

推理系统系列

推理引擎



ZOMI



Talk Overview

1. 推理系统介绍

- 推理系统与推理引擎
- 推理系统的工作流程
- 推理系统生命周期管理
- 推理引擎介绍

2. 模型小型化

- NAS神经网络搜索
- CNN小型化结构
- Transform小型化结构

3. 离线优化压缩

- 低比特量化
- 二值化网络
- 模型模型剪枝
- 模型模型蒸馏

4. 部署和运行优化

- 图转换优化（算子融合/重排/替换）
- 并发执行与内存分配
- 动态batch与bin Packing

Talk Overview

1. 推理系统与推理引擎
2. 推理系统的工作流程
3. 推理系统生命周期管理
4. 推理引擎介绍
 - 推理引擎特点
 - 技术挑战
 - 整体架构
 - 工作流程

整体架构

Architecture



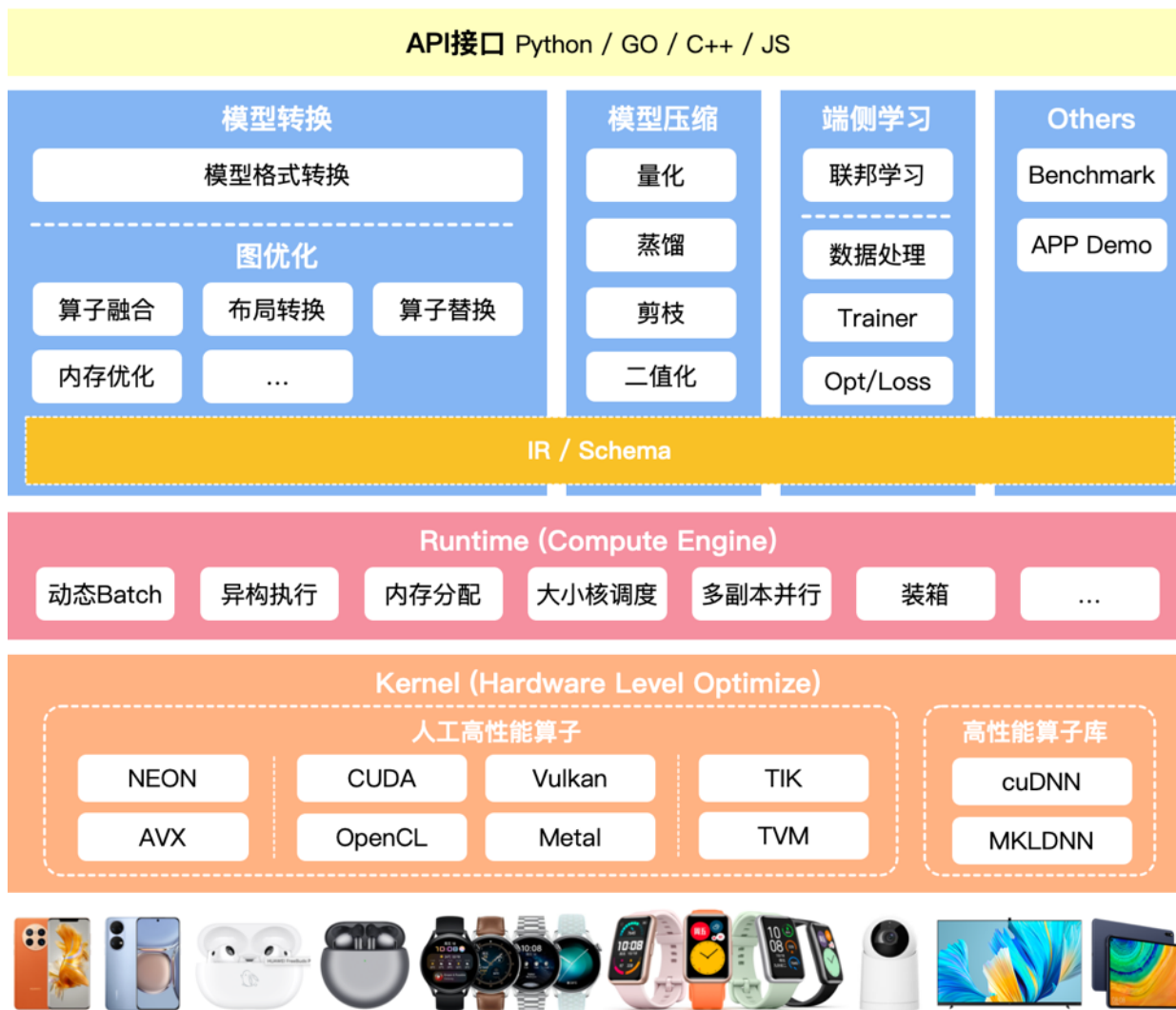
推理引擎架构

- **优化阶段**

模型转换工具，由转换和图优化构成；
模型压缩工具、端侧学习和其他组件组成。

- **运行阶段**

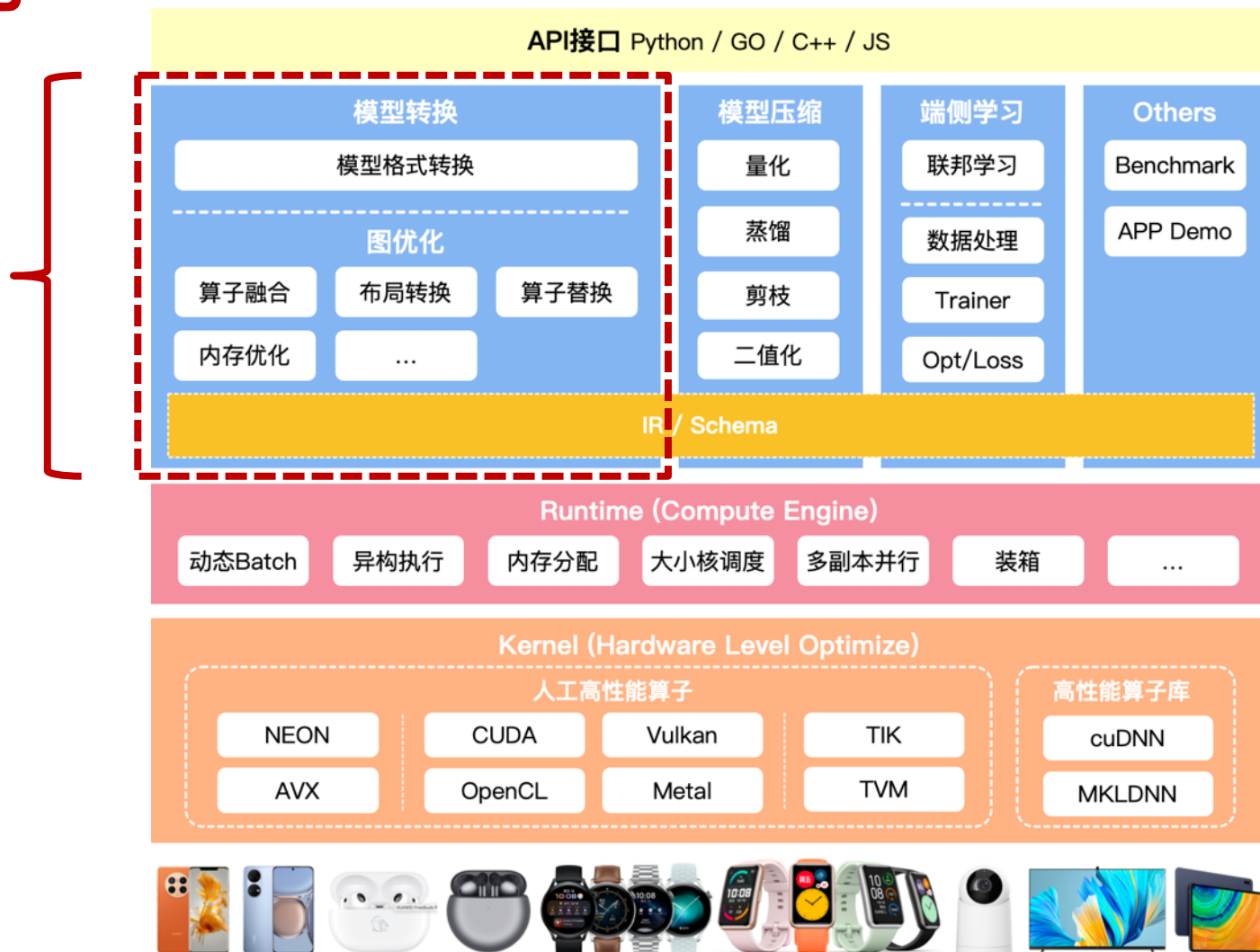
