NXP平台模型部署步骤及说明：

1. 下载onnxruntime，查看nxp支持版本及支持cpu或gpu、npu加速，版本、硬件类型、32或64位须对应正确；
2. 选择c++语言类型，完成交叉编译；
3. 调用onnxruntime的接口，完成模型调用。

推理过程：

1. 创建build
2. 创建engine
3. 传入数据
4. Inference

TensorRT的int8精度

1. weight进行使用不饱和的方式
2. activate使用饱和（saturation）进行量化

需要编写一个自己的calibrator类，然后通过builder->setInt8Calibrator(calibrator)告诉builder使用这个calibrator来做数据标定，从而减小量化误差。

Calibrator：

getBatchSize

getBatch

writeCalibrationCache(optional)

readCalibrationCache(optional)