



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0004415
(43) 공개일자 2025년01월08일

(51) 국제특허분류(Int. Cl.)

G06N 3/096 (2023.01) G06F 16/73 (2019.01)
G06F 16/732 (2019.01) G06F 16/783 (2019.01)
G06F 40/284 (2020.01) G06N 3/0455 (2023.01)
G06N 3/0499 (2023.01) G06N 3/09 (2023.01)

(52) CPC특허분류

G06N 3/096 (2023.01)
G06F 16/73 (2019.01)

(21) 출원번호 10-2023-0084060

(22) 출원일자 2023년06월29일

심사청구일자 2023년06월29일

(71) 출원인

이화여자대학교 산학협력단

서울특별시 서대문구 이화여대길 52 (대현동, 이화여자대학교)

(72) 발명자

강제원

서울특별시 마포구 독막로 145, 104동 904호(창전동, 서강쌍용예가아파트)

김현지

서울특별시 서대문구 신촌로7안길 76, 301호 (창천동)

이주희

인천광역시 부평구 길주남로10번길 21, 114동 1703호 (부평동, 래미안부평)

(74) 대리인

특허법인(유한)아이시스

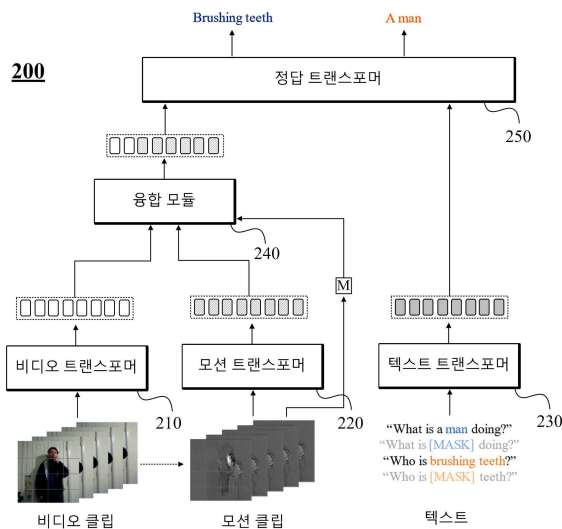
전체 청구항 수 : 총 8 항

(54) 발명의 명칭 모션 정보를 이용한 시각-언어 멀티모달 사전 학습 방법 및 비디오 텍스트 리트리벌을 수행하는 데이터 처리 장치

(57) 요약

모션 정보를 이용한 시각-언어 멀티모달 사전 학습 방법은 데이터 처리 장치가 비디오-텍스트 쌍을 입력받는 단계, 상기 데이터 처리 장치가 상기 비디오를 비디오 트랜스포머에 입력하여 비디오 토큰을 생성하는 단계, 상기 데이터 처리 장치가 상기 비디오에서 추출한 모션 정보를 모션 트랜스포머에 입력하여 모션 토큰을 생성하는 단계, 상기 데이터 처리 장치가 특정 단어가 마스킹된 상기 텍스트를 텍스트 트랜스포머에 입력하여 텍스트 토큰을 생성하는 단계, 상기 데이터 처리 장치가 상기 모션 정보에서 연산한 모션 가중치를 기준으로 상기 비디오 토큰과 상기 모션 토큰을 융합한 융합 특징을 생성하는 단계 및 상기 데이터 처리 장치가 상기 융합 특징 및 상기 텍스트 토큰을 정답 트랜스포머에 입력하여 상기 정답 트랜스포머가 상기 특정 단어를 예측하도록 학습하는 단계를 포함한다.

대표도 - 도2



(52) CPC특허분류

G06F 16/7328 (2019.01)
 G06F 16/7343 (2019.01)
 G06F 16/783 (2019.01)
 G06F 40/284 (2020.01)
 G06F 40/284 (2020.01)
 G06N 3/0455 (2023.01)
 G06N 3/0499 (2023.01)
 G06N 3/09 (2023.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193328
과제번호	2020-0-01460-004
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성
연구과제명	Simulated Reality 콘텐츠 구현을 위한 핵심기술 연구및 인력양성
기 여 율	1/2
과제수행기관명	이화여자대학교산학협력단
연구기간	2023.01.01 ~ 2023.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711190368
과제번호	2022R1A2C4002052
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)
연구과제명	거대 공간을 표현하는 대용량 몰입형 비디오의 압축도메인 학습 방법과 입체 공간
피처 레벨 신호처리에 관한	핵심 기술 개발
기 여 율	1/2
과제수행기관명	이화여자대학교
연구기간	2023.03.01 ~ 2024.02.29

명세서

청구범위

청구항 1

데이터 처리 장치가 비디오-텍스트 쌍을 입력받는 단계;

상기 데이터 처리 장치가 상기 비디오를 비디오 트랜스포머에 입력하여 비디오 토큰을 생성하는 단계;

상기 데이터 처리 장치가 상기 비디오에서 추출한 모션 정보를 모션 트랜스포머에 입력하여 모션 토큰을 생성하는 단계;

상기 데이터 처리 장치가 특정 단어가 마스킹된 상기 텍스트를 텍스트 트랜스포머에 입력하여 텍스트 토큰을 생성하는 단계;

상기 데이터 처리 장치가 상기 모션 정보에서 연산한 모션 가중치를 기준으로 상기 비디오 토큰과 상기 모션 토큰을 융합한 융합 특징을 생성하는 단계; 및

상기 데이터 처리 장치가 상기 융합 특징 및 상기 텍스트 토큰을 정답 트랜스포머에 입력하여 상기 정답 트랜스포머가 상기 특정 단어를 예측하도록 학습하는 단계를 포함하는 모션 정보를 이용한 시각-언어 멀티모달 사전 학습 방법.

청구항 2

제1항에 있어서,

상기 데이터 처리 장치는 상기 비디오를 옵티컬 플로우 모델에 입력하여 상기 모션 정보를 생성하는 모션 정보를 이용한 시각-언어 멀티모달 사전 학습 방법.

청구항 3

제1항에 있어서,

상기 데이터 처리 장치는 상기 모션 정보의 수직방향 벡터값과 수평방향 벡터값 각각에 소벨 필터를 적용하여 상기 모션 가중치를 생성하는 모션 정보를 이용한 시각-언어 멀티모달 사전 학습 방법.

청구항 4

제1항에 있어서,

상기 데이터 처리 장치는 상기 모션 가중치 기준으로 가중된 상기 모션 토큰과 상기 비디오 토큰을 결합(concatenation)한 값을 FFN(Feed-Forward Network)에 입력하여 상기 융합 특징을 생성하는 모션 정보를 이용한 시각-언어 멀티모달 사전 학습 방법.

청구항 5

제1항에 있어서,

상기 데이터 처리 장치는 상기 텍스트에서 질문의 유형을 기준으로 마스킹되는 상기 특정 단어를 결정하되,

상기 데이터 처리 장치는 상기 질문의 유형이 What 질의인 경우 상기 텍스트 중 명사를 상기 특정 단어로 결정하고, 상기 질문의 유형이 Who 질의인 경우 상기 텍스트 중 동사를 상기 특정 단어로 결정하는 모션 정보를 이용한 시각-언어 멀티모달 사전 학습 방법.

