NVIDIA TRANSFER LEARNING TOOLKIT &TensorRT

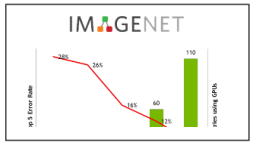
NVIDIA 开发者社区 何琨

POWERING THE DEEP LEARNING ECOSYSTEM

COMPUTER VISION

OBJECT DETECTION

IMAGE CLASSIFICATION



SPEECH & AUDIO

VOICE RECOGNITION

LANGUAGE TRANSLATION

NATURAL LANGUAGE PROCESSING

RECOMMENDATION ENGINES

SENTIMENT ANALYSIS

DEEP LEARNING FRAMEWORKS

Caffe

Chainer

DL4J
Deeplearning4j

Mocha.jl
julia

KERAS

Microsoft CNTK

mxnet

Purine

TensorFlow

torch

MatConvNet

MINERVA

OpenDeep

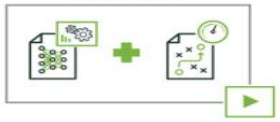
Pylearn2

theano

NVIDIA DEEP LEARNING SDK

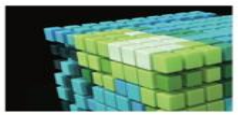
cuDNN

TensorRT



DeepStream SDK

cuBLAS



cuSPARSE

NCCL

developer.nvidia.com/deep-learning-software

Deep Learning Workflow Management

Deep Learning Challenges

Third Party Pre-Trained Models
- Lack accuracy
- Use case limitations
- Model size limitations
- Unoptimized for GPUs

Deep Learning Training
- Compute Resources
- Time spent training from scratch
- Learning DL frameworks

Deep Learning Inference
- Unclear workflows for production ready models
- Complex application pipeline

NVIDIA Deep Learning Solution

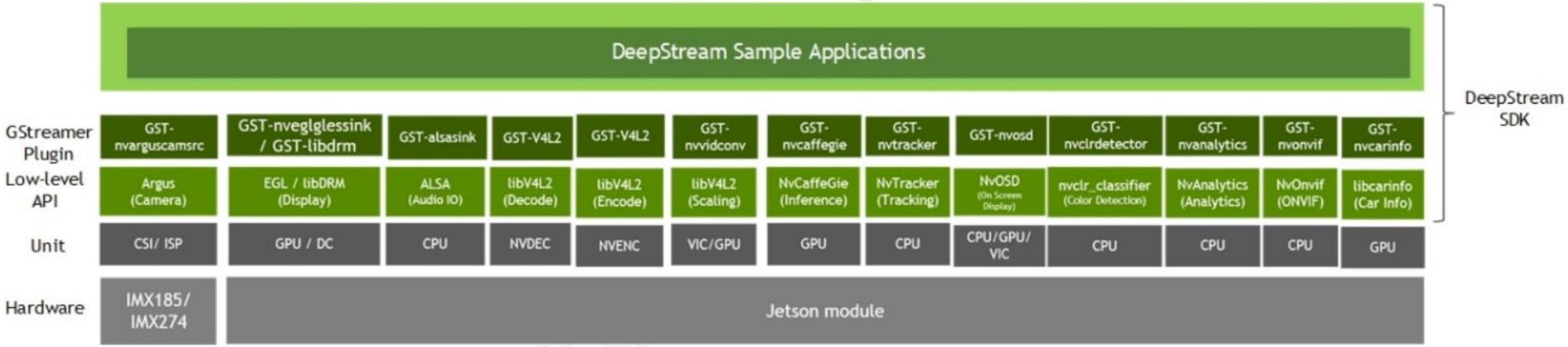
Transfer Learning Toolkit
- GPU accelerated pre trained models
- Incremental Training
- Pruning
- Easy to use
- Abstraction from learning DL frameworks

DeepStream SDK
- Faster intelligent insights
- Track inference
- End to end- easy AI deployment

