1. 任务描述：使用TensorRT（GIE）的单精度模型转换加速Openpose模型的inference过程达到可用级别。

2. TensorRT介绍：

i)对已训练的神经网络进行优化、验证和部署;

ii)C++库，官方API： TX1：/usr/share/doc/gie/doc/API/index.html；

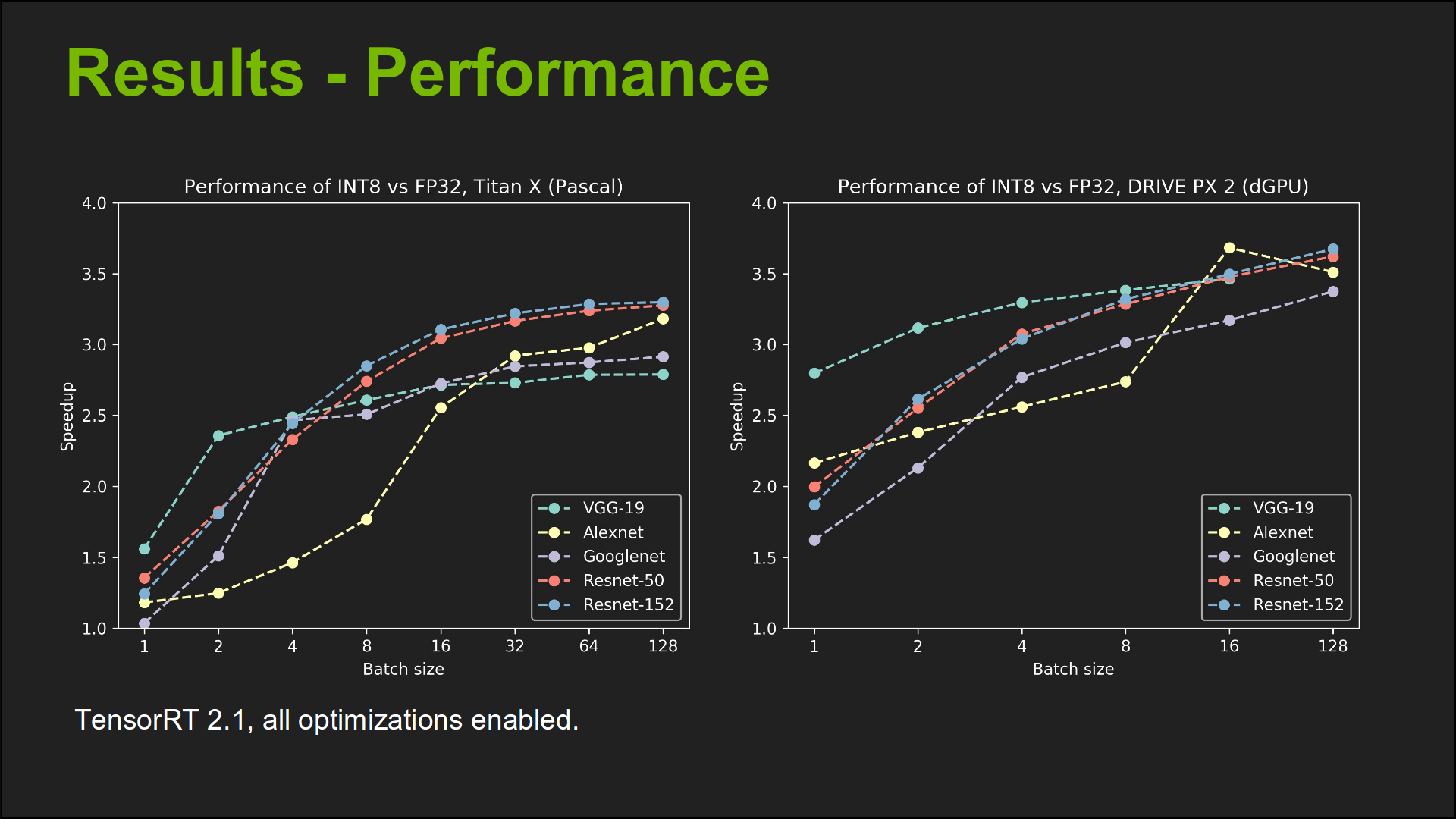
iii)适用于Jetson TX1和Pascal架构的显卡（Tesla P100, K80, M4 and Titan X等），支持fp16特性---半精度运算;