청구항 6

텍스트 쿼리 또는 비디오 쿼리를 입력받는 입력장치;

비디오 데이터베이스 또는 텍스트 데이터베이스를 저장하고, 상기 제1항 내지 상기 제5항 중 어느 한 항에 따른 시각-언어 멀티모달 사전 학습 방법으로 학습된 모델을 이용하여 구축된 비디오 텍스트 리트리벌 모델을 저장하

는 저장장치; 및

(i) 상기 입력장치가 상기 텍스트 쿼리를 입력받는 경우 상기 텍스트 쿼리를 상기 비디오 텍스트 리트리벌 모델의 텍스트 트랜스포머에 입력하여 텍스트 토큰을 생성하고, 상기 비디오 데이터베이스에 저장된 비디오들에 대한 융합 특징을 생성하고, 상기 비디오들 중 융합 특징이 상기 텍스트 쿼리의 텍스트 토큰과 가장 유사한 비디오를 결정하고, (ii) 상기 입력장치가 상기 비디오 쿼리를 입력받는 경우 상기 비디오 쿼리에 대한 융합 특징을 생성하고, 상기 텍스트 데이터베이스에 저장된 텍스트들 각각에 대한 텍스트 토큰을 생성하고, 상기 텍스트들 중 텍스트 토큰이 상기 비디오 쿼리의 융합 특징과 가장 유사한 텍스트를 결정하는 연산장치를 포함하는 비디오 텍스트 리트리벌을 수행하는 데이터 처리 장치.

청구항 7

제6항에 있어서,

상기 비디오 텍스트 리트리벌 모델은

대조 학습 과정에서 (i) 비디오-텍스트 쌍을 입력받고, (ii) 상기 비디오-텍스트 쌍 중 상기 비디오를 입력데이터로 삼고, 상기 사전 학습 방법에서 구축된 상기 비디오 트랜스포머 및 상기 모션 트랜스포머를 이용하여 생성되는 융합 특징과 (iii) 상기 비디오-텍스트 쌍 중 상기 텍스트를 상기 사전 학습 방법에서 구축된 상기 텍스트 트랜스포머에 입력하여 생성되는 텍스트 토큰이 임베딩 공간에서 일정 정도로근접하도록 대조 학습이 수행된 모델인 비디오 텍스트 리트리벌을 수행하는 데이터 처리 장치.

청구항 8

제7항에 있어서,

상기 비디오 텍스트 리트리벌 모델은

상기 대조 학습 과정에서 생성되는 비디오 토큰이 비디오 메모리뱅크에 저장되고, 모션 토큰이 모션 메모리뱅크에 저장되고, 텍스트 토큰이 텍스트 메모리뱅크에 저장되고,

상기 비디오 메모리뱅크, 상기 모션 메모리뱅크 및 상기 텍스트 메모리뱅크에 저장된 벡터들을 대조 샘플로 삼아 상기 대조 학습을 수행한 모델인 비디오 텍스트 리트리벌을 수행하는 데이터 처리 장치.

발명의 설명

기술 분야

[0001] 이하 설명하는 기술은 비디오 텍스트 리트리벌과 같은 애플리케이션 개발을 위한 시각-언어 멀티모달 사전 학습 방법에 관한 것이다.

배경 기술

[0002] 비디오 텍스트 리트리벌(retrieval)은 일반적으로 주어진 텍스트와 관련성이 높은 비디오를 데이터베이스에서 찾는 기법을 말한다. 한편, 비디오 텍스트 리트리벌은 비디오 쿼리에 대하여 관련성 높은 텍스트를 찾는 기법도 포함할 수 있다.

[0003] 딥러닝 기반 비디오 텍스트 리트리벌은 비디오 인코더(encoder)와 텍스트 인코더를 사용하여 산출되는 특징의 유사도를 기준으로 관련성 높은 비디오를 찾는다.

[0004] 비디오는 이미지와 달리 공간과 함께 시간적 특징을 갖는다. 비디오는 시간에 따라 다양한 내용(장면)을 포함할 수 있기에, 딥러닝 기반 모델이 텍스트와 매칭되는 표현을 생성하는데 한계가 있다.

선행기술문헌

비특허문헌

[0005] (비특허문헌 0001) Xiaohan Wang et al., T2VLAD: Global-Local Sequence Alignment for Text-Video Retrieval, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5079-5088.

발명의 내용

해결하려는 과제

- [0006] 비디오의 시간적 특징을 반영하기 위하여, 종래 기법은 공간 및 시간에서 주요 영역을 검출하기 위한 어텐션(attention)을 사용하였다. 또는, 종래 기법은 비디오 인코더와 텍스트 인코더를 사용하여 추출한 특징을 공동 임베딩(embedding) 공간에서 학습하는 구성을 사용하였다. 다만, 종래 기법은 비디오와 텍스트 사이의 정보 교환 없이 각 모달 내의 정보에 집중하여 여전히 표현 생성에 한계를 갖는다.
- [0007] 나아가, 최근 연구는 대량의 데이터에서 추출한 이미지와 텍스트 쌍에 대한 사전 학습을 진행하여 성능 향상을 보였다. 그러나, 이와 같은 사전 학습 모델은 셀프 어텐션을 통한 이미지 공간 정보에 집중하기 때문에 시간적 분석이 필요한 비디오 텍스트 리트리벌에 적합하지 않다.
- [0008] 이하 설명하는 기술은 비디오의 시간적 특징을 반영하기 위하여 모션 특징을 별도로 이용하는 사전 학습 기법을 제공하고자 한다. 이하 설명하는 기술은 비디오 특징과 모션 특징을 합성(fusion)하고, 시간적으로 주목할 영역을 마스킹하여 비디오의 시간적 특징을 충분히 반영한 사전 학습 기법을 제공하고자 한다.

과제의 해결 수단

- [0009] 모션 정보를 이용한 시각-언어 멀티모달 사전 학습 방법은 데이터 처리 장치가 비디오-텍스트 쌍을 입력받는 단계, 상기 데이터 처리 장치가 상기 비디오를 비디오 트랜스포머에 입력하여 비디오 토큰을 생성하는 단계, 상기 데이터 처리 장치가 상기 비디오에서 추출한 모션 정보를 모션 트랜스포머에 입력하여 모션 토큰을 생성하는 단계, 상기 데이터 처리 장치가 특정 단어가 마스킹된 상기 텍스트를 텍스트 트랜스포머에 입력하여 텍스트 토큰을 생성하는 단계, 상기 데이터 처리 장치가 상기 모션 정보에서 연산한 모션 가중치를 기준으로 상기 비디오 토큰과 상기 모션 토큰을 융합한 융합 특징을 생성하는 단계 및 상기 데이터 처리 장치가 상기 융합 특징 및 상기 텍스트 토큰을 정답 트랜스포머에 입력하여 상기 정답 트랜스포머가 상기 특정 단어를 예측하도록 학습하는 단계를 포함한다.

발명의 효과

- [0010] 이하 설명하는 기술은 사전 학습 과정에서도 시간적 특징을 반영한 모델을 구축하여 성능이 높은 비디오 텍스트 리트리벌 모델 등을 제공할 수 있다.

