayerNorm.bias,仅支持 weight/bias的尺寸:elementwise_shape与normalized_shape一 致;当为0时LayerNorm.weight为全1值,LayerNorm.bias为全0 值。 | | |
| | | | eps:double | eps/ 防止除法溢出的偏移 参数 | 无限制 | | |



| | | | | | 7.5506 TV 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 |
| | | | | batch/ 输入的batch | | | |
| Clip/ReLU6 | 支持 | int8 float16 | input_tensor [batch, channel, height, | channel/ 输入的channel | 无限制 | | per-layer |
| Сприсло | 文材 | noatro | width]:tensor | height/ 输入的height | 之上 PRR 市归 | | per-rayer |
| | | | | width/ 输入的width | प्रwidth | | |
| | | 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 | | |
| Did | ZW. | | | 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 | | |
| | | iloutio | width]:tensor | height/ 输入的height | スLPR fbg | | |
| | | | | width/ 输入的width | | | |



| o | 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 |
|---|----------|------|-----------------|-----------------------------------------------------|------------------------|---------------|------------|---------------------------|
| | | | | | 糊入参数含义 batch/ 输入的batch | | | |
| R | elu | 支持 | int8 float16 | input_tensor [batch, channel, height, | channel/ 输入的channel | 无限制 | | per-layer |
| | | Σ.Ν | | width]:tensor | height/ 输入的height | Zupany | | permyer |
| | | | | | width/ 输入的width | | | |
| | | | | | batch/ 输入的batch | | | |
| ī | eakyRelu | | int8 float16 | input_tensor [batch, channel, height, width]:tensor | channel/ 输入的channel | 无限制 | | per-layer |
| | | | | | height/ 输入的height | | | permyer |
| | | | | | width/ 输入的width | | | |
| | | | | | batch/ 输入的batch | | | |
| | | | | | channel/ 输入的channel | 无限制 | | |
| P | PRelu | 支持 | int8 float16 | input_tensor [batch, channel, height, width]:tensor | height/ 输入的height | | | per-layer/ per-channel |
| | | | | | width/ 输入的width | | | |
| | | | | 1 | slope/ PReLU系数 | 仅支持单个标量或C维度系数 | | |



| | | | | 111200 | 00/3308 NPU Operator List | · · · · · · · · · · · · · · · · · · · | 3 13 10 4 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 | | | | | | | | |
| | | | input_tensor [sequence, batch, input_size] :tensor | sequence/ 输入的sequence | 无限制,建议4对齐 | | | | | | | | |
| | | | | input_size/ 输入的input_size | 无限制,建议8对齐 | | | | | | | | |
| | | | direction:string | direction/ 指定GRU的运算方向 | forward: 指定GRU的运算方向为前向 reverse: 指定GRU的运算方向为反向 bidirectional: 指定GRU的运算方向为双向 | | | | | | | | |
| | | | batch_size:int64 (extern) | batch_size/ 指定GRU输入的 batchsize | 1 | | per-layer | | | | | | |
| GRU | 部分支持 GRU扩展以及变体 命名为exGRU算 | float16 | sequence_size :int64 (extern) | sequence_size/ 指定GRU输入的 seqsize | | | | | | | | | |
| GRU | 子,参数项中指明 (extern)的项为 exGRU独有的参数 项。 | | | | noat 16 | Hoat 16 | Hoat 16 | Hoatro | 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。 | | | | | | | | |
| | | c | 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] | | | | | | | | |
| | | | output_layout:string (extern) | | 致的layout | 要求填写指定的layout,同时要求填写该op实际对应的batch_size、sequence_size、hidden_size。directions>1时仅支持batches=1。 | | | | | | | |



| operator | 支持情况 | 输入数据类型 | 输入 | 设置项/ 输入参数含义 | 约束规格 | 广播支持(维度补齐) | 量化支持方式 |
|----------|-----------------------------------------------|------------------------------|---------------------------------------------------|-----------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------|------------|---------------------------|
| | | | | batch/ 输入的batch | batch>1时要求batch=4n,(n为正整数),建议n<=4。 注:LSTM单向:无限制,LSTM双向:不同时支持多batch。 | | |
| | | | input_tensor [sequence, batch, input_size]:tensor | sequence/ 输入的sequence | 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 | | |
| | | | hidden_size:int64 (extern) | hidden_size/ LSTM单元中的 hiddensize | 无限制,建议8对齐 | | |
| | 部分支持 LSTM 扩展以及变体命名为exLSTM算子,参数项中指明(extern)的项为 | 扩展以及变 法为exLSTM算 参数项中指明 | proj_size:int64 (extern) | proj_size/ LSTM单元存在 projection时的 | 0<=proj_size<=hiddensize 目前限定0,即尚不支持projection功能 | | |
| LSTM | | | input_forget:int64 | proj_size input_forget/ cifg变种的选择 | 1(T) or 0(F) 目前限定0,即尚不支持 | | per-layer/ per-channel |
| | exLSTM独有的参数 项。 | | has_dropout:int64 (extern) | has_dropout/ caffe框架下的 indicator功能的选择 | 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] | | |
| | | | | 2XH THIYOUT | 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。 | | |
| | | c | output_layout:string (extern) | output_layout/指定与 对应输出shape含义一 致的layout | 1、sbnc: 指定layout对应的输出shape为 [seqs,directions,batches, hidden_size] 2、(sn)c: 指定layout对应的输出shape为[seqs*batches,directions*input_size,1,1] | | |
| | | | | ±хллауош | 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。 | | |



| | ANTACAPAC | | | | THISS | 00/3308 NPU Operator List | | 3111104 13 |
|-------|-----------|------|-----------------|---------------------------------------|------------------------|---------------------------------------------------------------------|------------|------------|
| | operator | 支持情况 | 输入数据类型 | 输入 | 设置项/ 输入参数含义 | 约束规格 | 广播支持(维度补齐) | 量化支持方式 |
| | | | | input_tensor [batch, channel, height, | batch/ 输入的batch | | | |
| | | | | | 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 | | |
| | Concat | 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 | | 数 | |
| | | | | axis:int64 | aixs/ 拼接的维度 | 无限制 | | |
| | | | | | batch/ 输入的batch | | | |
| | Mish | | int8 float16 | input_tensor [batch, channel, height, | channel/ 输入的channel | | | |
| IVIIS | | ~.19 | | width]:tensor | height/ 输入的height | Zurvirj | | |
| | | | | | 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 | 1 | | |
| | | int8 float16 | input_tensor [batch, channel, height, | channel/ 输入的channel | T. 05 (b) | | |
| Pad | | noatro | width]:tensor | height/ 输入的height | 无限制 | | |
| | 支持 | | | width/ 输入的width | [1,8176] | | |
| rau | 义行 | 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 | | | |
| | f | 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 | | | |
| | 尚不支持 目前由 | int8 | width]:tensor | height/ 输入的height | [1,8192] | | |
| ReduceMean | 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 | | |



| | | | | | 00/3508 NPU Operator List | · · · · · · · · · · · · · · · · · · · | * 14111-1 | |
|-----------|------------------------------------------------|-----------------|---------------------------------------|-----------------------------|----------------------------------------------------|---------------------------------------|-------------|--|
| 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 | | | |
| | | | | | | | | |
| ReduceSum | 尚不支持 目前由 | int8 | | height/ 输入的height | [1,8192] | | per-layer/ | |
| | 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, | channel/ 输入的channel | KI 0.100 | | | |
| | 部分支持 | | width]:tensor | height/ 输入的height | [1,8192] | | | |
| Resize | 目前NPU仅支持宽 高方向不超过8倍的 整倍数的最邻近插 值缩放,其余不支 | int8 float16 | | width/ 输入的width | 1.[1,8176] 2.设放大倍数为s(s为正整数),width*s*(s-1)<=8192 | | per-layer | |
| | 持部分的会Fallback 到CPU上实现。 | | 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 | | roi/进行resize的输入 范围 | 仅支持全局([0,0,0,0,1,1,1,1]) | _ | |



| | 垢心似 | | | | KKSS | 66/3568 NPU Operator List | 墙心 似电于股份 | D H M A H |
|-------|------|------|-----------|------------------------------------------------------|--------------------------------------------------------------|---------------------------|-----------------|-----------|
| opera | ntor | 支持情况 | 输入数据类型 | 输入 | 设置项/ 输入参数含义 | 约束规格 | 广播支持 (维度补齐) | 量化支持方式 |
| | | | | | batch/ 输入的batch | 无限制 | | |
| | | | int8 | input_tensor [batch, channel, height, width]:tensor | channel/ 输入的channel | | | |
| | | | float16 | | height/ 输入的height | [1,8192] | | |
| | | | | | width/ 输入的width | [1,8176] | | |
| | | | 支持 | shape (batch_o, channel_o, height_o, width_o):tensor | batch_o/ 输出的batch_o | 无限制 | | |
| Resh | ape | 部分支持 | | | channel_o/ 输出的channel | [1,8192] | | |
| | | | | | height_o/ 输出的height | | | |
| | | | | | width_o/ 输出的width | [1,8176] | | |
| | | | | | [n,c,h1,w1]- >[n,c,h2,w2]/ (h1*w1=h2*w2) | 支持 | | |
| | | | | | [1,c,h,w]- >[c1,hw1,1,1]/ (c1=c/a, h*w=hw1/a,a 为整数) | | | |
| | | | | | [n,c,1,1]->[1,n1,h,w]/ (c=h*w/a, n1=n/a, a为 整数) | - 不支持 / / b | | |



| operator | 支持情况 | 输入数据类型 | 输 λ | 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 |
|-----------|------------|-----------------|-------------------------------------------------------------------------------------------|------------------------------|----------------|--------------|--------|
| operation | X IN IH IV | 和人双加大王 | | 输入参数含义 batch/ 输入的batch | 无限制 | / 1日人内《年区日月/ | 至亿久的万八 |
| | | | input_tensor [batch, channel, height, width]:tensor height 输 width]:tensor wid 输 loat16 | channel/ 输入的channel | -[1,8192] | | |
| | | | | height/ 输入的height | [1,0172] | | |
| Reverse | 尚不支持 | int8 float16 | | width/ 输入的width | [1,8176] | | |
| Sequence | | | batch_axis:int64 | batch_axis/ 指定是否为batch维度 | 1 | | |
| | | | time_axis:int64 | time_axis/ 指定是否为time维度 | 0 | | |
| | | | sequence_lens:int64[] | sequence_lens/ 指定序列翻转的数量 | 仅支持channel数 | | |
| | | | | batch/ 输入的batch | | | |
| Sigmoid | 支持 | int8 | input_tensor [batch, channel, height, | channel/ 输入的channel | 无限制 | | |
| Sigmoid | 人 时 | float16 | width]:tensor | height/ 输入的height | עין אָער וייַן | | |
| | | | | width/ 输入的width | | | |



| | | | | 111200 | 00/3308 NPU Operator List | | 3 13 10 4 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 |
| | | | <pre>input_tensor [batch, channel, height, width]:tensor</pre> | batch/ 输入的batch | | | |
| HardSigmoid | 支持 | int8 float16 | | channel/ 输入的channel | 无限制 | | per-layer |
| HaidSigilloid | | noatro | | height/ 输入的height | 之上 PRR 市切 | | per-rayer |
| | | | | width/ 输入的width | | | |
| | | | | batch/ 输入的batch | | | |
| Swish | 支持 | int8 float16 | input_tensor [batch, channel, height, width]:tensor | channel/ 输入的channel | - 无限制 | | |
| SWISH | 文 44 | | | 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 |
| Haluswish | × 14 | noatio | width]:tensor | height/ 输入的height | ע איני איני איני איני איני איני איני אינ | | per-tayer |
| | | | | width/ 输入的width | | | |



| or | erator | 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 channel/ 输入的channel input_tensor [batch, channel, height, width] tansor 无限制 | batch/ 输入的batch | | | |
| So | oftplus | 支持 | int8 float16 | | | per-layer | | |
| | | | | width]:tensor | height/ 输入的height | | | |
| | | | | | width/ 输入的width | | | |
| | | | input te | | batch/ 输入的batch | 无限制 | | |
| | | | | input_tensor [batch, channel, height, | channel/ 输入的channel | [1,8192] 建议4对齐 | | |
| So | ftmax | 尚不支持,目前由 CPU实现 | float16 | width]:tensor | height/ 输入的height | 1 | | per-layer |
| | | axis: | | 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/ | 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 | | | |
| | | | input_tensor [batch, channel, height, | channel/ 输入的channel | | | |
| | | | ## W | 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_f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。 | : 8对齐, 16bit数据: 4对齐。 | |
| | | | 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/ 选取切分对应轴的步 长 | | | | |
| | | | | 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 | | | |
| | | 5 | spil | 指定切分后维度的长 | 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。 | | |



