# 장기 동영상 이해를 위한 효율적인 메모리 메커니즘

조선희[1], 김종희[2*], 문진영[3]
[1]한밭대학교 전자공학과 학부생
[2]한국전자통신연구원 선임연구원
[3]한국전자통신연구원 책임연구원
josh201455@gmail.com, jhkim27@etri.re.kr, jymoon@etri.re.kr

# Efficient Memory Mechanism for Long-form Video Understanding

Sun hee Jo[1], Jonghee Kim[2*], Jinyoung Moon[2]
[1]Dept. of Electronics Engineering, Hanbat National University
[2]Electronics and Telecommunications Research Institute (ETRI)

요 약

본 논문에서는 장기 비디오 이해를 위한 새로운 메모리 메커니즘을 제안하였다. 제안된 메모리 메커니즘은 메모리 구성에 사용되는 시각 토큰을 압축하여, 메모리 사용량과 연산 비용을 줄이면서도 효율적인 비디오 처리를 목표로 한다. 다양한 시각 토큰 압축 방법을 적용 및 비교하였으며, MSVD-QA 데이터셋을 활용한 실험 결과, 제안된 메커니즘이 기존 방법에 비해 효율성과 성능 면에서 모두 우수함을 확인하였다. 본 연구는 장기 비디오 이해의 효율성을 높일 수 있는 새로운 접근 방식을 제시한다.

## 1. 서론

최근 인공지능 연구에서 대규모 언어 모델(LLM)을 기반으로 한 멀티모달 모델이 급속히 발전하고 있다. 이러한 모델은 언어뿐만 아니라 영상, 음성 등 다양한 데이터 형식을 이해하고 처리할 수 있으며, 특히 짧은 동영상에 대한 추론 능력이 크게 향상되었다. 그러나 긴 시간의 동영상을 처리하려면 많은 프레임을 분석해야 하기 때문에 여전히 계산 자원과 메모리 비용이 많이 소모되는 한계가 존재한다.

이 문제를 해결하기 위해, 최근 연구들은 인간의 기억 메커니즘을 모방한 방식을 채택하고 있다. 인간은 과거의 모든 정보를 기억하지 않고, 중요한 정보만을 압축해 장기 기억에 저장한 뒤 필요할 때 이를 참조한다. 이러한 개념을 바탕으로, 긴 비디오를 순차적으로 처리하면서 중요한 정보만을 메모리에 저장하고, 이후 이를 활용하는 방법이 제안되고 있다[1, 2]. 이는 기존에 모든 프레임을 개별적으로 처리하는 방식보다 훨씬 효율적이며, 긴 비디오 처리에 필요한 자원을 크게 절약할 수 있다.

본 논문에서는 기존 방법에서 제안된 메모리 메커니즘을 기반으로 효율적인 메모리 메커니즘을 제안하고자 한다. 제안하는 방법은 시각 정보를 압축하여 메모리 연산에 사용되는 데이터 양을 줄이는 방법으로, 이를 위해 다양한 시각 정보 압축 기법을 적용하여 장기 비디오 이해에서의 성능을 비교한다. 대표적인 동영상 질의응답 데이터셋인 MSVD-QA [3]에서의 실험을 통해 제안하는 방법이 기존 메모리 메커니즘에 비해 적은 연산량으로 더 높은 정확도를 달성했음을 입증하였다.

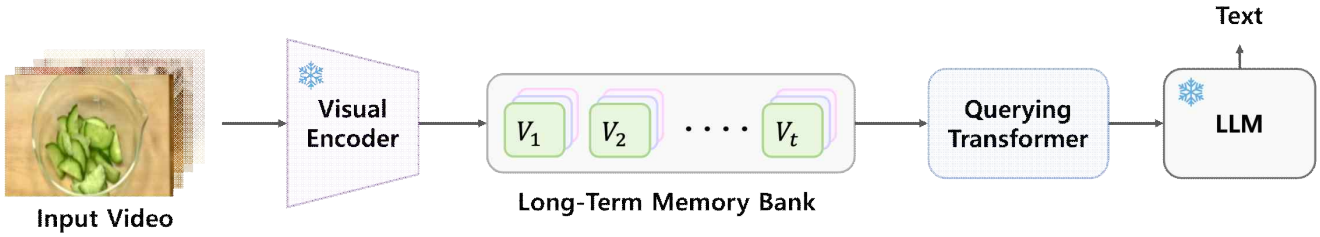
## 2. 제안하는 방법

본 논문에서는 메모리를 사용하는 장기 동영상 이해 모델인 MA-LMM [1]을 기반으로 다양한 시각 토큰 압축 기법을 활용하여, 효율적인 메모리 메커니즘을 제안한다.

