양자화의 장단점 비교

양자화 그 장점의 시작

- 신경망의 가중치와 활성화 값을 낮은 비트수에 매핑하는 작업
- 추론 속도도 향상시키고, 메모리 사용량을 줄이자
- 온디바이스 관련 산업군에서도 관심을 가지고 있음
    - 핵심적인 역할
    - CV 및 음성 인식기술에서 많이 활용되고 있음
    - 멀티 모달에서도 필요한 경우가 많이 있음

양자화의 장점: 메모리 및 에너지

- 모델의 구조를 바꾸지 않고 파라미터 값들의 분포를 조금만 활용하게(높은 메모리가 아닌) 하여 메모리 이점이 가장 큼
- 연산 에너지 절약

INT8 Operation	에너지 절약 (vs FP32)	저장공간 절약 (vs FP32)
Add 연산	30배	116배
Multiply 연산	18.5배	27배

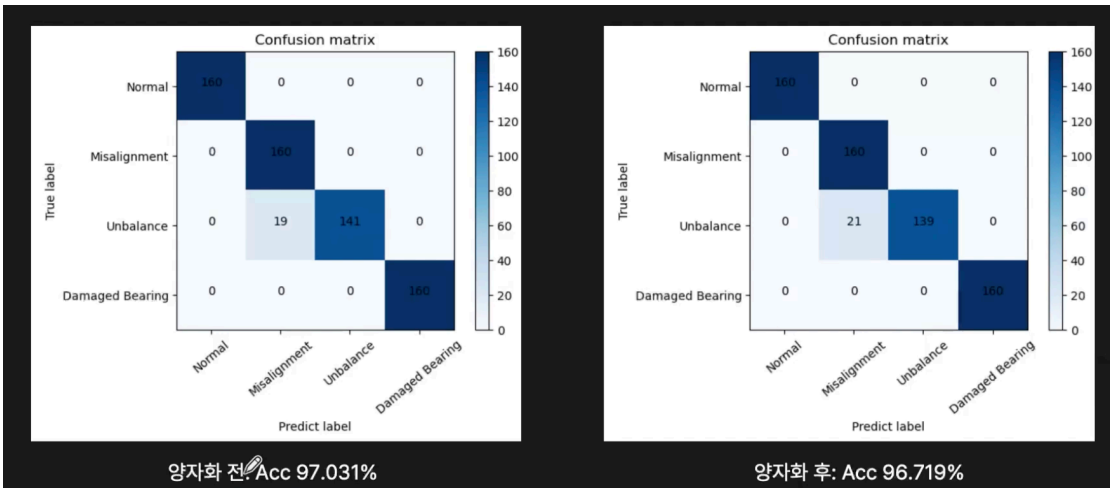
양자화 장점: 속도 및 소비 전력

- 추론속도 향상
    - 정수연산은 일반적으로 부동소수점 연산보다 더 적은 계산 복잡도를 가짐
- 전력 소비 감소
    - 단순해진 연산으로 전력소비가 줄어들게 된다

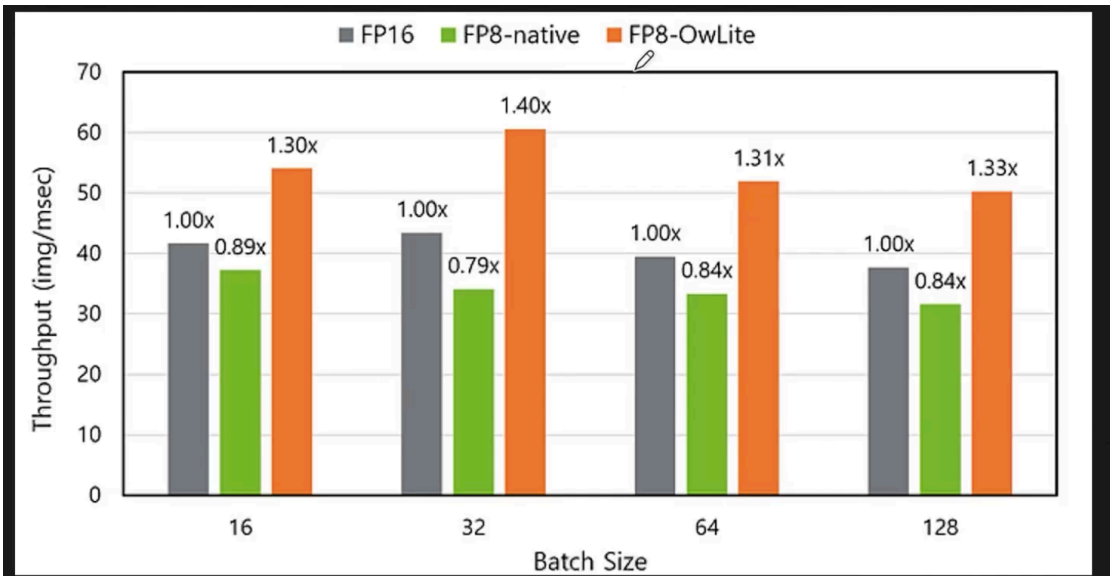
- 배터리 기반 장치에서도 더 오래 작동할 수 있음

성능이 막 떨어지지 않는가?

