Performance vs Size Tradeoffs

성능-크기 트레이드오프

압축 수준에 따른 모델 크기, 정확도, 속도의 균형점 찾기

파레토 최적 곡선 (Pareto Frontier)

Large Model

크기: 500MB
정확도: 95%
속도: 500ms

Medium Model

크기: 100MB
정확도: 93%
속도: 100ms

Small Model

크기: 20MB
정확도: 90%
속도: 20ms

Tiny Model

크기: 5MB
정확도: 85%
속도: 5ms

모델 선택 기준

클라우드 배포

Large/Medium 모델

정확도 최우선

모바일 앱

Medium/Small 모델

균형잡힌 성능

웨어러블

Small/Tiny 모델

크기 최소화

의사결정 가이드:

1. 최소 허용 정확도 결정
2. 목표 디바이스의 제약 조건 파악
3. 파레토 곡선 상에서 최적점 선택