Model Serving on Edge

Model Serving on Edge Devices

Infrastructure for efficiently deploying and running compressed models

Edge Inference Frameworks

TensorFlow Lite

- Provided by Google
- Android/iOS support
- Extensive hardware
- .tflite format

Core ML

- Apple exclusive
- iOS/macOS optimized
- Neural Engine utilization
- .mlmodel format

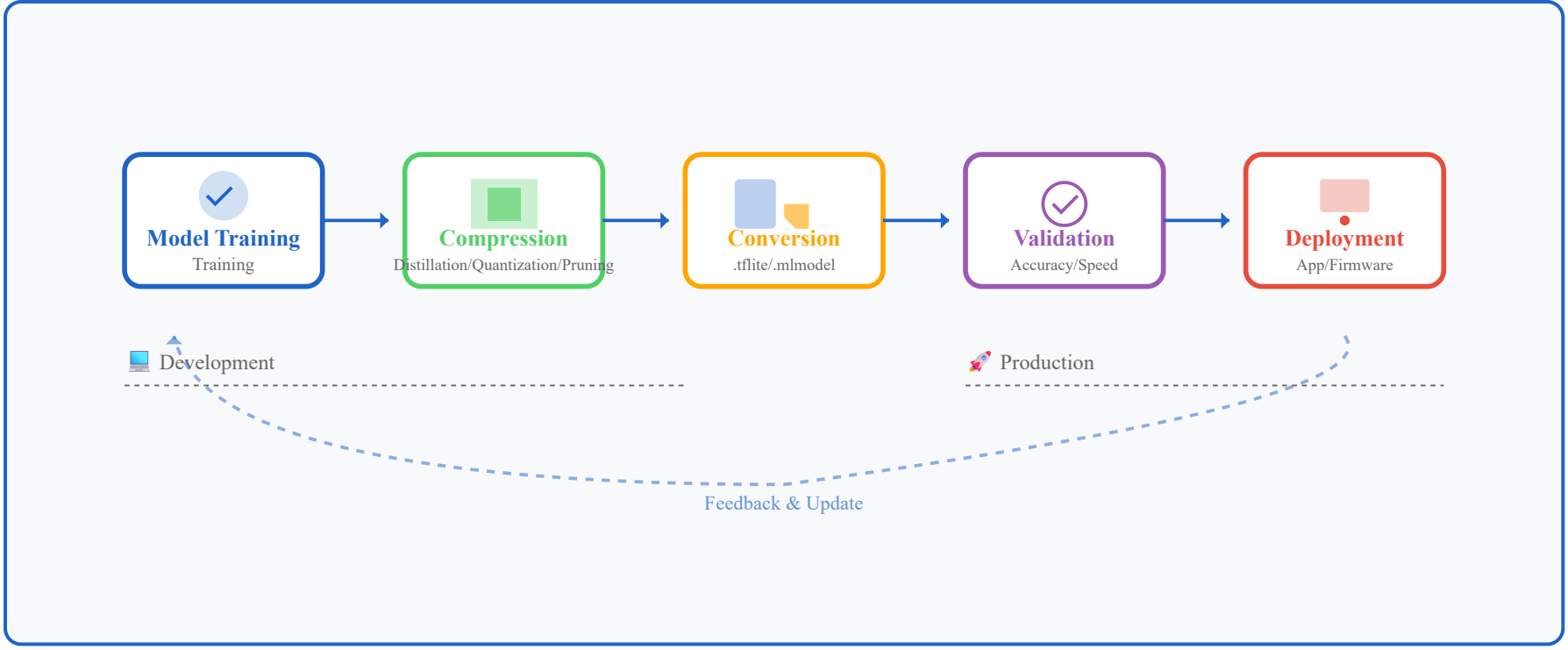
ONNX Runtime

- Cross-platform
- Various backends
- Model optimization tools
- .onnx format

PyTorch Mobile

- PyTorch based
- Android/iOS
- Easy conversion
- .ptl format

Deployment Pipeline



Best Practices

✓ Device-specific benchmarking required (test on actual hardware)

✓ OTA (Over-The-Air) update support

✓ Fallback mechanism (cloud inference)

✓ Monitoring and logging