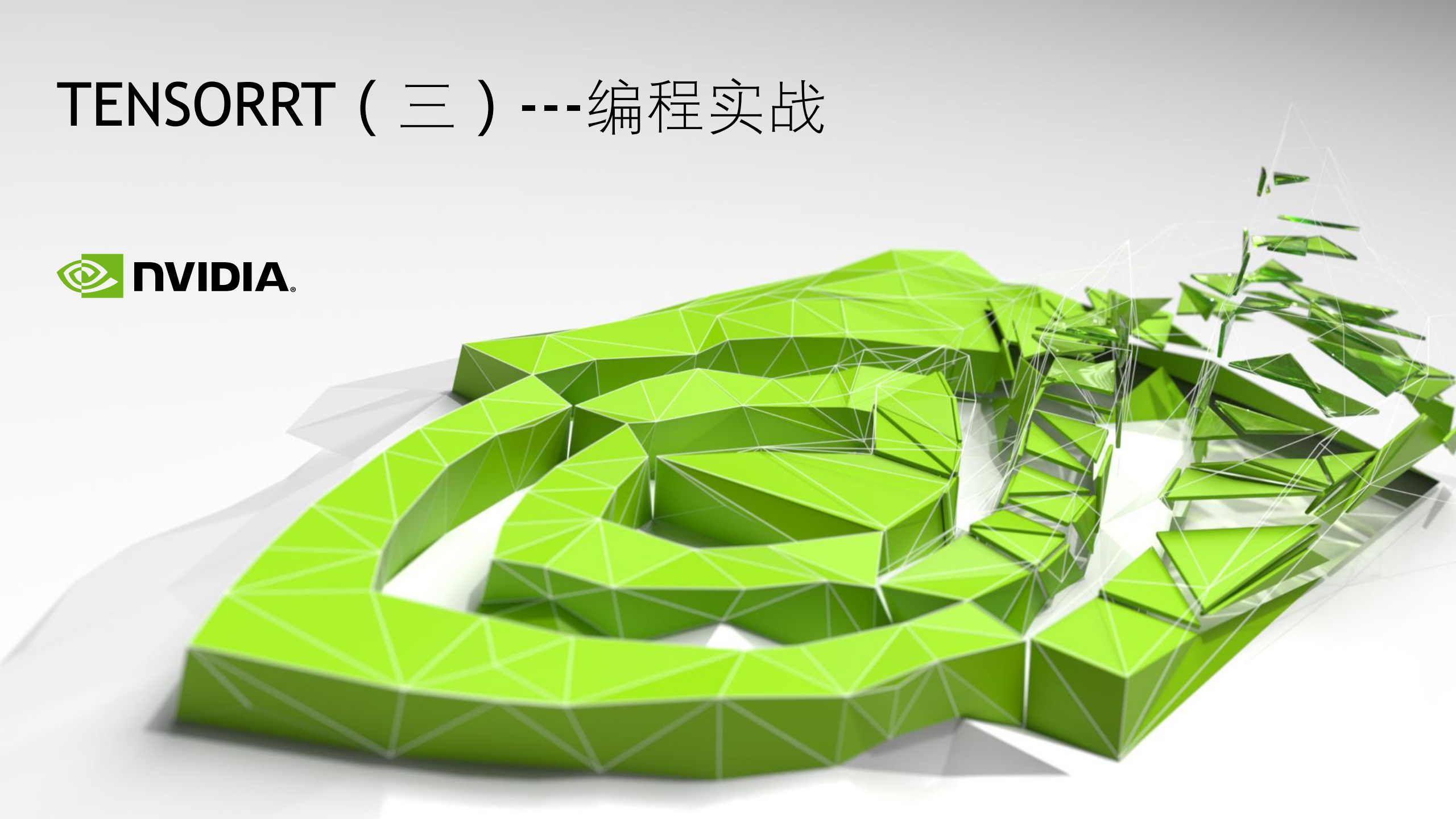
