Classification Level: Top Secret( ) Secret( ) Internal( ) Public(√)

Rockchip User Guide RKNN API

(Technology Department, Graphic Compute Platform Center)

Mark: [ ] Changing [√] Released	Version:	1.4.0
	Author:	Randall
	Completed Date:	10/Sep/2020
	Reviewer:	Randall
	Reviewed Date:	10/Sep/2020

Revision History

Version	Modifier	Date	Modify Description	Reviewer
v0.9.7	Yang Huacong	25/Jan/2019	Initial version	Zhuo Hongtian
V0.9.8	Yang Huacong	5/Jun/2019	1. Add RKNN API Library description 2. Update rknn example and rknn-toolkit path 3. Fix some mistake	Zhuo Hongtian
V1.3.3	Randall	2/June/2020	1. Support RV1109/RV1126 2. Remove the python support instructions, see the RKNN Toolkit Lite related documentation for python support.	Randall
V1.4.0	Randall	10/Sep/2020	1. Add some error code	Randall

Table of Contents

1 Overview

RKNN SDK provides a programming interface for platforms with NPU such as RK1808, which can help users deploy RKNN models exported using RKNN-Toolkit.

2 Supported hardware platforms

This document applies to the following hardware platforms:

1) RK1808, RK1806
2) RV1126, RV1109

The following description takes RK1808 as an example and also applies to other platforms mentioned above.

3 Instructions

3.1 RKNN SDK Development Process

Before using the RKNN SDK, users first need to use the RKNN-Toolkit tool to convert the user's model to the RKNN model. The user can obtain the tool's complete installation package and documentation at https://github.com/rockchip-linux/rknn-toolkit .

After successful conversion to the RKNN model, users can first connect to the RK1808 device via RKNN-Toolkit for online debugging to ensure that the accuracy and performance of the model meet the requirements.

After getting the RKNN model file, users can choose using C or Python interface to develop the application. The following chapters will explain how to develop application based on the RKNN SDK on RK1808 platform.

3.2 RKNN C API

3.2.1 RKNN API Library

The libraries and header files provided by the RKNN SDK are located at <sdk>/external/rknpu/rknn/ rknn_api/librknn_api directory, developers can use to develop applications.

It should be noted that the RKNN API of the RK1808 and RK3399Pro platforms is compatible, and applications developed by both can be easily ported. However, you need to pay attention to distinguish between the two platforms librknn_api.so, if the developer uses RK3399Pro's librknn_api.so will not be able to run on the RK1808 platform. Developers can use the following methods to distinguish the platform of librknn_api.so.

```
$ strings librknn_api.so |grep version
RK1808 librknn_api version 0.9.8 (58c7577 build: 2019-06-05 17:00:44)
```

3.2.2 EXAMPLES

The SDK provides MobileNet image classification, MobileNet SSD object detection, and Yolo v3 object detection demos. These demos provide reference for developer to develop applications based on the RKNN SDK. The emo code is located in the <sdk>/external/rknpu/rknn/rknn_api/examples directory. Let's take rknn_mobilenet_demo as an example to explain how to get started quickly.

1) Compile Demo Source Code

```
cd examples/rknn_mobilenet_demo
mkdir build && cd build
cmake ..
make && make install
cd –
```

2) Deploy to the RK1808 device

```
adb push install/rknn_mobilenet_demo /userdata/
```

3) Run Demo

```
adb shell
cd /userdata/rknn_mobilenet_demo/
./rknn_mobilenet_demo mobilenet_v1.rknn dog_224x224.jpg
```

3.2.3 API Reference

3.2.3.1 rknn_init

The *rknn_init* function will create a *rknn_context* object, load the RKNN model, and perform specific initialization behavior based on the flag.

API	rknn_init
Description	Initialize rknn
Parameters	*rknn_context *context*： Pointer to *rknn_context* object. After the function is called, the context object will be assigned.
	*void *model*：Binary data for the RKNN model.
	uint32_t size：Model size
	uint32_t flag：A specific initialization flag. Currently supports following flags: **RKNN_FLAG_COLLECT_PERF_MASK**： Open the performance collection debugging switch. After opening, you can query the running time of each layer of the network through the *rknn_query* interface. Note that the running time of *rknn_run* will be longer after this flag is set.
Return	int: Error code (See RKNN Error Code).