| | | | | | 00/3508 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 | | | | |
| Tanh | | int8 float16 | input_tensor [batch, channel, height, width]:tensor | input_tensor [batch, channel, height, | channel/ 输入的channel | 工作 41 | | per-layer |
| Tallii | 支持 | moatro | | height/ 输入的height | 无限制 | | per-rayer | |
| | | | | width/ 输入的width | | | | |
| | | | | batch/ 输入的batch | [1,1024] | | | |
| | | | input_tensor [batch, channel, height, width]:tensor | channel/ 输入的channel | [1,8192] | | | |
| | | | | height/ 输入的height | | | | |
| Transpose | 部分支持 | int8 float16 | | 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。 (5) perm=[0,3,1,2],in_shape=[n,c,h,w],且w要求8bit数据: 8 对齐, 16bit数据: 4对齐,并且c*h<8192。 | | | |



| | | | | | 700/3308 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 | 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 | J.L. PIX (PJ) | | |
| | | | num_input, kernel_h, kernel_w]:int64[] | kernel_h/ height方向的kernel大 小 | [1,31] | | per-layer/ per-channel |
| | | | | kernel_w/ width方向的kernel大 小 | | | |
| Convolution | 支持 | int8 float16 | otuidoo fotuidoo la otuidoo grilinteAfi | stride_h/ height方向的strides大 小 | [1,7] | | |
| | | | strides [strides_ii, strides_w]:Int04[] | 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 | [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 | | |
| | | dilations [dilations_h, dilations_w]:int64[] | | dilations_h/ height方向的dilations 大小 | [1, 32] | | |
| | | | dilations_w/ widtht方向的dilations 大小 | | | | |



| | | | | Miss | 000/3308 NPU Operator List | r 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 | 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 | [1,0] | | | | |
| Depthwise Convolution | 支持 | float16 | strides [strides_h, strides_w]:int64[] | stride_h/ height方向的strides大 小 | [1,7] | | | | |
| | | | suites [suites_ii, suites_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_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 | | | | | |
| | | | 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] | | | | |



| | | | | | 700/3306 IVI 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 | 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 | | |
| | | | | 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 | ri an | | |
| | | | | kernel_w/ width方向的kernel大 小 | [1,31] | | |
| | | int8 | | stride_h/ height方向的strides大 小 | {1,2,4,8} | | |
| ConvTranspose/ Deconvolution | 支持 | float16 | strides [strides_h, strides_w]:int64[] | stride_w/ width方向的strides大 小 | | | per-layer/ 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 | 支持0-15 · 设置pad时注意: 不支持 kernel_h * dilations_h - dilations_h - pads_top < 0 | | |
| | | | pads_bottom, 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 krride_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_h/ height方向的dilations 大小 | | | |
| | | dilations_w]:int64[] | 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/ | | | |
| | | | 输入数据的形状 input_tensor_2 [K,N]:tensor | | | | |
| | 不支持 由CPU实现 | int8 | alpha:double | alpha/ 矩阵A*B乘法的scale | | | per-layer/ |
| Gemm | 个支持 田CPU头现 | | 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/ B矩阵是否转置 | | | |
| | | | input_tensor_1 [batch, K, C]:tensor | batch/ 输入的batch | | | |
| MatMul | 不支持 由CPU实现 | int8 | | K/ 输入的K | | | per-layer/ |
| | , , , , , , , , , , , , , , , , , , , , | | input_tensor_2 [batch, C, H]:tensor | C/ 输入的C | | 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 | <pre>input_tensor [batch, channel, height,</pre> | channel/ 输入的channel | 无限制 | | |
| | | float16 | width]:tensor | height/ 输入的height | | | |
| Expand | 支持 | | | width/ 输入的width | | | |
| | XIV | | | batch_o/ 输出的batch_o | | | |
| | | int64 | 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 | 給 λ | 设置项/ | 约束规格 | 广播支持(维度补齐) | 量化支持方式 |
|----------|------|-------------------------------|----------------------------------------------------|--------------------------|-------|-------------|------------|
| operator | 又付旧仇 | - 個八数船天空 | 相八 | 输入参数含义 batch/ | 幻水风竹 | / 御义村(维及作介) | 里化又付刀八 |
| | | | | 输入的batch | | | |
| | | int8 float16 | x_tensor [batch, channel, height, width]:tensor | channel/ 输入的channel | 无限制 | | |
| | | int64 | | height/ 输入的height | | | |
| | | | | width/ 输入的width | | | |
| | | int8 | | batch/ 输入的batch | | | |
| | | | y_tensor [batch, channel, height, width]:tensor | channel/ 输入的channel | 工界机 | | per-layer/ |
| | | float16 int64 | | height/ 输入的height | - 无限制 | | |
| | | | | width/ 输入的width | | | |
| | +++ | | | batch/ 输入的batch | | | |
| Where | 支持 | h 1 | | channel/ 输入的channel | | | |
| | | bool | mask_tensor [batch, channel, height, width]:tensor | height/ 输入的height | 无限制 | | |
| | | | | width/ 输入的width | | | |
| | | | 1 | batch_o/ 输出的batch_o | | | |
| | | int8 shape (batch_o, channel_ | shape (batch_o, channel_o, height_o, | channel_o/ 输出的channel | | | |
| | | float16 | width_o):tensor | height_o/ 输出的height | 无限制 | | |
| | | | | width_o/ 输出的width | | | |





| | KKS500/3500 Ft 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 | |
| | int8 float16 | x_tensor [batch, channel, height, width]:tensor | batch/ 输入的batch channel/ 输入的channel height/ 输入的height width/ 输入的width | c*h*w满足如下限制 8bit数据:8对齐,16bit数据:4对齐 | | per-layer/ | | |
| exGlu | 支持 | Int64 | axis:int64 | axis/ 切分的维度 | axis ==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 |
|----------------------------------|------|--------------|----|----------------|------|------------|--------|
| 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 | | | | | | |



| | | | | | 00/3508 NI O Operator List | | U版电 1 放仍有限公司 |
|------------------------------------|------|--------------|-----|----------------|----------------------------|------------|--------------|
| operator | 支持情况 | 输入数据类型 | 输入 | 设置项/ 输入参数含义 | 约束规格 | 广播支持(维度补齐) | 量化支持方式 |
| 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 | ose | | | | |
| 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 | | | | | | |



| | L. Li. blaver | 44 > W 1 11 1 | | 设置项/ | V de la la | ala 196 Li Li (II) ala M ala N | |
|-----------------------------------------------|---------------|---------------|------------|----------------|------------|--------------------------------|--------|
| 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 | | | | | | |



| 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



| | | | | | \rt. EE -55 / | RRS500 FG C Operator Elst | | | _ |
|----------|--------------|------|-----------------|----------------------------------------|----------------------------------------------|---------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------|-----------------------------|
| operator | r | 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 |
| AJJÆ | Add/Bias | 支持 | int8 float16 | input_tensor | 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 : 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/ | |
| Add/bia | is | 又行 | noatro | [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。 例子见 <u>注释(1)</u> | per-channel | 行支 持 一 己 支 持 |
| | | | | width/ 输入的width | | | | | |
| | | | | | batch/ 输入的batch | | | | |
| Sub | | 支持 | int8 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 | per-layer/ | 尚不支持 |
| | | Z.N | | [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 | 尚不支持 |
| | | | | | width/ 输入的width | | | | |
| | | | | | batch/ 输入的batch | | | | |
| | Mul/Scale 支柱 | 1.16 | int8 | input_tensor | channel/ 输入的channel | | 支持ONNX規范的四维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/ | |
| Mul/Sca | | 文持 | 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 | 同小支持 |
| | | | | | 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 | 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/ | 以工士柱 |
| DIV | 部分又行 | moatro | [batch, channel, height, width]:tensor | height/ 输入的height | プレPR 申リ | 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 | 尚不支持 |
| | | | | 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: I、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 |
| | ~14 | | 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 | 17 7 2.19 |
| | | | | 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 | 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] | | | |



| | | | | | KK5588 NI O Operator List | | 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 | 1 | | | |
| Global | 支线 | int8 | channel/ 输入的channel input_tensor [batch, channel, height, | [1,8192] | | per-layer | 尚不支持 | |
| AveragePool | all 支持 float16 | width]:tensor | height/ 输入的height | [1,343](toolkit2支持范围) | | Per sayer | MALXIN | |
| | | | | width/ 输入的width | [1,945](WOIMIZ文刊记回) | | | |
| | | | | batch/ 输入的batch | 1 | | | |
| GlobalMaxPool | 支持 | int8 float16 | input_tensor [batch, channel, height, | channel/ 输入的channel | [1,8192] | | per-laver | 尚不支持 |
| GlobalMaxPool 3 | | | width]:tensor | height/ 输入的height | [1,343](toolkit2支持范围) | | per-layer | India VIV |
| | | | | width/ 输入的width | | | | |



| | | | | | KK3366 NI O Operator List | | 11以仍有限公 | | |
|-------------|------|-----------------|----------------------------------------|---------------------------------------------|---------------------------|-------------|---------------|----------------|------|
| 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] | | 年に入りの人 行う | | |
| | | | | width/ 输入的width | [1,6192] | | | | |
| | | | 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_h/ height方向的kernel大 小 | | pe | per-layer 尚不 | per-layer j | 尚不支持 |
| | | | kernel_w]:int64[] | kernel_w/ width方向的kernel大 小 | 无限制,NPU支持[1,7];其它由CPU支持。 | | | | |
| | | | | pads_left/ left方向的pads大小 | | | per-layer 尚不支 | | |
| | | | 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 | | | | | |
| | | | | stride_h/ height方向的strides大 小 | ri 0) | | | | |
| | | | strides [strides_h, strides_w]:int64[] | stride_w/ width方向的strides大 小 | [1,8] | | | | |



| | | | | | TARSSOOTAL C Operator East | | 2 1 1X 1/1 1/1 1/1 Z | | |
|----------|------|---------------------------|----------------------------------------|--------------------------------------------------|----------------------------|-------------|----------------------|----------------|--|
| 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 | [1,0192] | | | | |
| | | | 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 | | | | |
| | | | dilations [dilations_h, | dilations_h/ height方向的dilations 大小 | 1 | | | | |
| | | | dilations_w]:int64[] | dilations_w/ widtht方向的dilations 大小 | | | | | |
| MaxPool | 支持 | int8 float16 | kernel_shape [kernel_h, | kernel_h/ height方向的kernel大 小 | -无限制,NPU支持[1,7];其它由CPU支持。 | | per-layer | 尚不支持 | |
| | | | kernel_w]:int64[] | kernel_w/ width方向的kernel大 小 | 九敗帝明,NFO又将[1,7]; 共と田CFO又将。 | | | | |
| | | | | 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 | | | | | |
| | | | storage_order: int64 | storage_order/优先储 存方式 | 0 | | | | |
| | | strides [strides_h, stric | strides [strides_h, strides_w]:int64[] | stride_h/ height方向的strides大 小 stride_w/ | [1,8] | | | | |
| | | | | width方向的strides大 小 | | | | | |



| | | | | | KKS300141 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 |
| | | | epsilon:double | epsilon/ 除以标准差时加上防 止除0的实数 | 非0实数,参考值为1e-5 | | | |
| | | | momentum:double | momentum/ 训练时的滑动平均参 数 | 无限制 | | | |
| Batch Normalization | 支持 | int8 float16 | | batch/ 输入的batch | 1 | | per-layer/ per-channel | 尚不支持 |
| | | | input_tensor [batch, channel, height, | channel/ 输入的channel | 구떠사 | | | |
| | | | width]:tensor | height/ 输入的height | 无限制 | | | -layer/ -channel 尚不支持 |
| | | | | width/ 输入的width | 无限制 | | | |
| | | | | batch/ 输入的batch | 支持多batch | | | |
| | | | input_tensor [batch, channel, height, | channel/ 输入的channel | | | | |
| | | | width]:tensor | height/ 输入的height | 无限制 | | | |
| | | | | 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 | 即为文时 | noutro | width]:tensor(const) | width/ 输入的width | 等于input_width | | per layer | 网小文羽 |
| | | | 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_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 | | | | |
| Cli. ID 1116 | _late | int8 float16 | input_tensor [batch, channel, height, | channel/ 输入的channel | | | | |
| Clip/ReLU6 | 支持 | noatio | width]:tensor | height/ 输入的height | 无限制 | | per-layer | 多行 已 一 0 一 0 一 0 一 0 一 0 一 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 <td< td=""></td<> |
| | | | width/ 输入的width | | | | | |
| | | | | batch/ 输入的batch | | | | |
| Elu | 支持 | int8 float16 | input_tensor [batch, channel, height, | channel/ 输入的channel | - 无限制 | | | 冶不支持 |
| Eiu | XII | noatro | width]:tensor | height/ 输入的height | Zupk mj | | | 间小文符 |
| | | | | 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 |
| Sold | N.M. | NORTO . | width]:tensor | height/ 输入的height | | | | 國小文母 |
| | | | | 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 | +- 44: | int8 float16 | input_tensor [batch, channel, height, | channel/ 输入的channel | 无限制 | | per-layer | 二十件 |
| Keiu | 支持 | noatro | width]:tensor | height/ 输入的height | 元 陳神 | | per-rayer | C又持 |
| | | | | width/ 输入的width | | | | |
| | | | | batch/ 输入的batch | | | | |
| LeakyRelu | 支持 | int8 float16 | input_tensor [batch, channel, height, | channel/ 输入的channel | 无限制 | | per lever | 口去体 |
| EcakyReiu | 文 时 | noacro | width]:tensor | height/ 输入的height | Ушук пр | | per-rayer | 己支持 |
| | | | | width/ 输入的width | | | per-layer 己支持 | |
| | | | | batch/ 输入的batch | | | | |
| | | | | channel/ 输入的channel | 无限制 | | | |
| PRelu | 支持 | int8 float16 | input_tensor [batch, channel, height, width]:tensor | height/ 输入的height | | | per-layer/ per-channel | 已支持 |
| | | | | width/ 输入的width | | | | |
| | | | | slope/ PReLU系数 | 仅支持单个标量或C维度系数 | | | |



