iRAG: Advancing RAG for Videos with an Incremental Approach

Announcement Date	@2025년 3월 6일
Conference Name	CIKM 2024 Applied Research Paper
Keywords	LLM Multi-Modal RAG

초록 (Abstract)

1. 기존 RAG 시스템은 비디오 데이터를 텍스트로 변환하는 데 시간이 많이 걸리고, 중요한 세부 정보가 누락되는 한계를 가진다.
2. 또한, 사용자 질의를 사전에 알 수 없어 미리 텍스트를 생성하고 색인하는 방식은 비효율적이다.
3. 이를 해결하기 위해, 우리는 점진적 워크플로우를 도입한 iRAG 시스템을 제안한다.
4. iRAG는 대규모 비디오 데이터에 대해 빠르게 색인을 생성하고, 질의가 입력될 때 필요한 정보를 동적으로 추출한다.
5. 이를 통해 변환 속도를 기존보다 23~25배 향상시키면서도 응답 지연 시간과 품질을 유지하는 효율적인 질의응답 시스템을 구현할 수 있다.

1. 서론 (Introduction)

많은 응용 프로그램에서 비디오 데이터를 수집하고 저장하여 오프라인 분석에 활용하고 있습니다.

예를 들어, **감시 시스템**은 공공 장소, 공항, 교통 허브 및 주요 인프라에서 **보안 모니터링을 수행**하며, 감시 비디오를 저장하여 이후 수사 및 조사에 활용합니다.

또한, **병원 및 의료 시설에서도 환자 모니터링을 위해 비디오 분석 기술을 적극적으로 활용**하고 있습니다.

이 외에도, **교통 모니터링, 교통 정체 관리, 사건 감지 등 다양한 목적으로 비디오 분석이 사용**됩니다.

LLMs와 비디오 분석의 결합

대형 언어 모델(LLM)인 **ChatGPT**와 같은 기술이 등장하면서 **대규모 텍스트 데이터의 이해와 자연스러운 대화 능력이 크게 향상**되었습니다.

이러한 성공을 바탕으로, **LLM을 활용한 비디오 콘텐츠 분석**도 활발히 연구되고 있습니다.

1. **기존 접근 방식**
    - AI 기반 **비전 모델(Vision AI Model)**을 활용하여 개별 비디오 클립을 분석하고, **해당 비디오의 내용을 텍스트로 변환**합니다.
    - 예를 들어, **객체 탐지(Object Detection) 모델**은 **자동차, 트럭, 자전거, 사람 등의 객체를 인식**하고, 해당 위치를 포함한 설명을 텍스트로 변환할 수 있습니다.
    - 이러한 방식으로 **긴 비디오를 짧은 텍스트 조각들로 변환**하여, 대형 언어 모델(LLM)이 이를 활용할 수 있도록 합니다.
2. **RAG 기반 비디오 분석**
    - 최근 연구에서는 **검색 증강 생성(RAG, Retrieval-Augmented Generation)** 기술을 활용하여,**질의 응답 시스템을 강화**하는 방법이 제안되었습니다.
    - **RAG 시스템**은 **비디오 내용을 변환한 텍스트 데이터에서 사용자의 질의와 관련된 내용을 검색하여더 정확한 응답을 생성**할 수 있도록 합니다.

기존 연구의 주요 한계점

그러나, **기존 RAG 기반 비디오 분석 시스템에는 두 가지 주요 한계가 존재**합니다.

1. **긴 비디오를 처리하는 데 지나치게 오랜 시간이 소요됨**
    - 비디오를 사전에 분석하고 텍스트로 변환하려면 **수많은 AI 모델을 실행해야 하며, 이는 매우 높은 연산 비용과 시간을 요구**합니다.
    - 예를 들어, **24시간 분량의 감시 영상을 분석하는 데 최소 하루 이상의 시간이 소요될 수 있으며,**이는 범죄 수사와 같은 긴급한 상황에서 매우 치명적인 문제로 작용할 수 있습니다.
2. **비디오 데이터를 텍스트로 변환하는 과정에서 정보 손실 발생**
    - **비디오의 모든 시각적 정보를 텍스트로 변환할 수 있는 것은 아님**
        - 예를 들어, **중요한 장면에서 AI 모델이 특정 객체를 인식하지 못하면, 해당 정보는 텍스트 변환 과정에서 누락**됩니다.
    - 사용자의 질의를 사전에 알 수 없기 때문에, **어떤 AI 모델을 사용해야 할지 미리 결정하기 어렵다**

    - 특정 질의에 필요한 세부 정보가 기존 텍스트 데이터에 존재하지 않을 수 있으며,**이 경우 비디오를 다시 분석해야 하지만 기존 시스템에는 이를 처리하는 방법이 없음**

---

iRAG: 새로운 접근 방식

위 문제를 해결하기 위해, 우리는 **점진적(incremental) RAG 시스템인 iRAG를 제안합니다.**

iRAG는 다음과 같은 방식으로 동작합니다.

