Spring Project：Multi Backend Neural Network Auto Quantization and Deployment over ONNX

Fengwei Yu (SenseTime-China)

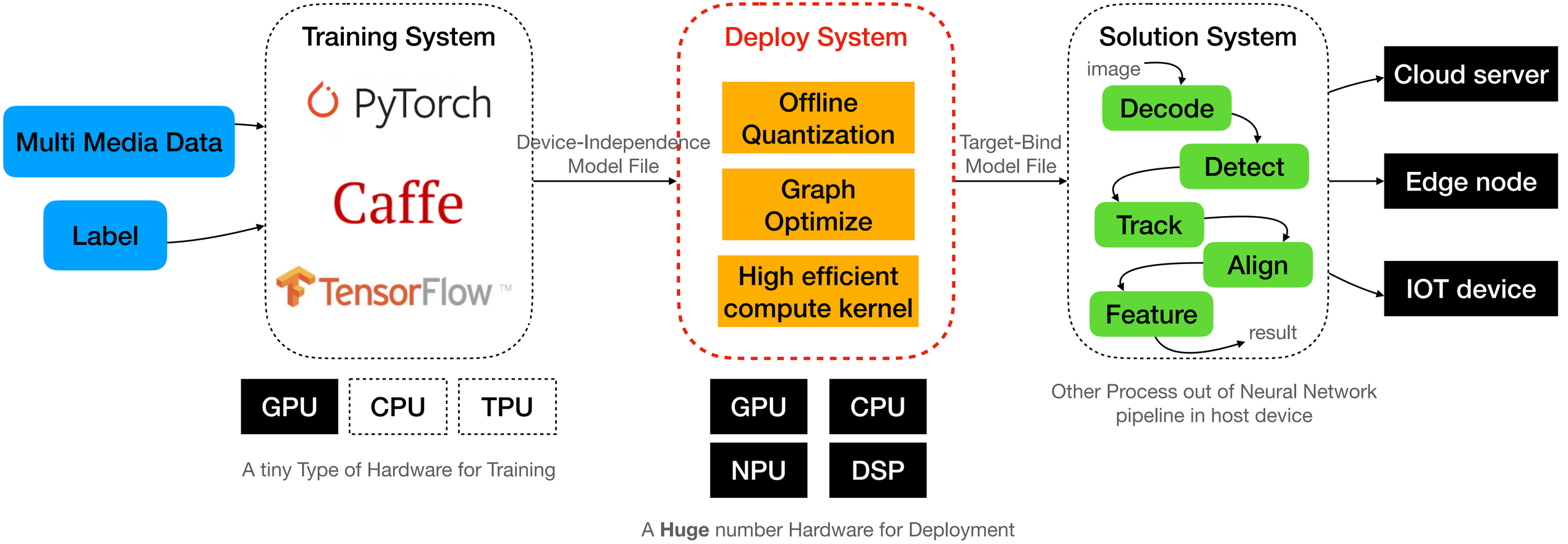
Contents

- What is Spring Project.
- Neural Network Deployment.
- Neural Network Quantization.
- From Caffe to Onnx.

What is Spring Project?

Spring

:Industrial Grade AI Model Production Framework



Neural Network Deployment：Features and Challenges

：**From** Device-Independence Model File to Target-Bind Model File

★3 key features of a Neural Network Deployment Framework

Automatic

more “end2end” in one model
we want “one-for-all”
we don’t want “compile failed”

Efficiency

high efficient on special target
support model compress like quantization and sparsity

Multi-Device

support more and more device with an unified framework
enable to run