| operator | 支持情况 | 输入数据类型 | 输入 | 设置项/ 输入参数含义 | 约束规格 | 广播支持(维度补齐) | 量化支持方式 | 多核协同运 行支持情况 |
|----------|-----------------------------------------------|---------|----------------------------------------------------|-----------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|--------------|-----------------|
| | | | | batch/ 输入的batch | 1 | | | 13 × 13 113 9 € |
| | | | 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 | | | |
| | 部分支持 | | 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 | | per-layer 尚. | |
| | 部分支持 GRU 扩展以及 变体命名为 | | sequence_size :int64 (extern) | sequence_size/ 指定GRU输入的 seqsize | 无限制,建议4对齐 | | | |
| GRU | exGRU算子,参 数项中指明 (extern)的项 为exGRU独有的 | float16 | hidden_size:int64(extern) | hidden_size/ GRU单元中的 hiddensize | 无限制,建议8对齐 | | | 尚不支持 |
| | 参数项。 | | linear_before_ reset:int64 | linear_before_ reset/ LBR变种的选择 | 1(T) or 0(F) | | | |
| | | | input_layout:string (extern) | input_layout/指定与对应输入shape含义一致的layout | 1、snc: 指定layout对应的输入shape为[seqs, batches, input_size] 2、(sn)c: 指定layout对应的输入shape为[seqs*batches, input_size,1,1] 要求填写指定的layout,同时要求填写该op实际对应的batch_size、sequence_size、hidden_size。 | | | |
| | | | output_layout:string (extern) | output_layout/指定与 对应输出shape含义一 致的layout | 1、sbnc: 指定layout对应的输出shape为[seqs,directions,batches, hidden_size] 2、(sn)c: 指定layout对应的输出shape为[seqs*batches, directions*input_size,1,1] 要求填写指定的layout,同时要求填写该op实际对应的batch_size、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 |
|----------|----------------------------------------------------------------------|----------------------------------|---------------------------------------------------|-------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|------------------------|------------|
| | 部分支持 LSTM 扩展以及 变体命名为 in | | | 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。 | | | 11 大小門ル |
| | | | 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 | | per-layer/ per-channel | |
| | | | 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:int6- | batch_size:int64 (extern) | batch_size/ 指定LSTM输入的 batchsize | 大于1时仅支持4的倍数 | | | |
| | LSTM 扩展以及 变体命名为 exLSTM算子, 参数项中指明 (extern)的项 为exLSTM独有 | sequence_size :int64 (extern) | sequence_size/ 指定LSTM输入的 seqsize | 无限制,建议4对齐 | | | | |
| | | | hidden_size:int64 (extern) | hidden_size/ LSTM单元中的 hiddensize | 无限制,建议8对齐 | | per-channel | |
| LSTM | | | 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 |
| | | nouri | input_forget:int64 | input_forget/ cifg变种的选择 | I(T) or 0(F) 目前限定0,即尚不支持 | | | 网小文玩 |
| | | | 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] 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。 | | | |



| 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 (維度补齐) | 量化支持方式 | 多核协同运 行支持情况 |
|----------|--------------|--------------------------------------------|--------------------------------------------------|------------------------|----------------------------------------------------------------------------|-------------|--------|----------------|
| | | | | batch/ 输入的batch | | | | |
| | | | <pre>input_tensor [batch, channel, height,</pre> | 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 | int8 float16 | width]:tensor | height/ 输入的height | 其他方向Concat无限制。 | | per-layer | 已支持 | |
| | | | | width/ 输入的width | | | | |
| | | | axis:int64 | aixs/ 拼接的维度 | 无限制 | | | |
| | | | | batch/ 输入的batch | | | | |
| Mish | 支持 | int8 input_tensor [batch, channel, height, | channel/ 输入的channel | 无限制 | | | 尚不支持 | |
| | | | width]:tensor | height/ 输入的height | | | | |
| | | | wid 输 <i>)</i> | width/ 输入的width | | | | |



| | | | | | TARS300 TA 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 | 1 | | | |
| | | int8 | input_tensor [batch, channel, height, | channel/ 输入的channel | | | | |
| | | float16 | width]:tensor | height/ 输入的height | 无限制 | | | |
| 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 ER制 | | | | | | |
| | | | mode:string | mode/pad模式 | Zupik ipi | | | |
| | | | | batch/ 输入的batch | 无限制 | | | |
| | | | input_tensor [batch, channel, height, | channel/ 输入的channel | | | 尚不 | |
| ReduceMean | 尚不支持 目前 | int8 float16 | width]:tensor | height/ 输入的height | [1,8192] | | | 光工工柱 |
| Reduceivican | 由CPU实现 | 110at10 | | width/ 输入的width | | | | 问小义行 |
| | | | 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 | | | |



| | | | KK9300 W C Operator List | | 和心脉七 1 以从 1 下五寸 | | |
|---------------------------------------------------------------------------------------------|----------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|------------------------|----------------|
| 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 | | | | |
| 问个又行 日削 | int8 | width]:tensor | height/ 输入的height | [1,8192] | | | ル ナート |
| 由CPU实现 | Hoat 16 | | width/ 输入的width | | | per-layer/ per-channel | 尚不支持 |
| | | axes:int64[] | axes/ 指定reduce的轴 | 单轴:无限制, 多轴:{2、3} | | | |
| | | keepdims:int64[] | keepdims/ 是否需要保持维度不 变 | 0 | | | |
| | input_tensor [batch, channel, height width]:tensor | | batch/ 输入的batch | 支持多batch | | | |
| | | input_tensor [batch, channel, height, width]:tensor channel/ 输入的channel height/ 输入的height | [1 8102] | | | | |
| | | | height/ 输入的height | [1,8192] | | | |
| 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 | | | |
| 部分支持 目前NPU仅支持 宽高方向不超过 8倍的整倍数的 最邻近插值缩 放,其余不支持 部分的会 Fallback到CPU 上实现。 | int8 | mode:string | mode/resize采用的模式 | 仅支持nearest | | | |
| | float16 | scales:int64[] | scales/尺寸放大倍数 | 仅支持1-8整数倍 | | per-layer | 尚不支持 |
| | 实现。 | roi:int64[] | roi/进行resize的输入范围 | 仅支持全局([0,0,0,0,1,1,1,1]) | | | |
| | 尚不支持理 間前所PU实现 部所PU有需的整插不支持 可能不要的不是 的,其分的到底,其分的到CPU | 尚不支持目前由CPU实现 int8float16 int8float16 int8 float16 int8 float16 int8 float16 int8 float16 int8 float16 int8 float16 float16 int8 float1 | int8 float16 int8 float16 axes:int64[] keepdims:int64[] input_tensor [batch, channel, height, width]:tensor int8 float16 input_tensor [batch, channel, height, width]:tensor int8 float16 int8 float16 int8 float16 gen part and int8 float16 scales:int64[] scales:int64[] | int8 float 16 int8 | | Mac | |



| | | | | Art milest's | KH3500 TVI C Operator Elist | | | |
|----------|------|---------|------------------------------------------------------|---------------------------------------------------------------|------------------------------------|------------|--------|----------------|
| operator | 支持情况 | 输入数据类型 | 输入 | 设置项/ 输入参数含义 | 约束规格 | 广播支持(维度补齐) | 量化支持方式 | 多核协同运 行支持情况 |
| | | | | batch/ 输入的batch | 无限制 | | | |
| | | int8 | input_tensor [batch, channel, height, | channel/ 输入的channel | -[1,8192] | | | |
| | | float16 | width]:tensor | height/ 输入的height | [1,8192] | | | |
| | | | | width/ 输入的width | [1,8176] | | | |
| | | | | batch_o/ 输出的batch_o | 无限制 | | | |
| Reshape | 部分支持 | | | channel_o/ 输出的channel | [1,8192] | | | 尚不支持 |
| | | | | height_o/ 输出的height | [| | | |
| | | int64 | shape (batch_o, channel_o, height_o, width_o):tensor | width_o/ 输出的width | [1,8176] | | | |
| | | | | [n,c,h1,w1]- >[n,c,h2,w2]/ (h1*w1=h2*w2) | 支持 | | 尚不 | |
| | | | | [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, | channel/ 输入的channel | (1.0100) | | | |
| | | | width]:tensor | height/ 输入的height | [1,8192] | | | |
| Reverse Sequence | | int8 float16 | | width/ 输入的width | [1,8176] | | | 尚不支持 |
| | | | batch_axis:int64 | batch_axis/ 指定是否为batch维度 | 1 | | | |
| | | | time_axis:int64 | time_axis/ 指定是否为time维度 | 0 | | | |
| | | | sequence_lens:int64[] | sequence_lens/ 指定序列翻转的数量 | 仅支持channel数 | | | |
| | | | | batch/ 输入的batch | | | | |
| | | int8 | input_tensor [batch, channel, height, | channel/ 输入的channel | | | | |
| Sigmoid | 支持 | float16 | width]:tensor | height/ 输入的height | 无限制 | | | 尚不支持 |
| | | | | width/ 输入的width | | | | |



| operator | 支持情况 | 输入数据类型 | 输入 | 设置项/ 输入参数含义 | 约束规格 | 广播支持(维度补齐) | | 多核协同运 行支持情况 |
|-------------|------|-----------------|---------------------------------------|------------------------|---------|------------|-----------|----------------|
| | | | | batch/ 输入的batch | | | | |
| M 101 | | 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 | Xit | noatro | width]:tensor | height/ 输入的height | Zupk mj | | | 同小文符 |
| | | | | 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 | 尚不支持 |
| | | | width]:tensor | height/ 输入的height | Zunin | | r m, o. | INV.T.X.IV |
| | | | | 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 |
| Sorphis | Z ₁₇ Hoatto | | width]:tensor | height/ 输入的height | ZGRANI | | por lego. | 101 219 |
| | | | | 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 | | withingtensor | 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 | | per-layer 尚不 | | |
| | | | | width/ 输入的width | | | | | |
| Slice | llice 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 | 尚不支持 | |
| | | | ends:int64[] | ends/ 切分的终止位置 | channel方向Slice时,channel_end要对齐。 对齐量: 8bit数据: 16对齐,16bit数据: 8对齐。 其他方向无限制。 | | | | |
| | | | axes:int64[] | axes/ 选取切分的轴 | 支持任意0~3轴, 支持同时多轴选择 | | | | |
| | | | steps:int64[] | steps/ 选取切分对应轴的步 长 | 1 | | | | |
| | | | | batch/ 输入的batch | | | | | |
| | | | input_tensor [batch, channel, height, | channel/ 输入的channel | 无限制 | | | | |
| | | | width]:tensor | height/ 输入的height | 人上呼に pij | | | | |
| Split | 部分支持 | int8 float16 | | width/ 输入的width | | | per-layer | 尚不支持 | |
| | | | axis:int64 | axis/ 切分的维度 | 无限制 | | per-layer 尚不支 | | |
| | | | 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 | +- b+ | int8 | input_tensor [batch, channel, height, | channel/ 输入的channel | 7.03.44 | | | WITH |
| rann | 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 | 无限制 | | | |
| | | | input_tensor [batch, channel, height, | channel/ 输入的channel | [1,8192] | | | |
| | | | width]:tensor | height/ 输入的height | [1,0192] | | | |
| Transpage | 如八十柱 | int8 float16 | | width/ 输入的width | [1,8176] | | | W.T.+.t. |
| Transpose | 部分支持 | повито | 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->NWHC。 3、以上四种转置无对齐要求。但在满足对齐要求时效率更高。对齐要求为:第1点中参数的c1、c2均要满足8bit数据: 16对齐, 16bit数据: 8对齐。 4、NPU限制项:(1)perm=[0,2,3,1]时,8bit数据时,h1*w1<8176,w1*c1<512; 16bit 数据时,h1*w1<8176,w1*c1<1023。(2)perm=[0,3,1,2]时,h1*w1<8176。(3)perm=[0,3,2,1]时,h1*w1<6184,h2*w2<8*8192,w1<1024。(4)perm=[0,3,1,2]时,in_shape=[n,c,h,w],且w要求8bit数据: 16 对齐,16bit数据:8 对齐,并且c*h<8192。 | | | 尚不支持 |