1. **빠른 색인(Indexing) 생성**
    - 기존 방식처럼 모든 비디오를 사전에 텍스트로 변환하는 것이 아니라,*가벼운 AI 모델(예: DETR, CLIP 등)을 활용하여 비디오의 주요 내용을 신속하게 색인(indexing)**합니다.
2. **질의가 들어올 때 추가 정보 추출**
    - 사용자가 질의를 입력하면,**색인된 정보를 바탕으로 비디오의 특정 부분에서 추가적인 세부 정보를 동적으로 추출**합니다.
    - 이를 위해, 필요한 경우에만 **무거운 AI 모델(예: GRiT)을 사용하여 추가 분석을 수행**합니다.
3. **정보 손실 최소화**
    - 기존 방식과 달리, **질의가 들어올 때만 비디오에서 추가 정보를 추출**하므로,**모든 정보를 미리 변환하는 기존 방식보다 정보 손실을 줄일 수 있음**.

---

2. 점진적 RAG (Incremental RAG)

2.1 기존 연구에서 비디오 맥락 활용 방식

기존 연구에서는 **비디오의 전체 내용을 먼저 텍스트로 변환**한 후, 해당 텍스트를 기반으로 검색과 질의 응답을 수행하는 방식을 사용했습니다.

이 과정은 다음과 같은 단계로 이루어집니다.

1. **비디오를 작은 클립으로 분할**
    - 긴 비디오는 여러 개의 짧은 클립으로 나뉩니다.
2. **각 클립을 AI 모델로 분석하여 텍스트 변환**

- 개별 비디오 클립을 **객체 탐지(Object Detection) 모델, 비디오 설명 모델** 등을 활용하여 텍스트로 변환합니다.
- 예를 들어, **차량, 사람, 물체 등의 정보를 인식하여 텍스트로 저장**하는 방식입니다.

3. **모든 클립의 텍스트를 하나의 문서로 결합**
   - 개별 클립에서 추출된 텍스트를 연결하여 **긴 문서 형태로 저장**합니다.
   - 이 문서는 향후 RAG 시스템에서 검색할 수 있도록 준비됩니다.
4. **질의가 들어오면, RAG 시스템이 관련 문서 검색 후 응답 생성**
   - 사용자의 질의에 맞는 **관련 텍스트를 검색**하여, **LLM을 활용해 최종 답변을 생성**합니다.

기존 RAG 방식의 한계

이 방식은 **RAG 시스템을 비디오 분석에 적용하는 기본적인 방법**이지만, 다음과 같은 **치명적인 단점**이 존재합니다.

1. **긴 비디오를 처리하는 데 지나치게 많은 시간이 필요**
   - 모든 비디오 클립을 **AI 모델을 통해 텍스트로 변환하는 데 오랜 시간이 걸림**
   - 특히 **고성능 AI 모델을 사용할 경우, 24시간 분량의 감시 영상을 분석하는 데 하루 이상 소요될 수 있음**
2. **정보 손실 문제**
   - **텍스트 변환 과정에서 중요한 세부 정보가 누락될 가능성 존재**
   - 사용자의 질의가 사전에 정해진 것이 아니므로,**어떤 정보를 미리 변환해야 할지 알기 어려움**
   - 따라서, 특정 질의에 대한 중요한 정보가 텍스트 변환 과정에서 포함되지 않으면, **정확한 답변을 제공할 수 없음**

2.2 iRAG의 사전 처리 (Preprocessing in iRAG)

기존 방식과 달리, iRAG는 **비디오의 모든 내용을 미리 변환하지 않고, 빠르게 색인을 생성하여 즉각적인 질의 응답을 가능하게 합니다.**

이 과정은 다음과 같은 방식으로 진행됩니다.

1. **빠른 색인(Indexing) 생성**
   - *경량 AI 모델 (DETR, CLIP 등)**을 사용하여 **비디오의 주요 내용을 빠르게 색인**합니다.