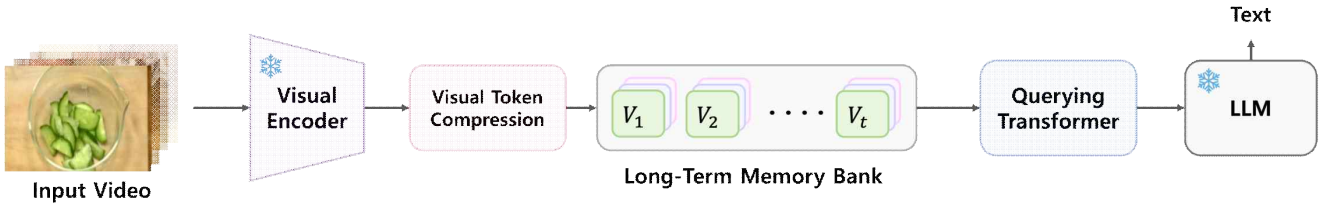
### 2.1 MA-LMM

MA-LMM은 장기 비디오 이해를 위해 설계된 모델로, 비디오 프레임을 순차적으로 처리하면서 과거 비디오 정보를 메모리 뱅크에 저장하여 장기적인 비디오 분석을 가능하게 한다. 이 모델은 크게 세 가지 주요 구성 요소로 이루어져 있다. 먼저 Visual Encoder는 각 비디오 프레임의 시각적 특징을 추출하여 메모리 뱅크로 저장하고, 비디오 프레임이 순

* 교신저자 (Corresponding Author)




(a) MA-LMM




(b) 제안하는 방법

(그림1) (a) MA-LMM 구조 및 (b) 제안하는 시각 토큰 압축 기법을 활용하는 LMM. 시각 토큰 압축 기법에는 시각 토큰 풀링, 계층적 시각 토큰 병합, ToMe 기반 시각 토큰 압축이 사용될 수 있음.

차적으로 입력될 때 시각적 정보를 점진적으로 축적한다. Long-Term Memory Bank는 장기 비디오 분석에서 핵심적인 역할을 수행하며, 비디오의 과거 정보를 저장하고 이를 나중에 참조한다. 이 메모리 뱅크는 Visual Memory Bank와 Query Memory Bank로 나뉘며, Visual Memory Bank는 시각적 특징을 저장하고 교차 주의(attention) 메커니즘에서 사용되며, Query Memory Bank는 Q-Former에서 생성된 학습된 쿼리를 저장하여 시간에 따른 비디오 정보를 축적한다.

Querying Transformer (Q-Former)는 시각적 정보와 텍스트 임베딩을 정렬하여 상호작용을 가능하게 하며, 과거 정보와 현재 정보를 함께 처리하여 장기적인 맥락을 반영한다. 또한, MA-LMM은 Memory Bank Compression 기술을 사용하여 메모리 뱅크의 길이가 지나치게 길어지지 않도록 유사한 프레임 간의 특징을 선택하고 평균화하여 메모리를 압축한다.

### 2.2 시각 토큰 풀링

시각 토큰 풀링은 비디오 프레임의 시각적 정보를 압축하여 처리 효율성을 높이는 방법이다. VideoGPT+ [4]에서 사용된 것처럼, 각 프레임에서 추출된 시각적 특징을 공간적으로 풀링해 정보의 차원을 축소한다. 이를 통해 중요한 시각적 특징은 유지하면서도 불필요한 세부 정보를 제거할 수 있다. 이 방법은 특히 긴 비디오 시퀀스를 처리할 때 연산 부담을 줄이고 메모리 사용량을 효율적으로 관리하는 데 중요한 역할을 한다.