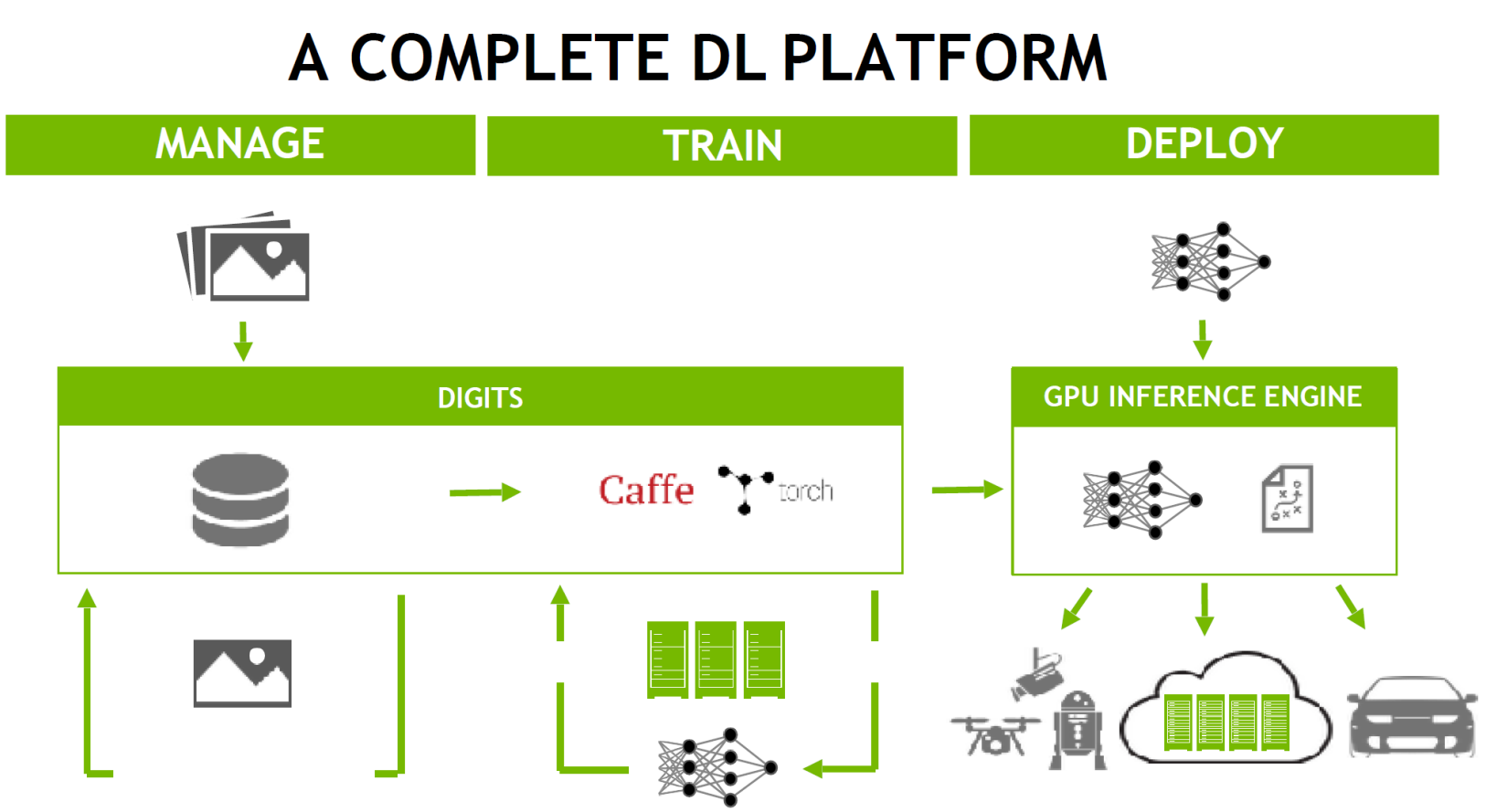
TENSORRT（三）---编程实战



