DRP-AI Extension Pack (Pruning Tool)

Sparse model processing speed check guide

Contents

1. Overview

DRP-AI has the function to accelerate the execution of the sparse model, but how much faster it actually is depending on the pruning rate and the structure of the AI model.

Therefore, this guide shows how to confirm how much faster a user's AI model can be by pruning it at different pruning rates.

Note: The procedure in this guide does not retrain the sparse model, so accuracy cannot be confirmed.

In this guide, the Sparse model will be created without retraining. Therefore, the processing performance can be quickly confirmed when pruning is applied.

Figure 1-1 shows the flow to confirm the Sparse model's performance. For more details on how to confirm it, refer to "How to confirm the processing speed of the Dense and Sparse models".

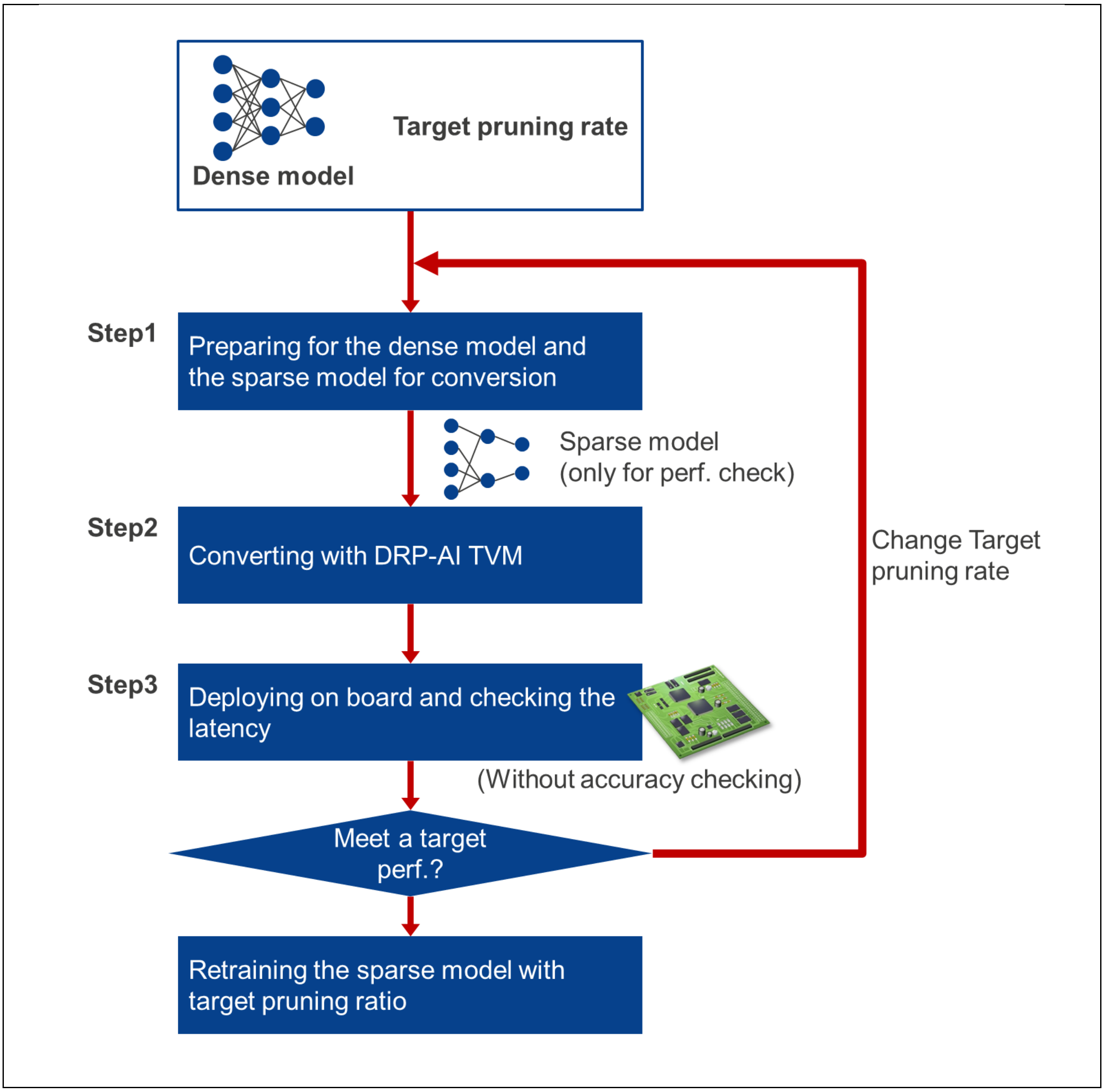

Dense model
Target pruning rate
Step1
Preparing for the dense model and the sparse model for conversion
Sparse model
(only for perf. check)
Step2
Converting with DRP-AI TVM
Change Target pruning rate
Step3
Deploying on board and checking the latency
(Without accuracy checking)
Meet a target perf.?
Retraining the sparse model with target pruning ratio