即推理引擎，负责AI模型的加载与执行，
可分为调度与执行两层。



推理引擎架构

模型转换工具

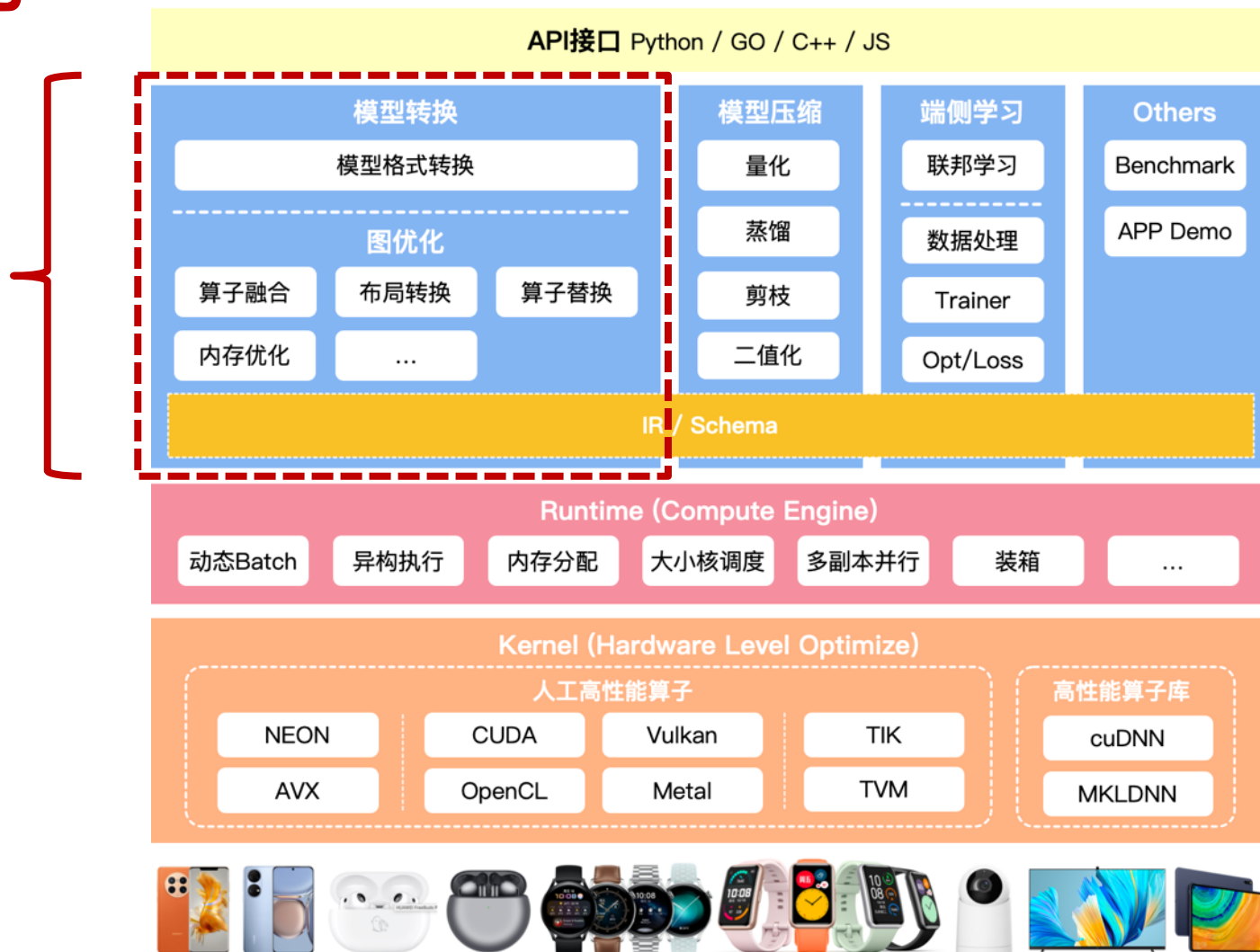
- 模型格式转换
- 计算图优化



推理引擎架构

模型转换工具

- 模型格式转换
- 计算图优化



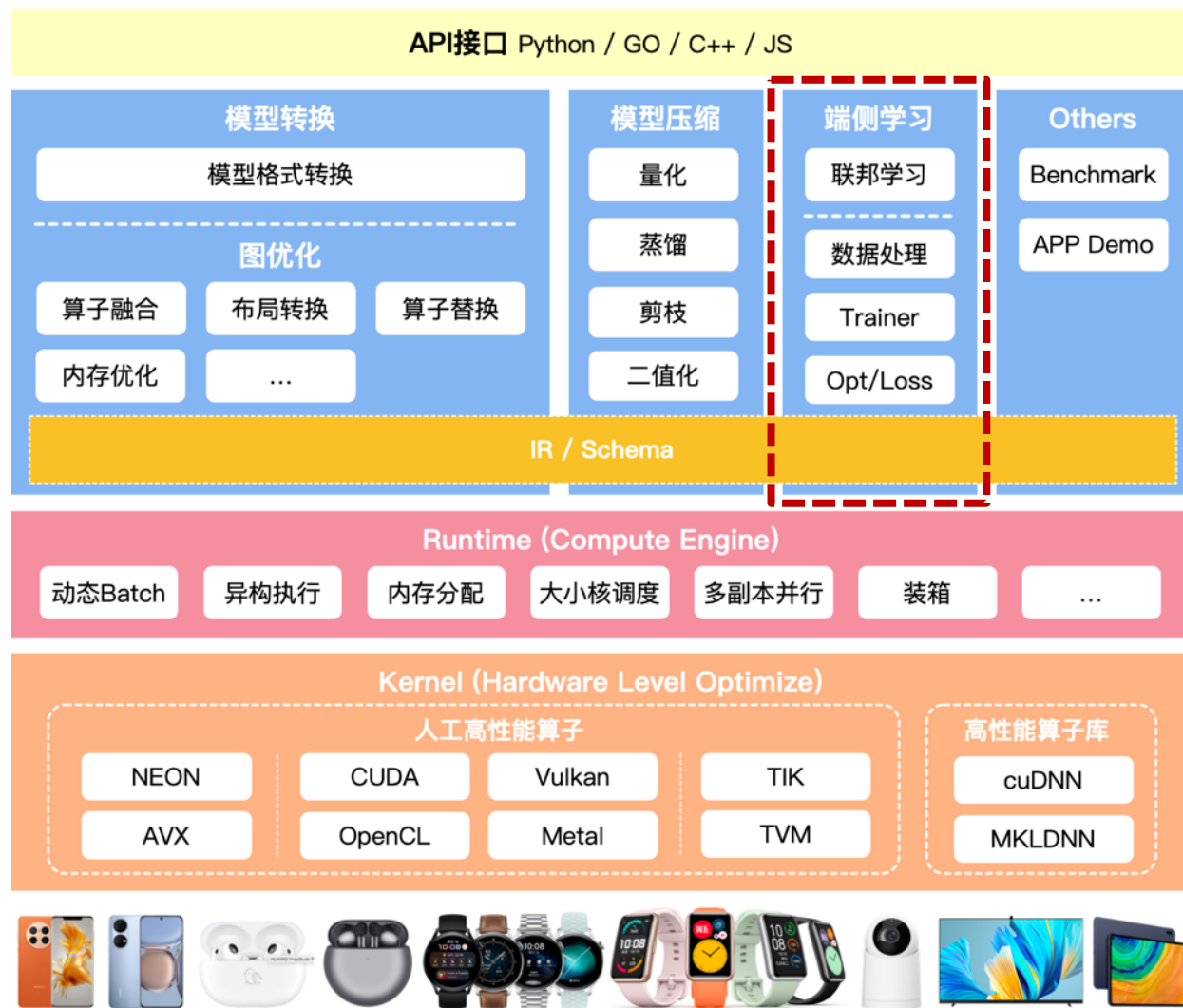
推理引擎架构

对模型进行压缩

- 减少模型大小
- 加快训练速度
- 保持相同精度