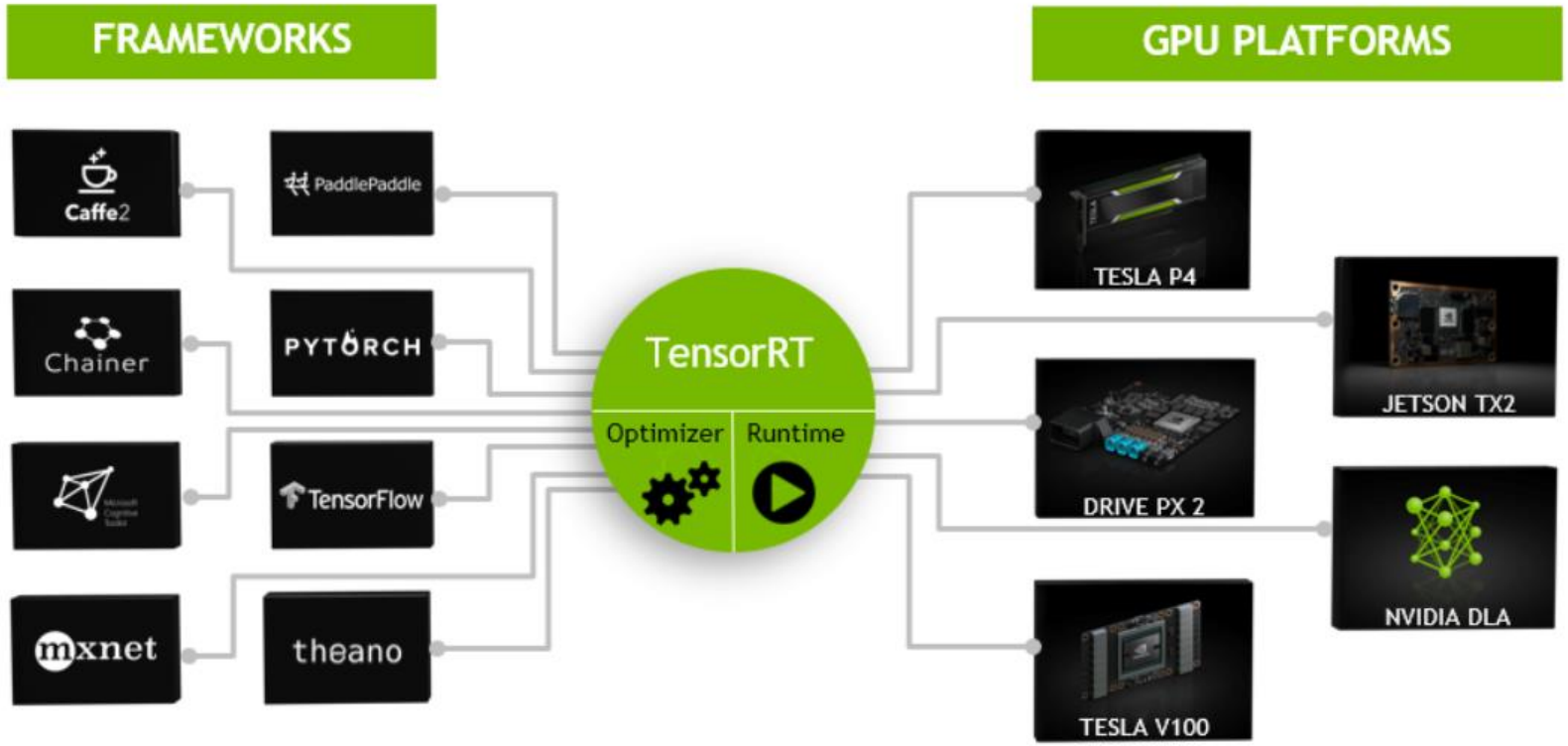
概述

- TensoRT回顾
- TensorRT—Plugin
- TensorRT编程模型—SSD
- TensoRT实例展示

TensorRT回顾

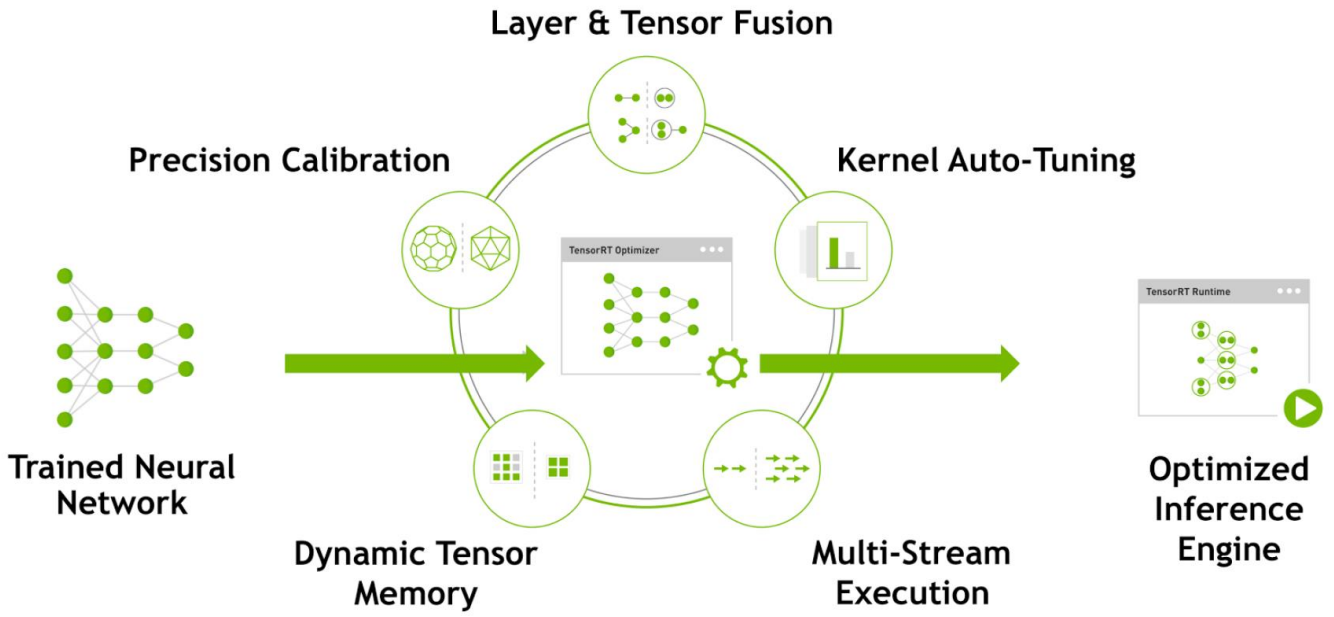

A COMPLETE DL PLATFORM
MANAGE
TRAIN
DEPLOY
DIGITS
Caffe
torch
GPU INFERENCE ENGINE

TensorRT回顾


FRAMEWORKS
Caffe2
PaddlePaddle
Chainer
PYTORCH
TensorFlow
mxnet
theano
TensorRT
Optimizer
Runtime
GPU PLATFORMS
TESLA P4
JETSON TX2
DRIVE PX 2
NVIDIA DLA
TESLA V100

TensorRT回顾

TENSORRT: WORK FLOW



TensorRT支持的网络层

Caffe

The following list describes the operations that are supported in a Caffe framework.

- Convolution
- Pooling
- InnerProduct
- SoftMax
- ReLU, TanH, Sigmoid
- LRN
- Power
- ElementWise
- Concatenation
- Deconvolution
- BatchNormalization
- Scale
- Crop
- Reduction
