

Mixed Precision Strategies

혼합 정밀도 전략

레이어별로 다른 정밀도를 사용하여 성능과 효율성의 균형 달성

전략 예시

입력층

INT8/FP16
빠른 처리 가능

중간층

INT8
대부분의 연산

출력층

FP16/FP32
정확한 확률 계산

레이어 민감도 분석

민감한 레이어 → FP16/FP32 유지

덜 민감한 레이어 → INT8/INT4 적용

도구

PyTorch Quantization, TensorRT, ONNX Runtime