졸업논문
소개

자기회귀 비디오 디퓨전 모델을 위한 학습 없는 KV 캐시 관리 기법 연구

이정 / KAIST 김재철AI대학원

최근 자기회귀(autoregressive) 비디오 디퓨전 모델은 인과적 어텐션(causal attention)과 KV(Key-Value) 캐시를 활용하여 프레임을 실시간으로 스트리밍 생성할 수 있게 되었으며, 사용자의 행동 입력에 따라 장면을 생성하는 인터랙티브 월드 모델(world model)의 핵심 구조로 자리 잡고 있다. 그러나 이러한 모델은 수 초 길이의 짧은 영상으로 학습되기 때문에 분 단위의 긴 영상을 생성하면 오차가 누적되어 화질이 저하되고, 슬라이딩 윈도우(sliding window) 밖으로 밀려난 과거 정보를 잃어 이미 지나온 장면을 일관되게 재현하지 못한다. 전체 KV 캐시를 유지하면 이를 일부 완화할 수 있으나 메모리 사용량과 어텐션 연산량이 생성 길이에 비례하여 증가하므로 실시간성을 잃게 된다. 본 연구에서는 추가 학습 없이(training-free) KV 캐시를 관리하여 장시간 비디오 생성의 품질 저하와 월드 모델의 장기 기억 문제를 해결하는 두 가지 기법을 제안한다.

첫 번째로, 장시간 비디오 생성에서 발생하는 화질 저하와 움직임 정체를 완화하기 위해 Deep Forcing을 제안한다. 기존 Self Forcing 계열 모델은 캐시가 가득 차면 가장 오래된 프레임부터 제거하는 FIFO(First-In, First-Out) 방식을 사용하여 중요한 문맥을 잃게 된다. 또한 대형 언어 모델에서 제안된 어텐션 싱크(attention sink)를 그대로 적용하여 초기 몇 프레임만 유지하면 오히려 화질 저하와 움직임 정체가 나타난다. 본 연구에서는 사전 학습된 Self Forcing의 어텐션 분포를 분석하여, 모델이 초기 프레임뿐만 아니라 중간 구간의 프레임에도 상당한 어텐션을 할당한다는 점을 확인하였다.

이를 바탕으로 제안하는 Deep Sink는 슬라이딩 윈도우의 약 절반을 지속적으로 유지되는 싱크 토큰에 할당하고, 싱크 토큰의 시간 축 RoPE(Rotary Positional Embedding) 인덱스만을 현재 시점에 맞게 재정렬한다. 이를 통해 공간 정보는 그대로 유지하면서 시간적 불연속으로 인한 깜빡임과 롤백(roll-back) 현상을 방지한다. 또한 Participative Compression은 최근 프레임의 쿼리가 각 후보 토큰에 부여하는 어텐션 점수를 합산하여 중요도를 계산하고 상위 C개의 토큰만 캐시에 남긴다. 이를 통해 불필요하거나 열화된 과거 토큰이 어텐션을 분산시켜 오차를 누적시키는 현상을 억제한다.

두 번째로, 인터랙티브 월드 모델에서 이전에 방문한 시점으로 되돌아왔을 때 동일한 장면을 재현하는 장기 기

억 문제를 다루기 위해 WorldKV를 제안한다. 분석 결과, 짧은 영상으로 학습된 Matrix-Game-2.0도 전체 KV 캐시에 접근할 수 있으면 과거 장면을 재현하였으며, 모델은 현재 시점과 겹치는 과거 시점의 KV 캐시에 높은 어텐션을 부여하였다. 즉 KV 캐시 자체가 이미 월드 메모리로 기능하며, 핵심은 이를 효율적으로 활용하는 방법에 있다.

제안하는 World Retrieval은 슬라이딩 윈도우에서 제거되는 KV 캐시 청크를 버리지 않고 생성 당시의 카메라 및 행동 정보와 함께 GPU/CPU 메모리에 저장한 뒤, 현재 카메라 상태와 가장 가까운 상위 k개의 청크를 재인코딩 없이 어텐션 윈도우에 다시 삽입한다. World Compression은 인접 프레임의 KV 캐시가 서로 거의 중복된다는 점에 착안하여, 각 청크의 첫 프레임을 앵커로 두고 나머지 프레임의 키 중 앵커 키와의 코사인 유사도가 높은 토큰을 제거한다. 이를 통해 청크 크기를 약 절반으로 줄여 동일한 메모리 예산에서 두 배의 과거 정보를 보존한다.

Deep Forcing은 5초 길이로 학습된 Self Forcing에 적용하여 12배 이상 긴 60초 영상을 생성하였으며, VBench-Long 평가에서 실시간 생성 속도(15.75 FPS)를 유지하면서 Dynamic Degree 57.19를 달성하여 Self Forcing(31.98)은 물론 학습 기반 기법인 LongLive(43.49)와 Rolling Forcing(31.35)을 상회하였다. WorldKV는 60개의 장면-궤적 재방문 벤치마크에서 평가하였으며, LingBot-World-Fast(14B)에서 전체 KV 캐시 어텐션에 근접한 기억 성능(PSNR 15.66 대 15.90)을 약 2배의 처리 속도로 달성하였다. Matrix-Game-2.0(1.3B)에서는 PSNR 14.10으로 슬라이딩 윈도우(11.42)와 전체 KV 캐시(13.75)를 모두 능가하였고, 메모리 모듈을 별도로 학습한 WorldPlay, Yume-1.5와 비교해서도 경쟁력 있는 성능을 보였다.

본 연구는 자기회귀 비디오 디퓨전 모델의 KV 캐시가 단순한 연산 버퍼가 아니라 장시간 생성 품질과 장기 기억을 좌우하는 핵심 요소임을 보이고, 추가 학습 없이 KV 캐시를 선택, 압축, 검색하는 것만으로 학습 기반 기법에 필적하는 성능을 달성하였다. 제안 기법들은 사전 학습된 모델에 바로 적용할 수 있으므로 실시간 인터랙티브 월드 모델과 장시간 비디오 생성 시스템의 효율적인 추론 기법으로 활용될 수 있을 것으로 기대된다.



이 정

- 2024년 8월 : 연세대학교 전기전자공학부 학사
- 현재 : KAIST 김재철AI대학원 석사과정
- 주관심분야 : 비디오 생성, 월드 모델, 효율적 추론