Figure 1-1 Flow of checking sparse model inference speed

1.1 Operating Environment

Thid document is based on the following operating environment.

- DRP-AI Extension Pack V1.0.0

Fore more details on how to setup DRP-AI Extension Pack, refer to the following document.

- DRP-AI TVM document: rzv_drp-ai_tvm/pruning/setup/README.md at main · renesas-rz/rzv_drp-ai_tvm (github.com)

1.2 Related documents

Related documents as follows.

- DRP-AI Extension Pack User’s Manual (R20UT5188)
- DRP-AI TVM document：renesas-rz/rzv_drp-ai_tvm: Extension package of Apache TVM (Machine Learning Compiler) for Renesas DRP-AI accelerators powered by Edgecortix MERA(TM) Based Apache TVM version: v0.11.1 (github.com)

2. How to confirm the processing speed of the Dense and Sparse models

To confirm how much faster the sparse model can be, use the flow shown in Figure 1-1. In this guide, the Sparse model will be created without retraining after pruning. Therefore, the processing performance can be quickly confirmed when pruning is applied.

Specifically, the processing performance when pruning is applied can be confirmed by executing the following three steps.

Step1. Prepare the Dense and Sparse models for conversion.

Step2. Convert the Dense and Sparse models with DRP-AI TVM.

Step3 Run the models on board and measure performance.

For more details about each step, see the following chapters.

2.1 Step1. Prepare the Dense and Sparse models for conversion

This chapter describes how to create the Dense and Sparse models to run on the board.

