

模型部署与推理

- 贾志刚



内容提纲

- 预训练模型库
- 模型推理框架



1.模型库



框架自带的模型库

Model name	Speed (ms)	COCO mAP[^1]	Outputs
ssd_mobilenet_v1_coco	30	21	Boxes
ssd_mobilenet_v1_0.75_depth_coco ☆	26	18	Boxes
ssd_mobilenet_v1_quantized_coco ☆	29	18	Boxes
ssd_mobilenet_v1_0.75_depth_quantized_coco ☆	29	16	Boxes
ssd_mobilenet_v1_ppn_coco ☆	26	20	Boxes
ssd_mobilenet_v1_fpn_coco ☆	56	32	Boxes
ssd_resnet_50_fpn_coco ☆	76	35	Boxes
ssd_mobilenet_v2_coco	31	22	Boxes
ssd_mobilenet_v2_quantized_coco	29	22	Boxes
ssdlite_mobilenet_v2_coco	27	22	Boxes
ssd_inception_v2_coco	42	24	Boxes
faster_rcnn_inception_v2_coco	58	28	Boxes
faster_rcnn_resnet50_coco	89	30	Boxes
faster_rcnn_resnet50_lowproposals_coco	64		Boxes
rfcn_resnet101_coco	92	30	Boxes

rfcn_resnet101_coco	92	30	Boxes
faster_rcnn_resnet101_coco	106	32	Boxes
faster_rcnn_resnet101_lowproposals_coco	82		Boxes
faster_rcnn_inception_resnet_v2_atrous_coco	620	37	Boxes
faster_rcnn_inception_resnet_v2_atrous_lowproposals_coco	241		Boxes
faster_rcnn_nas	1833	43	Boxes
faster_rcnn_nas_lowproposals_coco	540		Boxes
mask_rcnn_inception_resnet_v2_atrous_coco	771	36	Masks
mask_rcnn_inception_v2_coco	79	25	Masks
mask_rcnn_resnet101_atrous_coco	470	33	Masks
mask_rcnn_resnet50_atrous_coco	343	29	Masks



框架自带的模型库

- https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/detection_model_zoo.md
- <https://github.com/BVLC/caffe/wiki/Model-Zoo>
- https://github.com/opencv/open_model_zoo
- `torch.utils.model_zoo.load_url(url, model_dir=None, map_location=None, progress=True, check_hash=False)`