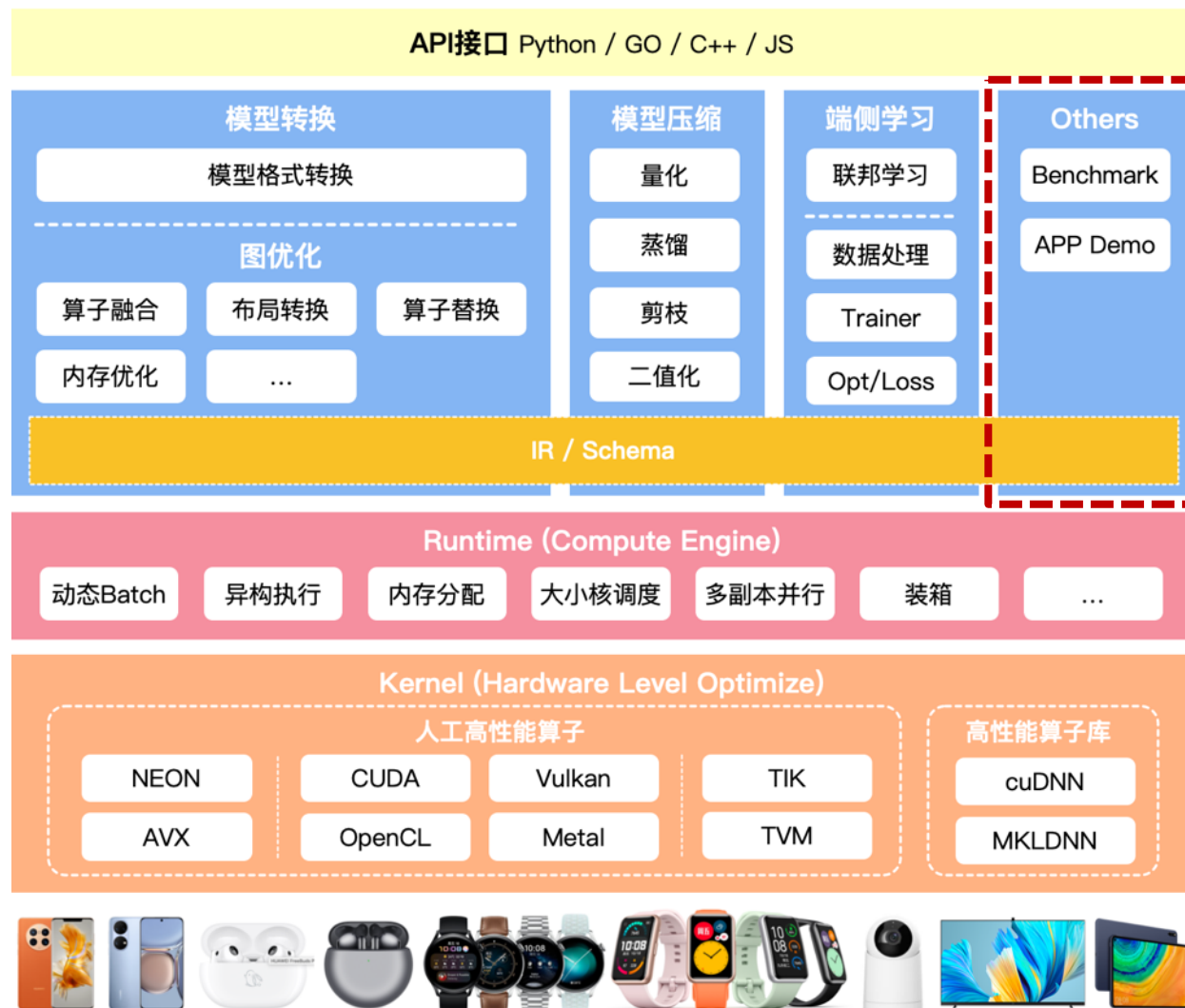
推理引擎架构



端侧学习

- 增量学习
- 联邦学习

推理引擎架构



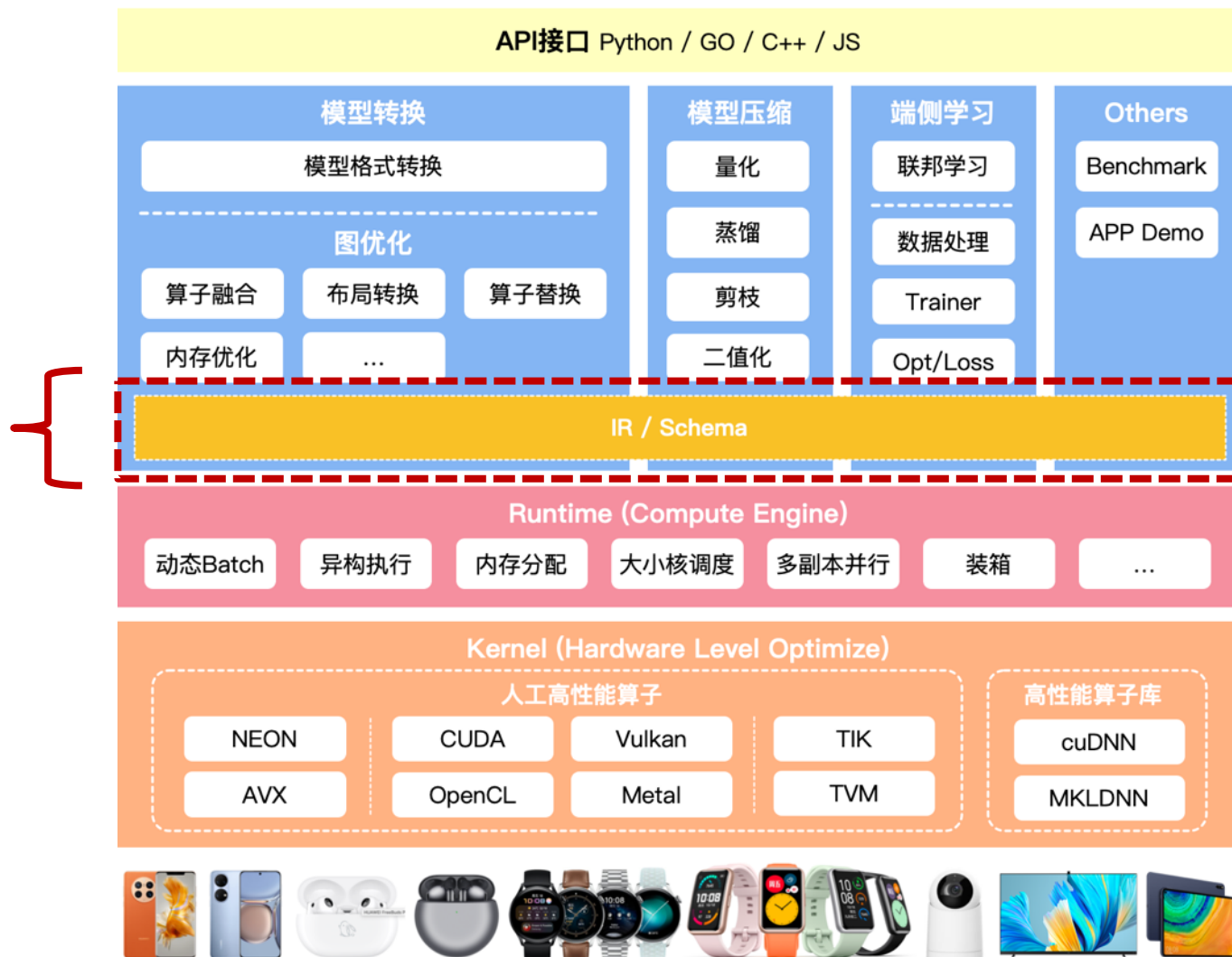
其他模块

- 性能对比
- 集成模块

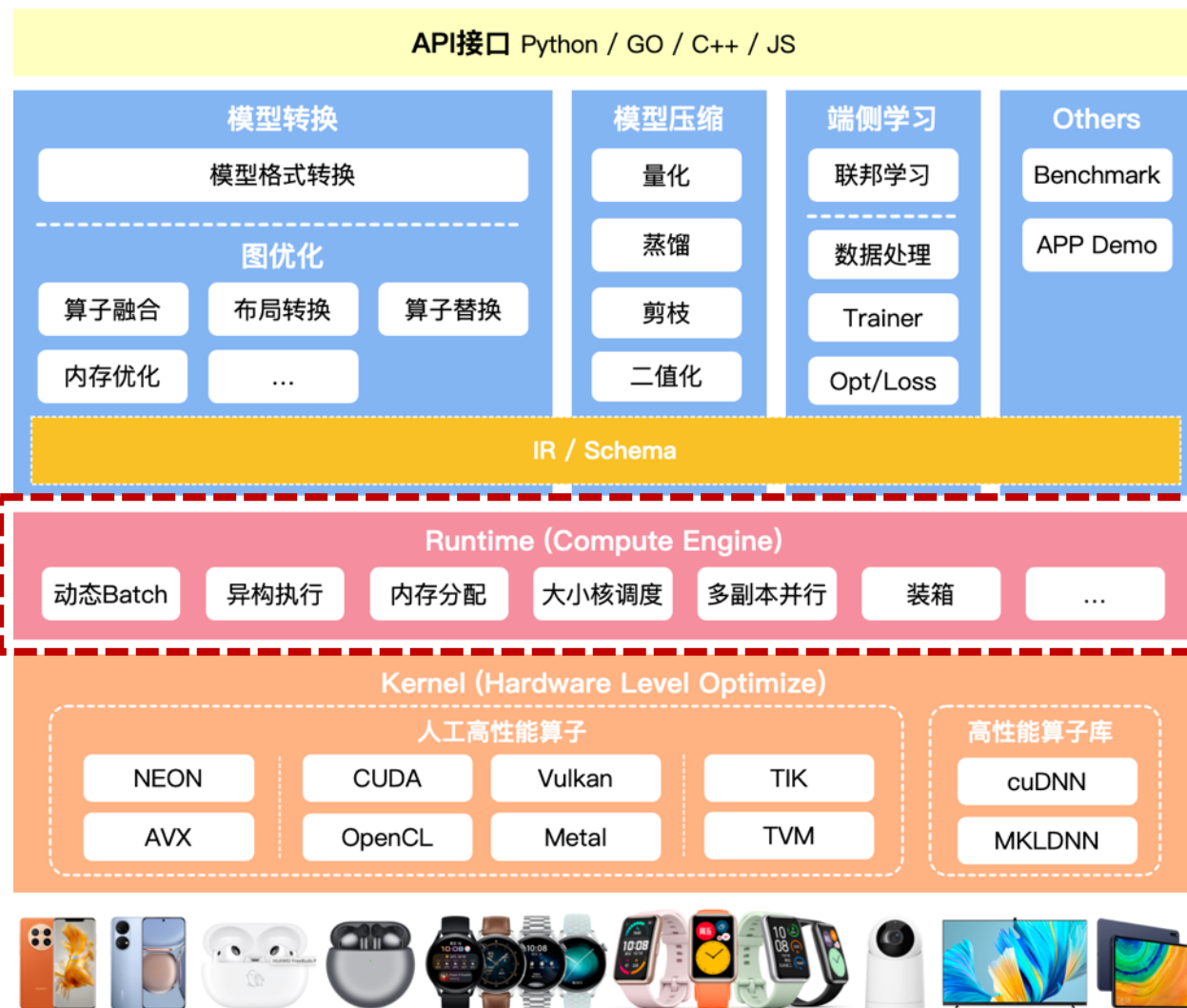
推理引擎架构

中间表达

- Schema
- 统一表达



推理引擎架构

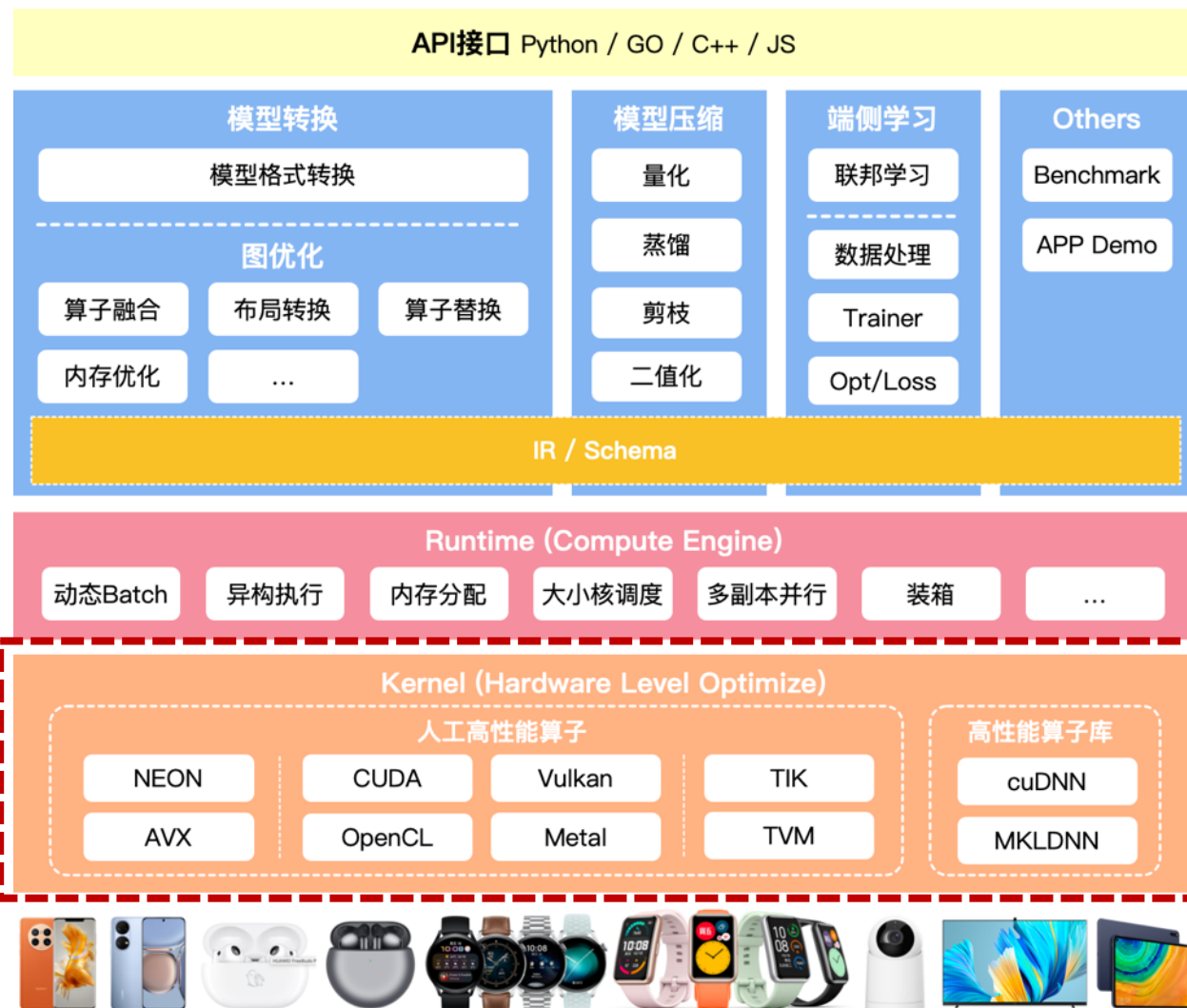