도면의 간단한 설명

- [0011] 도 1은 비디오 텍스트 리트리벌 모델을 구축하는 과정의 예이다.
- 도 2는 모션 정보를 이용하는 사전 학습 모델의 예이다.
- 도 3은 모션 특징과 외관 특징을 합성하는 합성 모듈의 예이다.
- 도 4는 사전 학습의 워크 플로우의 예이다.
- 도 5는 사전 학습된 트랜스포머를 이용하여 비디오 텍스트 리트리벌 모델을 구축하는 과정의 예이다.
- 도 6은 비디오 텍스트 리트리벌을 수행하는 데이터 처리 장치의 예이다.

발명을 실시하기 위한 구체적인 내용

- [0012] 이하 설명하는 기술은 다양한 변형을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세하게 설명하고자 한다. 그러나, 이는 이하 설명하는 기술을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 이하 설명하는 기술의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.
- [0013] 제1, 제2, A, B 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 해당 구성요소들은 상기 용어들에 의해 한정되지는 않으며, 단지 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다. 예를 들어, 이하 설명하는 기술의 권리 범위를 벗어나지 않으면서 제1 구성요소는 제2 구성요소로 명명될 수

있고, 유사하게 제2 구성요소도 제1 구성요소로 명명될 수 있다. 및/또는 이라는 용어는 복수의 관련된 기재된 항목들의 조합 또는 복수의 관련된 기재된 항목들 중의 어느 항목을 포함한다.

- [0014] 본 명세서에서 사용되는 용어에서 단수의 표현은 문맥상 명백하게 다르게 해석되지 않는 한 복수의 표현을 포함하는 것으로 이해되어야 하고, "포함한다" 등의 용어는 설명된 특징, 개수, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함을 의미하는 것이지, 하나 또는 그 이상의 다른 특징들이나 개수, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 배제하지 않는 것으로 이해되어야 한다.
- [0015] 도면에 대한 상세한 설명을 하기에 앞서, 본 명세서에서의 구성부들에 대한 구분은 각 구성부가 담당하는 주기능 별로 구분한 것에 불과함을 명확히 하고자 한다. 즉, 이하에서 설명할 2개 이상의 구성부가 하나의 구성부로 합쳐지거나 또는 하나의 구성부가 보다 세분화된 기능별로 2개 이상으로 분화되어 구비될 수도 있다. 그리고 이하에서 설명할 구성부 각각은 자신이 담당하는 주기능 이외에도 다른 구성부가 담당하는 기능 중 일부 또는 전부의 기능을 추가적으로 수행할 수도 있으며, 구성부 각각이 담당하는 주기능 중 일부 기능이 다른 구성부에 의해 전담되어 수행될 수도 있음은 물론이다.
- [0016] 또, 방법 또는 동작 방법을 수행함에 있어서, 상기 방법을 이루는 각 과정들은 문맥상 명백하게 특정 순서를 기재하지 않은 이상 명기된 순서와 다르게 일어날 수 있다. 즉, 각 과정들은 명기된 순서와 동일하게 일어날 수도 있고 실질적으로 동시에 수행될 수도 있으며 반대의 순서대로 수행될 수도 있다.
- [0018] 이하 설명하는 기술은 비디오 텍스트 리트리벌 기법이다. 이하 설명하는 기술은 비디오 텍스트 리트리벌을 위한 모델 구축 기법이다.
- [0019] 이하 설명하는 기술은 비디오와 텍스트의 특징을 추출하여 관련된 정보를 산출한다. 이하 설명하는 기술은 비전 트랜스포머(vision transformer) 기반의 모델을 사용한다. 비전 트랜스포머 기반의 모델은 비디오 텍스트 리트리벌 모델이라고 명명할 수도 있다.
- [0020] 이하 데이터 처리 장치가 비디오 텍스트 리트리벌 모델을 사용하여 비디오 텍스트 리트리벌을 수행한다고 설명한다. 데이터 처리 장치는 PC, 스마트기기, 네트워크의 서버, 데이터 처리 전용 칩셋 등의 형태를 가질 수 있다.
- [0021] 한편, 후술한 비디오 텍스트 리트리벌 모델은 사전 학습 등을 이용하여 사전에 구축되어야 한다. 해당 모델의 학습은 별도의 학습장치가 수행할 수 있다. 학습장치는 데이터 처리 및 모델 학습이 가능한 컴퓨팅 장치를 의미한다. 다만, 설명의 편의를 위하여 이하 설명에서는 데이터 처리 장치가 학습 과정도 수행한다고 설명한다.
- [0023] 비디오는 다양한 장면(다수의 프레임들 내지 프레임 구간)을 포함하며, 해당 장면에 따라 시간 정보의 중요성이 다르다. 예컨대, 움직임이 드문 비디오는 시간 중요도가 낮기 때문에 모션정보 보다 일반적인 외관정보가 중요하다. 그러나 움직이는 객체를 담은 비디오는 모션정보가 외관정보보다 더욱 중요하다고 할 수 있다. 한편, 시간 중요도의 차이는 질의응답을 통해 드러날 수 있다. 인간은 비디오에서 파악할 수 있는 중요한 정보를 질의응답을 통해 표현할 수 있다. 예컨대, 객체의 움직임이 두드러진 비디오에 대한 질의응답은 해당 움직임에 대해 질문하는 경향이 강하다. 결국, 비디오 텍스트 리트리벌은 비디오의 내용에 따라 모션 또는 외관에 집중한 표현을 사용하는 것이 바람직하고 할 수 있다.
- [0024] 이하 설명하는 기술은 모션 정보를 활용하되 비디오의 내용에 따라 모션 정보 또는 외관 정보에 가중을 두는 합성 모듈을 사용한다. 합성 모듈은 영상의 특징에 따른 가중을 달리하여 모션 정보 또는 외관 정보에 특화된 특징을 출력한다.
- [0025] 또한, 이하 설명하는 기술은 질의응답 데이터를 활용하여 사전 학습을 수행하되, 네트워크는 who 및/또는 what 질문을 통해 비디오의 시각적 요소에서 집중해야 할 부분을 사전 학습한다. 나아가, 이하 설명하는 기술은 사전 학습된 트랜스포머 기반한 비디오 텍스트 리트리벌 모델 학습에 효율적인 메모리 बैं크와 이를 이용한 대조 학습(contrastive learning)을 제안한다.
- [0027] 도 1은 비디오 텍스트 리트리벌 모델을 구축하는 과정(100)의 예이다.