DeepStream

NVIDIA DeepStream SDK是为大规模创建和部署基于AI的视频分析应用程序解决方案而设计的，它提供完整的框架和所有基本构建模块。

- Gstreamer
- TensorRT



NVIDIA TRANSFER LEARNING TOOLKIT

在指定的公共数据集上训练的图像分类和目标检测模型，可与Transfer Learning Toolkit一起使用。

Image Classification

- ResNet10/18/50
- VGG16/19
- MobileNet V1/V2
- AlexNet
- SqueezeNet
- GoogLeNet

Object Detection

Faster RCNN supporting backbones:

- ResNet10/18/50
- VGG16/19
- GoogLeNet
- MobileNet V1/V2

DetectNet_v2 supporting backbones:

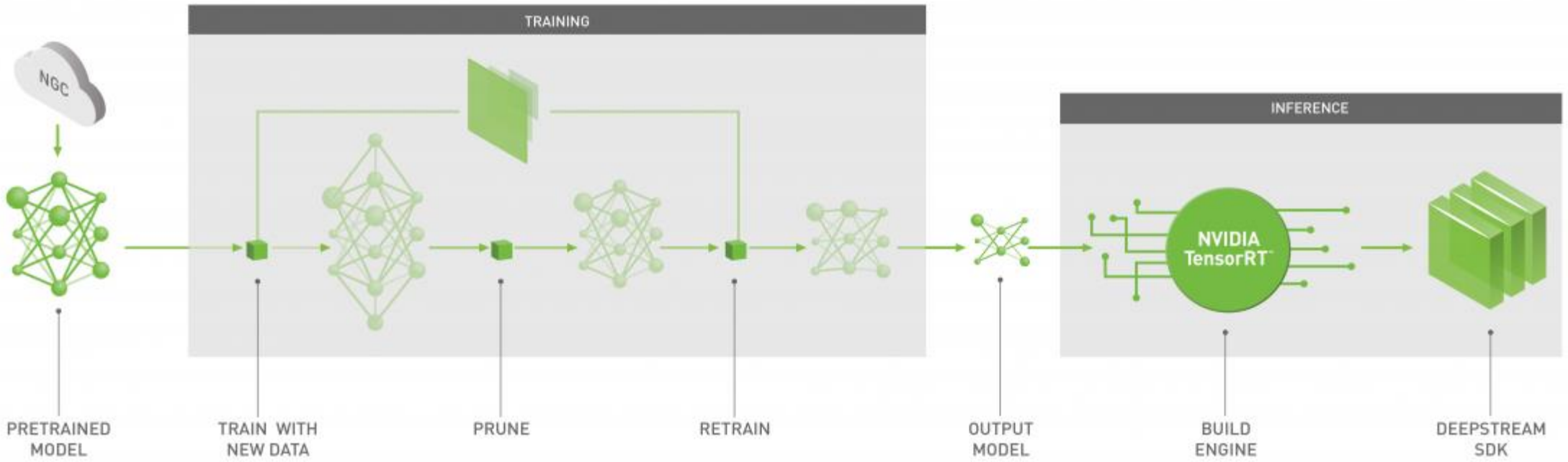
- ResNet10/18/50
- VGG 16/19
- GoogLeNet
- MobileNet V1/V2

SSD:

- ResNet10/18

NVIDIA TRANSFER LEARNING TOOLKIT

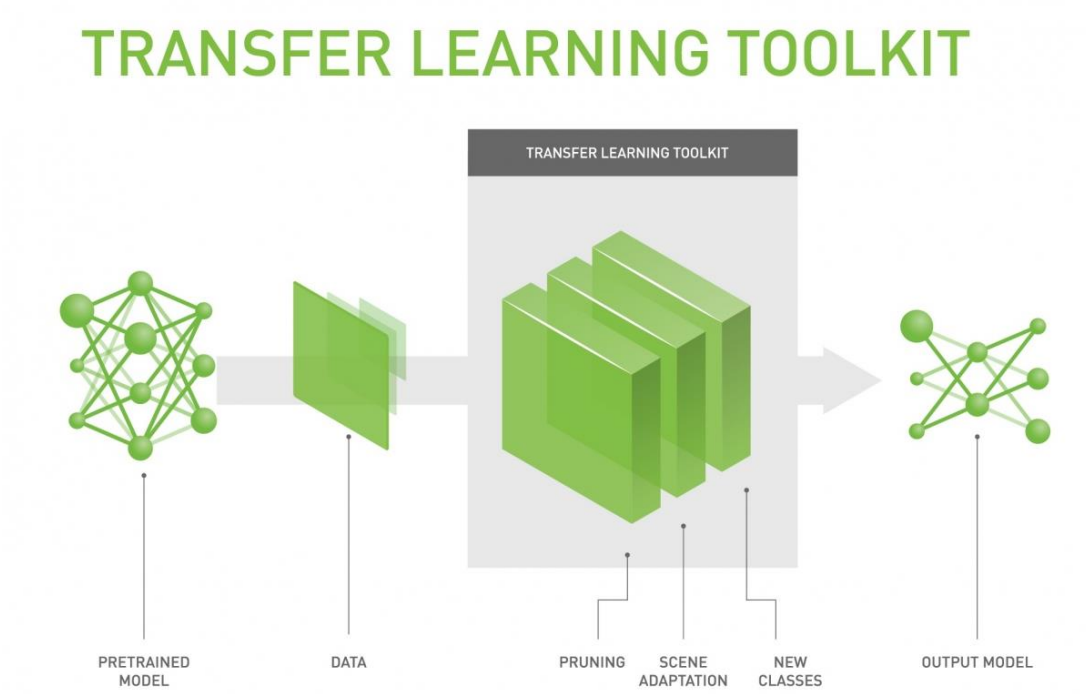
为应用在计算机视觉领域的深度学习工作流程，提供了全方位的便利工具



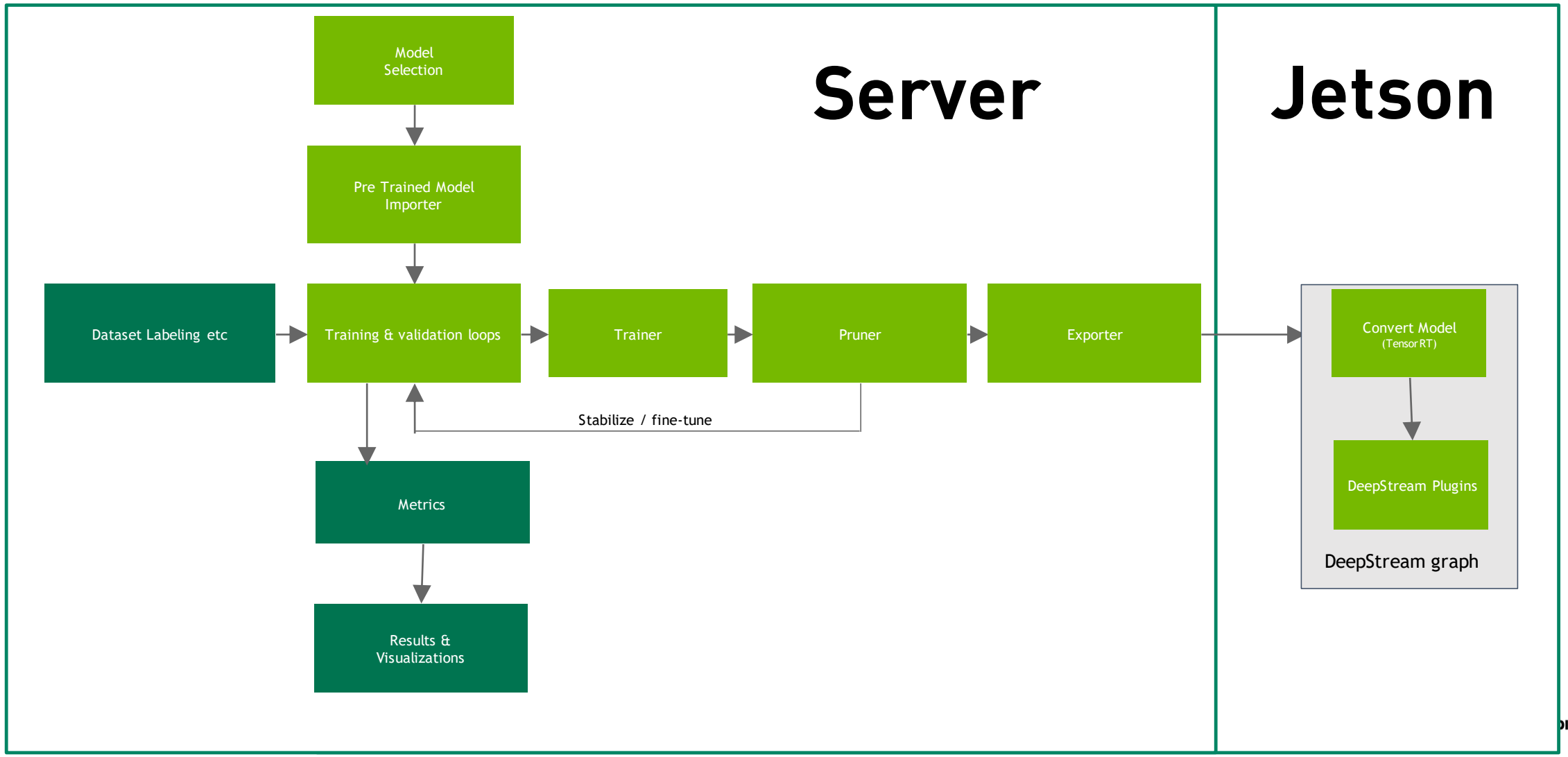
NVIDIA TRANSFER LEARNING TOOLKIT

Transfer Learning Toolkit是一个基于python的工具包，它使开发人员能够利用NVIDIA预先训练的模型，并为开发人员提供一系列的工具，使流行的网络架构适应他们自己的数据，并且能够训练、调整、修剪和导出模型以进行部署。它还拥有简单的接口和抽象API，提高了深度学习训练工作流的效率。

- GPU优化的预训练砝码，可用于计算机视觉任务
- 轻松修改配置文件以添加新类并使用自定义数据重新训练模型
- 在异构的多GPU环境中执行模型调整和重新训练
- 使用修剪功能缩小模型尺寸
- 模型导出API，可在具有NVIDIA Tesla和Jetson产品的NVIDIA DeepStream SDK上部署

