



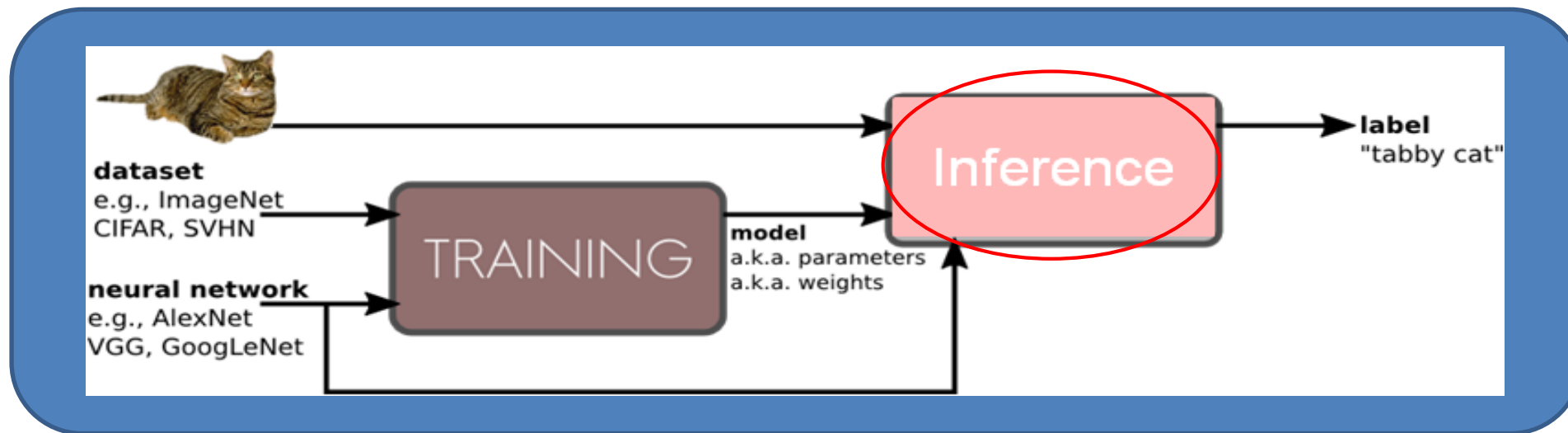
华为云异构计算私享会

--AI，在华为云FPGA加速服务上绽放

张佳

LEADING NEW ICT

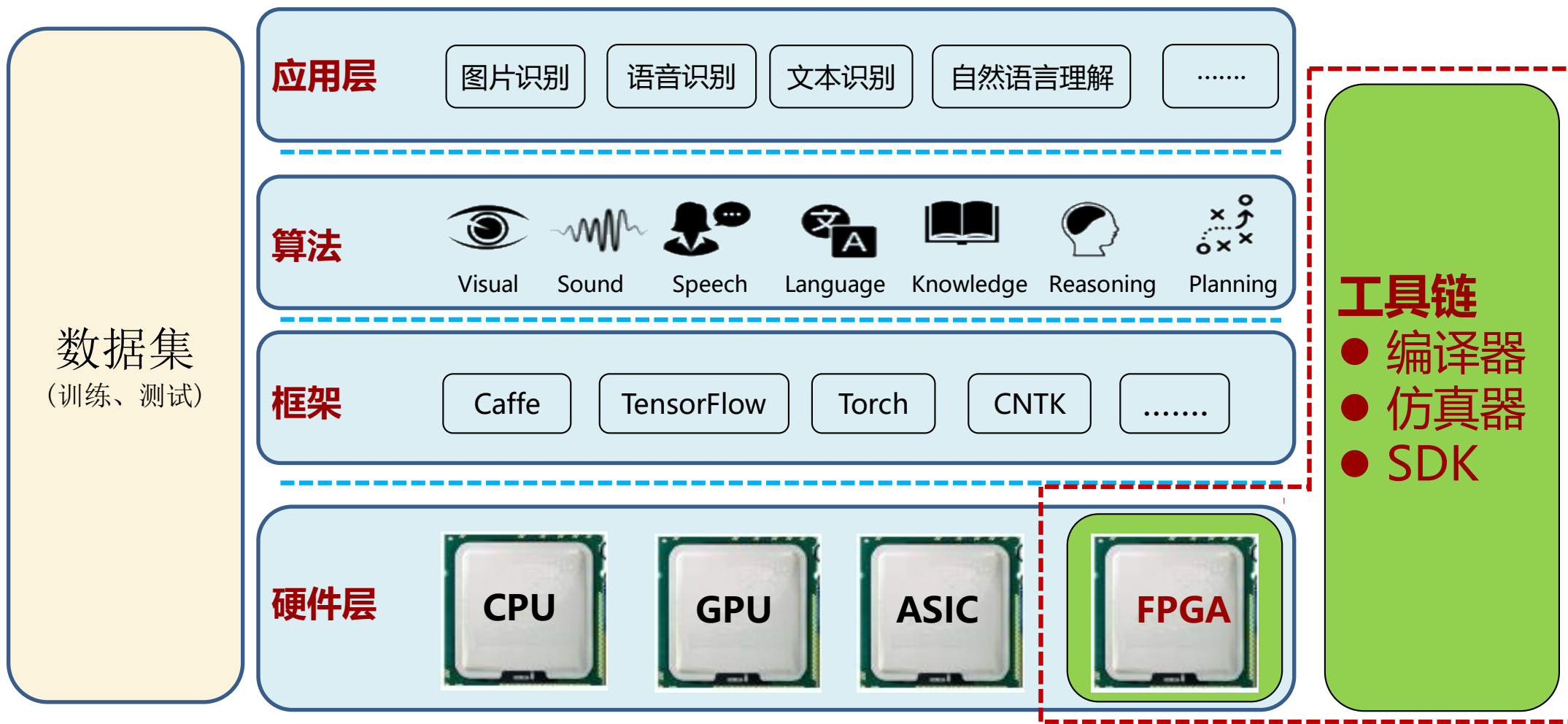
- **FPGA 深度学习推理加速解决方案**
- **FPGA 深度学习推理加速Feature**