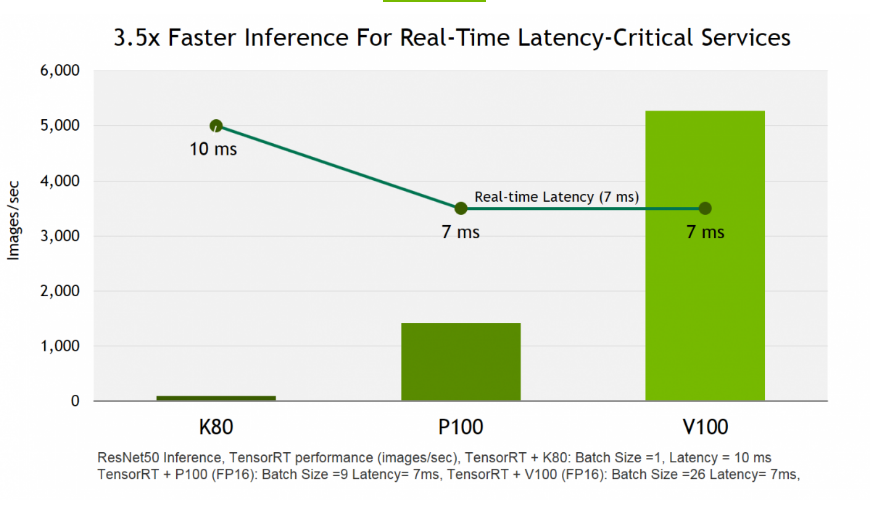
iv)对inference过程加速，只能用来做Inference，不能用来做train;

v)解析caffe模型-->NvCaffeParser,根据prototxt文件和caffemodel权值，转化为支持半精度的新的模型;

vi)支持caffe中大部分的层;

3.官方加速效果：





4.试用：

1)常见模型单精度模式的加速比

数据来源：TX1自带gie samples & caffe中的model

model\half2 普通t 单精度t 加速比

mnist 1.6 1.2 1.3(batch=12)