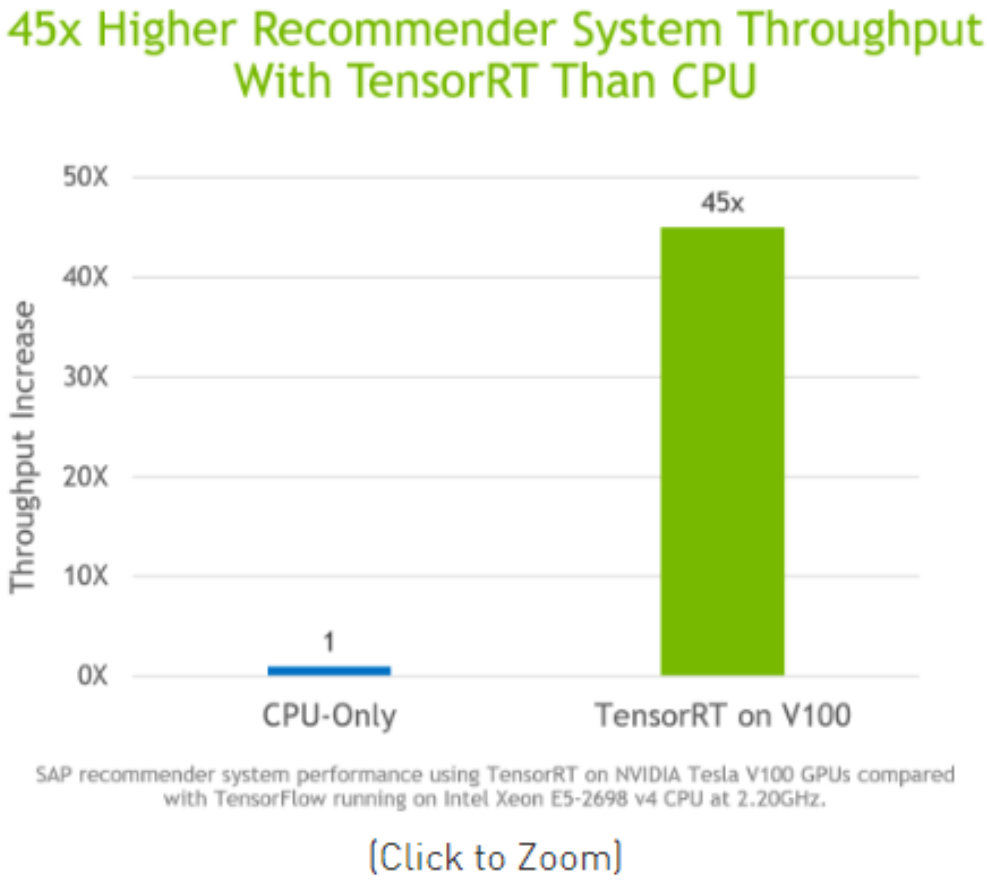


TRANSFER LEARNING TOOLKIT的工作流程



TensorRT

NVIDIA TensorRT™是一种高性能深度学习推理优化器和运行时加速库，可为深度学习推理应用程序提供低延迟和高吞吐量。 使用TensorRT，您可以优化神经网络模型，以高精度校准低精度，最后将模型部署到超大规模数据中心，嵌入式或汽车产品平台。



- 新编译器加速递归神经网络常用的语音和异常检测
- 支持超过20个新的ONNX操作，加速关键语音模型，如BERT，TacoTron 2和WaveRNN
- 扩展了对动态形状的支持，以支持关键的人工智能会话模型

ONNX: Added ConstantOfShape, DequantizeLinear, Equal, Erf, Expand, Greater, GRU, Less, Loop, LRN, LSTM, Not, PRelu, QuantizeLinear, RandomUniform, RandomUniformLike, Range, RNN, Scan, Sqrt, Tile, and Where

TENSORRT WORKFLOW

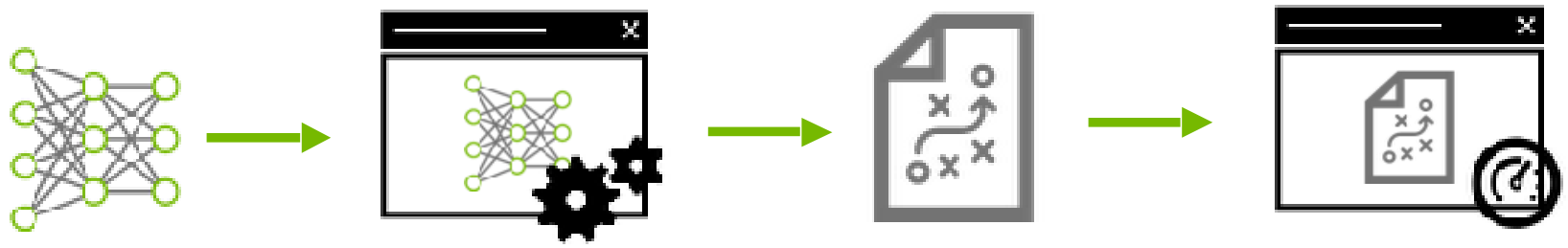
Input

- Pre-trained FP32 model and network

Output

- Optimized execution engine on GPU for deployment

Serialized a PLAN can be reloaded from the disk into the TensorRT runtime. There is no need to perform the optimization step again.