框架自带的模型库

- 迁移学习训练起始点
- 快速模型验证与应用演示开发
- 部分直接使用，节省开发时间



3.推理框架



推理框架

- OpenCV DNN
- OpenVINO
- Tensorflow Lite



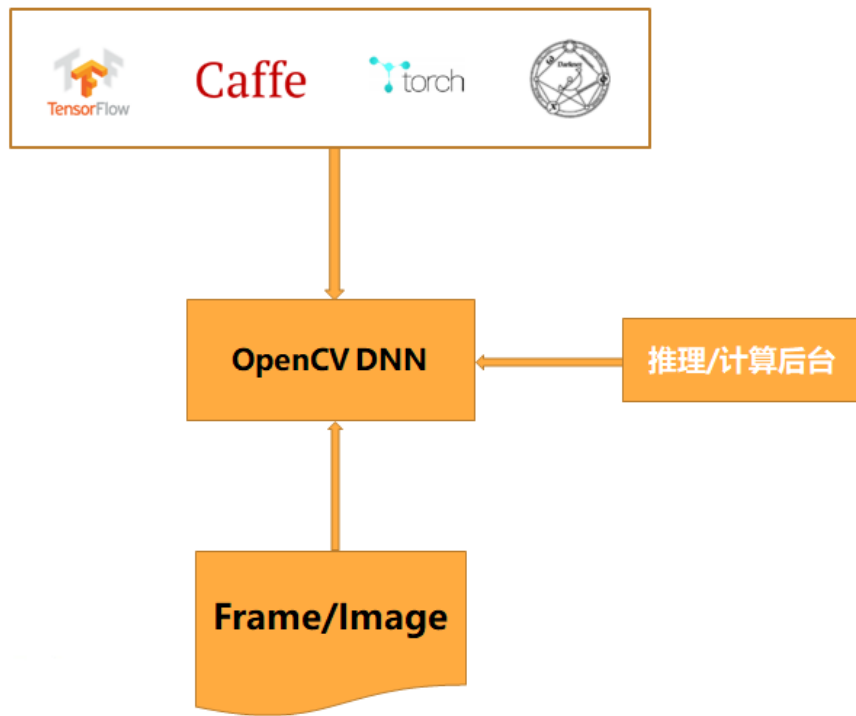
OpenCV DNN

- - 基于Tiny DNN改进，只支持推理
- - 支持多种深度学习框架导出模型读取 (Tensorflow/Caffe/torch/Darknet)
- - 深度整合tensorflow 对象检测框架
- - OpenCV4.x版本支持
- <https://github.com/opencv/opencv/tree/master/samples/dnn>

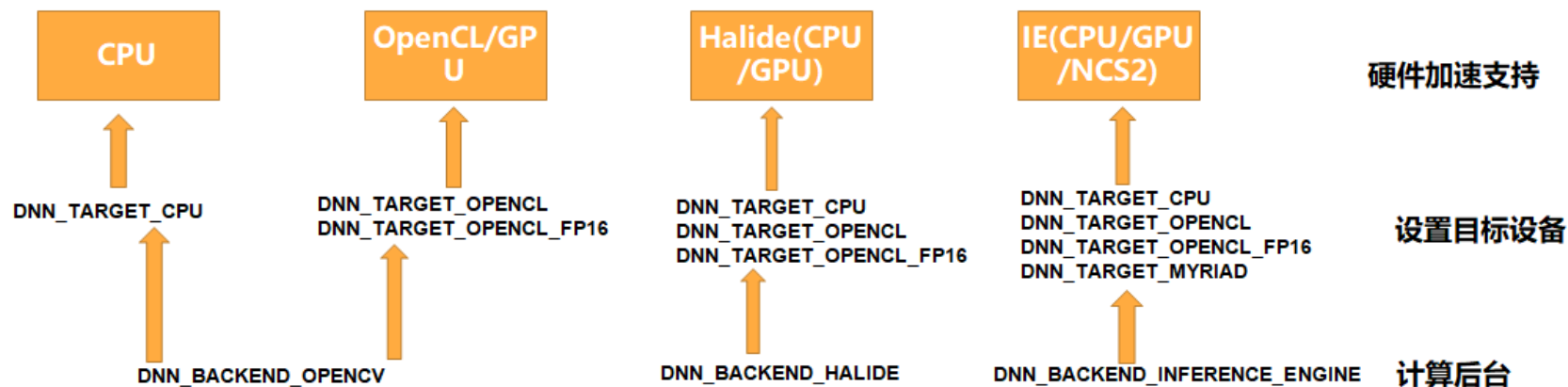


OpenCV DNN支持

- - 图像分类
- - 对象检测
- - 人脸检测
- - 图像分割
- - 姿态评估
- - 文本检测
- - 图像分割化



OpenCV DNN-计算后台与精度





// 设置计算后台


```
net.setPreferableBackend(DNN_BACKEND_INFERENCE_ENGINE);  
net.setPreferableTarget(DNN_TARGET_CPU);
```

OpenCV DNN-模型转换支持

- - 模型转换脚本
- 需要参数支持

 tf_text_graph_faster_rcnn.py

 tf_text_graph_mask_rcnn.py

 tf_text_graph_ssd.py

- -- config, 训练时候pipeline配置文件
- -- input 导出BP文件
- -- output 导出的pbtxt文件路径



OpenCV DNN - API使用

- 加载模型

// 加载Faster-RCNN

```
Net net = readNetFromTensorflow(model, config);  
net.setPreferableBackend(DNN_BACKEND_OPENCV);  
net.setPreferableTarget(DNN_TARGET_CPU);
```

- 设置输入:

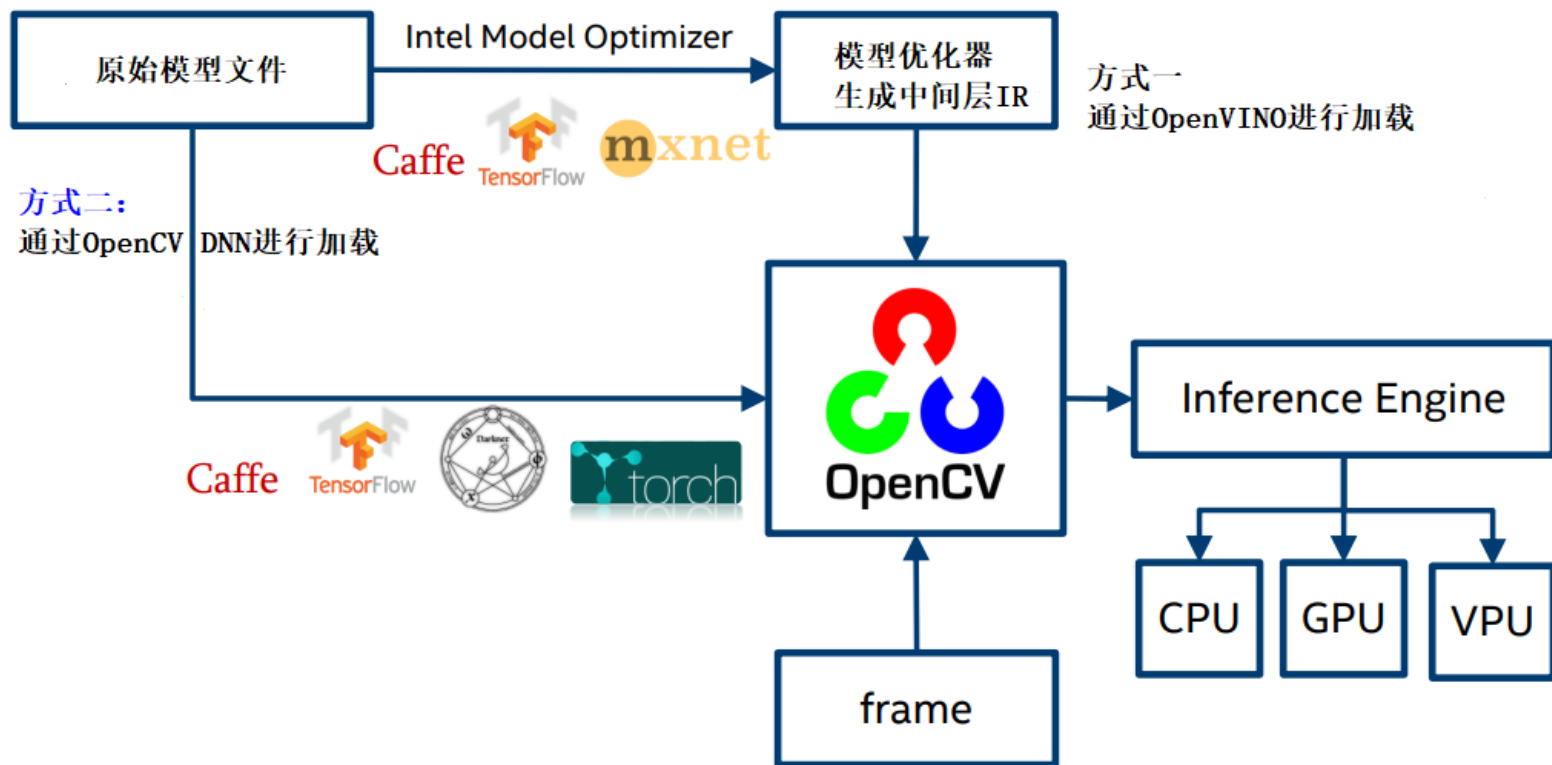
```
net.setInput(blob, "data");
```

- 推理预测:

```
Mat detection = net.forward("detection_out");
```



OpenCV DNN – 加速



OpenCV DNN-特点

- 离线部署
- 支持多系统平台
- 模型自动优化

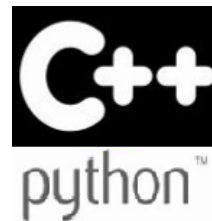


OpenVINO介绍

- 快速构建视觉应用原型系统与解决方案
- 在端侧设备上实时运行模型
- 当前最新版本-2019R03



DL/CV框架，系统，语言支持



Common API支持层 - OpenVINO™ Toolkit

插件支持层

MKLDNN

cIDNN

FPGA

Myriad



COREi5
COREi7
6th~8th
XEON

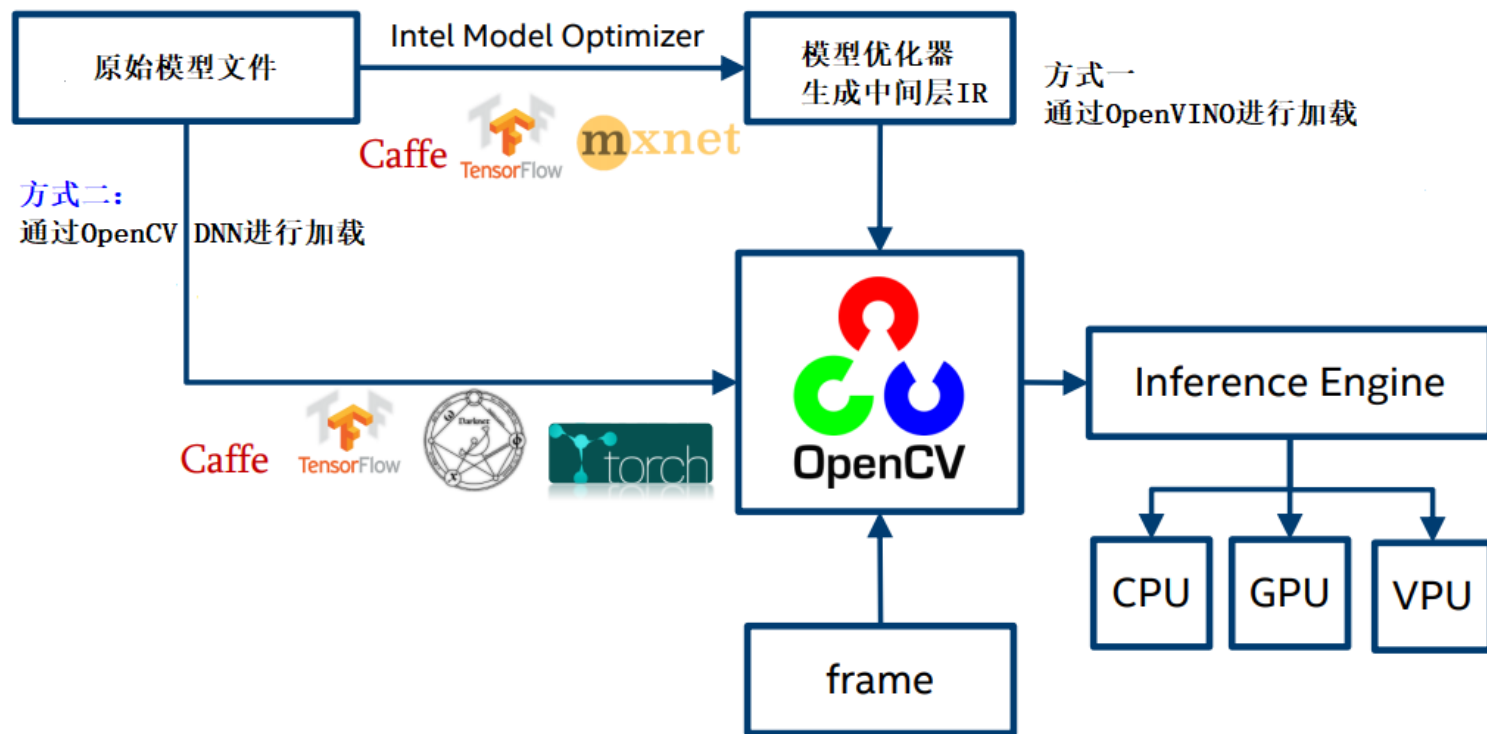
HD Graphics
UHD
Graphics

FPGA

