# SGTPU 软件指南

发行版本 0958503

SOPHGO

2025 年 01 月 24 日

# 目录

**法律声明**



**注意**

您购买的产品、服务或特性等应受算能商业合同和条款的约束, 本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定, 算能对本文档内容不做任何明示或默示的声明或保证。由于产品版本升级或其他原因, 本文档内容会不定期进行更新。除非另有约定, 本文档仅作为使用指导, 本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

**技术支持**

**地址**
北京市海淀区丰豪东路 9 号院中关村集成电路设计园 (ICPARK)1 号楼

**邮编**
100094

**网址**
https://www.sophgo.com/

**邮箱**
sales@sophgo.com

**电话**
010-57590723

**发布记录**

| 版本 | 发布日期 | 说明 |
| --- | --- | --- |
| v0.1.0 | 2025.01.24 | 初版发布, 介绍 SGTPUV8 的软件使用指南。 |

CHAPTER 1

# SGTPU 软件简介

提供算能开源芯片 SGTPU 配套软件程序，包括:

1. 神经网络模型编译器 TPU-MLIR
2. 神经网络模型编译示例 SGTPU-ModelZoo
3. 模型运行时库 libsophon
4. 神经网络算子库 SGTPU-KernelModule
5. 芯片仿真库 SGTPU-Cmodel

目前支持 SGTPUV8。

使用以上程序可实现将不同框架神经网络模型 (pytorch、ONNX 等) 转换为在 SGTPU 上高效运行的二进制文件-bmodel，并使用模型运行时库调用 SGTPU 来执行 bmodel，以实现 SGTPU 对神经网络模型的加速。

其中 TPU-MLIR 提供了一套完整的工具链用于转换不同框架神经网络模型，SGTPU-ModelZoo 提供了一些预训练好的开源模型和执行 TPU-MLIR 的编译脚本示例，libsophon 提供 bmodel 运行时库、驱动以及 bmodel 测试运行程序。

SGTPU-KernelModule 提供基于 SGTPU 指令编写的神经网络算子动态库，SGTPU-Cmodel 提供在 x86 机器上模拟 SGTPU 指令行为的芯片仿真动态库。神经网络算子动态库与芯片仿真动态库作为 TPU-MLIR 和 libsophon 的依赖库使用。

CHAPTER 2

# 模型转换

## 2.1 从已有网络模型快速开始

从 https://github.com/sophgo/SGTPU-ModelZoo 获取 SGTPU-ModelZoo 源码，并按照主页介绍下载源模型，执行模型导航中的模型转换脚本来获取 bmodel。

需要保留这一步产生包含 compilation.bmodel 的文件夹。

## 2.2 手动转换网络模型为 bmodel

参考 pytorch 或 onnx 导出教程，从神经网络框架中导出静态网络模型 (.pt 或.onnx 格式)。

从 https://github.com/sophgo/tpu-mlir/tree/sgtpuv8 获取 TPU-MLIR 源码，注意使用 sgtpuv8 分支，并按照主页介绍完成 model_transform、run_calibration、model_deploy 步骤，实现单步从神经网络模型到 bmodel 的转换。

需要保留这一步产生包含 compilation.bmodel 的文件夹。

CHAPTER 3

# 运行时库使用

## 3.1 运行时库安装

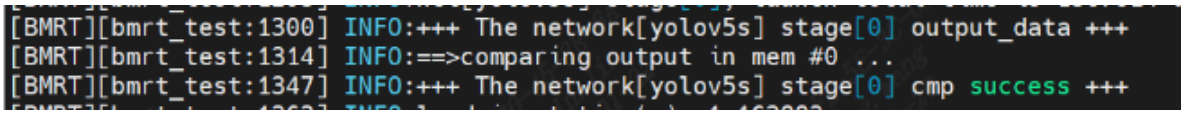
从 https://github.com/sophgo/libsophon/tree/SGTPUV8 获取 libsophon 源码，注意需要 SGTPUV8 分支。按照主页步骤编译、安装 libsophon。

安装完成后会在 {/path/to/libsophon}/bin 目录下会出现模型测试程序 bmrt_test。

将模型转换步骤中得到包含 compilation.bmodel 的模型目录作为输入，执行以下命令：

```
$ .{/path/to/libsophon}/bin/bmrt_test --context_dir {path_to_compilation_dir}
```

出现以下内容时表示运行模型且比对成功：




注意：这部分内容需要在SOC环境使用

## 3.2 推理程序开发

参考 https://github.com/sophgo/libsophon/tree/SGTPUV8/tpu-runtime tpu-runtime 仓库的主页介绍，使用 bmlib 和 BMRuntime 推理接口编写推理程序。

推理程序示例代码：https://github.com/sophgo/libsophon/blob/SGTPUV8/tpu-runtime/docs/reference/source_zh/bmruntime_sample/bmruntime_sample.rst。

注意：这部分内容需要在SOC环境使用

CHAPTER 4

# 源码链接

TPU-MLIR: https://github.com/sophgo/tpu-mlir/tree/sgtpuv8

SGTPU-ModelZoo: https://github.com/sophgo/SGTPU-ModelZoo

libsophon: https://github.com/sophgo/libsophon/tree/SGTPUV8

SGTPU-KernelModule: https://github.com/sophgo/tpu-mlir/blob/sgtpuv8/third_party/nntoolchain/lib/libsgtpuv8_kernel_module.so

SGTPU-Cmodel: https://github.com/sophgo/tpu-mlir/blob/sgtpuv8/third_party/nntoolchain/lib/libcmodel_sgtpuv8.so