TensorRT

1.创建Builder

2.创建Network

3.创建Parser

4.绑定输入、输出以及自定义组件

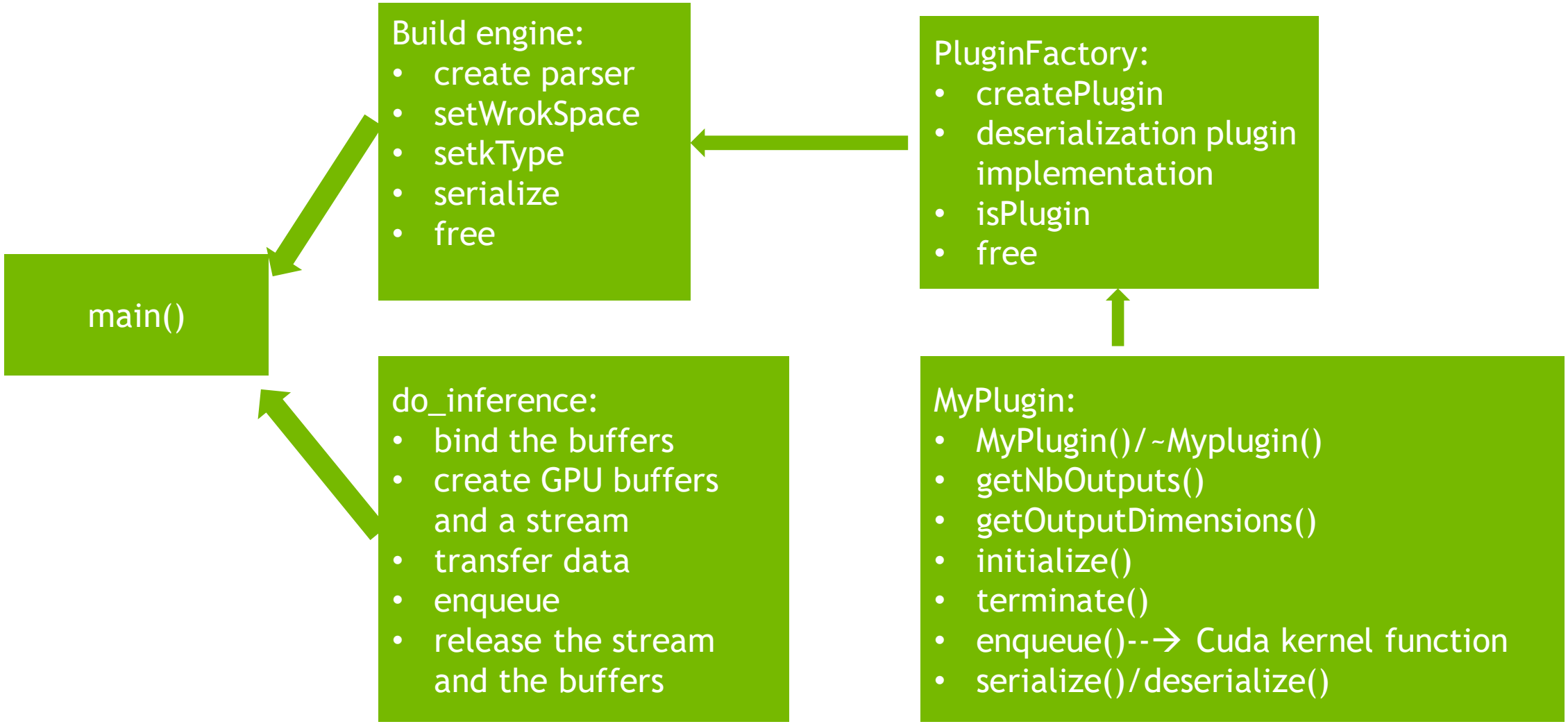
5.序列化或者反序列化

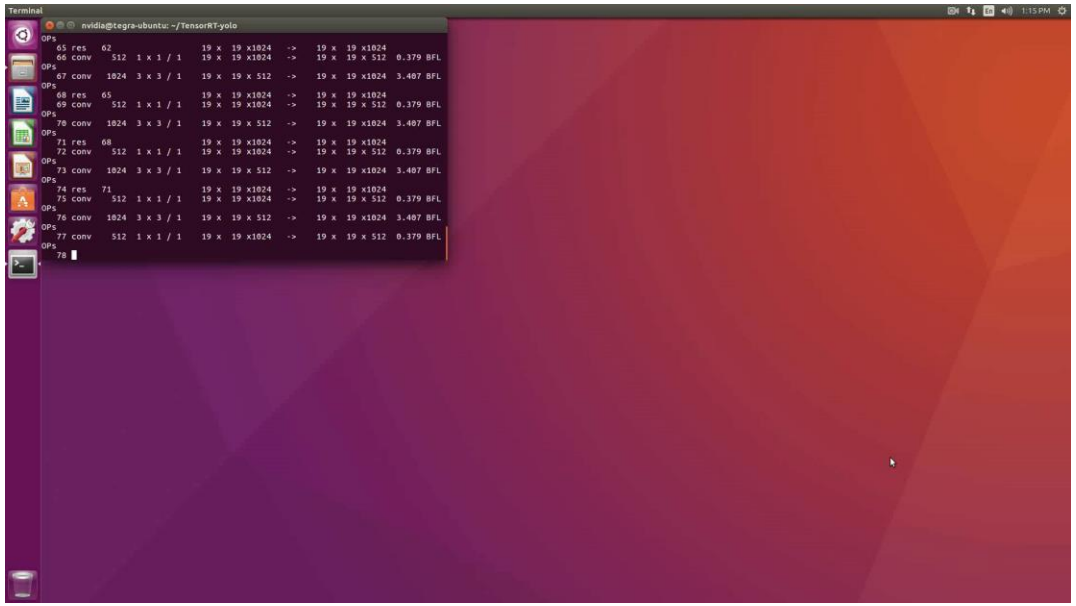
6.传输计算数据（host->device）

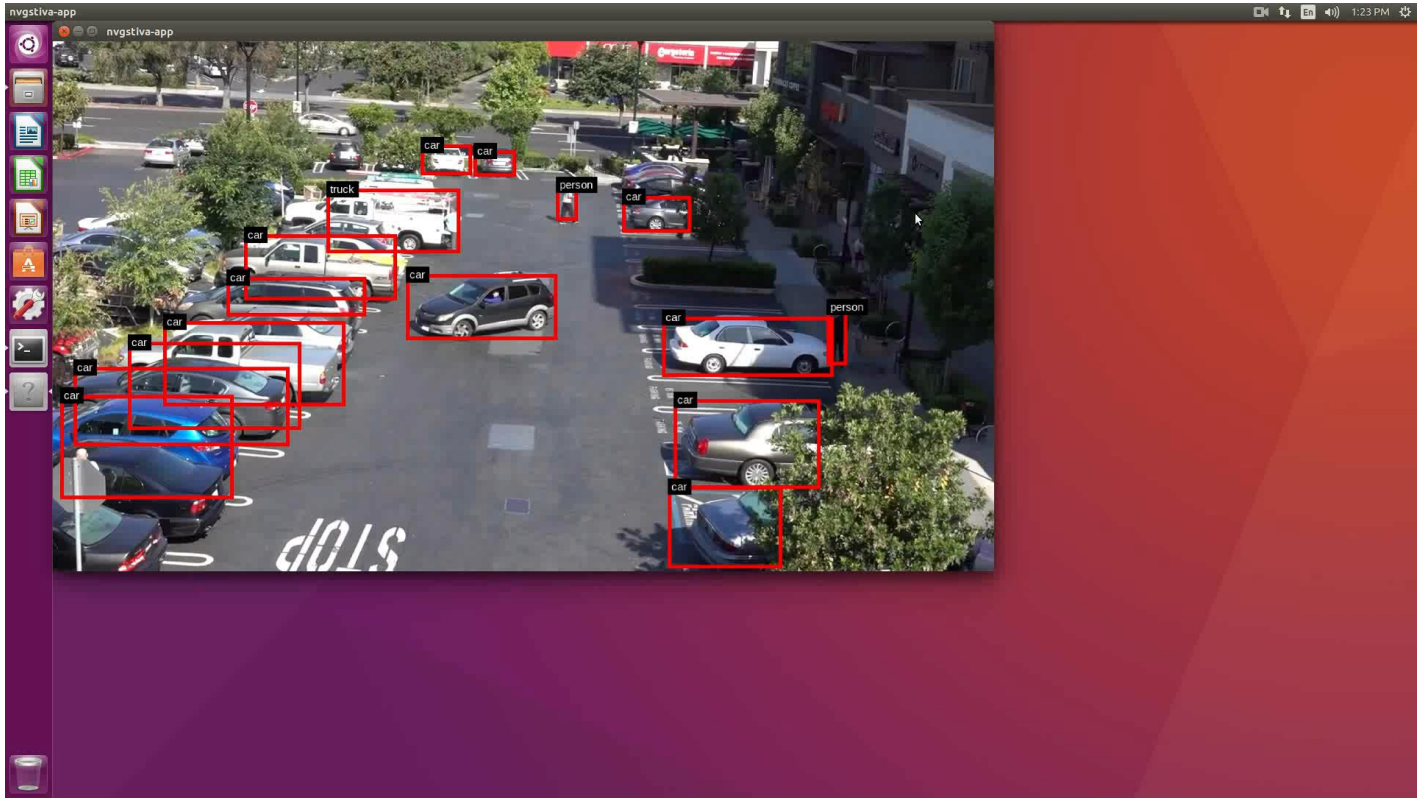
7.执行计算

8.传输计算结果（device->host）

TensorRT架构




Terminal
1:15 PM
nvidia@tegra-ubuntu: ~/TensorRT-yolo
OPs
65 res 62 19 x 19 x1024 -> 19 x 19 x1024
66 conv 512 1 x 1 / 1 19 x 19 x1024 -> 19 x 19 x 512 0.379 BFL
OPs
67 conv 1024 3 x 3 / 1 19 x 19 x 512 -> 19 x 19 x1024 3.407 BFL
OPs
68 res 65 19 x 19 x1024 -> 19 x 19 x1024
69 conv 512 1 x 1 / 1 19 x 19 x1024 -> 19 x 19 x 512 0.379 BFL
OPs