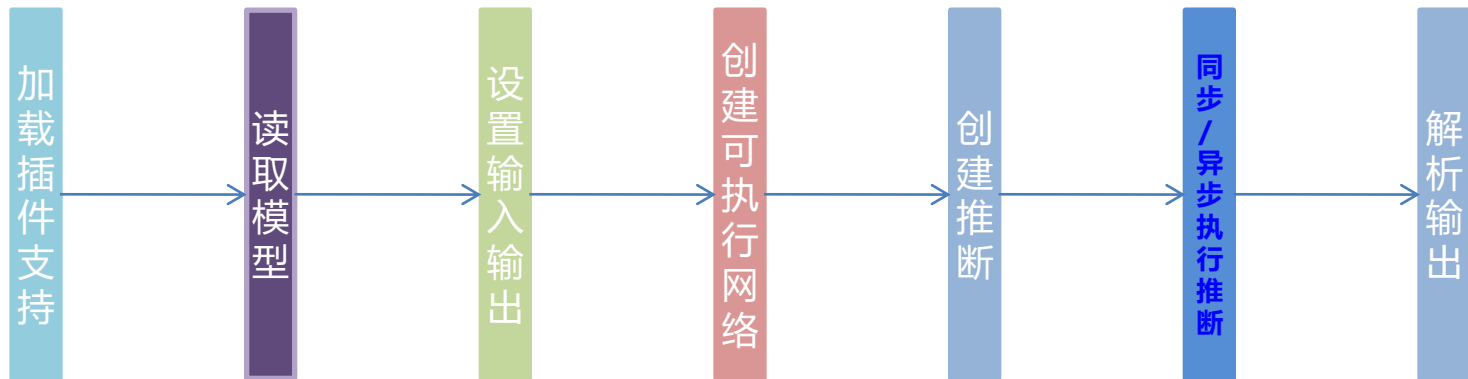
VPU/NCS

硬件支持

集成OpenCV DNN支持



程序流程



API代码

// 创建IE插件

```
InferenceEnginePluginPtr engine_ptr = PluginDispatcher(dirs).getSuitablePlugin(TargetDevice::eCPU);  
InferencePlugin plugin(engine_ptr);
```

// 加载CPU扩展库支持

```
plugin.AddExtension(std::make_shared<Extensions::Cpu::CpuExtensions>());
```

// SSD 车辆检测模型

```
CNNNetReader network_reader;  
network_reader.ReadNetwork(model_xml);  
network_reader.ReadWeights(model_bin);
```

// 请求网络输入与输出信息

```
auto network = network_reader.getNetwork();  
InferenceEngine::InputsDataMap input_info(network.getInputsInfo());  
InferenceEngine::OutputsDataMap output_info(network.getOutputsInfo());
```

/**设置输入精度与维度**/

```
for (auto &item : input_info) {  
    auto input_data = item.second;  
    input_data->setPrecision(Precision::U8);  
    input_data->setLayout(Layout::NCHW);  
}
```

/** 设置输出精度与内容**/

```
for (auto &item : output_info) {  
    auto output_data = item.second;  
    output_data->setPrecision(Precision::FP32);  
}
```