- Reshape
- Permute
- Dropout

TensorRT支持的网络层

TensorFlow

The following list describes the operations that are supported in a TensorFlow framework.

- Placeholder
- Const
- Add, Sub, Mul, Div, Minimum and Maximum
- BiasAdd
- Negative, Abs, Sqrt, Rsqrt, Pow, Exp and Log
- FusedBatchNorm
- ReLU, TanH, Sigmoid
- SoftMax
- Mean
- ConcatV2
- Reshape
- Transpose
- Conv2D
- DepthwiseConv2dNative
- ConvTranspose2D
- MaxPool
- AvgPool

TensorRT样例

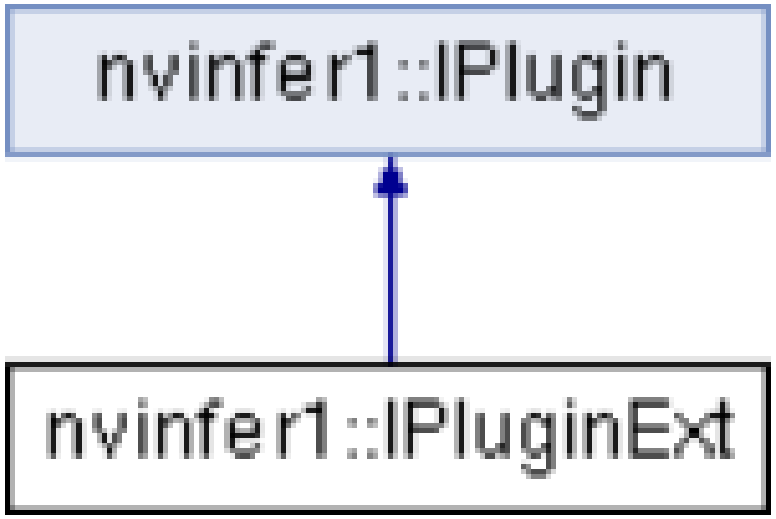
C++:
/usr/src/tensorrt/sample
Python:
{PYTHON_PACKAGE_DIR}/tensorrt/examples

使用文档：
https://docs.nvidia.com/deeplearning/sdk/tensorrt-developer-guide/index.html

Python api 文档：
https://docs.nvidia.com/deeplearning/sdk/tensorrt-api/python_api/index.html