Sample Code:

```
rknn_context ctx;
int ret = rknn_init(&ctx, model_data, model_data_size, 0);
```

3.2.3.2 rknn_destroy

The *rknn_destroy* function will release the *rknn_context* object and its associated resources.

API	rknn_destroy
Description	Destroy the rknn_context object and its related resources.
Parameters	*rknn_context context*: The rknn_context object to be destroyed.
Return	int: Error code (See RKNN Error Code).

Sample Code:

```
int ret = rknn_destroy (ctx);
```

3.2.3.3 rknn_query

The *rknn_query* function can query the information of model input and output tensor attribute, performance information and SDK version etc.

API	rknn_query
Description	Query the information about the model and the SDK.
Parameters	*rknn_context context*: The object of *rknn_contex*.
	rknn_query_cmd cmd: Query command.
	void info*: Structure object that stores the result of the query.
	uint32_t size: the size of the info Structure object.
Return	int: Error code (See RKNN Error Code).

Currently, the SDK supports the following query commands:

Query command	Return result structure	Function
RKNN_QUERY_IN_OUT_NUM	rknn_input_output_num	Query the number of input and output

		Tensor.
RKNN_QUERY_INPUT_ATTR	rknn_tensor_attr	Query input Tensor attribute.
RKNN_QUERY_OUTPUT_ATTR	rknn_tensor_attr	Query output Tensor attribute.
RKNN_QUERY_PERF_DETAIL	rknn_perf_detail	Query the running time of each layer of the network.
RKNN_QUERY_SDK_VERSION	rknn_sdk_version	Query the SDK version.

Next we will explain each query command in detail.

1） Query the number of input and output Tensor

The *RKNN_QUERY_IN_OUT_NUM* command can be used to query the number of model input and output Tensor. You need to create the *rknn_input_output_num* structure object first.

Sample Code:

```
rknn_input_output_num io_num;
ret = rknn_query(ctx, RKNN_QUERY_IN_OUT_NUM, &io_num,
                 sizeof(io_num));
printf("model input num: %d, output num: %d\n", io_num.n_input,
                 io_num.n_output);
```

2） Query input Tensor attribute

The *RKNN_QUERY_INPUT_ATTR* command can be used to query the attribute of the model input Tensor. You need to create the *rknn_tensor_attr* structure object first.

Sample Code:

```
rknn_tensor_attr input_attrs[io_num.n_input];
memset(input_attrs, 0, sizeof(input_attrs));
for (int i = 0; i < io_num.n_input; i++) {
    input_attrs[i].index = i;
    ret = rknn_query(ctx, RKNN_QUERY_INPUT_ATTR, &(input_attrs[i]),
                     sizeof(rknn_tensor_attr));
}
```

3） Query output Tensor attribute

The *RKNN_QUERY_OUTPUT_ATTR* command can be used to query the attribute of the model output Tensor. You need to create the *rknn_tensor_attr* structure object first.

Sample Code:

```
rknn_tensor_attr output_attrs[io_num.n_output];
memset(output_attrs, 0, sizeof(output_attrs));
for (int i = 0; i < io_num.n_output; i++) {
    output_attrs[i].index = i;
    ret = rknn_query(ctx, RKNN_QUERY_OUTPUT_ATTR, &(output_attrs[i]),
                     sizeof(rknn_tensor_attr));
}
```

4） Query the running time of each layer of the network

If the *RKNN_FLAG_COLLECT_PERF_MASK* flag has been set on the *rknn_init* function, the *RKNN_QUERY_PERF_DETAIL* command can be passed to query the runtime of each layer of the network after *rknn_run* function execution completed. You need to create the *rknn_perf_detail* structure object first.

Sample Code:

```
    rknn_perf_detail perf_detail;
    ret = rknn_query(ctx, RKNN_QUERY_PERF_DETAIL, &perf_detail,
                     sizeof(rknn_perf_detail));
    printf("%s", perf_detail.perf_data);
```

It should be noted that *rknn_perf_detail.perf_data* does not need to be released. The SDK will automatically manage this buffer memory.

5） Query the SDK version

The *RKNN_QUERY_SDK_VERSION* command can be used to query the version information of the RKNN SDK. You need to create the *rknn_sdk_version* structure object first.