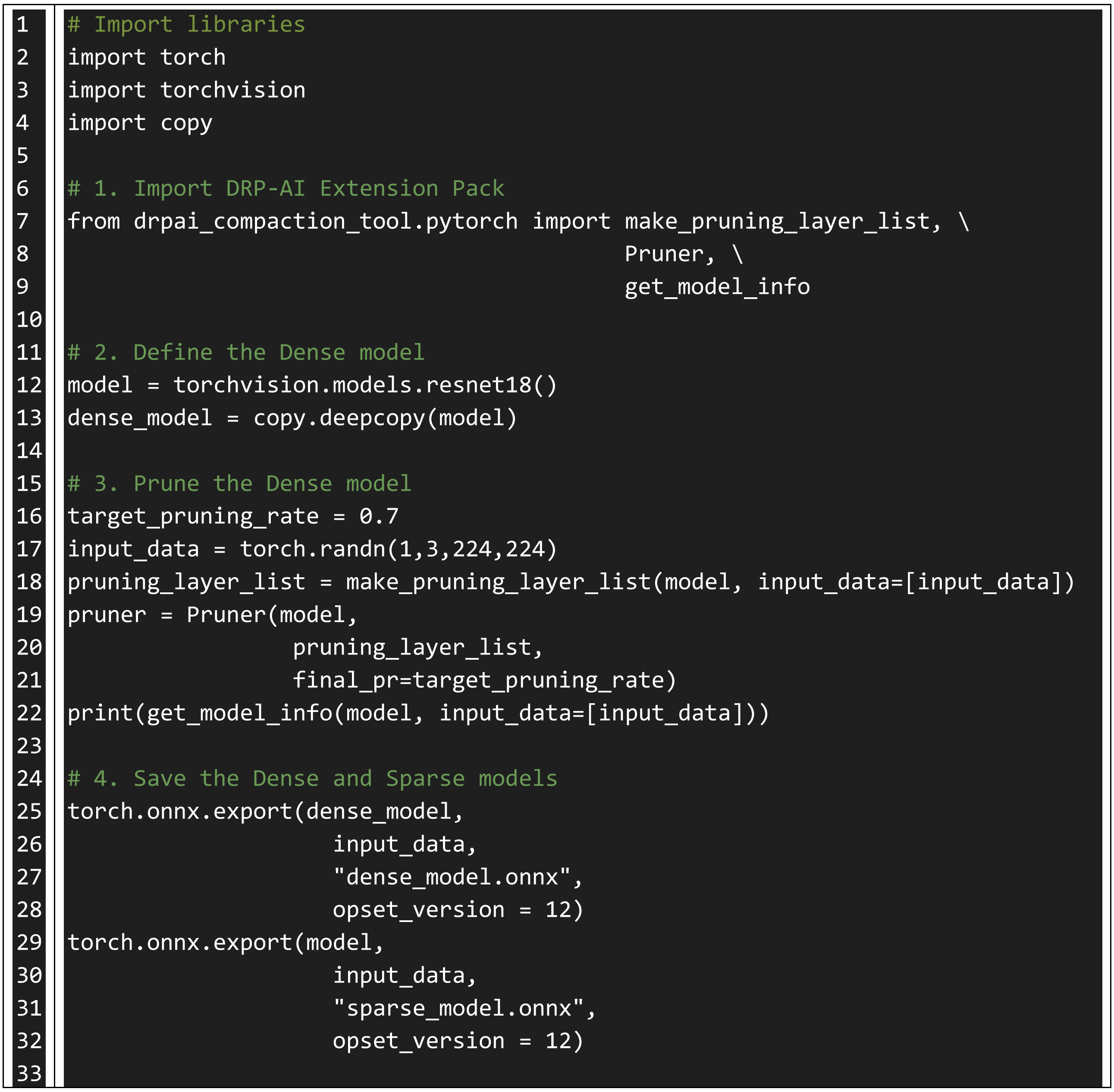
2.1.1 PyTorch

For PyTorch, follow the steps below to create the Dense and Sparse models for conversion.

1. Import DRP-AI Extension Pack
2. Define the Dense model
3. Prune the Dense model
4. Save the Dense and Sparse models

The following shows how to create it using torchvision resnet18.

Any Sparse model can be created by modifying the Dense model definition and the input data (input_data). The pruning rate can be changed by modifying the target_pruning_rate.



```
# Import libraries
import torch
import torchvision
import copy

# 1. Import DRP-AI Extension Pack
from drpai_compaction_tool.pytorch import make_pruning_layer_list, \
                                          Pruner, \
                                          get_model_info

# 2. Define the Dense model
model = torchvision.models.resnet18()
dense_model = copy.deepcopy(model)

# 3. Prune the Dense model
target_pruning_rate = 0.7
input_data = torch.randn(1,3,224,224)
pruning_layer_list = make_pruning_layer_list(model, input_data=[input_data])
pruner = Pruner(model,
                 pruning_layer_list,
                 final_pr=target_pruning_rate)
print(get_model_info(model, input_data=[input_data]))

# 4. Save the Dense and Sparse models
torch.onnx.export(dense_model,
                    input_data,
                    "dense_model.onnx",
                    opset_version = 12)
torch.onnx.export(model,
                    input_data,
                    "sparse_model.onnx",
                    opset_version = 12)

```

Figure 2-1 How to save the Dense and Sparse models (PyTorch)