Runtime

- 模型加载
- 模型执行



推理引擎架构



高性能算子层

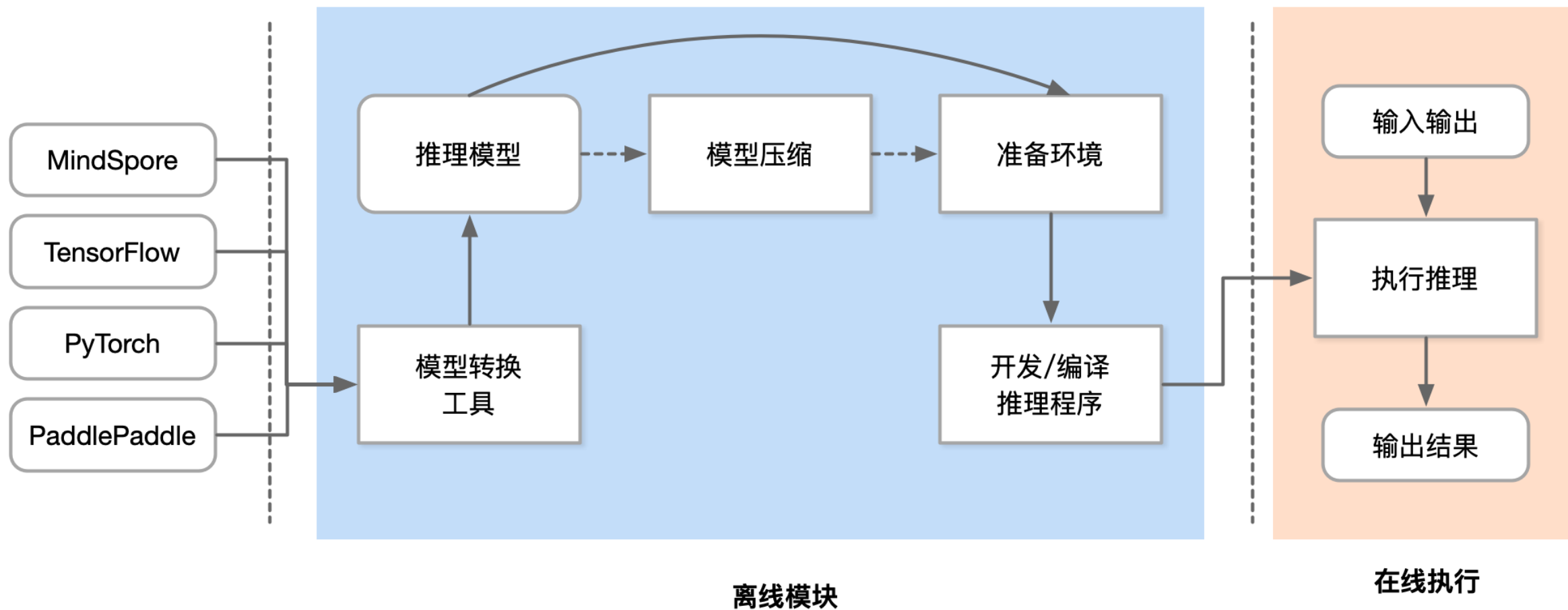
- 算子优化
- 算子执行
- 算子调度

工作流程

Workflow



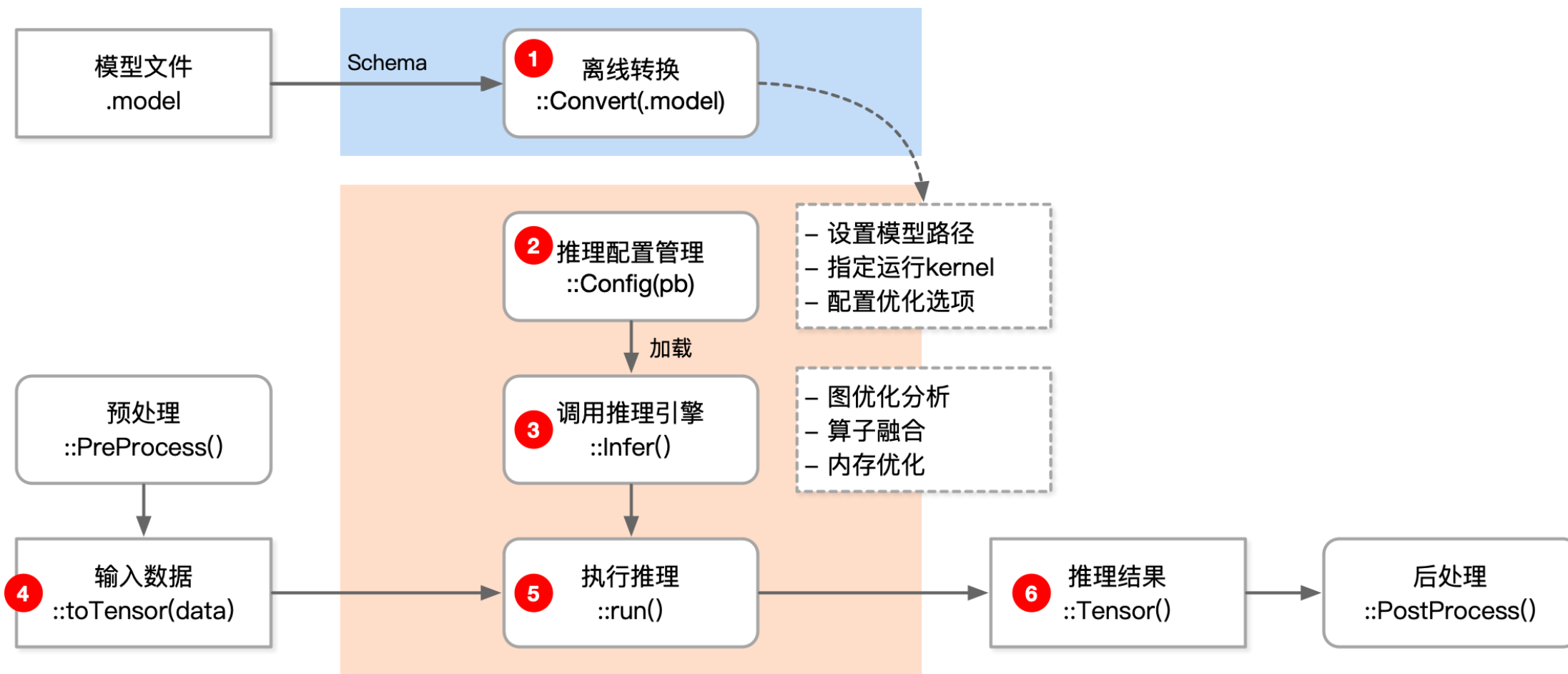
推理流程



开发推理程序

1. 配置推理选项 `::Config` , 包括设置模型路径、运行设备、开启/关闭计算图优化等
2. 创建推理引擎对象 `::Predictor(Config)` , 其中 `Config` 为配置推理选项
3. 准备输入数据
 - 将原始输入数据根据模型需要做相应的预处理 (比如减均值等标准化操作)
 - 先通过 `auto input_names = predictor->GetInputNames()` 获取模型所有输入 Tensor 名称
 - 通过 `auto tensor = predictor->GetInputTensor(input_names[i])` 获取输入 Tensor 的指针
 - 通过 `tensor->copy(data)` , 将预处理之后的数据 `data` 拷贝/转换到 `tensor` 中
4. 执行推理 , 运行 `predictor->Run()` 完成推理执行
5. 获得推理结果并进行后处理
 - 通过 `auto out_names = predictor->GetOutputNames()` 获取模型所有输出 Tensor 名称
 - 通过 `auto tensor = predictor->GetOutputTensor(out_names[i])` 获取输出 Tensor 指针
 - 通过 `tensor->copy(data)` , 将 `tensor` 数据拷贝/转换到 `data` 指针上
 - 批量推理验证数据集并计算模型精度判断推理结果的正确性
 - 将模型推理输出数据进行后处理 (根据检测框位置裁剪图像等)

开发推理程序



参考文献



BUILDING A BETTER CONNECTED WORLD

THANK YOU

Copyright©2014 Huawei Technologies Co., Ltd. All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.