

# Model Serving on Edge

## 엣지 디바이스에서의 모델 서빙

압축된 모델을 효율적으로 배포하고 실행하기 위한 인프라

### 엣지 추론 프레임워크



#### TensorFlow Lite

- Google 제공
- Android/iOS 지원
- 광범위한 하드웨어
  - .tflite 포맷



#### Core ML

- Apple 전용
- iOS/macOS 최적화
- Neural Engine 활용
  - .mlmodel 포맷



#### ONNX Runtime

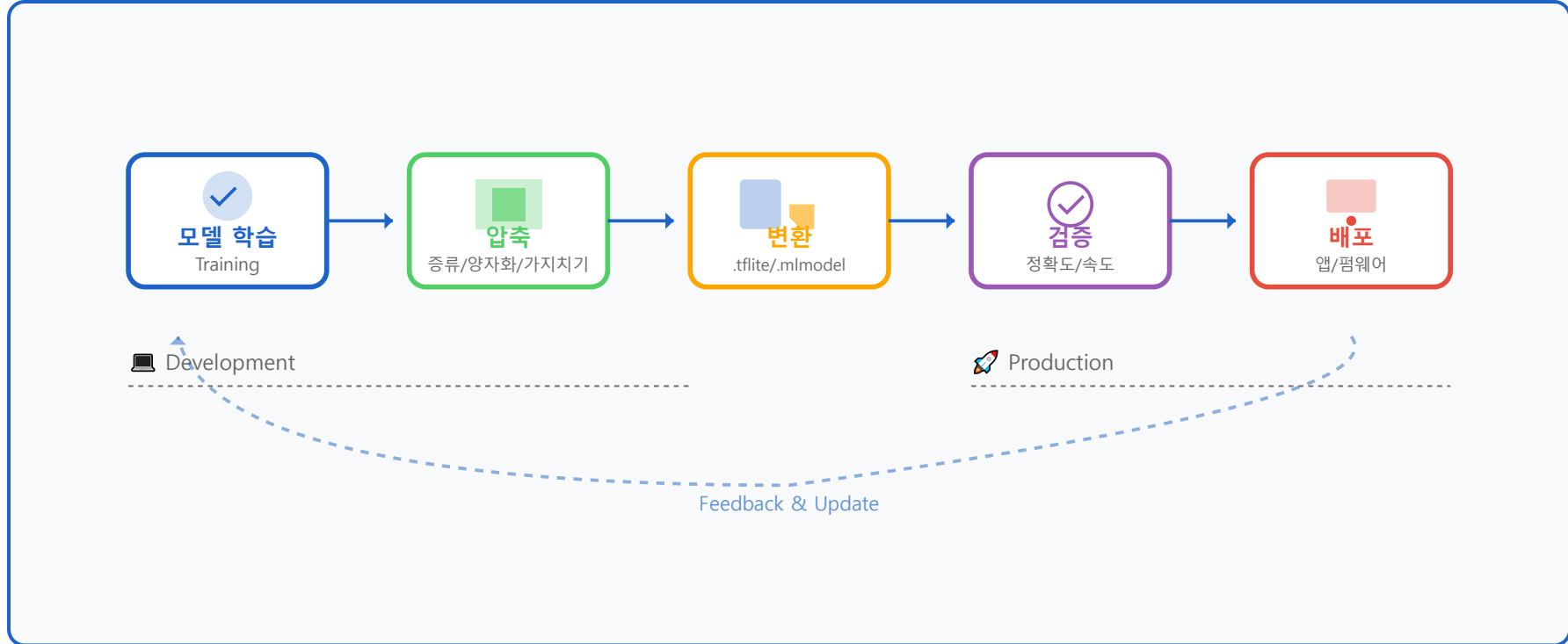
- 크로스 플랫폼
- 다양한 백엔드
- 모델 최적화 도구
  - .onnx 포맷



#### PyTorch Mobile

- PyTorch 기반
- Android/iOS
- 쉬운 변환
  - .ptl 포맷

### 배포 파이프라인



## 모범 사례

✓ 디바이스별 벤치마크 필수 (실제 하드웨어에서 테스트)

✓ OTA(Over-The-Air) 업데이트 지원

✓ Fallback 메커니즘 (클라우드 추론)

✓ 모니터링 및 로깅