// 创建可执行网络对象

```
auto executable_network = plugin.LoadNetwork(network, {});
```

// 请求推断图

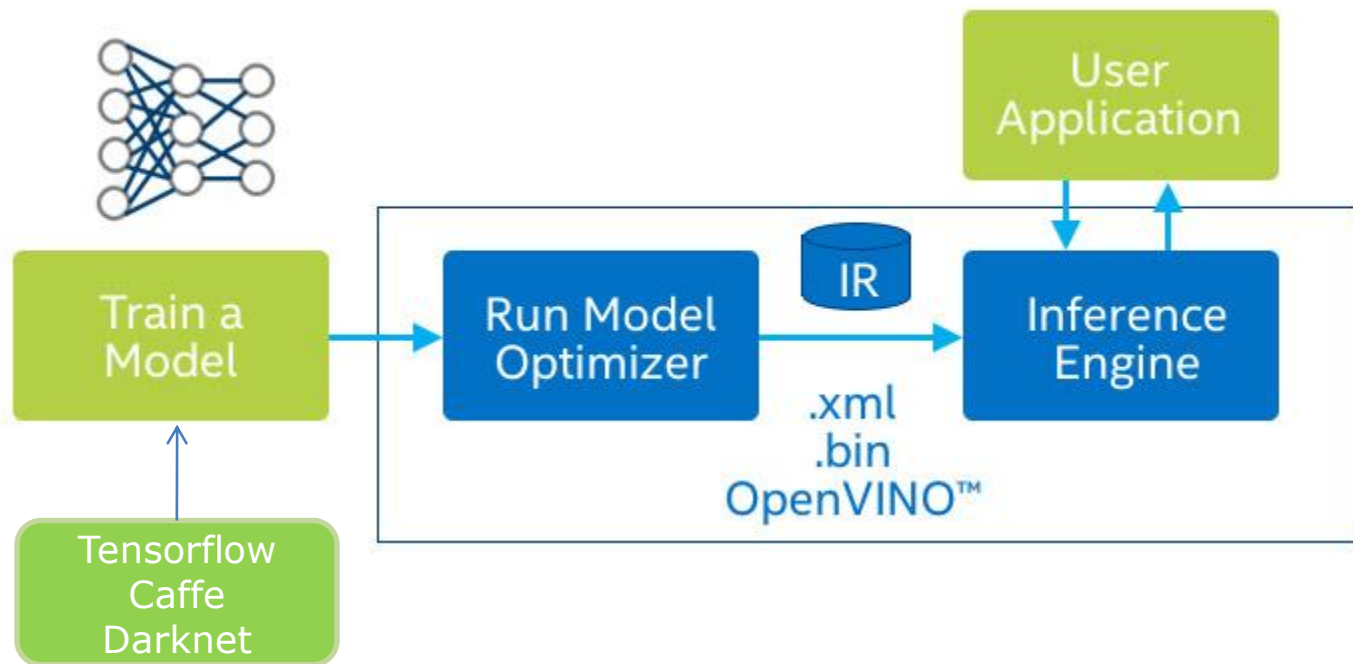
```
auto infer_request = executable_network.CreateInferRequest();
```