| | | | | | KK3388 NI U Operator List | 机心放电 1 放仍有限公司 | | | |
|-------------|------|-----------------|----------------------------------------|-------------------------------------------|------------------------------------------|---------------|---------------------------|----------------|--|
| operator | 支持情况 | 输入数据类型 | 输入 | 设置项/ 输入参数含义 | 约束规格 | 广播支持(维度补齐) | 量化支持方式 | 多核协同运 行支持情况 | |
| | | | | batch/ 输入的batch | 无限制 | | | | |
| | | | input_shape [batch, channel, height, | channel/ 输入的channel | 无限制 | | | | |
| | | | width]:tensor | height/ 输入的height | 无限制 | | | | |
| | | | | width/ 输入的width | 仅对首层输入width存在限制 详见 <u>首层输入说明</u> | | | | |
| | | | | num_output/ 输出的channel | | | | | |
| | | | kernel_shape [num_output, | num_input/ 输入的channel | 无限制 | | | | |
| | | nu ke | k | 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/ per-channel | 已支持 | |
| | | | strides [strides_h, strides_w]:int64[] | stride_w/ width方向的strides大 小 | [1,7] | | 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 | [0,13] | | | | |
| | | dill | | pads_bottom/ bottom方向的pads大小 | | | | | |
| | | | group:int64 | group/ group的大小 | 无限制 | | | | |
| | | | dilations (dilati | dila heig | dilations_h/ height方向的dilations 大小 | | | | |
| | | | | dilations_w]:int64[] dil | dilations_w/ widtht方向的dilations 大小 | [1, 32] | | | |
| | 1 | 1 | 1 | I | 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 | | | |
| | | | 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 | | | | |
| Depthwise | 支持 | int8 float16 | | kernel_w/ width方向的kernel大 小 | [1,8] | | per-layer/ per-channel | 已支持 |
| Convolution | 文付 | noatro | | stride_h/ height方向的strides大 小 | | | per-channer | 口文科 |
| | | | strides [strides_h, strides_w]:int64[] | 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 | [0,13] | | | |
| | | | | pads_bottom/ bottom方向的pads大小 | | | | |
| | | | dilations [dilations_h, | dilations_h/ height方向的dilations 大小 | | | | |
| | | | dilations_w]:int64[] | dilations_w/ widtht方向的dilations 大小 | [1, 32] | | | |



| | | | | | KKS500 IV C Operator Elst | | 1 1 10 10 11 10 Z | | |
|---------------------------------|------|---------------------------|-------------------------------------------|------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------|--|
| 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 | | (本文) (行支持) | | |
| | | kernel_shape [num_output, | num_input/ 输入的channel | 无限制 | | | | | |
| | | | num_input, kernel_h, kernel_w]:int64[] | kernel_h/ height方向的kernel大 小 | | | | | |
| | | | | kernel_w/ width方向的kernel大 小 | [1,31] | | | | |
| | | | | stride_h/ height方向的strides大 | | | | | |
| ConvTranspose/ Deconvolution | 支持 | int8 float16 | strides [strides_h, strides_w]:int64[] | stride_w/ width方向的strides大 小 | {1,2,4,8} | | | 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: 不支持 kernel_h * dilations_h - dilations_h - pads_top < 0 | | | | |
| | | | pads_bottom, pads_right]:int64[] | pads_top/ top方向的pads大小 | 不支持 kernel_w * dilations_w - dilations_w - pads_top < 0 不支持 kernel_w * dilations_w - dilations_w - pads_left < 0 不支持 stride_h *(height - 1) - pads_top + 1 < output_h 不支持 stride_w *(width - 1) - pads_left + 1 < output_w | | | | |
| | | | | pads_bottom/ bottom方向的pads大小 | | | | | |
| | | | group:int64 | group/ group的大小 | 支持1 当且仅当num_input=num_output时,支持num_output | | | | |
| | | dilations (dilations h | dilations [dilations_h, | dilations_h/ height方向的dilations 大小 | U 221 | | | | |
| | | | dilations_w]:int64[] | dilations_w/ widtht方向的dilations 大小 | [1, 32] | | | | |



| | La lab lebourt | 4A > W. HII M. 70 | <i>t</i> A.) | 设置项/ | N In In | 라 III 그 나 / / / / / / / / / / / / / / / / / / | | 多核协同运 |
|----------------|----------------------------------|----------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------|-----------------------------------------------|--------------------------|-------|
| operator | 支持情况 | 输入数据类型 | 输入 | 输入参数含义 | 约束规格 | 广播支持(维度补齐) | 重化 文持万式 | 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 | TO STANDARD STANDARD | | per-layer/ per-channel 尚 | |
| Gemm | 尚不支持 目前 | int8 | alpha:double | alpha/ 矩阵 A*B 乘法的scale | 无限制 | | per-layer/ | 尚不支持 |
| Gemm | 由CPU实现 | | beta:double | beta/ 输入C矩阵的scale | 无限制 | | per-channel | 同个文符 |
| | | | transA:int64 | transA/ A矩阵是否转置 | (7) 翰 | | | |
| | | | transB:int64 | (A) ACCORDING A PARTY OF THE PA | | | | |
| | | | input_tensor_1 [batch, K, C]:tensor | batch/ 输入的batch | 双feature时: batch、H无限制 K支持[8,8192],对齐要求为8bit数据: 16对齐, 16bit数据: 8对齐 C支持[32,19384],对齐要求: 32对齐 | | | |
| MatMul | 部分支持 目前该支持仅针 对双feature输入 | int8 float16 | imput_tensor_r [batch, K, C].tensor | K/ 输入的K | K*C <=65532 C*H <=65532 K*H <=65532 | | per-layer/ | 业工士柱 |
| Matwiti | 未来将支持输入 为 feature+constant | noatro | AG . L G . W. | 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 |
| | | | input_tensor_2 [batch, C, H]:tensor input_tensor_2 [batch, C, H]:tensor H/ weight[H,C,1,1]fyConv; 若input_tensor_2 为feature,则转为batch个feature[1,C,H,1] + weight[K,C,1,1]fyConv; 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) | 4) | int8 float16 | input_tensor_2 [batch, channel, N, | K/ 输入的K | K*M <=65532 M*N <=65532 feature+constant時: | | | 尚不支持 |
| | | M]:tensor 若input_tensor_1为feature,则转为batch*channel个feature[K,N,1,1]+ | | | | | | |
| | | | | 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 | | | |



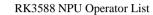
| | | | | | 300 IVI O Operator List | 机心脉电 1 放仍有限公司 | | |
|----------|---------|-----------------|---------------------------------------|--------------------------|-------------------------|---------------|--------|--|
| operator | 支持情况 | 输入数据类型 | 输入 | 设置项/ 输入参数含义 | 约束规格 | 广播支持(维度补齐) | 量化支持方式 | |
| | | | | batch/ 输入的batch | | | | |
| | | int8 float16 | input_tensor [batch, channel, height, | channel/ 输入的channel | | | | |
| | noati o | width]:tensor | height/ 输入的height | 无限制 | | | | |
| Expand | 支持 | | | width/ 输入的width | | | | |
| | ~., | | | batch_o/ 输出的batch_o | | | | |
| | | int64 | 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 | | | | |





| | | | | | 300 TH C Operator Elst | 200-0-08-0-1 08-0-1 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 | |
| | | | | batch/ 输入的batch | | | | |
| | | int8 float16 | x_tensor [batch, channel, height, width]:tensor | channel/ 输入的channel | 无限制 | | | |
| | | int64 | | height/ 输入的height | | | | |
| | | | | width/ 输入的width | | | | |
| | | | | batch/ 输入的batch | | | | |
| | | int8 | y_tensor [batch, channel, height, | channel/ 输入的channel | T: //H /h-l | | | |
| | | float16 int64 | width]:tensor | height/ 输入的height | - 无限制 | | per-layer/ | |
| | | | | width/ 输入的width | | | | |
| | 支持 | | | batch/ 输入的batch | | | | |
| Where | 文 14 | bool | mask_tensor [batch, channel, height, | channel/ 输入的channel | 无限制 | | | |
| | | 0001 | width]:tensor | height/ 输入的height | ル PR 中J | | | |
| | | | | width/ 输入的width | | | | |
| | | | | batch_o/ 输出的batch_o | | | | |
| | | int8 | shape (batch_o, channel_o, height_o, | channel_o/ 输出的channel | 无限制 | | | |
| | | float16 | width_o):tensor | height_o/ 输出的height | | | | |
| | | | | width_o/ 输出的width | | | | |

| | | | | | | 17 - 18 2 3 760 13 17 12 | · · |
|----------|------|-----------------|-------------------------------------------------|--------------------------------------------------------------------------------------------|--------------------------------------------|--------------------------|------------|
| 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 |
| | | int8 float16 | x_tensor [batch, channel, height, width]:tensor | batch/ 输入的batch channel/ 输入的channel height/ 输入的height width/ 输入的width | c*h*w满足如下限制 8bit数据: 16 对齐, 16bit数据: 8对齐 | | |
| exGlu | 支持 | Int64 | axis:int64 | axis/ 切分的维度 | axis ==1 | | per-layer/ |



瑞芯微电子股份有限公司



| operator | 支持情况 | 输入数据类型 | | 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 |
|----------|------|--------|--|----------------|------|------------|--------|----------------|
|----------|------|--------|--|----------------|------|------------|--------|----------------|

瑞芯微电子股份有限公司



| | | 111111111111111111111111111111111111111 | |
|--------------------------------------|------|-----------------------------------------|------|
| Convolution + Relu | 支持 | | 已支持 |
| Convolution + Clip | 支持 | | 已支持 |
| Convolution + PRelu/LeakyR elu | 支持 | | 已支持 |
| Convolution + Add | 支持 | | 已支持 |
| Convolution + Mul | 尚不支持 | | 尚不支持 |
| Convolution + Sigmoid | 尚不支持 | | 尚不支持 |
| Convolution + Tanh | 尚不支持 | 同Convolution | 尚不支持 |
| Convolution + Softplus | 尚不支持 | | 尚不支持 |
| Convolution + HardSigmoid | 尚不支持 | | 尚不支持 |
| Convolution + HardSwish | 尚不支持 | | 尚不支持 |
| Convolution + Elu | 支持 | | 尚不支持 |
| Convolution + Swish | 尚不支持 | | 尚不支持 |
| Convolution + Mish | 尚不支持 | | 尚不支持 |



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



| | | | | | KK3366 W C Operator List | 圳心脉 | 电 1 放历 有限 四 | 11 |
|-------------------------------------------------------------|------------------------------------------------------------------------------------------------|-------------------------------------------------------------------|--------------------------------------------------------|----------------|--------------------------|------------|-------------|----------------|
| operator | 支持情况 | 输入数据类型 | 输入 | 设置项/ 输入参数含义 | 约束规格 | 广播支持(维度补齐) | 量化支持方式 | 多核协同运 行支持情况 |
| Add+Relu | 支持 | 同Add | | | | | | |
| Mul+Relu | 支持 | 同Mul | | | | | | |
| 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)): 0 W),B(C,1,1)): OP(A W),B(scalar)): OP(A W),B(H,W)): OP(A 数是常量时,建议 | A(1,16,32 8),B(16)) A(1,16,32,8),B(1))= a(1,16,32,8),B(32x8 | =C(1,16,32,8) 3))=C(1,16,32,8) 约乘法。乘法在运算效率显著大于除; | 法。 | | | | |



第三章 RV1103/1106 NPU Operator List



| | | | | | 71100 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 | | 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 | 输入的channel height/ 输入的height | | 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。 例子见注释(1) | per-layer/ per-channel |
| | | | | width/ 输入的width | | PI 7.2.2.4.(17) | |
| | | | | batch/ 输入的batch | | | |
| Sub | 支持 | 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 | | | |
| M 1/6 1 | | | input_tensor [batch, channel, height, width]:tensor h | channel/ 输入的channel | | | per-layer/ per-channel |
| Mul/Scale | 支持 | | | height/ 输入的height | 元 [於] | | |
| | | | | width/ 输入的width | | | |



| | | | | | | 71100 NI O Operator List | | |
|-----|-------|-------------|-----------------|-----------------------------------------------------|------------------------|------------------------------|-------------------------------------------------------------------------------------------------------------------|---------------------------|
| 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 |
| | | | | | batch/ 输入的batch | | 支持两个tensor的广播操作,以ONNX默认排列NCHW做说 | |
| Div | | 部分支持 | float16 | input_tensor | channel/ 输入的channel | 工作和 | 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/ |
| | | HP77 (C.19) | | [batch, channel, height, width]:tensor | height/ 输入的height | | 3、OP((N,C,H,W),scalar),即以单个标量做broadcasting 说明:A或B都可以作为广播方。 例子见注释(1) | per-channel |
| | | | | | width/ 输入的width | | | |
| | | | | input_tensor [batch, channel, height, width]:tensor | batch/ 输入的batch | 1 | | per-layer/ per-channel |
| Ma | x | | | | channel/ 输入的channel | [1,8192] | | |
| | | | | | height/ 输入的height | | | |
| | | | | | width/ 输入的width | [1,8176] | | |
| | | | | | batch/ 输入的batch | 1 | | |
| Miı | 1 | 暂不支持 | int8 float16 | input_tensor [batch, channel, height, width)tensor | 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/ |
| | | | | width]:tensor | height/ 输入的height | 2.OP(A(N,C,H 3.OP(A(N,C,H | 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] | | |