- [0028] 데이터 처리 장치는 학습 데이터를 획득한다(110). 이때 학습 데이터는 사전 학습을 위한 비디오-질의응답 세트 및 비디오 텍스트 리트리벌 모델 학습을 위한 비디오-텍스트 세트를 포함할 수 있다. 비디오-질의응답 세트는 다수의 비디오와 각 비디오에 대한 질의응답 쌍들로 구성된다. 질의응답은 사전 학습을 위한 텍스트 데이터로 질의 및 질의에 대한 정답 쌍으로 구성된다. 학습 데이터는 관련 데이터를 저장하는 데이터베이스에서 추출될 수 있다. 또는 학습 데이터는 개발자가 별도로 구축할 수도 있다.
- [0029] 데이터 처리 장치는 비전트랜스포머 기반의 모델을 이용하여 사전 학습을 수행한다(120). 데이터 처리 장치는 비디오-텍스트 뿐만 아니라, 비디오에서 모션 정보를 더 사용하여 사전 학습을 수행한다. 데이터 처리 장치는 비디오 트랜스포머, 모션 트랜스포머 및 텍스트 트랜스포머를 이용하여 사전 학습을 수행할 수 있다.
- [0030] 데이터 처리 장치는 옵티컬 플로우(optical flow)와 같은 모델을 이용하여 비디오에서 모션 정보를 추출할 수 있다. 데이터 처리 장치는 현재 비디오의 모션 정도에 따라 모션 특징 또는 일반 비디오 특징(외관 특징)에 가중을 두어 융합 특징을 생성한다. 즉, 데이터 처리 장치는 비디오 내용에 따라 모션 정보에 가중값을 부여하여 모션 특징과 비디오 특징을 융합하여 융합 특징을 생성한다. 데이터 처리 장치는 모션 정도에 따라 모션 특징과 비디오 특징의 비율을 달리하여 융합 특징을 생성한다. 따라서, 데이터 처리 장치는 비디오 내용에 따른 적응적인 모션 정보를 이용하여 사전 학습을 수행한다고 할 수 있다. 구체적인 사전 학습 과정은 후술한다.
- [0031] 나아가, 데이터 처리 장치는 사전 학습 과정에서 구축된 트랜스포머들을 이용하여 비디오와 텍스트를 입력 데이터로 사용하는 특정 애플리케이션 모델을 학습할 수 있다(130). 예컨대, 데이터 처리 장치는 사전 학습 과정에서 구축된 트랜스포머들을 이용하여 비디오 텍스트 리트리벌 모델을 구축할 수 있다.
- [0033] 비전-언어 사전학습은 대규모 이미지-텍스트 데이터를 통해 멀티모달 데이터의 일반적인 공동 표현을 학습하는 것을 목적으로 한다. 도 2는 모션 정보를 이용하는 사전 학습 모델(200)의 예이다. 사전 학습 모델(200)은 비디오 트랜스포머(210), 모션 트랜스포머(220), 텍스트 트랜스포머(230), 융합 모듈(240) 및 정답 트랜스포머(250)를 포함한다. 학습 데이터는 비디오-질의응답 세트를 포함한다. 질의응답은 해당 비디오에 대하여 사전에 마련된 질의 및 그에 대한 응답(정답)으로 구성된다.
- [0034] 비디오 트랜스포머(210) 및 모션 트랜스포머(220)는 비전트랜스포머 모델이다. 비전트랜스포머는 자연어 처리 분야에서 소개된 어텐션 기반 트랜스포머 모델을 컴퓨터 비전 분야에 적용한 모델이다. 텍스트 트랜스포머(230)는 전통적으로 텍스트의 단어 위치 정보를 인코딩하는 어텐션 기반의 트랜스포머 모델이다. 트랜스포머의 구조에 대한 상세한 설명은 생략한다.
- [0035] 비디오 트랜스포머(210)는 비디오 클립을 입력받아 비디오 벡터(비디오 토큰)를 출력한다. 비디오는 여러 내용을 포함하기 때문에, 데이터 처리 장치는 시간 영역에서 샘플링 된 N개의 클립 중 질의와 유사도가 가장 높은 클립을 선택하여 인코딩을 진행할 수 있다.
- [0036] 비디오 트랜스포머(210)는 다양한 비전트랜스포머 중 어느 하나로 구현될 수 있다. 예컨대, 비디오 트랜스포머(210)는 ViT 구조의 모델로 구현될 수 있다. 비디오 클립 $V \in \mathbb{R}^{M \times 3 \times H \times W}$ 가 비디오 트랜스포머(210)입력으로 주어질 수 있다. 이때 M은 프레임 개수, $H \times W$ 는 비디오의 해상도이다. 입력된 프레임은 $M \times N$ 패치로 나누어지며, 각각의 패치는 $P \times P$ 의 해상도를 지닌다. 이때 $N = HW/P$ 이다. 비디오 트랜스포머(210)는 비디오 패치 $v \in \mathbb{R}^{M \times N \times P \times P}$ 에 선형 프로젝션 레이어를 적용하여 비디오 토큰 $z_v \in \mathbb{R}^{M \times N \times D}$ 을 생성한다.
- [0037] 모션 트랜스포머(220)는 비디오 트랜스포머(210)에 입력되는 비디오 클립에 대응하는 모션 클립을 입력받는다. 모션 트랜스포머(220)는 모션 클립을 입력받아 모션 벡터(모션 토큰)를 출력한다. 모션 클립은 다양한 방법을 통해 비디오 클립으로부터 획득될 수 있다. 예컨대, 데이터 처리 장치는 비디오 클립을 옵티컬플로우 예측 모델에 입력하여 모션 클립을 생성할 수 있다. 옵티컬플로우 예측 모델은 RAFT(Zachary Teed, Jia Deng, RAFT: Recurrent All-Pairs Field Transforms for Optical Flow, 16th European Conference on Computer Vision, ECCV 2020, pages 402-419, 2020.참고)를 이용할 수 있다.
- [0038] 모션 트랜스포머(220)도 다양한 비전트랜스포머 중 어느 하나로 구현될 수 있다. 예컨대, 모션 트랜스포머(220)는 ViT 구조의 모델로 구현될 수 있다. 모션 클립 $V_m \in \mathbb{R}^{M \times 2 \times H \times W}$ 가 모션 트랜스포머(220)입력으로 주어질 수 있다. 모션 클립은 옵티컬플로우로 획득한 모션맵으로 구성될 수 있다. $V_m = \{m_{12}, m_{23}, m_{34}, m_{45}, m_{56}, \dots\}$ 이다.

이때, m_{ij} 는 비디오 프레임 f_i 와 f_j 로부터 얻은 옵티컬플로우를 통해 재구성된 모션맵이다. 옵티컬플로우는 수직 및 수평 방향으로 추출된다. 수직방향 벡터 값과 수평방향 벡터 값을 x, y 라고 할 때, 데이터 처리 장치는 모션 정보를 강화하기 위해 (x, y) 에 소벨 필터(Sobel filter)를 적용하여 G_x, G_y 를 추출한다. 최종적으로 m_{ij} 는 G_x, G_y 의 픽셀단위 덧셈을 통해 얻어져 모션 엠티지가 강조될 수 있다.

[0039] 모션 클립 V_m 은 비디오 인코더와 같이 $M \times N$ 패치로 나누어지며, 모션 트랜스포머(220)는 모션 배치에 선형 프로젝션 레이어를 적용하여 모션 토큰 $z_m \in \mathbb{R}^{M \times N \times D}$ 을 생성한다.

[0040] 텍스트 트랜스포머(230)는 다양한 트랜스포머 모델 중 어느 하나로 구현될 수 있다. 예컨대, 텍스트 트랜스포머(230)는 DistilBERT 등과 같은 모델로 구현될 수 있다. 텍스트 트랜스포머(230)는 비디오 클립에 대한 질문 텍스트이다. 텍스트 트랜스포머(230)는 입력되는 질문 텍스트를 토큰화하여 토큰 y 를 생성한다.