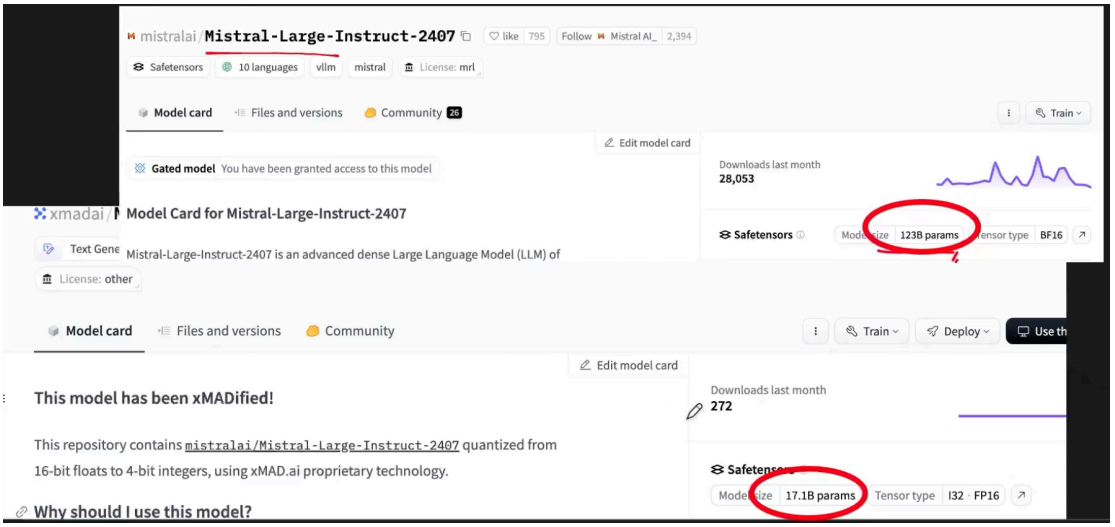
- CNN 모델 기준 (CV)



- 처리량 (Throughtput)



- 모델 크기 차이



- 메모리 사용량 감소

Model	Size	Version	Batch Size	Prefill Length	Decode Length	Prefill tokens/s	Decode tokens/s	Memory (VRAM)
Mixtral	46.7B	GEMM	1	32	32	149.742	93.406	25.28 GB (53.44%)
Mixtral	46.7B	GEMM	1	64	64	1489.64	93.184	25.32 GB (53.53%)
Mixtral	46.7B	GEMM	1	128	128	2082.95	92.9444	25.33 GB (53.55%)
Mixtral	46.7B	GEMM	1	256	256	2428.59	91.5187	25.35 GB (53.59%)
Mixtral	46.7B	GEMM	1	512	512	2633.11	89.1457	25.39 GB (53.67%)
Mixtral	46.7B	GEMM	1	1024	1024	2598.95	84.6753	25.75 GB (54.44%)
Mixtral	46.7B	GEMM	1	2048	2048	2446.15	77.0516	27.98 GB (59.15%)
Mixtral	46.7B	GEMM	1	4096	4096	1985.78	77.5689	34.65 GB (73.26%)

양자화의 단점

1. 성능 하락

- quantize error를 줄이면서 양자화를 하여도 성능적 하락이 크다는 문제점이 있음

- 특히, 8비트까지는 양호하지만, 4비트로 양자화하는 경우 코드 쉬이칭이나 모델이 망가지는 경우가 많음
- 특히, calibrating 과정에서 carlibrate data를 고려하지 않을시, 성능이 상당히 망가지는 경우가 생김
    - calibrating : outlier를 어떻게 제거할 것인가?
- 가중치의 outlier를 적절히 처리해도, 기존의 성능보다 하락하는 문제를 해결할 수 있음

2. 연산속도

- 보통의 커널은 float 연산에 최적화가 잘 되었지만, int 연산에서는 그렇지 않다.

Precision	Number of parameters	Hardware	Time per token in milliseconds for Batch Size 1	Time per token in milliseconds for Batch Size 8	Time per token in milliseconds for Batch Size 32
bf16	176B	8xA100 80GB	239	32	9.9
int8	176B	4xA100 80GB	282	37.5	10.2
bf16	176B	14xA100 40GB	285	36.5	10.4
int8	176B	5xA100 40GB	367	46.4	oom
fp16	11B	2xT4 15GB	11.7	1.7	0.5
int8	11B	1xT4 15GB	43.5	5.3	1.3
fp32	3B	2xT4 15GB	45	7.2	3.1
int8	3B	1xT4 15GB	312	39.1	10.2