| | KV1105/1100 NPO 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 | 1 | | | |
| Global | 支持 | int8 | input_tensor [batch, channel, height, | channel/ 输入的channel | [1,8192] | | per-laver | |
| AveragePool | XII | | width]:tensor | height/ 输入的height | [1,343] (toolkit2支持范围) | | per-layer | |
| | | | | width/ 输入的width | | | | |
| | | | | batch/ 输入的batch | 1 | | | |
| GlobalMaxPool | 支持 | channel/ 输入的channel [1,8192] | | | | | | |
| GiobaliwaxPooi | | int8 | width]:tensor | height/ 输入的height | -[1,343] (toolkit2支持范围) | | per-layer | |
| | | | | width/ 输入的width | [1,545](t001k1t2又打花四) | | | |



| | | | | | 71100 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 |
| | | | | 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 | | | |
| | | | | 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 | | |
| | | | 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 | 10.73 | | |
| | | | pads_right]:int64[] | pads_top/ top方向的pads大小 | [0,7] | | |
| | pads_bottom/bottom方向的pads大小 strides [strides_h, strides_w]:int64[] stride_w/ width方向的strides大小 [1,8] | | | | | | |
| | | stride_l height/ strides_h, strides_w]:int64[] | height方向的strides大 | | | | |
| | | | surues [strides_n, strides_w]:into4[] | stride_w/ width方向的strides大小 | | | |



| | | | | | 7/1100 IVI C Operator List | 机心族电 7 放[| |
|----------|------|-----------------------------------------|-----------------------------------------------------|---------------------------------------------|----------------------------|------------|-----------|
| operator | 支持情况 | 输入数据类型 | 输入 | 设置项/ 输入参数含义 | 约束规格 | 广播支持(维度补齐) | 量化支持方式 |
| | | | | 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 | [1,9102] | | |
| | | | | 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 | | | |
| MaxPool | 支持 | int8 kernel_shape [kerkernel_w]:int64[] | kernel_shape [kernel_h, | kernel_h/ height方向的kernel大小 | - 无限制,NPU支持[1,7];其它由CPU支持。 | | per-layer |
| | | | kernel_w]:int64[] | kernel_w/ width方向的kernel大小 | | | |
| | | | | pads_left/ left方向的pads大小 | | | |
| | | | pads [pads_top, pads_left, pads_bottom, | pads_right/ right方向的pads大小 | [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 | | |
| | | | stride_h/ height方向的strides大 小 | | | | |
| | | stri | strides [strides_h, strides_w]:int64[] | stride_w/ width方向的strides大小 | [1,8] | | |
| | |] | | I | | | |



| | | | | | 71100 NI C Operator List | 和心族电 7 放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 |
| | | | 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 | XIII | lioatro | input_tensor [batch, channel, height, | channel/ 输入的channel | 无限制 | | per-channel |
| | | | width]:tensor | height/ 输入的height | √7 hi≤ thì | | |
| | | | | 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 | | | |
| | | | | height/ 输入的height | 无限制 | | |
| | | | | width/ 输入的width | | | |
| | | | layernorm_weight [channel, height, | channel/ 输入的channel | 等于input_channel | | |
| Layer | 业工士柱 | float16 | width]:tensor(const) layernorm_bias [channel, height, | height/ 输入的height | 等于input_height | | |
| Normalization | 尚不支持 | noatro | width]:tensor(const) | width/ 输入的width | 等于input_width | | per-layer |
| | | normalized_sha | normalized_shape:int64[] | normalized_ shape /参与每一批归一化的 Feature的尺寸 | NPU仅支持,包含除第0维(batch维)以外的其他所有维度, 如input_shape[n,c,h,w],仅支持normalized_shape[c,h,w], 如input_shape[n,c,h],仅支持normalized_shape[c,h], 如input_shape[n,c],仅支持normalized_shape[c], 其余情况会转到CPU执行。 | | |
| | | | elementwise_affine:int64 | elementwise_affine/ 是否具有可学习数 | 0 或 1(默认为 0)。 当为1时拥有LayerNorm.weight与LayerNorm.bias,仅支持weight/bias 的尺寸: elementwise_shape与normalized_shape一致;当为0时 LayerNorm.weight为全1值,LayerNorm.bias为全0值。 | | |
| | | | eps:double | eps/ 防止除法溢出的偏移参 数 | 无限制 | | |



| operator | 支持情况 | 输入数据类型 | 输入 | 设置项/ 输入参数含义 | 约束规格 | 广播支持(维度补齐) | 量化支持方式 |
|------------|------|--------------------------------------------------|--------------------------------------------------|------------------------|------|------------|-----------|
| | | batch/ 输入的batch | | | | | |
| Clip/ReLU6 | ++4. | int8 float16 | input_tensor [batch, channel, height, | channel/ 输入的channel | | | nor lover |
| Сприссеоб | 支持 | rioati 6 | width]:tensor | height/ 输入的height | 无限制 | | per-layer |
| | | | | width/ 输入的width | | | |
| | | | | batch/ 输入的batch | 无限制 | | |
| Elu | 支持 | int8 float16 | input_tensor [batch, channel, height, | channel/ 输入的channel | | | |
| Eiu | XH | width]:tensor height/ 输入的height width/ 输入的width/ | | | | | |
| | | | | 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 | | |
| Solu | XIV. | | width]:tensor height/ 输入的height width/ 输入的width/ | height/ 输入的height | | | |
| | | | | | | | |



| | | | | 11,1100 | 71100 NPU Operator List | | 3 13 17 17 17 |
|------------|------|-----------------|-----------------------------------------------------|------------------------|-------------------------|-------------|---------------------------|
| operator | 支持情况 | 输入数据类型 | 输入 | 设置项/ 输入参数含义 | 约束规格 | 广播支持 (维度补齐) | 量化支持方式 |
| | | | input_tensor [batch, channel, height, | batch/ 输入的batch | 1 | | |
| Relu | 支持 | int8 float16 | | channel/ 输入的channel | | | per-layer |
| Refu | XIT | 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 | | | per-layer |
| 26 my rota | ZN | | | height/ 输入的height | 无限制 | | |
| | | | | width/ 输入的width | | | |
| | | | | batch/ 输入的batch | 1 | | |
| | | | | channel/ 输入的channel | [1,8192] | | |
| PRelu | 支持 | int8 float16 | input_tensor [batch, channel, height, width]:tensor | 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 | | |



| | | | | | 71100 NFO 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 | 1 | | | |
| | | | | 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 | I | | | |
| | 尚不支持 | | sequence_size :int64 (extern) | sequence_size/ 指定GRU输入的seqsize | 无限制,建议4对齐 | | | |
| GRU | 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。 | IU算子,参 明(extern) GRU独有的 | hidden_size:int64 (extern) | hidden_size/ GRU单元中的 hiddensize | 无限制,建议8对齐 | | per-layer | |
| | | | linear_before_ reset:int64 | linear_before_ reset/ LBR变种的选择 | 1(T) or 0(F) | | | |
| | | | input | input_layout:string(extern) | input_layout/指定与对 应输入shape含义一致 的layout | 1、snc: 指定layout对应的输入shape为[seqs, batches, input_size] 2、(sn)c: 指定layout对应的输入shape为[seqs*batches, input_size,1,1] 要求填写指定的layout,同时要求填写该op实际对应的batch_size、sequence_size、hidden_size。 | | |
| | | | output_layout:string (extern) | output_layout/指定与对应输出shape含义一致的layout | 1、sbnc: 指定layout对应的输出shape为[seqs,directions,batches, hidden_size] 2、(sn)c: 指定layout对应的输出shape为[seqs*batches, directions*input_size,1,1] 要求填写指定的layout,同时要求填写该op实际对应的batch_size、sequence_size、hidden_size。directions>1时仅支持batches=1。 | | | |



| | | | | 100 | 71106 NPU Operator List | · · · · · · · · · · · · · · · · · · · | 7 11 10 4 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 channer | | | | | | | | | | | | | |
| LSTM | 部分支持 LSTM 扩展以及变体 | 0 | proj_size:int64 (extern) | proj_size/ LSTM单元存在 projection时的proj_size | 0<=proj_size<=hiddensize | | | | | | | | | | | | | | | |
| LSTM | 命名为exLSTM算子, 参数项中指明 (extern)的项为 exLSTM独有的参数项 。 | 1 | 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 | | | | | | | | | | | | | | | |
| | | | 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 | I(T) or 0(F) 目前限定0,即尚不支持 | | | | | | | | | | | | | | | |
| | | | | | | | | į | | | | | | | inp | input_layout:string (extern) | input_layout/指定与对应输入shape含义一致 | 1、snc: 指定layout对应的输入shape为[seqs, batches, input_size] 2、(sn)c: 指定layout对应的输入shape为[seqs*batches, input_size,1,1] | | |
| | | | | 的layout | 要求填写指定的layout,同时要求填写该op实际对应的batch_size、sequence_size、hidden_size。 | | | | | | | | | | | | | | | |
| | | output_layout:string (extern) | output_layout: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] | | | | | | | | | | | | | | | |
| | | | | 应输出shape含义一致的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。 | | | | | | | | | | | | | | | |



| | | | | | 71100 IVI O Operator List | 河心厥七 1 双 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channe1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 | 限制 | | |
| | | | | batch/ 输入的batch | | | |
| Mish | 支持 | int8 float16 | 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 | | | |



| | | | | | | 71100 141 C Operator Elst | 和心脉毛 1 及[| | |
|----|-----------|------|-----------------|---------------------------------------|-------------------------------------------------------------------------------------------|-----------------------------------------|------------|--------|--|
| op | erator | 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 | | | | |
| | | | | width]:tensor | height/ 输入的height | 无限制 | | | |
| | | | | | width/ 输入的width | [1,8176] | | | |
| Pa | d | 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 | end, 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] | | | |
| Re | educeMean | | int8 float16 | | width/ 输入的width | | | | |
| | | | | axes:int64[] | axes/ 指定reduce的轴 | 单轴:无限制,多轴:{2,3} | | | |
| | | | keep | | keepdims:int64[] | keepdims/ 是否需要保持维度不变 | 0 | | |



| | | | | | 71100 NI O Operator Elst | 机心放电 1 放儿 | |
|--------------|----------|-----------------|-----------------------------------------------------------------------------------------|--------------------------|--------------------------------------------------------------------------------------------------------------------------------------------|------------|-------------|
| 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 | | | |
| Parking Com- | N T + 4: | int8 | width]:tensor | height/ 输入的height | [1,8192] | pe | per-layer/ |
| ReduceSum | 尚不支持 | float16 | | width/ 输入的width | | | per-channel |
| | | | axes:int64[] | axes/ 指定reduce的轴 | 单轴:无限制,多轴:{2,3} | | |
| | | | keepdims:int64[] | keepdims/ 是否需要保持维度不变 | 0 | | |
| | | int8 float16 | input_tensor [batch, channel, height, 6 width]:tensor (input_tensor的维度为4维时看作nchw) | batch/ 输入的batch | 约束规格: (1) height * width * type_bytes <= 130816; (2) input_tensor非四维时,shape无限制 | | |
| | | | | channel/ 输入的channel | | | |
| | | routio | | height/ 输入的height | | | |
| | | | | width/ 输入的width | | | |
| Reshape | 部分支持 | | | batch_o/ 输出的batch_o | | | |
| | | | shape (batch_o, channel_o, height_o, | channel_o/ 输出的channel | 计算量: alignment=16/type_bytes; | | |
| | | int64 | width_o):tensor (输出shape指定维度为4维时看作nchw) | height_o/ 输出的height | 约束规格: (1) height_o * width_o * type_bytes <= 65535; (2) Align(height_o * width_o, alignment) <= 8192; (3) 输出shape非四维时,shape无限制 | | |
| | | | | width_o/ 输出的width | で 間にいいのなーサドユミE中リ 、SHAPCノレドK中リ | | |



| | | | | | 7 1700 TAT C Operator East | | T T T T T T T T T T T T T T T T T T T | | | | |
|---------------------|-------------------------------------|-----------------|-----------------------------------------------------|-----------------------------|----------------------------------------------------|------------|---------------------------------------|--------------------|-----|--|--|
| operator | 支持情况 | 输入数据类型 | 输入 | 设置项/ 输入参数含义 | 约束规格 | 广播支持(维度补齐) | 量化支持方式 | | | | |
| | | | | batch/ 输入的batch | 支持多batch | | | | | | |
| | | | input_tensor [batch, channel, height, width]:tensor | channel/ 输入的channel | H 9192 | | | | | | |
| | 部分支持 目前NPU仅支持宽高 | | | 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 | | |
| | | | input_tensor [batch, channel, height, | channel/ 输入的channel | H 9192 | | | | | | |
| | | int8 float16 | width]:tensor | 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 | | | | | | | |
| Sigmoid | + 4± | 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 | | | | | | | |
| | | | wie | width/ 输入的width | | | | | | | |