- 예를 들어, **객체 탐지 모델을 사용하여 비디오에서 등장하는 주요 사물과 위치 정보를 저장**합니다.

2. **질의 발생 시 필요한 정보만 추가 분석**
   - 사용자의 질의가 입력되면, **색인된 정보를 활용하여 특정 비디오 클립을 선택**하고, 필요한 경우에만 **고성능 AI 모델(GRiT 등)을 사용하여 추가 분석**을 수행합니다.

예제 시나리오

사용자가 다음과 같은 질의를 입력했다고 가정해 보겠습니다.

💬 **"비디오에서 FedEx 트럭이 등장하는 부분을 찾을 수 있을까요?"**

1. **기존 RAG 방식**
   - 전체 비디오를 사전에 분석하여 모든 내용을 텍스트로 변환해야 하므로,**처리 시간이 매우 길어질 가능성이 큼**
2. **iRAG 방식**
   - **색인 정보에서 '트럭'이 등장한 구간을 빠르게 검색**
   - 해당 클립에서 **추가 정보를 추출하여 FedEx 트럭이 맞는지 확인**
   - 필요한 경우, **텍스트 변환을 추가 수행하여 LLM이 답변 생성 가능하도록 보강**
   - 이를 통해 **빠르고 정확한 응답 생성 가능**

2.3 질의 응답 (Query-Response)

iRAG는 RAG 시스템을 활용하여 사용자의 질의에 대한 응답을 생성하는 과정에서 **기본적인 질의 응답(기존 RAG 방식)과 점진적 추출 방식(iRAG 방식)을 혼합**합니다.

- 우선, 기존 RAG 방식처럼 **질의와 관련된 색인된 정보를 검색하여 응답을 생성**합니다.
- 만약 **질의 응답이 불가능한 경우**, iRAG는 **비디오의 특정 부분을 추가로 분석하는 "점진적 추출(incremental extraction)"을 수행**합니다.

💬 **기본 질의 응답 프롬프트 예시**

```
당신은 RAG 시스템을 기반으로 동작하는 챗봇입니다.
주어진 컨텍스트를 활용하여 사용자의 질의에 답변하세요.
만약 충분한 정보를 찾을 수 없으면,
"추가 모델 실행 필요"라고 답변하세요.
```

```
컨텍스트: {색인 정보}
질의: {사용자 입력}
```

✔️ 만약 색인 정보만으로 응답이 가능하면, 바로 답변 제공

✔️ 만약 부족한 경우, 추가적인 정보 추출 수행 후 응답 생성

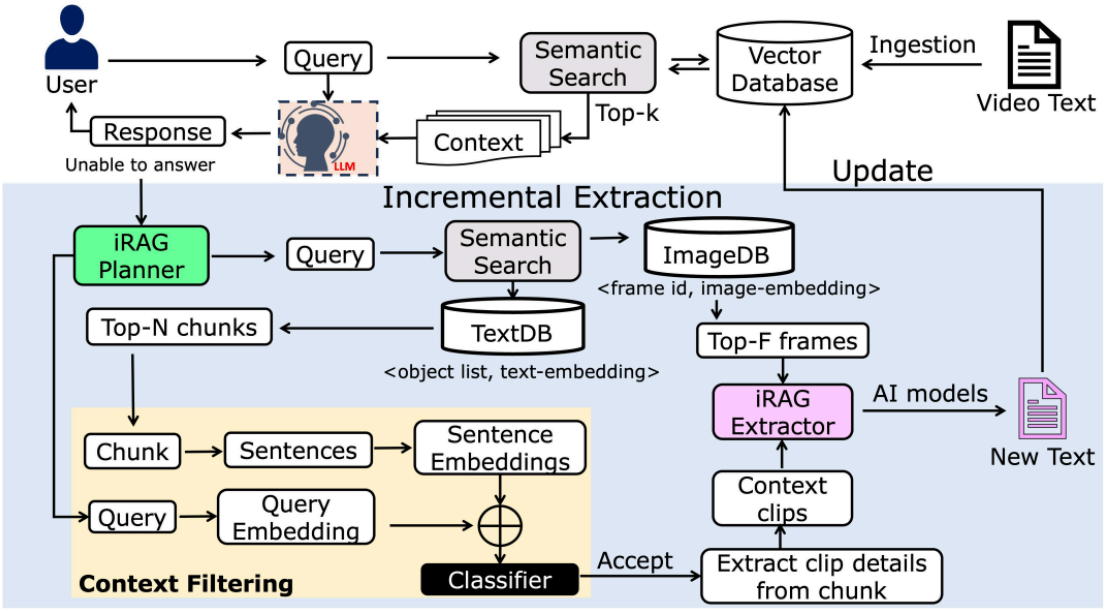

Figure 4: iRAG overview: The light blue rectangle denotes additional workflow compared to a conventional RAG.