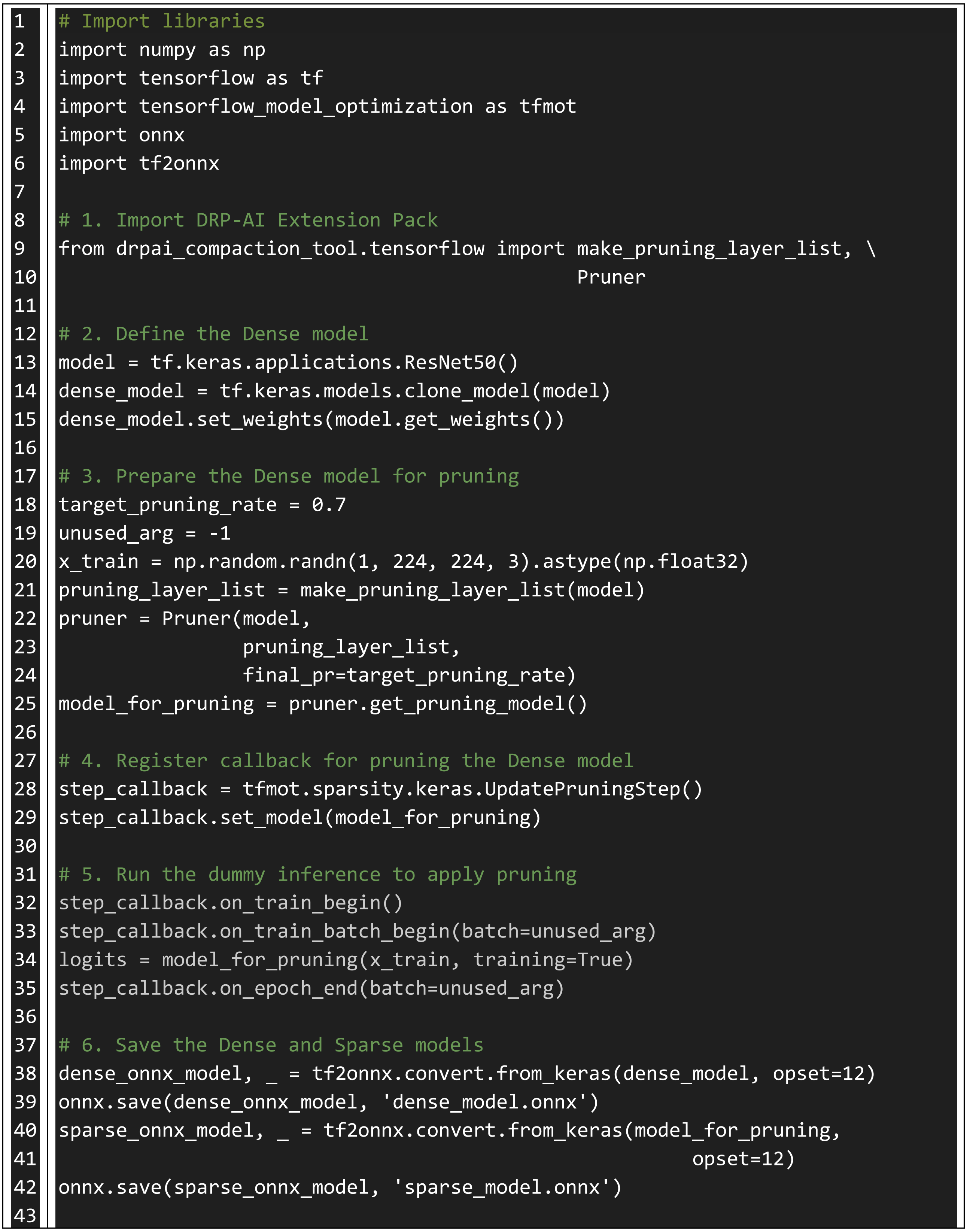
2.1.2 TensorFlow

For TensorFlow, follow the steps below to create the Dense and Sparse models for conversion.

1. Import DRP-AI Extension Pack
2. Define the Dense model
3. Prepare the Dense model for pruning
4. Register callback for pruning the Dense model
5. Run the dummy inference to apply pruning
6. Save the Dense and Sparse models

The following shows how to create it using TensorFlow applications resnet50.

Any Sparse model can be created by modifying the Dense model definition and the input data (x_train). The pruning rate can be changed by modifying the target_pruning_rate.



```python
# Import libraries
import numpy as np
import tensorflow as tf
import tensorflow_model_optimization as tfmot
import onnx
import tf2onnx

# 1. Import DRP-AI Extension Pack
from drpai_compaction_tool.tensorflow import make_pruning_layer_list, \
                                             Pruner

# 2. Define the Dense model
model = tf.keras.applications.ResNet50()
dense_model = tf.keras.models.clone_model(model)
dense_model.set_weights(model.get_weights())

# 3. Prepare the Dense model for pruning
target_pruning_rate = 0.7
unused_arg = -1
x_train = np.random.randn(1, 224, 224, 3).astype(np.float32)
pruning_layer_list = make_pruning_layer_list(model)
pruner = Pruner(model,
                pruning_layer_list,
                final_pr=target_pruning_rate)
model_for_pruning = pruner.get_pruning_model()

# 4. Register callback for pruning the Dense model
step_callback = tfmot.sparsity.keras.UpdatePruningStep()
step_callback.set_model(model_for_pruning)

# 5. Run the dummy inference to apply pruning
step_callback.on_train_begin()
step_callback.on_train_batch_begin(batch=unused_arg)
logits = model_for_pruning(x_train, training=True)
step_callback.on_epoch_end(batch=unused_arg)

# 6. Save the Dense and Sparse models
dense_onnx_model, _ = tf2onnx.convert.from_keras(dense_model, opset=12)
onnx.save(dense_onnx_model, 'dense_model.onnx')
sparse_onnx_model, _ = tf2onnx.convert.from_keras(model_for_pruning,
                                                       opset=12)
onnx.save(sparse_onnx_model, 'sparse_model.onnx')

```

Figure 2-2 How to save the Dense and Sparse models (TensorFlow)

2.2 Step2. Convert the Dense and Sparse models with DRP-AI TVM

Use DRP-AI TVM to convert the Dense and Sparse models created in Step 1 for execution on the board.