Sample Code:

```
rknn_sdk_version version;
ret = rknn_query(ctx, RKNN_QUERY_SDK_VERSION, &version,
                 sizeof(rknn_sdk_version));
printf("sdk api version: %s\n", version.api_version);
printf("driver version: %s\n", version.drv_version);
```

3.2.3.4 rknn_inputs_set

The input data of the model can be set by the *rknn_inputs_set* function. This function can support multiple inputs, each one is a *rknn_input* structure object. Developers needs to set these object field before passing in.

API	rknn_inputs_set
Description	Set the model input data.
Parameter	*rknn_context context*: The object of rknn_contex.
	uint32_t n_inputs: Number of inpus.
	rknn_input inputs[]: Array of rknn_input.
Return	int: Error code (See RKNN Error Code).

Sample Code:

```
rknn_input inputs[1];
memset(inputs, 0, sizeof(inputs));
inputs[0].index = 0;
inputs[0].type = RKNN_TENSOR_UINT8;
inputs[0].size = img_width*img_height*img_channels;
inputs[0].fmt = RKNN_TENSOR_NHWC;
inputs[0].buf = in_data;

ret = rknn_inputs_set(ctx, 1, inputs);
```

3.2.3.5 rknn_run

The *rknn_run* function will perform a model reasoning. The input data need to be set by the *rknn_inputs_set* function before *rknn_run is* called.

API	rknn_run
Description	Perform a model reasoning.
Parameter	*rknn_context context*: The object of rknn_contex.
	rknn_run_extend extend*: Reserved for extension, currently not used, you can pass NULL.
Return	int: Error code (See RKNN Error Code).

Sample Code:

```
ret = rknn_run(ctx, NULL);
```

3.2.3.6 rknn_outputs_get

The *rknn_outputs_get* function can get the output data of the model reasoning. This function can get multiple output data. Each of these outputs is a *rknn_output* structure object, which needs to be created and set in turn before the function is called.

There are two ways to store buffers for output data:

1) Developer allocate and release buffers themselves. At this time, the *rknn_output.is_prealloc* needs to be set to 1, and the *rknn_output.buf* points to users' allocated buffer;
2) The other is allocated by SDK. At this time, the *rknn_output .is_prealloc* needs to be set to 0. After the function is executed, *rknn_output.buf* will be created and store the output data.

API	rknn_outputs_get
Description	Get model inference output data.
Parameter	*rknn_context context*: The object of rknn_context.
	uint32_t n_outputs: Number of output.
	rknn_output outputs[]: Array of rknn_output.
	rknn_run_extend extend*: Reserved for extension, currently not used, you can pass NULL.
Return	int: Error code (See RKNN Error Code).

Sample Code:

```
rknn_output outputs[io_num.n_output];
memset(outputs, 0, sizeof(outputs));
for (int i = 0; i < io_num.n_output; i++) {
    outputs[i].want_float = 1;
}
ret = rknn_outputs_get(ctx, io_num.n_output, outputs, NULL);
```

3.2.3.7 rknn_outputs_release

The *rknn_outputs_release* function will release the relevant resources of the *rknn_output* object.

API	rknn_outputs_release
Description	Release the rknn_output object
Parameter	*rknn_context context*: rknn_context object
	uint32_t n_outputs: Number of output.
	rknn_output outputs[]: rknn_output array to be release.
Return	int: Error code (See RKNN Error Code).

Sample Code:

```
ret = rknn_outputs_release(ctx, io_num.n_output, outputs);
```

3.2.4 RKNN DataStruct Define

3.2.4.1 rknn_input_output_num

The structure *rknn_input_output_num* represents the number of input and output Tensor， The following table shows the definition:

Field	Type	Meaning
n_input	uint32_t	The number of input tensor
n_output	uint32_t	The number of output tensor

3.2.4.2 rknn_tensor_attr

The structure *rknn_tensor_attr* represents the attribute of the model's Tensor. The following table shows the definition:

Field	Type	Meaning
index	uint32_t	Indicates the index position of the input and output Tensor.
n_dims	uint32_t	The number of Tensor dimensions.
dims	uint32_t[]	Values for each dimension.
name	char[]	Tensor name.
n_elems	uint32_t	The number of Tensor data elements.
size	uint32_t	The memory size of Tensor data.
fmt	rknn_tensor_format	The format of Tensor dimension, has the following format: **RKNN_TENSOR_NCHW** **RKNN_TENSOR_NHWC**
type	rknn_tensor_type	Tensor data type, has the following data types: **RKNN_TENSOR_FLOAT32** **RKNN_TENSOR_FLOAT16** **RKNN_TENSOR_INT8** **RKNN_TENSOR_UINT8** **RKNN_TENSOR_INT16**
qnt_type	rknn_tensor_qnt_type	Tensor Quantization Type, has the following types of quantization: **RKNN_TENSOR_QNT_NONE:** Not quantified; **RKNN_TENSOR_QNT_DFP**: Dynamic fixed point

		quantization; **RKNN_TENSOR_QNT_AFFINE_ASYMMETRIC**: Asymmetric quantification.
fl	int8_t	RKNN_TENSOR_QNT_DFP quantization parameter
zp	uint32_t	RKNN_TENSOR_QNT_AFFINE_ASYMMETRIC quantization parameter.
scale	float	RKNN_TENSOR_QNT_AFFINE_ASYMMETRIC quantization parameter.

3.2.4.3 rknn_input

The structure *rknn_input* represents a data input to the model, used as a parameter to the *rknn_inputs_set* function. The following table shows the definition:

Field	Type	Meaning
index	uint32_t	The index position of this input.
buf	void*	Point to the input data Buffer.
size	uint32_t	The memory size of the input data Buffer.
pass_through	uint8_t	When set to 1, buf will be directly set to the input node of the model without any pre-processing.
type	rknn_tensor_type	The type of input data.
fmt	rknn_tensor_format	The format of input data.

3.2.4.4 rknn_output

The structure *rknn_output* represents a data output of the model, used as a parameter to the *rknn_outputs_get* function. The following table shows the definition:

Field	Type	Meaning
want_float	uint8_t	Indicates if the output data needs to be converted

		to float type.
is_prealloc	uint8_t	Indicates whether the Buffer that stores the output data is pre-allocated.
index	uint32_t	The index position of this output.
buf	void*	Pointer which point to the output data Buffer.
size	uint32_t	Output data Buffer memory size.

3.2.4.5 rknn_perf_detail

The structure *rknn_perf_detail* represents the performance details of the model. The following table shows the definition:

Field	Type	Meaning
perf_data	char*	Performance details include the run time of each layer of the network.
data_len	uint64_t	The Length of perf_data.

3.2.4.6 rknn_sdk_version

The structure *rknn_sdk_version* is used to indicate the version information of the RKNN SDK. The following table shows the definition:

Field	Type	Meaning
api_version	char[]	SDK API Version information.
drv_version	char[]	Driver version information.

3.2.5 RKNN Error Code

The return code of the RKNN API function is defined as shown in the following table.

Error Code	Message
RKNN_SUCC（0）	Execution is successful

RKNN_ERR_FAIL (-1)	Execution error
RKNN_ERR_TIMEOUT (-2)	Execution timeout
RKNN_ERR_DEVICE_UNAVAILABLE (-3)	NPU device is unavailable
RKNN_ERR_MALLOC_FAIL (-4)	Memory allocation is failed
RKNN_ERR_PARAM_INVALID (-5)	Parameter error
RKNN_ERR_MODEL_INVALID (-6)	RKNN model is invalid
RKNN_ERR_CTX_INVALID (-7)	rknn_context is invalid
RKNN_ERR_INPUT_INVALID (-8)	rknn_input object is invalid
RKNN_ERR_OUTPUT_INVALID (-9)	rknn_output object is invalid
RKNN_ERR_DEVICE_UNMATCH (-10)	Version does not match
RKNN_ERR_INCOMPATILE_PRE_COMPILE_MODEL (-11)	This RKNN model use pre_compile mode, but not compatible with current driver.
RKNN_ERR_INCOMPATILE_OPTIMIZATION_LEVEL_VERSION (-12)	This RKNN model use optimization level mode, but not compatible with current driver.
RKNN_ERR_TARGET_PLATFORM_UNMATCH (-13)	This RKNN model don’t compatible with current platform.