Chapter4. Quantization의 최신 트렌드

Spin Quant

https://arxiv.org/pdf/2405.16406

- LLM에서 가중치 및 activation 값의 outlier는 양자화 범위를 확장시켜
  정밀도 손실을 유발, 이러한 이상치를 제거하기 위해 가중치와 activation에 rotation tarnsformation을 적용해 기존 분포의 이상치를 제거함
- Cayley SGD를 활용하여 회전행렬을 학습하며,
  이 회전 변환은 모델의 출력에 영향을 미치지 않으면서 outlier를 제거
- 두 가지의 Rotation을 제시
  - SpinQuant no had
    - 회전 행렬과 R1, R2 가중치를 행렬에 병합하여 양자화를 수행
    - 이 방식은 저비트 양자화가 필요하지 않은 경우에 사용되며
    - 모델 아키텍처를 변경하지 않고 성능 향상
  - SpinQuant had
    - 4비트처럼 매우 낮은 비트 양자화가 필요할 때 추가적으로 Hadamard 회전 행렬 R3, R4를 적용
    - 이는 MLP 블록의 activation 값과 KV 캐시의 이상치를 줄여 극단적인 양자화 조건에서도 성능을 유지
- Rotation Transformation은 데이터나 값들을 특정 각도로 회전시키며
  LLM에서 회전 행렬을 적용하면 특정 축을 기준으로 데이터 포인트를 새로운 위치로 이동시켜, 데이터 분포를 새롭게 조정
- 만약, 특정 축에서 활성화 값이나 가중치 값이 평균에서 멀리 떨어져 있다면(즉, 이상치가 있다면), 이 값은 그 축의 양자화 범위를 넓혀, 전체 분포를 균일하게 양자화하기 어려운 상태로 만듦
- 회전 행렬을 적용하여 이 값을 다른 축으로 이동시키면,
  각 축에 대해 고르게 분포가 재조정되며, 극단적인 값들이 평균에 더 가깝게 모이게 됨으

로 전체 분포의 폭이 줄어들어 양자화가 쉬워질 수 있음

Spin Quant - Cayley SGD

- Cayley SGD
  - Stiefel Manifold에서 회전 행렬을 최적화하기 위해 설계된 효율적인 기법으로
  - Cayley 변환이라는 특별한 매핑을 사용하여, skew-symmetric matrix를 통해 직교 행렬을 갱신
  - 이는 직교성을 유지하면서 행렬의 방향을 조정할 수 있게 해주는 기법
  - Stiefel Manifold는 특정 조건을 만족하는 직교행렬들의 집합을 나타내는 공간으로 n*p 크기의 행렬 V가 있을 때, 각 V의 각 열 벡터가 서로 직교하고, 각각의 벡터가 단위 길이를 가질 때 이 행렬이 Stiefel 다양체에 속한다고 함 (서로 직교하며, 길이가 1)
  - 특징
    - othogonality
      - Stiefel 다양체에 속하는 행렬의 열 벡터들은 서로 직교하고
        단위 길이를 가지기 때문에 데이터의 특정 패턴을 유지하면서 변환하는 성질
    - 제약된 최적화 문제 해결
      - Stiefel 다양체는 직교 제약을 포함한 최적화 문제를 해결하는 데 자주 사용
    - Stefel 다양체는 LLM에서 회전 행렬을 최적화하여 활성화나 가중치의 분포를 고르게 만드는 데 유용
    - 즉, 회전 행렬은 othogonal하기 때문에 Stiefel 다양체에 속하며
      이러한 특징 때문에 Cayley SGD를 활용하여 직교성을 유지하는 최적의 회전 행렬을 찾음