Thank You

강의 요약

Part 1

Knowledge Distillation

- Teacher-Student 프레임워크
- Soft targets & Temperature
- Feature/Attention 전달

Part 2

Quantization & Pruning

- INT8/INT4 양자화
- 구조적/비구조적 가지치기
- 동적 희소성

Part 3

Edge Deployment

- 모바일/웨어러블 최적화
- POC 시스템
- 성능-크기 트레이드오프

핵심 베스트 프랙티스

✓ 압축 전 정확도 목표 및 제약사항 명확히 정의

✓ 여러 압축 기법 조합으로 시너지 효과

✓ 실제 디바이스에서 벤치마크 필수

✓ 의료 AI는 정확도 하락에 특히 민감 - 임상 검증 필수

✓ 배포 후 지속적 모니터링 및 업데이트

추천 도구 모음

압축: PyTorch Quantization, TensorFlow Model Optimization

변환: ONNX, TensorFlow Lite, Core ML Tools

추론: TensorRT, ONNX Runtime, TFLite

벤치마킹: MLPerf Mobile, AI Benchmark

Thank You!

Lecture 12: Knowledge Distillation and Model Compression

Questions? Contact via course forum or office hours