| _ | | | | | | 71100 141 C Operator Eist | | |
|---|-----------------|------------|-----------------|---------------------------------------|------------------------|---------------------------|------------|-----------|
| (| 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 |
| | | | | 1 | 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 |
| | rai u.s.iginoid | 义 ੰ | Hoatro | width]:tensor | height/ 输入的height | | | per-rayer |
| | | | | | width/ 输入的width | | | |
| | | | | | batch/ 输入的batch | | | |
| 9 | Swish | 支持 | int8 float16 | input_tensor [batch, channel, height, | channel/ 输入的channel | 无限制 | | |
| | W 1511 | Д П | TOMETO | width]:tensor | height/ 输入的height | | | |
| | | | | | width/ 输入的width | | | |



| KVIIIIII.O.III O Opera | | | | | | MODIFIED THE | | |
|------------------------|-----------|------------|-----------------|---------------------------------------|------------------------|----------------------------------------------------|------------|-----------|
| (| 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 | | | |
| , | HardSwish | 支持 | int8 float16 | input_tensor [batch, channel, height, | channel/ 输入的channel | 无限制 | | per-layer |
| 1 | | 文 持 | Hoatro | 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 | Hoatro | width]:tensor | height/ 输入的height | 无限制 | | per-layer |
| | | | | | width/ 输入的width | | | |
| | | | | | batch/ 输入的batch | 无限制 | | |
| | | | | input_tensor [batch, channel, height, | channel/ 输入的channel | 硬件支持[1,4096] 建议8对齐 | | |
| 5 | oftmax | 部分支持 | float16 | width]:tensor | height/ 输入的height | · 1 T. P. P. A. | | per-layer |
| | | | | | width/ 输入的width | axis=1, 无限制 axis=3/-1,width[1, 8192], height无限制 | | |
| | | | | axis:int64 a | axis/ 做softmax的轴 | 1,即channel方向 | | |



| | | | | | 71100 NI O Operator Elst | 河心灰屯 1 及[| |
|----------|------|-----------------|---------------------------------------|---------------------------------------|---------------------------------------------------------------------------------------------|------------|-----------|
| 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/ fibatch, channel, height. | 77 177 441 | | |
| | | | 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 | | |
| | | | | 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 | 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 | | | |
| T. 1 | bt | int8 | input_tensor [batch, channel, height, | channel/ 输入的channel | 无限制 | | , |
| Tanh | 支持 | float16 | width]:tensor | height/ 输入的height | 7 CPR (P) | | 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 | 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8bit数据: 16对齐,16bit数据: 8对齐。 4. NPU限制项:(1)perm=[0,2,3,1]时,8bit数据时,h1*w1<8176,w1*c1<1023。(2)perm=[0,3,1,2]时,h1*w1<8176。(3)perm=[0,3,1,2]时,h1*w1<8176。(3)perm=[0,3,2,1]时,h1*w1<1024。 | | |



| | | | | | 71100 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 | 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, | 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] | | |
| | | int8 | | kernel_w/ width方向的kernel大小 | [1,51] | | per-layer/ per-channel |
| Convolution | 支持 | | 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_bottom, | pads_right/ right方向的pads大小 | [0,15] | | |
| | | | pads_right]:int64[] | pads_top/ top方向的pads大小 | [0,15] | | |
| | | | | pads_bottom/ bottom方向的pads大小 | | | |
| | | | group:int64 | group/ group的大小 | 无限制 | | |
| | | | 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 |
|-----------------------|------|--------|-----------------------------------------|------------------------------------------|-------------------------------------|------------|---------------------------|
| | | | | 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, | num_input/ 输入的channel | 无限制 | | |
| | 支持 | int8 | kernel_h, kernel_w]:int64[] | kernel_h/ height方向的kernel大小 | [1,8] | | per-layer/ per-channel |
| Depthwise Convolution | | | | kernel_w/ width方向的kernel大小 | [1+0] | | |
| Departition . | | | | stride_h/ height方向的strides大 小 | [1,7] | | per-enamer |
| | | | sates [sates_ii, sates_n]iii.c i | 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 | [over] | | |
| | | | | pads_bottom/ bottom方向的pads大小 | | | |
| | | d d | dilations [dilations_h, | dilations_h/ height方向的dilations大 小 | [1, 32] | | |
| | | | dilations_w]:int64[] d | dilations_w/ widtht方向的dilations大 小 | | | |



| | | | | | • | 圳心城屯 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 | 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, | num_input/ 输入的channel | 无限制 | | |
| | | | kernel_h, kernel_w]:int64[] | kernel_h/ height方向的kernel大小 | | | |
| | 支持 | int8 | | kernel_w/ width方向的kernel大小 | [1,31] | | per-layer/ per-channel |
| ConvTranspose/ Deconvolution | | | 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 | | | |
| | | | | 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 | | |
| | | | 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 | 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 | | | |



| | RV1103/1106 NPU Operator List | | | | | | | וי באיוווי | |
|---|-------------------------------|-------------|-----------|-------------------------------------|-------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|-------------|--|
| ď | 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_1 [M, K]:tensor | M,K,N/ | | | | |
| | | | | input_tensor_2 [K,N]:tensor | 输入数据的形状 | おうがaunui大死,を3米円がaunui | | | |
| | Lamma | 尚不支持 目前由CPU | | alpha:double | alpha/ 矩阵A*B乘法的scale | 无限制 | | per-layer/ | |
| | | 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 | | | | |
| | | | 由CPU int8 | | | batch/ 输入的batch | | | |
| | 1atMul | 尚不支持 目前由CPU | | input_tensor_1 [batch, K, C]:tensor | K/ 输入的K | 双feature时: batch、H无限制 K支持[8,8192],对齐要求为8bit数据: 16对齐,16bit数据:8对齐 C支持[32,19384],对齐要求为32对齐 feature+constant时: | | per-layer/ | |
| | rauviui | 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 | |
| | | | | I - | 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 | |
|----------|------|---------|-----------------------------------------------------|--------------------------|------------------------|------------|--------|--|
| | | | input_tensor [batch, channel, height, width]:tensor | batch/ 输入的batch | | | | |
| | | int8 | | 输 | channel/ 输入的channel | | | |
| | | float16 | | height/ 输入的height | 无限制 | | | |
| Expand | 支持 | | | width/ 输入的width | | | | |
| | | int64 | | 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 |
|----------|------|-----------------|-------------------------------------------------|-------------------------------------------|-------------|------------|------------|
| | | | | batch/ 输入的batch | | | |
| | | int8 | x_tensor [batch, channel, height, width]:tensor | channel/ 输入的channel | 无限制 | | |
| | | int64 | | height/ 输入的height | | | |
| | | | | width/ 输入的width | | | |
| | | | | batch/ 输入的batch | | | |
| | | int8 | y_tensor [batch, channel, height, | channel/ 输入的channel | 无限制 | | |
| | | int64 | width]:tensor | height/ 输入的height | 7/L PKC 1PU | | per-layer/ |
| | | | | width/ 输入的width | | | |
| | 支持 | | | batch/ 输入的batch | - 无限制 | | |
| Where | ΧN | | | channel/ 输入的channel | | | |
| | | boo1 | width]:tensor | height/ 输入的height | 7/L PPC 1PU | | |
| | | | | width/ 输入的width | | | |
| | | | | batch_o/ 输出的batch_o | | | |
| | | int8 | shape (batch_o, channel_o, height_o, | channel_o/ 输出的channel | 无限制 | | |
| in | | width_o):tensor | height_o/ 输出的height | ען אין אין אין אין אין אין אין אין אין אי | | | |
| | | | | width_o/ 输出的width | | | |

| | | | | | | 117 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 127 - 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 |
| | | int8 float16 | x_tensor [batch, channel, height, width]:tensor | batch/ 输入的batch channel/ 输入的channel height/ 输入的height width/ 输入的width | c*h*w 满足如下限制 8bit数据: 16 对齐,16bit数据: 8对齐 | | per-layer/ |
| exGlu | 支持 | Int64 | axis:int64 | axis/ 切分的维度 | axis ==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 |
|----------------------------------|------|--------------|----|----------------|------|-------------|--------|
| 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 | | | | | | |



| | | | | | 71100 Til C Operator Elist | >III.G. IW. G. 1 W. | |
|------------------------------------|------|----------------|----|----------------|----------------------------|---------------------|--------|
| operator | 支持情况 | 输入数据类型 | 输入 | 设置项/ 输入参数含义 | 约束规格 | 广播支持(维度补齐) | 量化支持方式 |
| 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 | | | | | | |



| | | | | K V 1103 | | | |
|--------------------------------------------|------|----------------|-----------|----------------|------|------------|--------|
| 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 | nvolution | | | | |
| Depthwise Convolution + Softplus | 尚不支持 | | | | | | |
| Depthwise Convolution + HardSigmoid | 尚不支持 | | | | | | |
| Depthwise Convolution + HardSwish | 尚不支持 | | | | | | |
| Depthwise Convolution + Elu | 支持 | | | | | | |
| Depthwise Convolution + Swish | 尚不支持 | | | | | | |
| Depthwise Convolution + Mish | 尚不支持 | | | | | | |



| operator | 支持情况 | 输入数据类型 | 输入 | 设置项/ 输入参数含义 | 约束规格 | 广播支持(维度补齐) | 量化支持方式 |
|----------------------|------|--------------|----|----------------|------|------------|--------|
| Add+Relu | 支持 | 同Add | | | | | |
| Mul+Relu | 支持 | 同Mul | | | | | |
| 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 | | * . * * |
|---|--------------|------|-----------------|--------------------------------------------------------|------------------------------|------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------|
| | 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 |
| | Add/Bias | | | | batch/ 输入的batch | | 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 : | |
| | | 支持 | int8 float16 | input_tensor [batch, channel, height, width]:tensor | 输入的channel height/ 输入的height | 无限制 | 1.OP(A(N,C,H,W),B(N,C,H,W)), 即两个维度相同的tensor进 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 4.OP(A(N,C,H,W),B(H,W)), 即HW维度做broadcasting | per-layer/ per-channel |
| | | | | | width/ 输入的width | | 说明: A或B都可以作为广播方。 例子见注释 (1) | |
| • | | | | | batch/ 输入的batch | | | |
| | Sub | 支持 | 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: l.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。 4.OP(A(N,C,H,W),B(H,W)),即HW维度做broadcasting | per-layer/ per-channel |
| | | | | | height/ 输入的height | | | |
| | | | | | width/ 输入的width | | | |
| | | | | | batch/ 输入的batch | | | |
| | Mul/Scale 支持 | | int8 | input_tensor | channel/ 输入的channel | | | per-layer/ |
| | | 文符 | | [batch, channel, height, width]:tensor heig 输入 | height/ 输入的height | 7元[PR] 市) | | per-channel |
| | | | | | width/ 输入的width | | | |



| | | | | | | 302 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 | | THE A STATE OF THE PROPERTY OF | |
| | Div | 部分支持 | float16 | input_tensor | channel/ 输入的channel | 工作和 | 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/ |
| | JIV | 即 刀又付 | Hoatro | [batch, channel, height, width]:tensor | height/ 输入的height | | 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 | | | |
| | | | int8 float16 | input_tensor [batch, channel, height, width]:tensor | batch/ 输入的batch | 1 | | per-layer/ per-channel |
| | Иах | 暂不支持 | | | 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。 | |
| | THE STATE OF THE S | | | | height/ 输入的height | | | |
| | | | | | width/ 输入的width | [1,8176] | | |
| | | | | | batch/ 输入的batch | 1 | | |
| | Min 暂不 | 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.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/ |
| | | | Hoat16 | width]:tensor h 第 | height/ 输入的height | 2.OP(A(N,C,H,W),B(C,1,1)), 即 C 维 度 做 broadcasting | 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 | [1,8176] | | |



| | | KKC502 IV C Operator List | | | | | | |
|--------|-----------|---------------------------|-----------------|-----------------------------------------------------|------------------------|------------------------------|------------|-----------|
| operat | 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 |
| | | | | | batch/ 输入的batch | 1 | | |
| Globa | ıl | 支持 | int8 | input_tensor [batch, channel, height, | channel/ 输入的channel | [1,8192] | | per-layer |
| Avera | gePool | X IV | | width]:tensor | height/ 输入的height | [1,343] (toolkit2支持范围) | | per layer |
| | | | | | width/ 输入的width | [1,543] (001K162 X] [4,613] | | |
| | | | | input_tensor [batch, channel, height, width]:tensor | batch/ 输入的batch | 1 | _ | |
| Globa | ılMaxPool | | int8 float16 | | channel/ 输入的channel | [1,8192] | | per-layer |
| | | 支持 | float16 | | height/ 输入的height | [1,343] (toolkit2支持范围) | | |
| | | | | | width/ 输入的width | [Li, 0+4](COUINICZX 特化四) | | |



| | | | | 302 NPO Operator List | 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 | | |
| | | | 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 | | |
| | 支持 | int8 | 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 | | | 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 | 10.71 | | |
| | | | pads_right]:int64[] | pads_top/ top方向的pads大小 | [0,7] | | |
| | | | | pads_bottom/ bottom方向的pads大小 | | | |
| | | | 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] | | |