[0041] 융합 모듈(240)은 비디오 트랜스포머(210)과 모션 트랜스포머(220)가 각각 출력하는 비디오 토큰과 모션 토큰을 일정한 비율로 융합하여 융합 특징을 생성한다. 융합 모듈(240)은 비디오의 내용에서 모션 정도라는 특징에 따라 비율을 달리하여 비디오 토큰과 모션 토큰을 합성한다.

[0042] 도 3은 모션 특징과 외관 특징을 합성하는 합성 모듈의 예이다. 데이터 처리 장치는 모션 클립(옵티컬플로우)을 이용하여 모션 가중치 M 을 산출할 수 있다. 모션 가중치 M 은 입력 비디오의 특징을 반영하기 위해 출력한 스칼라 값이다. 모션 가중치 M 은 아래 수학식 1과 같이 정의할 수 있다.

수학식 1

[0044]
$$M = \sqrt{\text{sum}(G_u * u)^2 + \text{sum}(G_v * u)^2}$$

[0045] u 와 v 는 각각 옵티컬플로우의 수직 방향 벡터값 및 수평방향 벡터 값이다. G_u 와 G_v 는 각각 수직방향 소벨 필터 및 수평방향 소벨 필터이다.

[0046] 모션 가중치 M 은 0 ~ 1사이의 값으로 정규화될 수 있다. 모션 가중치 M 은 합성 특징 생성 시 비디오 토큰 z_v 과 모션 토큰 z_m 의 비율을 결정한다. 예컨대, 융합 모듈(240)은 모션 토큰 z_m 을 M 만큼의 차원 축소하고, 비디오 토큰 z_v 을 $1-M$ 만큼의 차원 축소를 할 수 있다.

[0047] 융합 모듈(240)은 서로 다른 가중치를 차원 축소한 모션 토큰 z_m 과 차원 축소한 비디오 토큰 z_v 에 부가하여 결합(concatenation)할 수 있다. 융합 모듈(240)은 가중치 w_v 를 곱한 비디오 토큰 z_v 과 가중치 w_m 를 곱한 모션 토큰 z_m 을 결합($[w_v z_v, w_m z_m]$)한 토큰을 FFN(Feed-Forward Network)에 입력하여 융합 특징 v_{out} 을 출력한다. v_{out} 은 아래 수학식 2와 같이 표현할 수 있다.

수학식 2

[0049]
$$v_{out} = \text{FFN}(W_o [w_v z_v, w_m z_m])$$

[0050] 융합 특징 $v_{out} \in \mathbb{R}^{M \times N \times D}$ 이다. 가중치 w_m , w_v 및 w_o 는 학습 과정을 통해 결정될 수 있다.

[0052] 데이터 처리 장치는 입력되는 질문 텍스트 중 일부 단어를 마스킹하여 텍스트 트랜스포머(230)에 입력한다. 텍스트 트랜스포머(230)가 출력하는 텍스트 토큰 y 는 정답 트랜스포머(250)에 입력된다.

[0053] 정답 트랜스포머(250)는 텍스트 트랜스포머(230)의 출력을 쿼리(query)로 사용하고, 융합 모듈(240)의 출력인

융합 특징(융합 토큰)을 키(key)와 값(value)으로 사용하여 크로스 모달 어텐션을 수행한다. 정답 트랜스포머(250)는 멀티 헤드 어텐션을 통해 상호작용된 토큰을 출력한다. 정답 트랜스포머(250)는 이전 단계의 트랜스포머의 정답 토큰과 멀티 헤드 어텐션을 통해 산출된 상호작용된 토큰에 대한 셀프 어텐션을 통해 새로운 정답 토큰을 산출한다. 정답 토큰은 마지막 블록에서 출력된다.

- [0054] 데이터 처리 장치는 정답 트랜스포머(250)가 출력하는 정답과 알고 있는 실제 정답을 비교하면서 사전 학습 모델(200)의 파라미터를 업데이트한다. 데이터 처리 장치는 다수의 비디오-질의응답 쌍을 이용한 사전 학습 과정을 수행할 수 있다.
- [0056] 도 3에서 살펴본 바와 같이, 데이터 처리 장치는 질문 텍스트에서 일정한 영역(단어)를 마스킹하여 텍스트 트랜스포머에 입력하고 정답 트랜스포머가 정답을 출력하도록 사전 학습을 수행한다. 이때 질문 텍스트에서 마스킹되는 영역을 질문의 성질에 따라 적응적으로 결정할 수 있다. 전술한 바와 같이 질문은 비디오의 시간적 특징을 나타내는 단어를 포함할 수 있고, 데이터 처리 장치는 질문에서 시간적 특징과 관련성 높은 단어를 마스킹하여 사전 학습 과정에서 비디오의 시간적 특징이 잘 반영되게 유도할 수 있다.
- [0057] 도 4는 사전 학습의 워크 플로우의 예이다. 도 4는 질문 텍스트의 유형에 따라 마스킹되는 영역을 달리하여 사전 학습을 수행하는 예이다.
- [0058] 데이터 처리 장치는 비디오-질의응답 쌍을 입력받는다.
- [0059] 데이터 처리 장치는 전술한 바와 같이 비디오 데이터를 각각 모션 트랜스포머 및 비디오 트랜스포머에 입력하여 융합 특징을 생성한다.
- [0060] 데이터 처리 장치는 질의응답에서 질문의 유형을 분류한다. 질문의 유형은 What 질의(주체 또는 객체의 행동에 대한 질의)와 Who 질의(주체에 대한 질의)로 구분할 수 있다. (1) 데이터 처리 장치는 What 질의에 대해서는 명사를 마스킹하여 텍스트 트랜스포머에 입력할 수 있다. 예컨대, "Who is cooking?" 이라는 질의에 대하여, 데이터 처리 장치는 동사 cooking을 [MASK] 토큰으로 변경하고, 모델은 [MASK]토큰과 질의에 대한 답변을 예측하도록 학습된다.
- [0061] (2) 데이터 처리 장치는 Who 질의에 대해서는 동사를 마스킹하여 텍스트 트랜스포머에 입력할 수 있다. 예컨대, "What is a man doing?" 이라는 질의에 대하여, 데이터 처리 장치는 명사 man을 [MASK] 토큰으로 변경하고, 모델은 [MASK]토큰과 질의에 대한 답변을 예측하도록 학습된다.
- [0062] 한편, 데이터 처리 장치는 랜덤하게 질의의 마스킹 여부를 결정하고, 마스킹 되지 않은 질의는 입력된 쿼리와 함께 정답 트랜스포머를 거쳐 답변 예측하도록 이용한다.
- [0064] 도 2 내지 도 3은 시각-언어 멀티모달 데이터에 대한 사전 학습에 해당한다. 데이터 처리 장치는 사전 학습된 트랜스포머들을 이용하여 비디오 및/또는 텍스트를 입력 데이터로 사용하는 다양한 애플리케이션 모델을 구축할 수 있다. 예컨대, 데이터 처리 장치는 비디오에서 일정 시간 구간에 대한 예측 태스크, 비디오에서 특정 객체의 행동 분류, 텍스트 쿼리를 영상에 표시(video grounding) 등에 전술한 사전 학습 모델을 활용할 수 있다.
- [0066] 다만, 이하 데이터 처리 장치가 사전 학습된 트랜스포머들을 이용하여 비디오 텍스트 리트리벌 모델을 구축하는 예를 중심으로 설명한다.
- [0067] 도 5는 사전 학습된 트랜스포머를 이용하여 비디오 텍스트 리트리벌 모델을 구축하는 과정(300)의 예이다. 도 5에서 모션 트랜스포머, 비디오 트랜스포머 및 텍스트 트랜스포머는 모두 도 2의 사전 학습을 수행한 인코더이다.
- [0068] 데이터 처리 장치는 비디오 텍스트 리트리벌 학습을 위한 비디오-텍스트 쌍을 입력받는다(310).
- [0069] 데이터 처리 장치는 RAFT와 같은 옵티컬플로우 예측 모델을 이용하여 입력되는 비디오에 대한 옵티컬플로우를 생성한다(320).
- [0070] 데이터 처리 장치는 옵티컬플로우를 기준으로 전술한 모션 가중치 M을 연산한다(330).

