졸업논문
소개

Coarse-Fine LM Head 근사화 기반 고성능 LPU 설계

김정우 / 광운대학교 Intelligent Computing Lab.

대규모 언어 모델의 추론 과정은 입력 토큰을 한 번에 처리하여 KV cache를 형성하는 prefill 단계와, 이후 한 시점에 하나의 토큰을 순차적으로 생성하는 decoding 단계로 구분된다. Prefill 단계는 동일 가중치를 다수 토큰에 공유할 수 있어 병렬 처리 효율이 높지만, decoding 단계는 매 시점 단일 토큰을 처리하므로 행렬 연산의 병렬성이 제한되고 사용자 체감 지연을 지배한다. 따라서 실제 응답 속도를 개선하기 위해서는 decoding 단계의 지연을 유발하는 주요 병목을 직접 완화하는 접근이 필요하다.

Decoding 단계에서 마지막에 수행되는 LM Head는 최종 hidden state를 어휘 크기 차원의 logit으로 변환하는 단일 선형 연산이다. 이때 LM Head 가중치 행렬의 크기는 모델 차원과 어휘 크기에 비례하여 매우 커지며, 어휘가 10만 단위를 넘는 설정에서는 수백 MB에서 수 GB 규모에 이른다. Decoding은 토큰을 하나 생성할 때마다 LM Head의 전 열에 해당하는 가중치를 반복적으로 참조해야 하므로, 연산량뿐 아니라 외부 메모리에서 가중치를 읽어오는 비용이 지연과 대역폭을 지배하는 구조가 된다. 이러한 특성 때문에 LM Head는 decoding 단계에서 큰 비중의 시간과 메모리 트래픽을 차지하며, 단순한 계산 재배치만으로는 병목을 근본적으로 제거하기 어렵다.

본 논문은 LM Head 병목이 전체 어휘에 대한 정밀 logit을 매 시점 완전히 계산하는 구조에서 기인한다고 보고, LM Head를 두 단계로 재구성하는 coarse-fine 구조를 제안한다. Coarse 단계에서는 전체 어휘에 대한 점수를 저비용으로 근사 계산하여 상위 후보 집합을 빠르게 추출하고, fine 단계에서는 이 후보에 대해서만 원래의 정밀 계산을 수행하여 최종 argmax 토큰을 결정한다. 결과적으로 외부 메모리에서 읽어야 하는 정밀 LM Head 가중치의 양이 전체 어휘 규모에서 후보 집합 크기만큼으로 감소하고, 고비용 GEMV 수행도 후보 집합에 제한된다.

Coarse 단계는 곱셈을 최소화하기 위해 부호 기반 근사나 LUT 조회와 덧셈 중심의 계산으로 구성된다. 다만 부호만 남기는 단순 근사는 logit 순위를 왜곡할 수 있어, top-1 토큰을 coarse 후보에 포함시키려면 k가 커지는 문제가 생긴다. 이를 완화하기 위해 본 연구는 타일 단위로 LM Head 가중치를 분할하고, 각 타일에서 k-means 클러스터링으로 대표 벡터 코드북과 인덱스 테이블을 구성하는 타일 단위 벡터 양자화를 적용한다. 디코딩 시점에는 활성화 타일과 코드북의 내적을 타일별 LUT로 만들고, 어

휘별 인덱스를 스트리밍하여 LUT 조회값을 누적함으로써 coarse logit을 계산한다. 이때 외부 메모리에서 전송되는 것은 고정소수점 가중치가 아니라 소형 인덱스이므로 대역폭이 크게 줄고, 클러스터링으로 순위 왜곡이 감소하여 동일 정확도를 더 작은 k로 달성한다.

제안한 coarse-fine LM Head를 효율적으로 실행하기 위해 HTE-LPU 아키텍처를 설계한다. Multi-head attention의 head 단위 독립성을 활용해 head-wise 엔진을 기본 단위로 구성하고, 엔진 내부에 reduction과 최대값 탐색에 적합한 트리 구조 연산을 수행하는 TCU, online softmax 기반 attention 스케줄링을 파이프라인으로 처리하는 OSU, 그리고 prefill의 GEMM과 decoding의 GEMV를 모두 효율적으로 지원하는 multi-pod VPU를 배치한다. 이 구조는 decoding 단계에서 LM Head와 attention, feed-forward를 동일한 엔진 패턴으로 처리 가능하게 하며, 엔진 수를 늘리는 방식으로 성능을 확장할 수 있다.

또한 고정소수점 시뮬레이션 방법론을 적용하여 부동소수점 의존성을 제거한다. 노드 단위로 정수부와 소수부 비트폭을 결정하고, SiLU 및 softmax 관련 비선형 연산은 LUT 및 수치 근사로 구현하되 입력 범위와 해상도를 실험적으로 설정하여 수치 안정성을 확보한다. 최종적으로 WikiText-2, C4, PTB에서 perplexity 변화가 매우 작은 수준임을 확인하여, 고정소수점 기반 구현에서도 정확도가 유지됨을 보인다.

마지막으로 AMD Alveo V80 FPGA를 대상으로 구현하여 실효성을 검증한다. Coarse-fine 구조는 decoding 단계에서 LM Head의 정밀 GEMV를 후보 집합으로 제한하여 연산과 메모리 트래픽을 크게 줄이고, 클러스터링 기반 coarse 계산은 필요한 k를 감소시켜 top-k 선택과 fine 계산 부담을 동시에 완화한다. 또한 HTE-LPU는 LUT와 온칩 메모리를 적극 활용해 데이터 이동 병목을 줄이고, head-wise 병렬성으로 처리량을 향상시킨다. 이는 decoding 중심 병목을 대상으로 알고리즘 구조와 하드웨어 구조를 함께 재설계하면, 정확도 유지와 처리량 개선을 동시에 달성할 수 있음을 보여준다.

김 정 우

- 2024년 : 광운대학교 전자재료공학과 학사
- 2026년 2월 : 광운대학교 전자재료공학과 석사졸업(예정)
- 주관심분야 : LLM 최적화, AI 가속기 설계, FPGA 및 ASIC 설계