### 2.3 계층적 시각 토큰 병합

계층적 시각 토큰 병합은 비디오의 각 프레임에서 추출된 시각적 특징을 단계적으로 병합하는 방법이다. 각 단계에서는 프레임 내 시각 토큰 간의 유사도를 계산하고, 상위 $n$쌍의 유사한 토큰들을 병합하여 토큰 수를 점진적으로 줄인다. 토큰을 단계적으로 병합하면서 유사하지 않은 토큰 쌍이 유사도 기준 상위 $n$쌍에 포함되어 병합되는 것을 방지한다. 점진적인 병합 과정을 통해 중복된 정보를 제거하면서도 중요한 시각적 특징을 보존할 수 있으며, 동시에 연산 복잡도를 효과적으로 낮출 수 있다.

### 2.4 ToMe 기반 시각 토큰 압축

ToMe(Token Merging) [5] 기반 시각 토큰 압축은 긴 비디오 시퀀스에서 발생하는 연산 비용을 줄이기 위한 방법이다. ToMe는 유사한 시각 토큰들을 병합하여 중복된 정보를 압축하며, 이를 통해 비디오 프레임의 시각적 특징을 더 적은 수의 토큰으로 표현할 수 있다. 이 방식은 가장 유사한 인접한 두 개의 토큰을 반복적으로 병합함으로써 중복된 데이터를 제거하면서 그 순서를 유지할 수 있는 방법이다. ToMe 기반 시각 토큰 압축은 긴 비디오 시퀀스의 처리에서 매우 유용하며, 메모리 사용량과 연산 속도를 크게 개선할 수 있다.

또한, MA-LMM에서 사용된 방법과 달리, 먼저 시각 정보를 압축하여 시각 메모리 뱅크를 구성하기

<표1> MSVD 데이터셋에서 Top-1 Accuracy에 대한 비교

| | Epoch 1 | Epoch 2 | Epoch 3 | Epoch 4 | Epoch 5 | 최고 성능 | 학습 시간 |
|---|---|---|---|---|---|---|---|
| MA-LMM [1] | 58.14 | 60.23 | 59.92 | 60.06 | 59.81 | 60.23 | 5시간 14분 |
| 시각 토큰 풀링 (4×4) | 55.65 | 57.48 | 57.89 | 57.14 | 57.34 | 57.89 | 4시간 31분 |
| 시각 토큰 풀링 (6×6) | 55.58 | 56.10 | 57.91 | 57.70 | 57.46 | 57.91 | 4시간 46분 |
| 계층적 시각 토큰 병합 (최종 토큰 개수 65개) | 57.85 | 59.61 | 59.26 | 58.91 | 58.65 | 59.61 | 4시간 31분 |
| 계층적 시각 토큰 병합 (최종 토큰 개수 129개) | 58.15 | 60.03 | 59.41 | 59.53 | 59.91 | 60.03 | 4시간 47분 |
| 계층적 시각 토큰 병합 (최종 토큰 개수 193개) | 58.21 | 59.77 | 59.62 | 59.34 | 58.95 | 59.77 | 5시간 0분 |
| ToMe 기반 시각 토큰 압축 | 58.24 | 60.58 | 59.93 | 60.18 | 59.34 | 60.58 | 4시간 18분 |
| ToMe 기반 시각 토큰 압축 (Q-former 1회 적용) | **58.48** | **61.15** | **60.58** | **60.27** | **60.18** | **61.15** | **3시간 53분** |

때문에 기존 방법보다 Q-Former를 적게 적용할 수 있으며, 이를 통해 메모리 사용량뿐만 아니라 연산량도 감소시킬 수 있다. 또한, 압축된 시각 메모리 뱅크는 동영상의 대표적인 표현들을 포함하고 있어, 시각 메모리 뱅크 전체를 Q-Former에 한 번만 적용하는 방식으로 연산량을 극적으로 줄일 수 있다.

## 3. 실험 결과

본 논문에서는 제안된 방법의 성능을 입증하기 위해 MSVD-QA 데이터셋을 사용하여 질문에 대한 예측 정확도를 성능 지표로 삼았으며, 성능뿐만 아니라 효율성 확인을 위해 학습 시간도 비교하였다. 표 1에서는 기존 방법과 제안하는 방법 간의 정확도와 학습 시간을 비교하여, 제안하는 방법의 성능과 효율성을 확인하였다.