70 conv 1024 3 x 3 / 1 19 x 19 x 512 -> 19 x 19 x1024 3.407 BFL
OPs
71 res 68 19 x 19 x1024 -> 19 x 19 x1024
72 conv 512 1 x 1 / 1 19 x 19 x1024 -> 19 x 19 x 512 0.379 BFL
OPs
73 conv 1024 3 x 3 / 1 19 x 19 x 512 -> 19 x 19 x1024 3.407 BFL
OPs
74 res 71 19 x 19 x1024 -> 19 x 19 x1024
75 conv 512 1 x 1 / 1 19 x 19 x1024 -> 19 x 19 x 512 0.379 BFL
OPs
76 conv 1024 3 x 3 / 1 19 x 19 x 512 -> 19 x 19 x1024 3.407 BFL
OPs
77 conv 512 1 x 1 / 1 19 x 19 x1024 -> 19 x 19 x 512 0.379 BFL
OPs
78


nvgstiva-app
1:23 PM
nvgstiva-app
car
car
truck
person
car
car
car
car
person
car
car
car
car
car
car
STOP

总结

- NVIDIA Transfer Learning Toolkit为深度学习训练部署流程提供了完整的工具链
- 把训练和剪裁好的模型部署在边缘设备(Jetson 平台)上时，需要在边缘设备上转换成TRT的格式
- 使用TLT导出模型时要注意参数的设置，特别是数据精度以及Tensor维度

链接：https://pan.baidu.com/s/1TWT5PMYn-VkYoxbVCWKoWA
提取码：cdaq

NVIDIA 深度学习学院 (DLI)

人工智能和加速计算实践培训

- 面向开发者、数据科学家和研究人员
- 权威机构和专家强强联合打造专业培训
- 运用前沿技术的端到端、多行业应用开发课程
- 真实经验案例分享，获取现实可用的专业知识
- 完全配置的GPU实时开发环境
- 由具有学科专业知识的DLI认证讲师授课
- 全球开发者培训证书

查看课程 www.nvidia.cn/dli

停课不停学 高校师生免费领

第一项 《在线自主培训课程》

- 高校学生和教师可领取
- 任意一门30美元“在线自主培训”课程

  www.nvidia.cn/dli
- 此活动每人仅限参加一次

第二项 《人工智能教学教材》

- 高校教师可领取
- 含讲义文档和视频，实验室方案，编程项目，电子书和更多免费 DLI 课程

第三项 《DLI校园大使及培训平台》

- 高校教师可参加
- 认证后可免费用DLI平台给学生授课学习

更多资源：

https://developer.nvidia-china.com


THANK YOU
NVIDIA®