// 执行预测

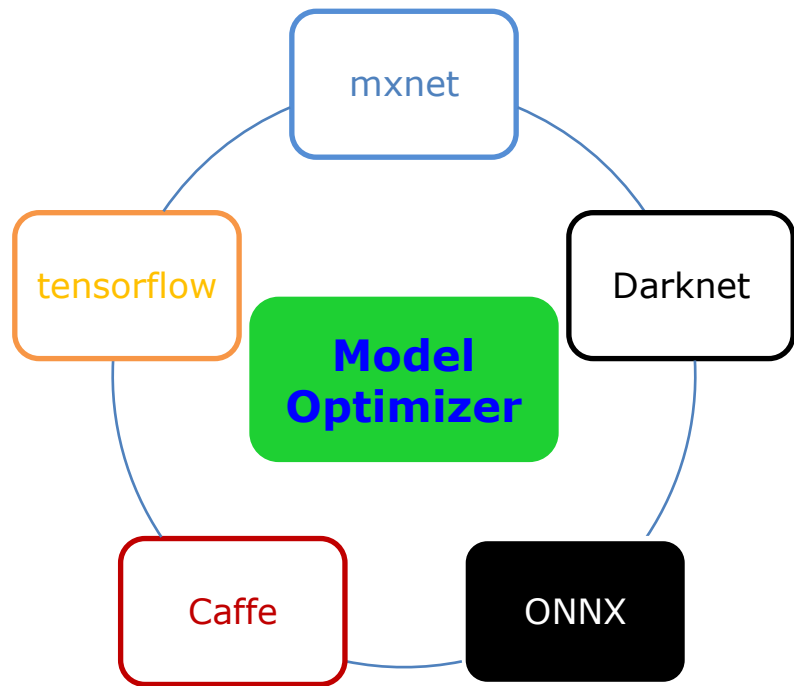
```
infer_request.Infer();
```



模型转换



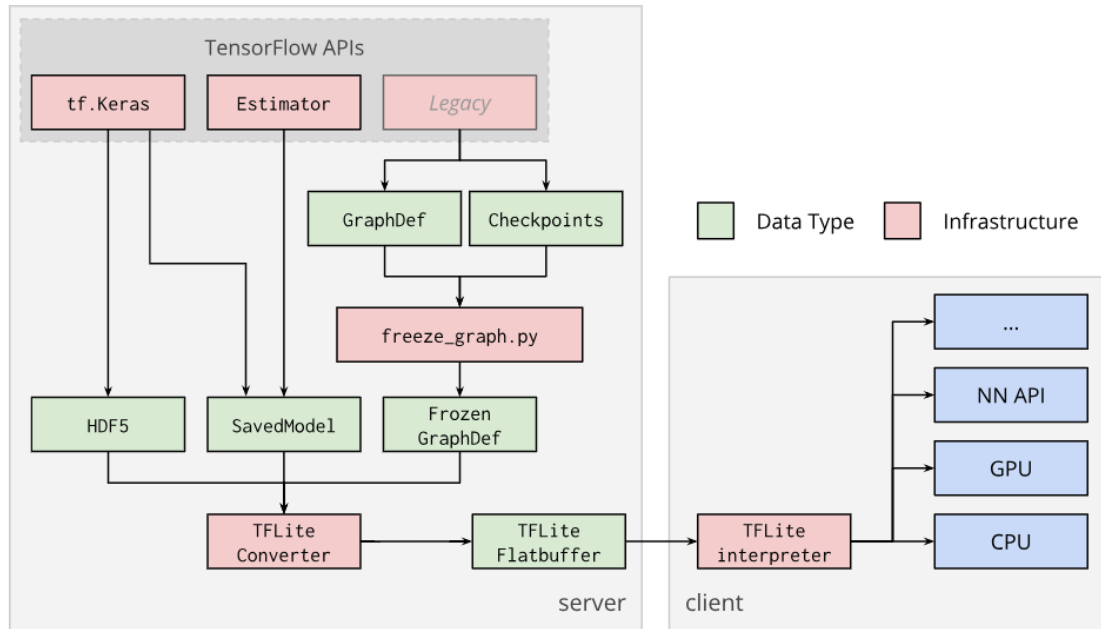
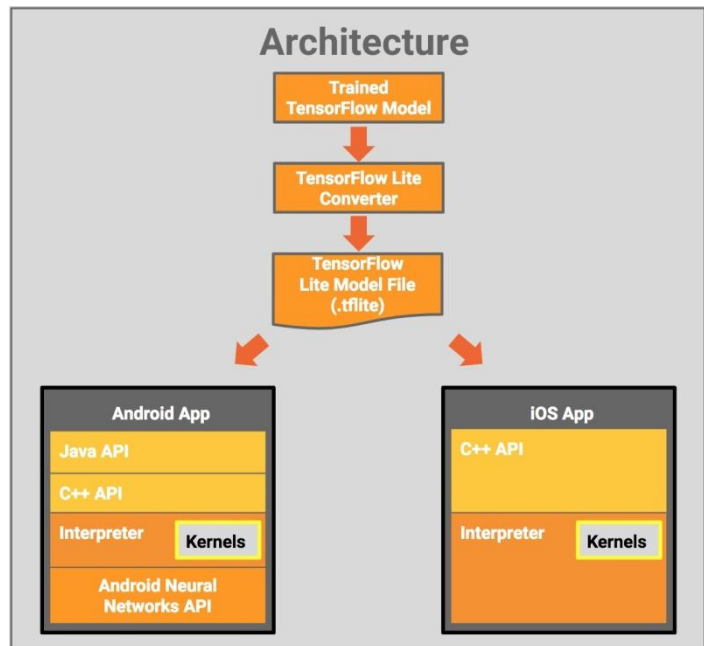
不同DL框架模型支持



- 脚本支持转换
- 自定义层解析
- 支持不同精度
- 支持任意尺寸



Tensorflow Lite



Tensorflow Lite

- FP32/FP16/INT8

Network	Model Parameters	Top-1 Accuracy on ImageNet (fp32)
Mobilenet_V1_0.25_128	0.47M	0.415
Mobilenet_V2_1_224	3.54M	0.719
Mobilenet_V1_1_224	4.25M	0.709
Nasnet_Mobile	5.3M	0.74
Mobilenet_V2_1.4_224	6.06M	0.749
Inception_V3	23.9M	0.78
Resnet_v1_50	25.6M	0.752
Resnet_v2_50	25.6M	0.756
Resnet_v1_152	60.4M	0.768
Resnet_v2_152	60.4M	0.778



ANY QUESTION?

