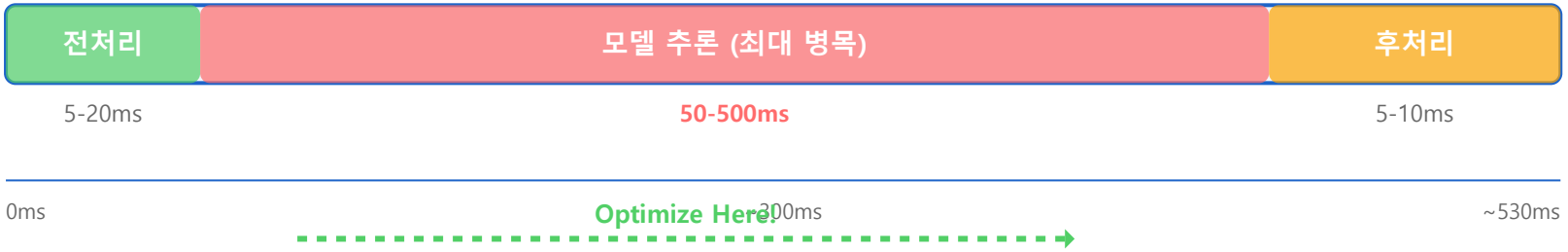
Latency Optimization

지연시간 최적화 (Latency Optimization)

실시간 의료 응용을 위한 추론 속도 향상 기법

Inference Latency Breakdown



지연시간 구성 요소

데이터 전처리
이미지 리사이징, 정규화
5-20ms

모델 추론
가장 큰 비중
50-500ms

후처리
결과 해석, 시각화
5-10ms

최적화 기법

모델 압축

- Quantization
- Pruning
- Distillation

2-10x 속도 향상

추론 엔진

- TensorRT
- ONNX Runtime
- TFLite

1.5-3x 속도 향상

하드웨어 가속

- GPU
- NPU/TPU
- FPGA

5-100x 속도 향상

실시간 처리 기준

Soft Real-time

< 100ms

일반 진단 보조

Hard Real-time

< 10ms

수술 로봇, 응급 시스템