| | KK5502 TV U Opciator 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 | | |
| | | | input_tensor [batch, channel, height, | channel/ 输入的channel | [1,8192] | | |
| | | | width]:tensor | height/ 输入的height | H 0102 | | |
| | | | | 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 | | |
| | 支持 | int8 float16 | dilations [dilations_h, | dilations_h/ height方向的dilations大小 | | | per-layer |
| | | | dilations_w]:int64[] | dilations_w/ widtht方向的dilations大 小 | | | |
| MaxPool | | | 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。 | | |
| | | | | 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 | | | |
| | | | 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 | [1 0] | | |
| | | | | stride_w/ width方向的strides大小 | [1,8] | | |



| | | | | 机心脉毛 1 从 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 |
| | | | 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 | 文 44 | noatro | input_tensor [batch, channel, height, | channel/ 输入的channel | 工明和 | | 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 | | |
| | | float16 | input_tensor [batch, channel, height, width]:tensor | channel/ 输入的channel | | | |
| | | | | height/ 输入的height | 无限制 | | |
| | | | | width/ 输入的width | | | |
| | | | layernorm_weight [channel, height, width]:tensor(const) layernorm_bias [channel, height, width]:tensor(const) normalized_shape:int64[] | channel/ 输入的channel | 等于input_channel | | |
| Layer | +·kt | | | height/ 输入的height | 等于input_height | | |
| Normalization | 支持 | | | width/ 输入的width | 等于input_width | | per-layer |
| | | | | normalized_ shape /参与每一批归一化的 Feature的尺寸 | NPU仅支持,包含除第0维(batch维)以外的其他所有维度,如input_shape[n,c,h,w],仅支持normalized_shape[c,h,w],如input_shape[n,c,h],仅支持normalized_shape[c,h],如input_shape[n,c],仅支持normalized_shape[c],其余情况会转到CPU执行。 | | |
| | | | elementwise_affine:int64 | elementwise_affine/ 是否具有可学习数 | 0 或 1(默认为 0)。 当为1时拥有LayerNorm.weight与LayerNorm.bias,仅支持weight/bias 的尺寸: elementwise_shape与normalized_shape一致; 当为0时 LayerNorm.weight为全1值,LayerNorm.bias为全0值。 | | |
| | | | eps:double | eps/ 防止除法溢出的偏移参 数 | 无限制 | | |



| | • | | | | 5502 IVI 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 | | |
| Clip/ReLU6 | +++ | int8 float16 | input_tensor [batch, channel, height, | channel/ 输入的channel | | | per-layer |
| Спр/кедоо | 支持 | noatro | width]:tensor | height/ 输入的height | 无限制 | | per-rayer |
| | | | | width/ 输入的width | | | |
| | | | input_tensor [batch, channel, height, width]:tensor | batch/ 输入的batch | | | |
| Elu | 支持 | int8 float16 | | channel/ 输入的channel | 无限制 | | |
| Elu | X19 | rioatro | | 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 | | | |



| | | | | | | 302 IVI O Operator List | 机心族电 7 放仍有限公司 | |
|---|-------------|--------------------|-----------------|-----------------------------------------------------|------------------------|-------------------------|---------------|---------------------------|
| (| 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 | | |
| | Relu | +- t4. | int8 float16 | input_tensor [batch, channel, height, | channel/ 输入的channel | | | per-layer |
| 1 | | 支持 | Hoatro | width]:tensor | height/ 输入的height | 无限制 | | per-rayer |
| | | | | | width/ 输入的width | | | |
| - | | int8 支持 float16 | | batch/ 输入的batch | 1 | | | |
| | andra Darla | | | input_tensor [batch, channel, height, width] tensor | channel/ 输入的channel | 无限制 | | |
| , | .eakyRelu | | | | height/ 输入的height | | | per-layer |
| | | | | | width/ 输入的width | | | |
| | | | | | batch/ 输入的batch | 1 | | |
| | | | | | channel/ 输入的channel | | | |
| 1 | Relu | 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 | | |



| | | | | 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 | 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 | | per-layer |
| | 尚不支持 GRU 扩展以及变体命 名为exGRU算子,参 数项中指明(extern) | float16 | 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,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。 | | |



| | | | | | 302 IVI 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 | 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。 | | | | | |
| | 部分支持 LSTM 扩展以及变体 | | 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 | | | | | |
| | LSTM 扩展以及变体 命名为exLSTM算子, 参数项中指明 (extern)的项为 | | 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/ per-channel | | | |
| | | | hidden_size/ LSTM单元中的 hiddensize 无限制,建议8对齐 | | per-channel | | | | | |
| LCTM | | int8 | 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 | | 明 input_forget:int64 | | | | | | | | |
| | exLSTM独有的参数项。 | | caffe框架下的indicator | Caffe框架下,启用该功能要求输入indicator,工具端自动配置,无 | | | | | | |
| | | | has_projection:int64 (extern) | | | | | | | |
| | | | | | | | input_layout:string (extern) | input_layout/指定与对应输入shape含义一致 | 2、(sn)c: 指定layout对应的输入shape为[seqs*batches, | |
| | | | | 的layout | 要求填写指定的layout,同时要求填写该op实际对应的batch_size、sequence_size、hidden_size。 | | | | | |
| | | | output_layout:string (extern) | output_layout/指定与对 应输出shape含义一致 | 1、sbnc: 指定layout对应的输出shape为[seqs,directions,batches, hidden_size] 2、(sn)c: 指定layout对应的输出shape为[seqs*batches, directions*input_size,1,1] | | | | | |
| | | | | 的layout | 要求填写指定的layout,同时要求填写该op实际对应的batch_size、sequence_size、hidden_size。 directions>1时仅支持batches=1。 | | | | | |



| | | | | | South C Sporter East | - IN-C DX E 1 7X D | |
|----------|------|-----------------|---------------------------------------|------------------------|-----------------------------------------------------------------------------------|--------------------|-----------|
| 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channe1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 | | | |
| Mish | 支持 | int8 float16 | 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 | | | |



| | | | | | | 302 TVI C Operator Eist | | 7 7 M A A |
|----|-----------|------------------|-----------------------|---------------------------------------|-------------------------------------------------------------------------------------------|------------------------------------|------------|-----------|
| op | erator | 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 | | | |
| | | | | width]:tensor | height/ 输入的height | 无限制 | | |
| | | | | | width/ 输入的width | [1,8176] | | |
| Pa | d | 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] | | |
| Re | educeMean | 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:int64[] | keepdims/ 是否需要保持维度不变 | 0 | | | |



| operator | 支持情况 | 输入数据类型 | 输入 | 设置项/ 输入参数含义 | 约束规格 | 广播支持 (维度补齐) | 量化支持方式 |
|-----------|----------|-----------------|--------------------------------------------|--------------------------|----------------------------------------------------------------------------------------------------------------------------------|-------------|---------------------------|
| | | | | batch/ 输入的batch | 无限制 | | |
| | | | input_tensor [batch, channel, height, | channel/ 输入的channel | | | |
| ReduceSum | 尚不支持 | int8 float16 | width]:tensor | height/ 输入的height | [1,8192] | | per-layer/ per-channel |
| reducesum | H4/1-X10 | nouro | | width/ 输入的width | | | per-channel |
| | | | axes:int64[] | axes/ 指定reduce的轴 | 单轴:无限制,多轴:{2,3} | | |
| | | | keepdims:int64[] | keepdims/ 是否需要保持维度不变 | 0 | | |
| | | int8 float16 | input_tensor [batch, channel, height, | batch/ 输入的batch | | | |
| | | | | channel/ 输入的channel | | | |
| | | | (input_tensor的维度为4维时看作nchw) | height/ 输入的height | | | |
| | | | | width/ 输入的width | | | |
| Reshape | 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 | 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 <= INT32_MAX; (2) Align(height_o * width_o, alignment) <= 8192*8192; (3) 输出shape非四维时,shape无限制 | | |
| | | | w M | | 188 LISHAPO-TEF-ESELLI J. SHAPO-ZUPKIRI | | |



| | | | | | 1302 IVI O Operator List | 机心脉毛 1 从 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 | 支持多batch | | | |
| | | | | channel/ 输入的channel | (1.0102) | | | |
| | | | width]:tensor | height/ 输入的height | [1,8192] | | | |
| Resize | | int8 float16 | 116 | 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 | |
| | | | 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]) | | | |
| | | | | batch/ 输入的batch | 无限制 | | | |
| | | int8 | input_tensor [batch, channel, height, | channel/ 输入的channel | [1,8192] | | | |
| | | | width]:tensor | width]:tensor | 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 | | | | |
| G: · · · | -to let- | int8 | input_tensor [batch, channel, height, | channel/ 输入的channel | 无限制 | | | |
| Sigmoid | 文持 | 去技 float16 | width]:tensor | height/ 输入的height | | | | |
| | | | 新 明 新 | | | | | |



| | | | | | | 302 NFO 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 |
| | | ΧN | TIOMET O | width]:tensor | height/ 输入的height | | | por aryon |
| | | | | | width/ 输入的width | | | |
| | | | | | batch/ 输入的batch | 无限制 | | |
| | Swish | 支持 | int8 float16 | input tensor [batch, channel, height, | channel/ 输入的channel | | | |
| | 交 | ΧN | | width]:tensor | height/ 输入的height | | | |
| | | | | | width/ 输入的width | | | |



| | | | | | | 302 TVI C Operator List | | 1 D IN A 11 |
|----------|---------------|-----------------|---------------------------------------|---------------------------------------|------------------------|-------------------------------------|------------|-------------|
| 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 | | | |
| HardSw | t. | +4: | int8 float16 | input_tensor [batch, channel, height, | channel/ 输入的channel | 无限制 | | |
| HardSw | visn | 支持 | float16 | width]:tensor | height/ 输入的height | | | per-layer |
| | | | | | width/ 输入的width | | | |
| | | | | | batch/ 输入的batch | | | |
| Softplus | int8 | int8 float16 | input_tensor [batch, channel, height, | channel/ 输入的channel | | | man layer | |
| Sortpius | is | 支持 | noatro | width]:tensor | height/ 输入的height | 无限制 | per-layer | |
| | | | | | width/ 输入的width | | | |
| | | | | | batch/ 输入的batch | 无限制 | | |
| | | | | input_tensor [batch, channel, height, | channel/ 输入的channel | 硬件支持[1,8192] | | |
| Softmax | Softmax 支持 | | int8 float16 | width]:tensor | height/ 输入的height | axis=1, 无限制 | | per-layer |
| | | | a | | width/ 输入的width | axis=3/-1,width[1, 8192], height无限制 | | |
| | | | | | axis/ 做softmax的轴 | 1,3,即channel和width方向 | | |



| | | | | | 302 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 |
| | | | | 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 | 入上ド に申り | | |
| | | | | 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: 终止位置 8bit数据: 16对齐,16bit数据: 8对齐。 其他方向无限制。 | | |
| | | | axes:int64[] | axes/ 选取切分的轴 | 支持任意0~3轴, 支持同时多轴选择 | | |
| | | | steps:int64[] | steps/ 选取切分对应轴的步长 | 1 | | |
| | | | | batch/ 输入的batch | | | |
| | | | input_tensor [batch, channel, height, | channel/ 输入的channel | 无限制 | | |
| | | | width]:tensor | height/ 输入的height | Д. рк. ф) | | |
| 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 | 支持 | int8 float16 | input_tensor [batch, channel, height, | channel/ 输入的channel | 无限制 | | por lavor |
| 1 ann | 文 村 | Hoatro | width]:tensor | height/ 输入的height | ZUNFI | | per-layer |
| | | | 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 | 192] 5说明如下: | | |
| | | | | width/ 输入的width | [1,8176] | W- | |
| 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 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, w1*c1<2048*2048。 | | |



| | | | | | 1302 NI O Operator List | 圳心城电 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 | 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, | 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 | [1,31] | | |
| Convolution | Convolution 支持 | int8 float16 | strides [strides_h, strides_w]:int64[] | stride_h/ height方向的strides大 小 | [1,7] | | per-layer/ per-channel |
| | | | strides [strides_n, strides_w]:into4[] | 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 | [0,15] | | |
| | | | | pads_bottom/ bottom方向的pads大小 | | | |
| | | | group:int64 | group/ group的大小 | 无限制 | | |
| | | | 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 | | | |



| | | | | | 302 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 | |
| | | | | 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> | per-la per-c | | | |
| | | | num_output/ 输出的channel | 无限制 | | | | |
| | | | kernel_shape [num_output, num_input, | num_input/ 输入的channel | 无限制 | | | |
| | | | kernel_h, kernel_w]:int64[] | kernel_h/ height方向的kernel大小 | [1,8] | | | |
| Depthwise Convolution | | int8 | | kernel_w/ width方向的kernel大小 | [1,0] | | per-layer/ per-channel | |
| Deputwise Convolution | 又付 | float16 | strides [strides_h, strides_w]:int64[] | stride_h/ height方向的strides大 小 | [1,7] | | per-channer | |
| | | | | 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 | [over | | | |
| | | | pads_bottom/ bottom方向的pads大小 | | | | | |
| | | dilations_h/ height方向的dilations大 dilations_w]:int64[] | | | | | | |
| | | | dilations_w]:int64[] | widtht方向的dilations大 | | | | |