2.4 iRAG의 주요 구성 요소 (iRAG Components)

iRAG는 **효율적인 질의 응답을 위해 두 가지 주요 구성 요소를 포함**합니다.

1. **iRAG Planner (계획 수립 모듈)**
   - 색인된 정보를 활용하여 **사용자 질의에 적합한 비디오 클립을 선택**
   - 필요할 경우, **해당 클립에 대한 추가 분석을 수행할지 결정**
   - 빠르고 효율적인 질의 응답을 위한 핵심 모듈
2. **iRAG Extractor (세부 정보 추출 모듈)**
   - Planner에서 선택한 **비디오 클립에서 추가적인 텍스트 변환 수행**
   - 무거운 AI 모델(GRiT 등)을 **질의 기반으로 선택적으로 실행**
   - 필요 시, 기존 색인 정보를 업데이트하여 **향후 질의 응답 성능 향상**

예제 흐름

📌 **질의**: “비디오에서 노란색 택시가 어디에서 등장하나요?”

🔷 **Planner**:

- 색인된 데이터에서 **"택시"가 등장하는 부분을 검색**
- 이 정보만으로 답변이 가능하면, 즉시 응답 생성
- 부족할 경우, **추가 분석 필요 여부를 판단**

🔷 **Extractor**:

- 특정 클립에서 **"택시 색상"을 분석할 수 있는 AI 모델 실행**
- **텍스트 변환 수행 후 색인 정보 업데이트**
- 새로운 정보와 함께 **최종 질의 응답 수행**

✔️ **이와 같은 점진적 분석을 통해 빠르고 정확한 응답 제공 가능**

3. 실험 결과 (Experimental Results)

본 섹션에서는 iRAG의 성능을 기존 RAG 시스템과 비교하여 평가합니다.

실험에서는 다음과 같은 **세 가지 주요 요소**를 검증합니다.

1. **비디오 사전 처리 속도 (Preprocessing Time)**
2. **질의 응답 성능 (Query Processing Performance)**
3. **점진적 추출 방식의 효과성 (Incremental Extraction Efficiency)**

3.1 데이터셋 (Datasets)

iRAG의 성능 평가를 위해 **실제 환경에서 수집된 다양한 데이터셋**을 사용하였습니다.

데이터셋	유형	영상 길이	키프레임 수	테스트 질의 수
VQA-v2	전통적인 데이터셋	2시간 10분	7,799	1,000
MSR-VTT	전통적인 데이터셋	3시간 30분	6,291	318
StreetAware	실제 감시 영상	46분	2,363	40
Tokyo MODI	실제 도시 영상	2시간	1,444	14

- **VQA-v2**와 **MSR-VTT**는 **비디오 기반 질의 응답(Video Question Answering, VQA) 연구에 자주 사용되는 데이터셋**입니다.
- **StreetAware**는 **뉴욕시의 도로 교통 영상을 포함한 감시 비디오 데이터셋**으로, 실제 교통 상황 분석에 활용됩니다.
- **Tokyo MODI**는 **일본 도심의 거리에서 촬영된 영상**으로, 실시간 감시 및 도시 분석에 활용될 수 있습니다.

3.2 질의 데이터 생성 (Query Generation)

- **VQA-v2**와 **MSR-VTT** 데이터셋에는 기존에 정의된 질의가 포함되어 있습니다.
- 하지만 **StreetAware**와 **Tokyo MODI** 데이터셋에는 사전 정의된 질의가 없기 때문에, LLM을 활용하여 질의를 생성하였습니다.

💬 **질의 생성 프롬프트 예시**

> 주어진 비디오 설명을 기반으로 질의를 생성하세요.
> 당신은 수사관이며, 비디오의 텍스트 설명을 바탕으로 조사 질문을 작성해야 합니다.
> 주어진 맥락:
> {비디오 캡션}

→ 이를 통해 StreetAware에서 40개, Tokyo MODI에서 14개의 질의를 생성하였습니다.

