

# INT8/INT4 Quantization

## 정수 양자화 (Integer Quantization)

부동소수점(FP32/FP16)을 정수(INT8/INT4)로 변환하여 모델 크기 감소

### 비트 수 비교

#### FP32

32 bits

4 bytes

기본 학습 정밀도

#### INT8

8 bits

1 byte

75% 메모리 절감

#### INT4

4 bits

0.5 bytes

87.5% 메모리 절감

### 양자화의 효과



메모리 사용량 감소



추론 속도 향상



전력 소비 감소

**정확도 트레이드오프:** INT8은 일반적으로 1% 미만 정확도 감소  
INT4는 Quantization-Aware Training 필요