For more details on how to convert, please refer to the DRP-AI TVM documentation.

- renesas-rz/rzv_drp-ai_tvm: Extension package of Apache TVM (Machine Learning Compiler) for Renesas DRP-AI accelerators powered by Edgecortix MERA(TM) Based Apache TVM version: v0.11.1 (github.com)

2.3 Step3. Run the models on board and measure the performance

In order to confirm the processing performance when pruning is applied, let the AI model run on the board. Measure and compare the execution speed of the Dense model and the Sparse model.

For more details on how to run the converted model on the board, please refer to the following DRP-AI TVM documentation.

- renesas-rz/rzv_drp-ai_tvm: Extension package of Apache TVM (Machine Learning Compiler) for Renesas DRP-AI accelerators powered by Edgecortix MERA(TM) Based Apache TVM version: v0.11.1 (github.com)

Revision History

Rev.	Date	Description	
		Page	Summary
1.00	Apr.10.24	-	First edition issued

Notice

1. Descriptions of circuits, software and other related information in this document are provided only to illustrate the operation of semiconductor products and application examples. You are fully responsible for the incorporation or any other use of the circuits, software, and information in the design of your product or system. Renesas Electronics disclaims any and all liability for any losses and damages incurred by you or third parties arising from the use of these circuits, software, or information.
2. Renesas Electronics hereby expressly disclaims any warranties against and liability for infringement or any other claims involving patents, copyrights, or other intellectual property rights of third parties, by or arising from the use of Renesas Electronics products or technical information described in this document, including but not limited to, the product data, drawings, charts, programs, algorithms, and application examples.
3. No license, express, implied or otherwise, is granted hereby under any patents, copyrights or other intellectual property rights of Renesas Electronics or others.
4. You shall be responsible for determining what licenses are required from any third parties, and obtaining such licenses for the lawful import, export, manufacture, sales, utilization, distribution or other disposal of any products incorporating Renesas Electronics products, if required.
5. You shall not alter, modify, copy, or reverse engineer any Renesas Electronics product, whether in whole or in part. Renesas Electronics disclaims any and all liability for any losses or damages incurred by you or third parties arising from such alteration, modification, copying or reverse engineering.
6. Renesas Electronics products are classified according to the following two quality grades: “Standard” and “High Quality”. The intended applications for each Renesas Electronics product depends on the product’s quality grade, as indicated below.

   "Standard": Computers; office equipment; communications equipment; test and measurement equipment; audio and visual equipment; home electronic appliances; machine tools; personal electronic equipment; industrial robots; etc.

   "High Quality": Transportation equipment (automobiles, trains, ships, etc.); traffic control (traffic lights); large-scale communication equipment; key financial terminal systems; safety control equipment; etc.