- [0071] 데이터 처리 장치는 옵티컬플로우(모션 정보)를 모션 트랜스포머에 입력하여 모션 토큰을 생성한다. 이때 데이터 처리 장치는 모션 트랜스포머가 출력하는 초기 모션 토큰에 모션 가중치 M 을 곱하여 가중된 모션 토큰을 생성할 수 있다(340).
- [0072] 데이터 처리 장치는 입력 비디오를 비디오 트랜스포머에 입력하여 비디오 토큰을 생성한다(350). 이때 데이터 처리 장치는 도 3과 같이 $M-1$ 을 비디오 토큰에 곱하여 가중된 비디오 토큰을 생성할 수도 있다.
- [0073] 데이터 처리 장치는 가중된 모션 토큰과 비디오 토큰(또는 가중된 비디오 토큰)을 융합하여 진술한 융합 특징을 생성한다(360).
- [0074] 데이터 처리 장치는 입력 텍스트를 텍스트 트랜스포머에 입력하여 텍스트 토큰을 생성한다(370).
- [0075] 이후 데이터 처리 장치는 임베딩 공간에서 최종 융합 특징과 텍스트의 유사도를 계산하여 대조학습을 수행할 수 있다(380).
- [0076] 대조학습은 비지도 학습 방법 중 하나로 딕셔너리 룩업(dictionary lookup)을 통해 표현을 학습하는 방법이다. 딕셔너리는 여러 키 벡터로 구성되어 있고 이 키 벡터는 인코더에 의해 산출된 특징 값을 의미한다. 대조학습은 매우 많은 키 벡터가 존재하는 사전에서 쿼리로 들어온 데이터와의 유사도 비교를 통해 양성 샘플(positive sample)과 입력 쿼리 사이의 유사도를 높게, 입력 쿼리와 음성 샘플(negative sample) 사이의 유사도를 낮도록 표현을 생성하는 것을 목표로 한다. 최근 쿼리 메모리 बैं크를 이용한 대조학습이 제안되었다. 도 5도 메모리 बैं크를 이용한 대조학습을 예시한다.
- [0077] 데이터 처리 장치는 비디오 트랜스포머의 출력 z_v 와 모션 트랜스포머의 출력 z_m 에 모션클립으로부터 얻은 모션 가중치 M 을 적용하여 융합 특징 v_{out} 을 생성한다. 이 과정은 융합 모듈을 통해 수행될 수 있다.
- [0078] 데이터 처리 장치는 아래 수학적식 3과 같이 융합 특징 v_{out} 과 텍스트 토큰 y 에 대한 유사도 s 를 연산할 수 있다.

수학적식 3

[0080]
$$s_i = v_{iout}^T y_i$$

[0081] 수학적식 3에서 i 는 학습 데이터 세트 중 비디오-텍스트 쌍의 인덱스를 의미한다.

[0083] 도 5에서 메모리 बैं크는 비디오, 모션 및 텍스트 각각에 대하여 존재한다(비디오 메모리 बैं크, 모션 메모리 बैं크 및 텍스트 메모리 बैं크). 각 메모리 बैं크는 큐(queue) 구조로 학습 과정에서 생성되는 비디오 특징 벡터(비디오 토큰), 모션 특징 벡터(모션 토큰), 텍스트 특징 벡터(텍스트 토큰)를 저장한다.

[0084] 이후 데이터 처리 장치는 융합 특징 v_{out} 과 텍스트 토큰 y 의 유사도를 연산할 때마다 동일한 인덱스로 각 큐에서 특징 벡터를 추출하여 대조 샘플로 삼는다. 학습 과정은 같은 내용에 대한 융합 특징 v_{out} 과 텍스트 토큰 y 는 양성 샘플로 삼고, 메모리 बैं크 내에서 쿼리하여 유사도가 낮은 K 개 샘플을 음성 샘플로 삼는다. 따라서, 데이터 처리 장치는 도 5와 같이 같은 내용에 대한 융합 특징 v_{out} 과 텍스트 토큰 y 는 임베딩 공간에서 거리가 (점차) 가깝게 위치하면서(일정 정도 근접하도록), 유사도가 낮은 대조 샘플과는 각각 거리가 (점차) 멀게 위치하도록 학습을 진행한다. 구체적으로 데이터 처리 장치는 융합 특징 v_{out} 과 텍스트 메모리 बैं크에서 추출한 대조 샘플이 멀리 위치하고, 텍스트 토큰 y 와 모션 메모리 बैं크 및 비디오 메모리 बैं크에서 추출한 대조 샘플이 멀리 위치하도록 학습을 진행한다.

[0085] 도 5의 대조 학습은 유사도 학습을 진행할 때 종전 기법과 같이 배치 내에서 음성 샘플을 선택하지 않고, 메모리 बैं크에서 업데이트되는 음성 샘플을 선택한다. 이를 통해 더 일반화된 학습이 가능하다. 메모리 बैं크 큐의 크기는 $K \times D$ 이다. 비디오 텍스트 리트리벌을 위한 대조학습의 목적 함수는 아래 수학적식 4와 같이 L_{vt} (video-to-text 리트리벌) 및 L_{tv} (text-to-video 리트리벌)라고 정의할 수 있다.

수학식 4

$$L_{v2t} = -\frac{1}{B} \sum_i^B \log \frac{\exp\left(\frac{v_{i\ out}^T y_i}{\sigma}\right)}{\sum_{j=1}^{K_t} \exp\left(\frac{x_{j\ out}^T y_j}{\sigma}\right)}$$

$$L_{t2v} = -\frac{1}{B} \sum_i^B \log \frac{\exp\left(\frac{y_i^T v_{i\ out}}{\sigma}\right)}{\sum_{j=1}^{K_v} \exp\left(\frac{y_j^T v_{j\ out}}{\sigma}\right)}$$

[0087]

[0088]

수학식 4에서 $v_{i\ out}$ 및 y_j 는 각각 i 번째 융합 특징과 j 번째 텍스트를 정규화한 임베딩을 나타낸다. B 는 배치사이즈이다. K_v 및 K_t 는 각각 해당 메모리뱅크의 크기이고, σ 는 상수 파라미터이다.