먼저, 시각 토큰 풀링 방법의 경우, 4×4, 6×6 두 가지 풀링 기법을 적용하였다. 두 방법 모두 MA-LMM에 비해 학습 시간이 적게 소요되었으나, 정확도는 낮게 나타났다. 이는 시각 토큰을 단순히 풀링함으로써 계산량을 줄일 수는 있었으나, 정보 손실이 발생해 장기 비디오 이해 성능에 악영향을 끼쳤음을 확인할 수 있었다.

다음으로, 계층적 시각 토큰 병합 방법은 최종적으로 병합된 각 프레임별 토큰의 개수를 달리하여 세 가지 실험을 수행하였다. 최종 토큰 개수는 각각 193개, 129개, 65개로, 한 프레임에서 발생되는 257개의 토큰 중 [CLS] 토큰은 유지하고 나머지 256개의 토큰들을 순차적으로 줄인 결과이다. 표 1에서 확인할 수 있듯이, 토큰 개수가 적을수록 학습 시간이 짧았으며, 정확도는 MA-LMM에 비해 약간 낮아지는 경향을 보였다. 이는 계산 효율성은 증가하지만, 성능은 약간 저하된 결과를 나타낸다.

마지막으로, ToMe 기반 시각 토큰 압축 방법에서는 시각 토큰을 압축하여 시각 메모리 뱅크를 구성한 후, 두 가지 방식으로 실험을 진행하였다. 첫 번째 방법은 시각 메모리를 순차적으로 Q-Former에 입력하는 방식이며, 두 번째 방법은 전체 메모리 뱅크를 한 번에 Q-Former에 입력하는 방식이다. 두 방법 모두 기존 MA-LMM에서 Q-Former를 20회 적용하는 것에 비해 각각 10회, 1회만 적용하여 계산량을 크게 줄였다. 이는 학습 시간에서 그 차이가 명확히 드러났으며, 특히 Q-Former를 한 번만 적용하는 방식은 학습 시간이 약 25% 감소했다. 또한, 질의응답 정확도에서도 두 방법 모두 기존 방식보다 향상된 성능을 보였다. 특히, Q-Former를 한 번만 사용하는 경우에는 학습 시간이 25%가량 단축되었으며, 질의응답 정확도는 약 0.9%p 향상되었다. 이를 통해 제안하는 메모리 메커니즘이 기존 MA-LMM의 Q-Former 기반 메모리 메커니즘에 비해 성능과 효율성 모두에서 우수한 결과를 나타냈음을 확인할 수 있었다.

## 4. 결론 및 향후 연구 방향

본 논문에서는 장기 비디오 이해를 위한 효율적인

메모리 메커니즘을 제안하였다. 제안된 메커니즘은 시각 토큰을 압축하여 메모리 뱅크를 효과적으로 활용함으로써 계산량과 메모리 사용량을 줄이는 동시에 성능을 향상시키고자 하였다. MSVD-QA 데이터셋을 활용한 실험 결과, 제안한 메모리 메커니즘은 기존 방법에 비해 연산 비용을 크게 절감하면서도 질의응답 정확도를 향상시키는 성과를 보였다. 이는 제안된 메커니즘이 장기 비디오 분석에서 효율성과 성능 면에서 모두 우수함을 입증한다. 향후 연구에서는 다양한 데이터셋에 대한 확장 실험과 다른 멀티모달 모델과의 결합을 통해, 제안한 메커니즘의 범용성을 더욱 검증할 예정이다.

## 사 사

이 논문은 과학기술정보통신부의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임. (No.2020-0-00004, 장기 시각 메모리 네트워크 기반의 예지형 시각지능 핵심기술 개발).

## 참고문헌

[1] Bo He et al., “MA-LMM: Memory-augmented large multimodal model for long-term video understanding,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.

[2] Enxin Song et al., “Moviechat: From dense token to sparse memory for long video understand
ing,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024.

[3] Dejing Xu et al., “Video Question Answering via Gradually Refined Attention over Appearance and Motion,” in Proceedings of the ACM International Conference on Multimedia, 2017.

[4] Muhammad Maaz et al., “VideoGPT+: Integrating Image and Video Encoders for Enhanced Video Understanding,” arXiv preprint arXiv:2406.09418, 2024.

[5] Daniel Bolya et al., “Token Merging: Your ViT But Faster,” in Proceedings of the International Conference on Learning and Representation, 2023.