自定义网络层

如何实现自定义网络层？



terminate
  terminate

自定义网络层

四个重要阶段：

1. configureWithFormat
2. Initialize
3. enqueue
4. terminate

```
terminate
  terminate
```

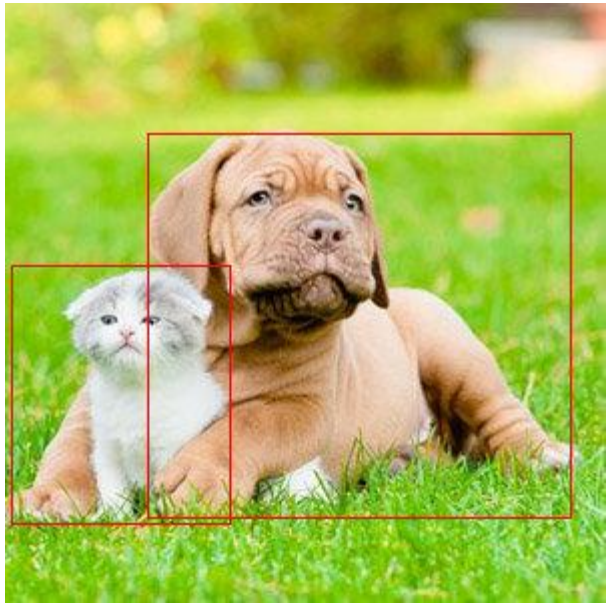
自定义网络层

三个重要方法：

1. getNbOutputs
2. getOutputDimensions
3. supportsFormat

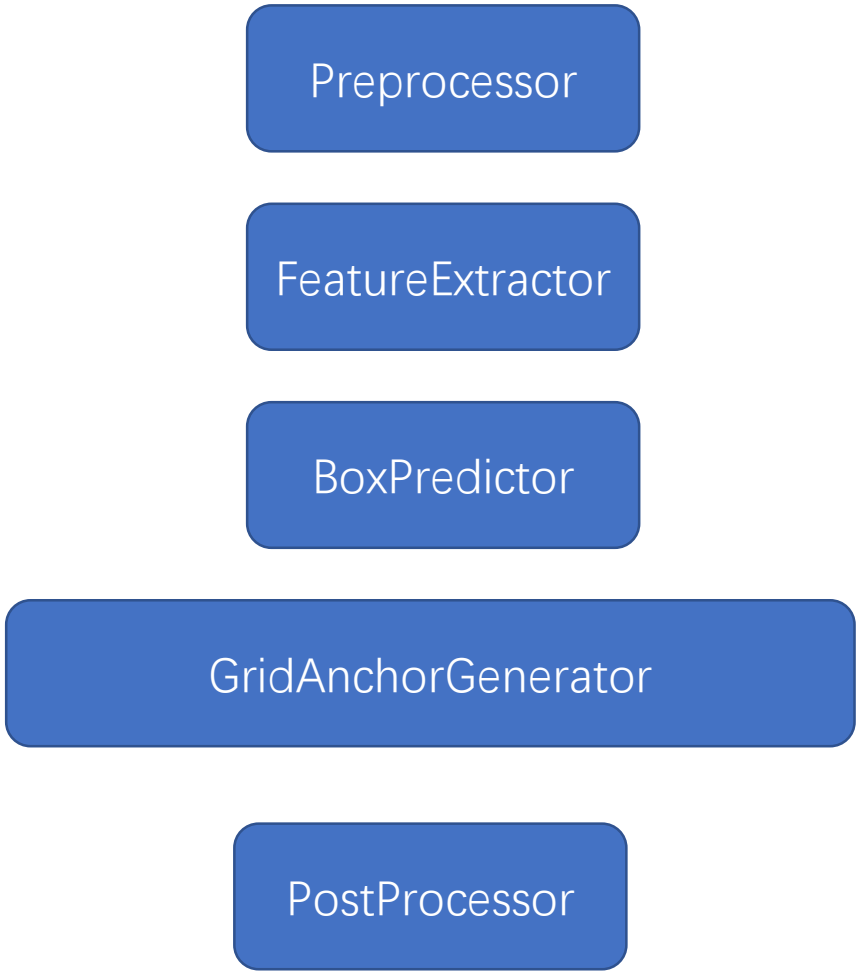
TensorRT编程模型

SampleUffSSD



TensorRT编程模型

Main Components



TensoRT实例展示