[0090]

도 6은 비디오 텍스트 리트리벌을 수행하는 데이터 처리 장치(400)의 예이다. 데이터 처리 장치(400)는 사전학습된 인코더들을 이용하여 구축된 비디오 텍스트 리트리벌을 보유한다고 전제한다. 즉, 도 5의 과정을 통해 비디오 텍스트 리트리벌 모델이 학습된 상태라고 가정한다.

[0091]

데이터 처리 장치(400)는 물리적으로 다양한 형태로 구현될 수 있다. 예컨대, 데이터 처리 장치(400)는 PC와 같은 컴퓨터 장치, 네트워크의 서버, 데이터 처리 전용 칩셋 등의 형태를 가질 수 있다.

[0092]

데이터 처리 장치(400)는 저장장치(410), 메모리(420), 연산장치(430), 인터페이스 장치(440), 통신장치(450) 및 출력장치(460)를 포함할 수 있다.

[0093]

저장장치(410)는 학습된 비디오 텍스트 리트리벌 모델을 저장한다. 비디오 텍스트 리트리벌 모델은 도 5의 과정으로 학습된 모션 트랜스포머, 비디오 트랜스포머 및 텍스트 트랜스포머를 포함한다.

[0094]

저장장치(410)는 비디오에서 모션 정보를 추출하는 옵티컬플로우 모델을 저장할 수 있다.

[0095]

저장장치(410)는 다양한 입력 텍스트 쿼리들에 대응하는 비디오들을 저장할 수 있다. 즉, 저장장치(410)는 대량의 비디오 데이터베이스에 해당할 수 있다.

[0096]

저장장치(410)는 다양한 비디오 쿼리에 대응하는 텍스트들을 저장할 수 있다. 이 경우, 저장장치(410)는 대량의 텍스트 데이터베이스에 해당할 수 있다.

[0097]

메모리(420)는 비디오 텍스트 리트리벌 과정에서 생성되는 데이터 및 정보 등을 저장할 수 있다.

[0098]

인터페이스 장치(440)는 외부로부터 일정한 명령 및 데이터를 입력받는 장치이다.

[0099]

인터페이스 장치(440)는 입력 데이터인 텍스트 또는 비디오를 입력받일 수 있다.

[0100]

인터페이스 장치(440)는 리트리벌 결과인 비디오 또는 텍스트를 다른 객체에 전달할 수 있다.

[0101]

통신장치(450)는 유선 또는 무선 네트워크를 통해 일정한 정보를 수신하고 전송하는 구성을 의미한다.

[0102]

통신장치(450)는 입력 데이터인 텍스트 또는 비디오를 수신할 수 있다.

[0103]

통신장치(450)는 리트리벌 결과인 비디오 또는 텍스트를 사용자 단말 또는 서비스 서버와 같은 외부 객체에 송신할 수도 있다.

[0104]

인터페이스 장치(440)는 통신장치(450)를 통해 수신된 정보를 내부로 전달하는 구성일 수 있다.

[0105]

텍스트를 입력하여 비디오를 검색하는 서비스(text-to-video)를 설명한다. 연산 장치(430)는 입력 데이터인 텍스트 쿼리를 텍스트 트랜스포머에 입력하여 텍스트 토큰을 생성한다. 연산 장치(430)는 저장장치(410)에 저장된 비디오 데이터베이스에 속한 비디오 각각에 대하여 (i) 비디오 트랜스포머를 이용하여 비디오 토큰을 생성하고, (ii) 비디오에 옵티컬플로우 모델을 적용하여 모션 정보를 생성하고, (iii) 모션 정보에서 모션 가중치 M 을 연

산하고, (iv) 모션 정보를 모션 트랜스포머에 입력하여 모션 토큰을 생성하고, (v) 각 비디오에 대한 모션 가중치를 이용하여 비디오 토큰과 모션 토큰을 융합하여 융합 특징을 생성한다. 이 과정 자체는 도 2, 도 3 내지 도 5에서 설명한 바와 같다. 연산 장치(430)는 융합 특징과 텍스트 토큰의 유사도를 연산하고, 비디오 데이터베이스에서 입력 텍스트 쿼리에 가장 유사한 비디오를 검색 결과로 결정할 수 있다.

[0106] 비디오를 입력하여 연관된 텍스트 검색하는 서비스(video-to-text)를 설명한다. 연산 장치(430)는 입력 데이터인 비디오에 대하여 (i) 비디오 트랜스포머를 이용하여 비디오 토큰을 생성하고, (ii) 비디오에 옵티컬플로우 모델을 적용하여 모션 정보를 생성하고, (iii) 모션 정보에서 모션 가중치 M 을 연산하고, (iv) 모션 정보를 모션 트랜스포머에 입력하여 모션 토큰을 생성하고, (v) 각 비디오에 대한 모션 가중치를 이용하여 비디오 토큰과 모션 토큰을 융합하여 융합 특징을 생성한다. 이 과정 자체는 도 2, 도 3 내지 도 5에서 설명한 바와 같다. 연산 장치(430)는 저장장치(410)에 저장된 텍스트 데이터베이스에 속한 텍스트 각각을 텍스트 트랜스포머에 입력하여 텍스트 토큰을 생성한다. 연산 장치(430)는 융합 특징과 텍스트 토큰의 유사도를 연산하고, 텍스트 데이터베이스에서 입력 비디오 쿼리에 가장 유사한(연관성 높은) 텍스트를 검색 결과로 결정할 수 있다.

[0107] 연산 장치(430)는 데이터를 처리하고, 일정한 연산을 처리하는 프로세서, AP, 프로그램이 임베디드된 칩과 같은 장치일 수 있다.

[0108] 출력장치(460)는 일정한 정보를 출력하는 장치이다. 출력장치(460)는 비디오 텍스트 리트리벌 과정에 필요한 인터페이스, 비디오 텍스트 리트리벌 결과 등을 출력할 수 있다.

[0110] 또한, 상술한 바와 같은 비디오 텍스트 리트리벌 모델 구축 방법 내지 비디오 텍스트 리트리벌 방법은 컴퓨터에서 실행될 수 있는 실행가능한 알고리즘을 포함하는 프로그램(또는 어플리케이션)으로 구현될 수 있다. 상기 프로그램은 일시적 또는 비일시적 판독 가능 매체(non-transitory computer readable medium)에 저장되어 제공될 수 있다.

[0111] 비일시적 판독 가능 매체란 레지스터, 캐쉬, 메모리 등과 같이 짧은 순간 동안 데이터를 저장하는 매체가 아니라 반영구적으로 데이터를 저장하며, 기기에 의해 판독(reading)이 가능한 매체를 의미한다. 구체적으로는, 상술한 다양한 어플리케이션 또는 프로그램들은 CD, DVD, 하드 디스크, 블루레이 디스크, USB, 메모리카드, ROM(read-only memory), PROM(programmable read only memory), EPROM(Erasable PROM, EPROM) 또는 EEPROM(Electrically EPROM) 또는 플래시 메모리 등과 같은 비일시적 판독 가능 매체에 저장되어 제공될 수 있다.