googlenet 27.6 16.8 1.6

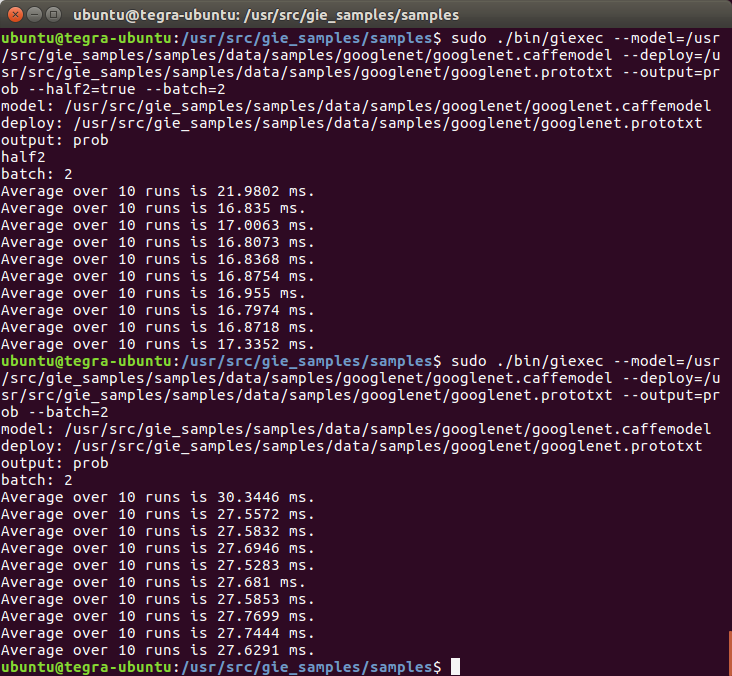
alexnet 22.6 14.7 1.5

Vgg16

Vgg19

ResNet

例图：googlenet加速比



2）官方数据显示batch越大则差异越显著，对于特定模型的影响测试如下：

数据源：caffe自带bvlc\_reference\_caffenet模型

batch\half2 普通 单精度 加速比

1 18.6 14.4 1.3

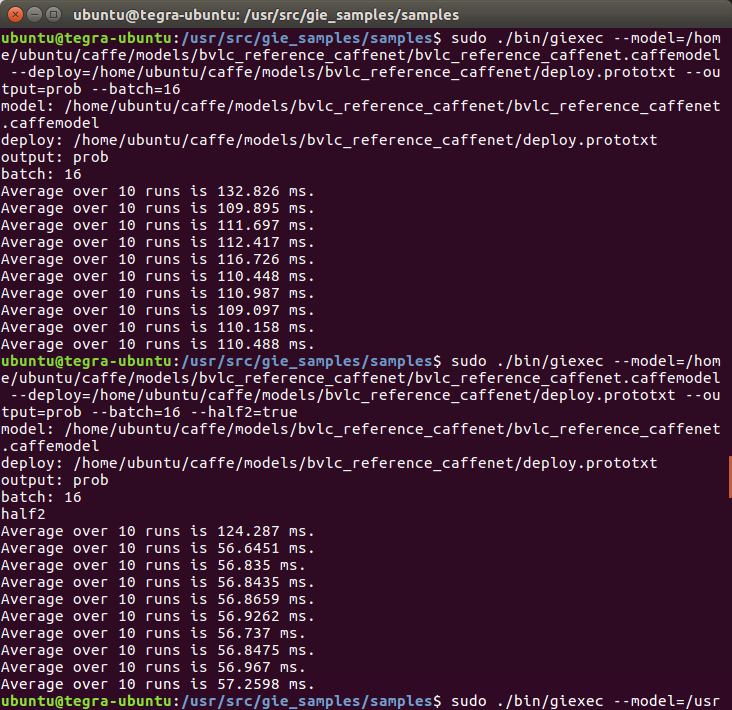
2 21.7 14.2 1.5

4 29.7 18.7 1.6`

8 58.1 36.2 1.6

16 112.4 56.8 2.0

日志提示：half2 requires even maximum batch size; rounding up to 2



5.分析：

加速比低于官方影响因素分析：

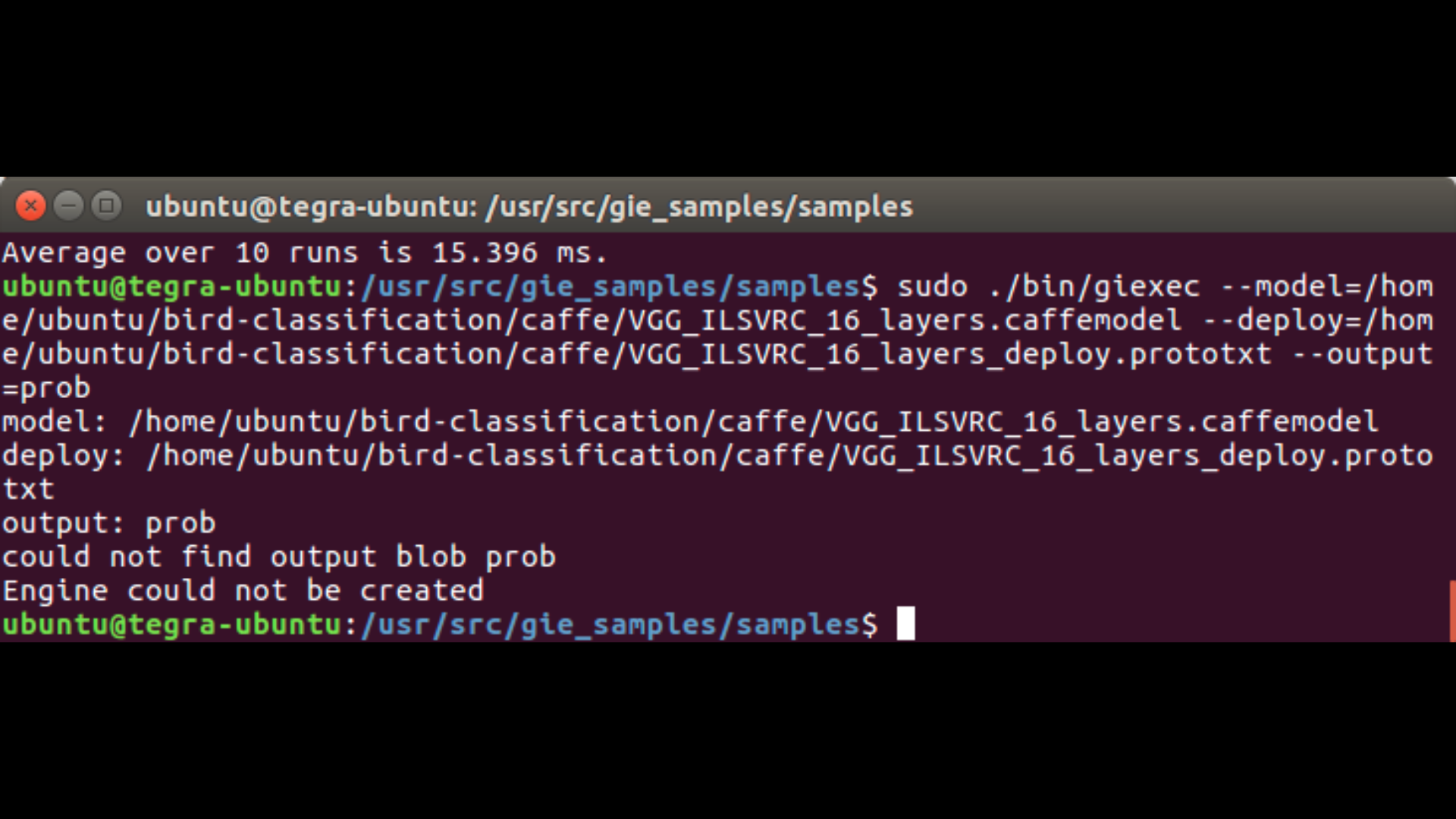
i)模型简单，层数少，无法体现单精度优势；

ii)参数设置原因，批处理数(取决于精度需求和硬件条件，决定梯度下降的方向和收敛的效果和速率，以及内存的利用率）；

一些模型（vgg16,vgg19)无法顺利运行因素分析：

i)参数设置不当，输出类型参考prototxt最后一层配置，或需加一层输出层；

ii)模型含有gie不支持的层；



6.方向：

1.转换任意的CNN/RNN（非.caffemodel的模型使其可以被TensorRT执行&&用TensorRT转换CNN/RNN的模型为支持单精度的模型）

2.分析openpose的prototxt使可运行在TensorRT上。

7.参考：

<https://github.com/dusty-nv/jetson-inference>TensorRT部署推理网络（TX1)

<http://on-demand.gputechconf.com/gtc/2017/presentation/s7458-chris-gottbrath-deploying-unique-microservices-tensorrt.pdf>TensorRT 8bit推理

<http://on-demand.gputechconf.com/gtc/2017/presentation/s7310-8-bit-inference-with-tensorrt.pdf>TensorRT部署深度学习网络

<https://handong1587.github.io/deep_learning/2015/10/09/object-detection.html>TensorRT目标探测