| | | | | | 362 NPU Operator List | · · · · · · · · · · · · · · · · · · · | 711111111 |
|---------------------------------|------|-----------------|-----------------------------------------|------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------|---------------------------|
| operator | 支持情况 | 输入数据类型 | 输入 | 设置项/ 输入参数含义 | 约束规格 | 广播支持(维度补齐) | 量化支持方式 |
| | | | | batch/ 输入的batch | 无限制 | | |
| | | | input_shape [batch, channel, height, | channel/ 输入的channel | 无限制 | | |
| | | | width]:tensor | height/ 输入的height | 无限制 | | |
| | | | | width/ 输入的width | 仅对首层输入width存在限制 详见 <u>首层输入说明</u> | | |
| | | | | num_output/ 输出的channel | 无限制 | | |
| | | | kernel_shape [num_output, num_input, | 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 | [1,31] | | |
| ConvTranspose/ Deconvolution | 支持 | int8 float16 | dride feride be dride advised ID | stride_h/ height方向的strides大 小 | [1,8] | | per-layer/ per-channel |
| | | | strides [strides_h, strides_w]:int64[] | 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: | | |
| | | | 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 | 7. 27 suide_w (widin-1)-paus_leit + 1 Coulput_w | | |
| | | | 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 | | | |



| | | | | | 1302 IVI O Operator Eist | 柳心厥电 7 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 |
| | | | 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 | マグiviaunui天光,ミルドpiviaunui | | |
| Gemm | 尚不支持 目前由CPU | first 目前由CPU int8 float16 beta:double transA:int64 | alpha:double | alpha/ 矩阵A*B乘法的scale | 无限制 | | per-layer/ |
| Genni | 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 | | | |
| | | | transB/ B矩阵是否转置 | → 仅静态tensor支持转置 | | | |
| 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 |



| 144 E 0 1 1 1 E 0 1 1 1 1 E 1 1 E 1 1 E 1 1 E 1 1 E 1 1 E 1 1 E 1 E 1 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E 1 E | | | | | | | |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|------------------------------------------------------------|--------------------------------------------------------------------|--------------------------|------|-------------|--------|
| 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 | int8 | | channel/ 输入的channel | | | |
| | | float16 mput_tensor [batch, channer, neight, width]:tensor | input_tensor [batch, channel, height, width]:tensor | height/ 输入的height | 无限制 | | |
| | | | width/ 输入的width | | | per-layer/ | |
| Expand | 支持 | int64 shape (batch_o, channel_o, height_o, width_o):tensor | batch_o/ 输出的batch_o | | | g | |
| | | | shape (batch_o, channel_o, height_o, width_o):tensor height_e 输出的t | channel_o/ 输出的channel | 无限制 | | |
| | | | | height_o/ 输出的height | Дукц | | |
| | | | | width_o/ 输出的width | | | |



| | | | | | 302 TH O Operator Elist | > M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - C - M - M | |
|----------|------|-------------------------------------------------------------------|-------------------------------------------------------|--------------------------|-------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
| 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 | x_tensor [batch, channel, height, width]:tensor | channel/ 输入的channel | 无限制 | | |
| | | int64 | | height/ 输入的height | | | |
| | | | | width/ 输入的width | | | |
| | | float16 wid int64 | | batch/ 输入的batch | | | |
| | | | | channel/ 输入的channel | - 无限制 | pe | per-layer/ |
| | | | | height/ 输入的height | J.L. PIX (P) | | |
| | | | | width/ 输入的width | | | |
| | 支持 | | ol mask_tensor [batch, channel, height, width]:tensor | batch/ 输入的batch | 无限制 | | |
| Where | XN | | | channel/ 输入的channel | | | |
| | | 5001 | | height/ 输入的height | | | |
| | | | | width/ 输入的width | | | |
| | | int8 shape (batch_o, channel_o, height_o, float16 width_o):tensor | | batch_o/ 输出的batch_o | | | |
| | | | shape (batch_o, channel_o, height_o, | channel_o/ 输出的channel | 无限制 | | |
| | | | 1 | height_o/ 输出的height | | | |
| | | | width_o/ 输出的width | | | | |



| | | | | Terro | 502 NPU Operator List | | K Z · J | |
|---------------|------|----------------------------------------|-------------------------------------------------------|-----------------------------------------|----------------------------------------------------|--------------|-----------|--|
| operator | 支持情况 | 输入数据类型 | 输入 | 设置项/ 输入参数含义 | 约束规格 | 广播支持(维度补齐) | 量化支持方式 | |
| | | c singut toncor I (batch abound beight | batch/ 输入的batch | 无限制 | | | | |
| | | | input_tensor_1 [batch, channel, height, width]:tensor | channel/ 输入的channel | 硬件支持[1,8192] | | | |
| exSoftmaxMask | 部分支持 | int8 float16 | | height/ 输入的height | axis=1, 无限制 axis=3/-1,width[1, 8192], height无限制 | | per-layer | |
| | | | width/ 输入的width | | | | | |
| | | | | batch/ 输入的batch | 1 | | | |
| | | | | input tensor 2 [batch, channel, height, | channel/ 输入的channel | 硬件支持[1,8192] | | |
| | | | width]:tensor | height/ 输入的height | 1 | | | |
| | | | | width/ 输入的width | axis=1: [1] axis=3/-1,[1, 8192] | | | |
| | | | axis:int64 | axis/ 做softmax的轴 | 1,3,即channel和width方向 | | | |
| | | | mask_value:int64 | mask/ 需要mask 的值 | 0或1 | | | |

| | | THE BOLT TO Operation Links | | | | 3177 | |
|----------|---------------|-----------------------------|-------------------------------------------------|--------------------------------------------------------------------------------------------|--------------------------------------------|-------------|------------|
| 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 |
| | int8 float | int8 float16 | x_tensor [batch, channel, height, width]:tensor | batch/ 输入的batch channel/ 输入的channel height/ 输入的height width/ 输入的width | c*h*w满足如下限制 8bit数据: 16 对齐, 16bit数据: 8对齐 | | |
| exGlu | 支持 | Int64 | axis:int64 | axis/ 切分的维度 | axis ==1 | | per-layer/ |





| 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 | | | | | | |



| | | | | | 302111 C Operator Elist | 710 -0-10K-1 1 //X | |
|------------------------------------|------|----------------|----|----------------|-------------------------|--------------------|--------|
| 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 | 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 | | | | | |
| Convolution+Add+Relu | 支持 | 同Convolution | | | | | |

注释:

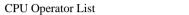
- (1) 广播支持举例:
- 1. OP(A(N,C,H,W),B(N,C,H,W)): OP(A(1,16,32,8),B(1,16,32,8))=C(1,16,32,8)
- 2、OP(A(N,C,H,W),B(C,1,1)): OP(A(1,16,32 8),B(16))=C(1,16,32,8)
- 3、OP(A(N,C,H,W),B(scalar)): OP(A(1,16,32,8),B(1))=C(1,16,32,8)
- 4. OP(A(N,C,H,W),B(H,W)): OP(A(1,16,32,8),B(32x8))=C(1,16,32,8)

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

Rockchip ^{瑞芯微电子}

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 无限制 取最小值的index ArgMin 取最大值的index ArgMax 无限制 BatchNormalization 无限制 批量归一化 SRC 支 持: float32/bool/int8/float16/int32/int64 Cast 数据类型转换 DST支持: float32/int8/int32/float16 数据截断激活层 无限制 Clip axis仅支持{0,1,2,3} Concat 合并操作 卷积操作 无限制 Convolution ConvTranspose/Deconvolution 转置卷积 无限制 Cos 余弦函数 无限制 DataConvert 仅支持 bool/int8/float类型转换 数据类型转换 通道方向空间方向转换 无限制 DepthToSpace 无限制 Div 除法操作 等于 Equal 无限制 Exp 指数函数 无限制

瑞芯微电子股份有限公司



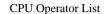
描述 规格约束 说明 Operator 拉平操作 无限制 Flatten 聚集操作 无限制 Gather 无限制 Greater 大于 **GreaterOrEqual** 大等于 无限制 无限制 GRU 门控循环单元 无限制 GRU (extern) 门控循环单元 ONNX扩展算子 HardSwish (extern) 激活函数 无限制 ONNX扩展算子 无限制 InstanceNormalization 单例归一化 LayerNorm (extern) 层归一化 无限制 ONNX扩展算子 无限制 小于 Less LessOrEqual 小等于 无限制 batchsize 仅支持 1 LogSoftmax 激活函数 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 无限制 多维矩阵相乘 取最大值 无限制 Max MaxPool 最大池化 无限制



Rockchip 瑞芯微电子

瑞芯微电子股份有限公司

| 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

| Rockchip |
|----------|
| |
| |

描述 规格约束 说明 Operator 仅支持Avg Pool Mode.batchsize 仅支持 1 RoiAlign 区域对齐池化 无限制 ScatterND N维索引取数 无限制 Sin 正弦函数 batchsize 仅支持 1 切片操作 Slice batchsize 仅支持 1 与ONNX OPSET 11规范一致 Softmax 激活函数 ONNX扩展算子,与ONNX OPSET 13规 batchsize 仅支持 1 Softmax (extern) 激活函数 范一致 SpaceToDetph 空间方向向通道方向转换 无限制 拆分数据 无限制 Split 求平方根 无限制 Sqrt 压缩数据维度 无限制 Squeeze Sub 无限制 减法 无限制 Tanh 双曲正切函数 batchsize 仅支持 1,不支持broadcasting 扩充拷贝数据 Tile 转置计算 Transpose 无限制 支持插值方式 bilinear; nearest2d Upsample 上采样 无限制 Not 按元素取非 无限制 通过mask取数 where



第六章 GPU Operator List



| Operator | 描述 | 规格约束 | 说明 |
|----------|--------|-----------|---------------------------------------------------------------------|
| MatMul | 多维矩阵相乘 | ניון אַקט | 只支持float16,需设置GPU优先(参考 《Rockchip_RKNPU_User_Guide_RK NN_API》) |



第七章 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 | 横刑 | ÷ | 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 | | |
|-------------------|-----------|----------|------------------------|------------------------------|------------------------------|---------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|--|
| | 输入维度 | 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 | NPU | | | 各卷积类型的width/kernel_h/kernel_w需要满足以下两式: 1. width * dilation kernel h < 1024*N | | |
| | | | int8 | | | | 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) | | |
| | int8 | | float16 | 8 | 8 | 1 | | 无限制 | |
| | | . 10. 24 | 其他类型(* <u>注释</u> 8) | CPU | | | | | |
| RK3566/3568 | | 4维度 | uint8 | | | 1 | 各卷积类型的width/kernel_h/kernel_w需要满足以下两式: | | |
| | | | int8 | | | | 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 | |
| | float16 | | float16 | CPU | 4 | | | | |
| | | | 其他类型(* <u>注释</u> 8) | | | | 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 | 非4维 | 无限制 | CPU | 1 | 1 | 无限制 | 无限制 | |
| | int8 | 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_h * dilation_kernel_w) / 128), 3) | | |
| | | | int8 | NPU | | | | | |
| | | | float16 | any | -16 | | | 无限制 | |
| | | | 其他类型(* <u>注释</u> 8) | CPU | | | | | |
| RK3588 | | | uint8 | | | 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 | | |
| | | | int8 | | | | 1. width * dhiation_kernei_n <= 1024 * N 2. width <= 8192 1 其中N必须为1到7的整数,超出范围的卷积不受支持,各卷积类型N的计算方式如下: Convolution: N = 12 - MAX(CEIL((dilation_kernel_h * dilation_kernel_w) / 128), 3) | | |
| | float16 | | float16 | CPU | 8 | | | 无限制 | |
| | | | 其他类型(* <u>注释</u> 8) | | | | 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) |) | |
| | 无限制 | 非4维 | 无限制 | CPU | 1 | 1 | 无限制 | 无限制 | |
| | | | uint8 | | | | 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 | | |

| | | | | | | 174771111 | 7 VIII CI 90 71 | 1 11/1/11/11/11 |
|-------------------|---------|--------------------------------|--------------------------------|------|----|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|
| RV1103/ RV1106 | int8 | 4维 | int8 | NPU | 16 | 1 | 其中N必须为1到7的整数,超出范围的卷积不受支持,各卷积类型N的计算方式如下: Convolution: N = 8 - MAX(CEIL((dilation_kernel_h * dilation_kernel_w) / 128), 2) Depthwise Convolution: N = 8 - CEIL((dilation_kernel_h * dilation_kernel_w) / 4096) ConvTranspose/Deconvolution: N = 8 - MAX(CEIL((dilation_kernel_h * dilation_kernel_w) / 128), 2) | 无限制 |
| int8 | | uint8 int8 float16 其他类型(*注释 8) | NPU -CPU | 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: 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) / 4096) | 无限制 | |
| RK3562 | float16 | - 4维度 | uint8 int8 float16 其他类型(*注释 8) | -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(channe1)〉 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 |
| | G | | 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 |



| | | | | | 110 × 1110 ET 80 71 |
|-------------------|---------|-----|----------|-----------------------------------|---------------------|
| RV1103/ RV1106 | | | 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 |
| | int8 | 4维 | NHWC | 无 | |
| | | | NCHW | 无 | 无 |
| | :40 | | 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 |
| | G16 | | NHWC | 无 | 无 |
| | float16 | | 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;