   Unless expressly designated as a high reliability product or a product for harsh environments in a Renesas Electronics data sheet or other Renesas Electronics document, Renesas Electronics products are not intended or authorized for use in products or systems that may pose a direct threat to human life or bodily injury (artificial life support devices or systems; surgical implantations; etc.), or may cause serious property damage (space system; undersea repeaters; nuclear power control systems; aircraft control systems; key plant systems; military equipment; etc.). Renesas Electronics disclaims any and all liability for any damages or losses incurred by you or any third parties arising from the use of any Renesas Electronics product that is inconsistent with any Renesas Electronics data sheet, user’s manual or other Renesas Electronics document.
7. No semiconductor product is absolutely secure. Notwithstanding any security measures or features that may be implemented in Renesas Electronics hardware or software products, Renesas Electronics shall have absolutely no liability arising out of any vulnerability or security breach, including but not limited to any unauthorized access to or use of a Renesas Electronics product or a system that uses a Renesas Electronics product. RENESAS ELECTRONICS DOES NOT WARRANT OR GUARANTEE THAT RENESAS ELECTRONICS PRODUCTS, OR ANY SYSTEMS CREATED USING RENESAS ELECTRONICS PRODUCTS WILL BE INVULNERABLE OR FREE FROM CORRUPTION, ATTACK, VIRUSES, INTERFERENCE, HACKING, DATA LOSS OR THEFT, OR OTHER SECURITY INTRUSION (“Vulnerability Issues”). RENESAS ELECTRONICS DISCLAIMS ANY AND ALL RESPONSIBILITY OR LIABILITY ARISING FROM OR RELATED TO ANY VULNERABILITY ISSUES. FURTHERMORE, TO THE EXTENT PERMITTED BY APPLICABLE LAW, RENESAS ELECTRONICS DISCLAIMS ANY AND ALL WARRANTIES, EXPRESS OR IMPLIED, WITH RESPECT TO THIS DOCUMENT AND ANY RELATED OR ACCOMPANYING SOFTWARE OR HARDWARE, INCLUDING BUT NOT LIMITED TO THE IMPLIED WARRANTIES OF MERCHANTABILITY, OR FITNESS FOR A PARTICULAR PURPOSE.
8. When using Renesas Electronics products, refer to the latest product information (data sheets, user’s manuals, application notes, “General Notes for Handling and Using Semiconductor Devices” in the reliability handbook, etc.), and ensure that usage conditions are within the ranges specified by Renesas Electronics with respect to maximum ratings, operating power supply voltage range, heat dissipation characteristics, installation, etc. Renesas Electronics disclaims any and all liability for any malfunctions, failure or accident arising out of the use of Renesas Electronics products outside of such specified ranges.
9. Although Renesas Electronics endeavors to improve the quality and reliability of Renesas Electronics products, semiconductor products have specific characteristics, such as the occurrence of failure at a certain rate and malfunctions under certain use conditions. Unless designated as a high reliability product or a product for harsh environments in a Renesas Electronics data sheet or other Renesas Electronics document, Renesas Electronics products are not subject to radiation resistance design. You are responsible for implementing safety measures to guard against the possibility of bodily injury, injury or damage caused by fire, and/or danger to the public in the event of a failure or malfunction of Renesas Electronics products, such as safety design for hardware and software, including but not limited to redundancy, fire control and malfunction prevention, appropriate treatment for aging degradation or any other appropriate measures. Because the evaluation of microcomputer software alone is very difficult and impractical, you are responsible for evaluating the safety of the final products or systems manufactured by you.
10. Please contact a Renesas Electronics sales office for details as to environmental matters such as the environmental compatibility of each Renesas Electronics product. You are responsible for carefully and sufficiently investigating applicable laws and regulations that regulate the inclusion or use of controlled substances, including without limitation, the EU RoHS Directive, and using Renesas Electronics products in compliance with all these applicable laws and regulations. Renesas Electronics disclaims any and all liability for damages or losses occurring as a result of your noncompliance with applicable laws and regulations.
11. Renesas Electronics products and technologies shall not be used for or incorporated into any products or systems whose manufacture, use, or sale is prohibited under any applicable domestic or foreign laws or regulations. You shall comply with any applicable export control laws and regulations promulgated and administered by the governments of any countries asserting jurisdiction over the parties or transactions.
12. It is the responsibility of the buyer or distributor of Renesas Electronics products, or any other party who distributes, disposes of, or otherwise sells or transfers the product to a third party, to notify such third party in advance of the contents and conditions set forth in this document.
13. This document shall not be reprinted, reproduced or duplicated in any form, in whole or in part, without prior written consent of Renesas Electronics.
14. Please contact a Renesas Electronics sales office if you have any questions regarding the information contained in this document or Renesas Electronics products.

(Note1) “Renesas Electronics” as used in this document means Renesas Electronics Corporation and also includes its directly or indirectly controlled subsidiaries.

(Note2) “Renesas Electronics product(s)” means any product developed or manufactured by or for Renesas Electronics.

(Rev.5.0-1 October 2020)

Corporate Headquarters

TOYOSU FORESIA, 3-2-24 Toyosu,
Koto-ku, Tokyo 135-0061, Japan
www.renesas.com

Trademarks

Renesas and the Renesas logo are trademarks of Renesas Electronics Corporation. All trademarks and registered trademarks are the property of their respective owners.

Contact information

For further information on a product, technology, the most up-to-date version of a document, or your nearest sales office, please visit: www.renesas.com/contact/.