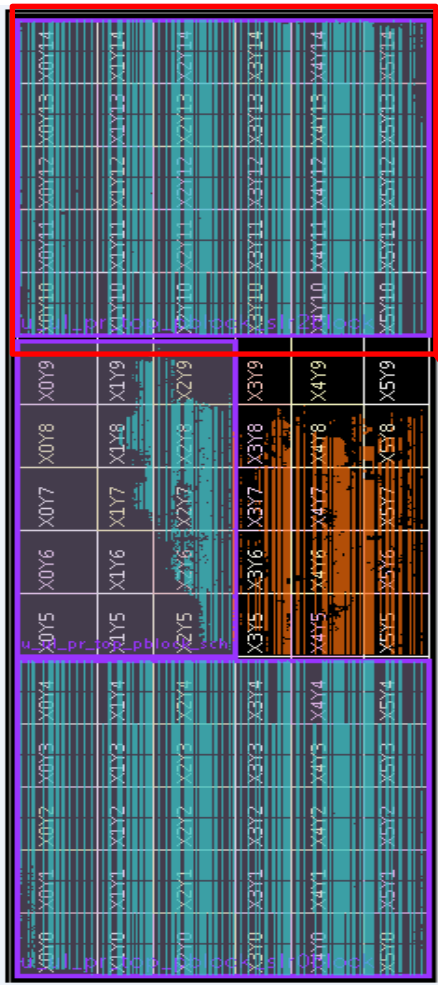
Training: 通过大量的数据输入，或采取增强学习等非监督学习方法，训练出一个复杂的深度神经网络模型。

Inference: 利用训练好的模型，使用新的数据去“推理”出各种结论。

FPGA 深度学习推理加速 -- 解决方案



提供FPGA深度学习推理端到端加速能力，配套完整工具链



VU9P FPGA Floorplan

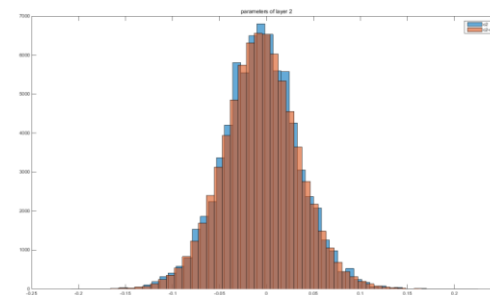
DL Inference IP

特性	描述
器件	Virtex UltraScale+ FPGA(VU9P)
数据/参数	INT8
网络	VGG16、Resnet50、Googlenet、Faster-RCNN、LSTM
框架	Caffe
配套(可选)	AXI接口 JPEG解码、缩放、加解密、压缩
场景	图片分类、目标检测、语音识别

1

高效比：高性能、低精度损失

- 8bit高性能量化方案，精度降低千分之五以下
- 神经网络加速专用逻辑架构、计算单元
- 支持Batch模式，节省DDR带宽
- 自研高性能深度学习加速专用压缩算法



Linear Quantization (INT8)

2

通用性：ISA（指令集架构），完备的工具链

- 支持CNN\DNN\Faster RCNN\RNN(LSTM)等主流神经网络
- 支持ISA指令集编程，网络结构灵活扩展
- 支持硬件加速算法灵活升级
- 提供完整的工具链（编译器、仿真器）

```
//下达计算指令
L1:  CALC_CONV    #32    #3ff    #0x3
      GPR_MOVE    $16    10
      WAIT_UNTIL  #0x01
      SS_SUB      $16    #1
      GPR_MOVE    $1      #1
      GPR_MOVE    $2      #500
      PARA_LOAD   $1      $3      $2      #0
      PARA_LOAD   $1      $3      $2      #1
      PARA_LOAD   $1      $3      $2      #2
      PARA_LOAD   $1      $3      $2      #3
```

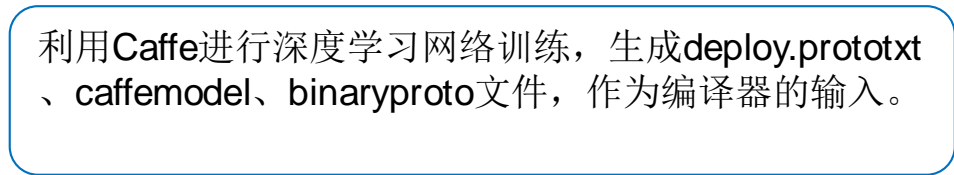
3

无缝支持Caffe、TensorFlow等平台

- 支持Caffe平台
- 支持TensorFlow

Caffe





利用编译器进行编译，将Caffe生成的文件生成FPGA AI Core所需的配置文件。

提供调用FPGA AI Core加速功能函数接口。

支持通过MPI接口调用，提供与FPGA AI Core相同执行结果。

可执行程序，支持对用户网络进行精度、性能仿真。

深度学习加速解决方案特点：

- 提供深度学习推理FPGA端到端加速能力.
- 专用算法优化，推理精度、性能具有独特优势.
- 配套完整的工具链，灵活易用.

THANK YOU

Copyright©2016 Huawei Technologies Co., Ltd. All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.