3.3 구현 세부 사항 (Implementation Details)

- **iRAG**는 **LangChain 프레임워크**를 활용하여 구현되었습니다.
- **백엔드 데이터베이스**: **FAISS (Facebook AI Similarity Search)** 기반 벡터 데이터베이스를 활용하여 비디오 색인을 구축했습니다.
- **언어 모델(LLM)**: **OpenAI GPT-3.5-turbo API**를 사용하여 질의 응답을 처리했습니다.
- **하드웨어 환경**:
  - **CPU**: AMD Ryzen 5950X
  - **GPU**: NVIDIA GeForce RTX 3090

3.4 평가 지표 (Evaluation Metrics)

- 📌 **recall@k**:

    - 질의에 대해 **검색된 비디오 클립이 기존 RAG 시스템에서 검색한 클립과 얼마나 일치하는지** 평가합니다.
    - k는 검색된 상위 k개 클립을 의미합니다.
    - **recall@k가 1에 가까울수록, 기존 방식과 비교하여 검색 성능이 우수함을 의미**합니다.
- 📌 **질의 처리 시간 (Query Processing Time)**:
    - iRAG와 기존 RAG 시스템 간의 **질의 처리 속도를 비교**합니다.

3.5 실험 결과 (System Evaluation)

3.5.1 비디오 사전 처리 속도 (Preprocessing Time)

iRAG는 기존 RAG 방식과 비교하여 **비디오 사전 처리 속도를 23배~25배 더 빠르게 수행**할 수 있습니다.

데이터셋	시스템	사용 모델	사전 처리 시간
VQA-v2	**iRAG**	DETR + CLIP	**48분 29초**
VQA-v2	**기존 RAG**	GRiT	1093분 36초 (18시간 13분)
MSR-VTT	**iRAG**	DETR + CLIP	**8분 23초**
MSR-VTT	**기존 RAG**	GRiT	199분 50초 (3시간 19분)

✅ **iRAG는 기존 RAG 대비 비디오를 사전 처리하는 속도가 대폭 향상**되었습니다.

3.5.2 점진적 추출 방식의 효과 (Incremental Extraction Efficiency)

k 값	VQA-v2에서 추가 추출된 비디오 비율	MSR-VTT에서 추가 추출된 비디오 비율
2	43.7%	32.3%
10	77.1%	45.8%
20	90.7%	49.3%

✅ **iRAG는 모든 비디오 클립을 처리하지 않고도, 적절한 비율로 필요한 추가 분석만 수행**합니다.

3.5.3 질의 처리 시간 (Query Processing Time)

iRAG는 질의 응답에서 **추가 분석을 수행해야 할 경우, 시간이 증가하지만 여전히 기존 RAG보다 훨씬 빠르게 응답할 수 있음**을 확인했습니다.

k 값	첫 200개 질의 평균 응답 시간 (초)	200~1000개 질의 평균 응답 시간 (초)
2	25~30초	5초 이하
4	70~75초	5초 이하

✅ 초기 질의 응답에서는 시간이 더 걸리지만, 이후에는 색인이 강화되면서 **응답 속도가 빠르게 단축됨**을 확인하였습니다.

---

4. 결론 및 논문의 주요 기여 (Contributions Summary)

✅ **iRAG의 핵심 기여는 다음과 같습니다.**

1. **기존 RAG 방식과 달리, 사전 변환 없이 점진적(incremental) 방식으로 비디오를 분석하는 시스템을 최초로 제안**
   - 기존 방식은 **비디오 전체를 미리 텍스트로 변환해야 하지만**, iRAG는 **경량 모델을 활용하여 빠르게 색인(indexing)**을 생성
   - 질의가 입력되었을 때만, **추가 정보를 필요할 때만 동적으로 추출**하는 방식
2. **효율적인 질의 응답을 위해 iRAG Planner와 Extractor 설계**
   - Planner는 **색인 데이터를 활용하여 사용자 질의와 관련된 비디오 클립을 빠르게 검색**
   - Extractor는 **필요한 경우에만 추가적인 AI 모델을 사용하여 세부 정보를 추출**
3. **실제 데이터셋을 활용한 실험에서 기존 RAG 대비 23배~25배 빠른 처리 속도를 달성**
   - 비디오 분석 시간을 획기적으로 줄이면서도 **질의 응답 품질을 유지**하는 성능 검증

✅ **결론적으로, iRAG는 비디오 기반 질의 응답 시스템의 효율성을 극대화하며, 다양한 실시간 응용 사례에서 활용될 수 있음**을 확인했습니다.