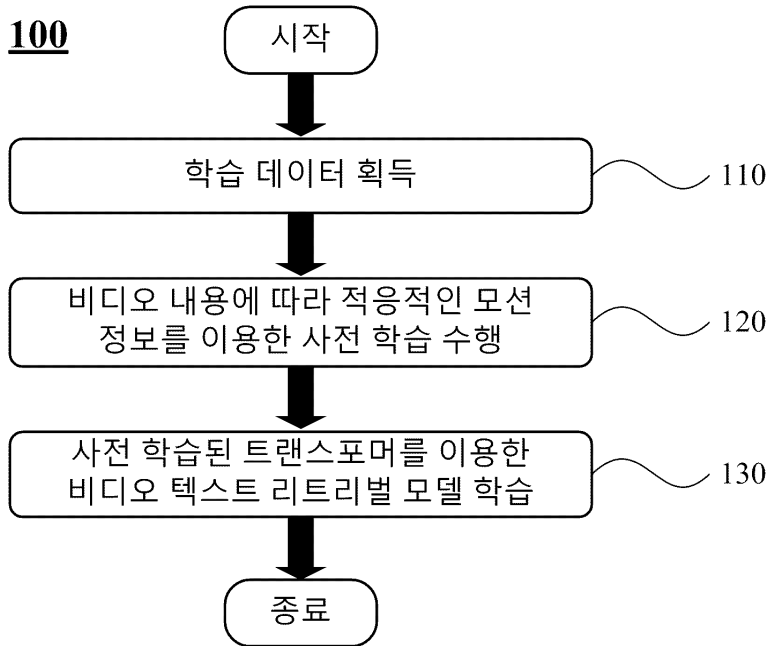
[0112] 일시적 판독 가능 매체는 스태틱 램(Static RAM, SRAM), 다이내믹 램(Dynamic RAM, DRAM), 싱크로너스 디램(Synchronous DRAM, SDRAM), 2배속 SDRAM(Double Data Rate SDRAM, DDR SDRAM), 증강형 SDRAM(Enhanced SDRAM, ESDRAM), 동기화 DRAM(SyncLink DRAM, SLDRAM) 및 직접 램버스 램(Direct Rambus RAM, DRRAM) 과 같은 다양한 RAM을 의미한다.

[0113] 본 실시예 및 본 명세서에 첨부된 도면은 전술한 기술에 포함되는 기술적 사상의 일부를 명확하게 나타내고 있는 것에 불과하며, 전술한 기술의 명세서 및 도면에 포함된 기술적 사상의 범위 내에서 당업자가 용이하게 유추할 수 있는 변형 예와 구체적인 실시례는 모두 전술한 기술의 권리범위에 포함되는 것이 자명하다고 할 것이다.

도면

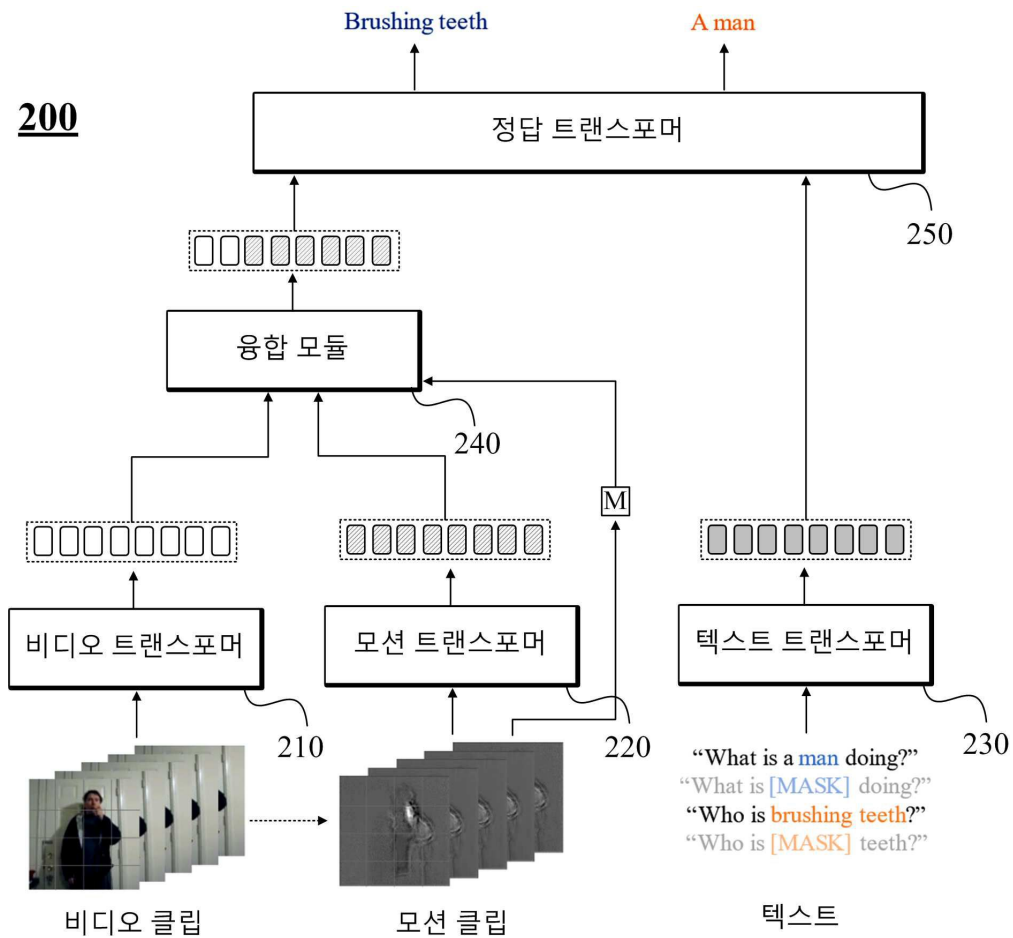
도면1

100

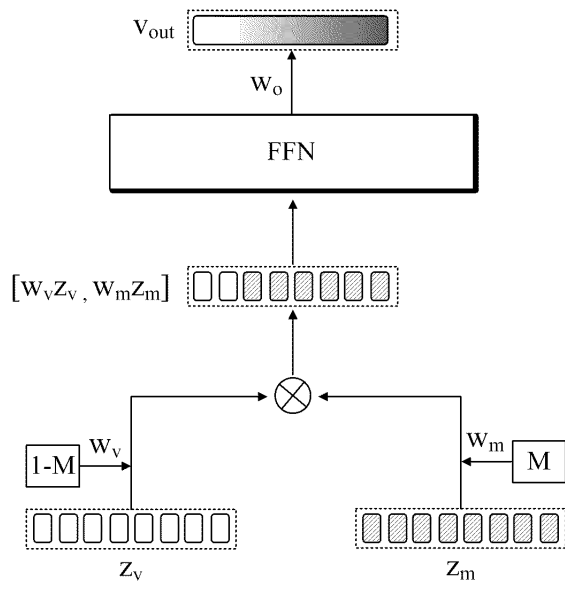


도면2

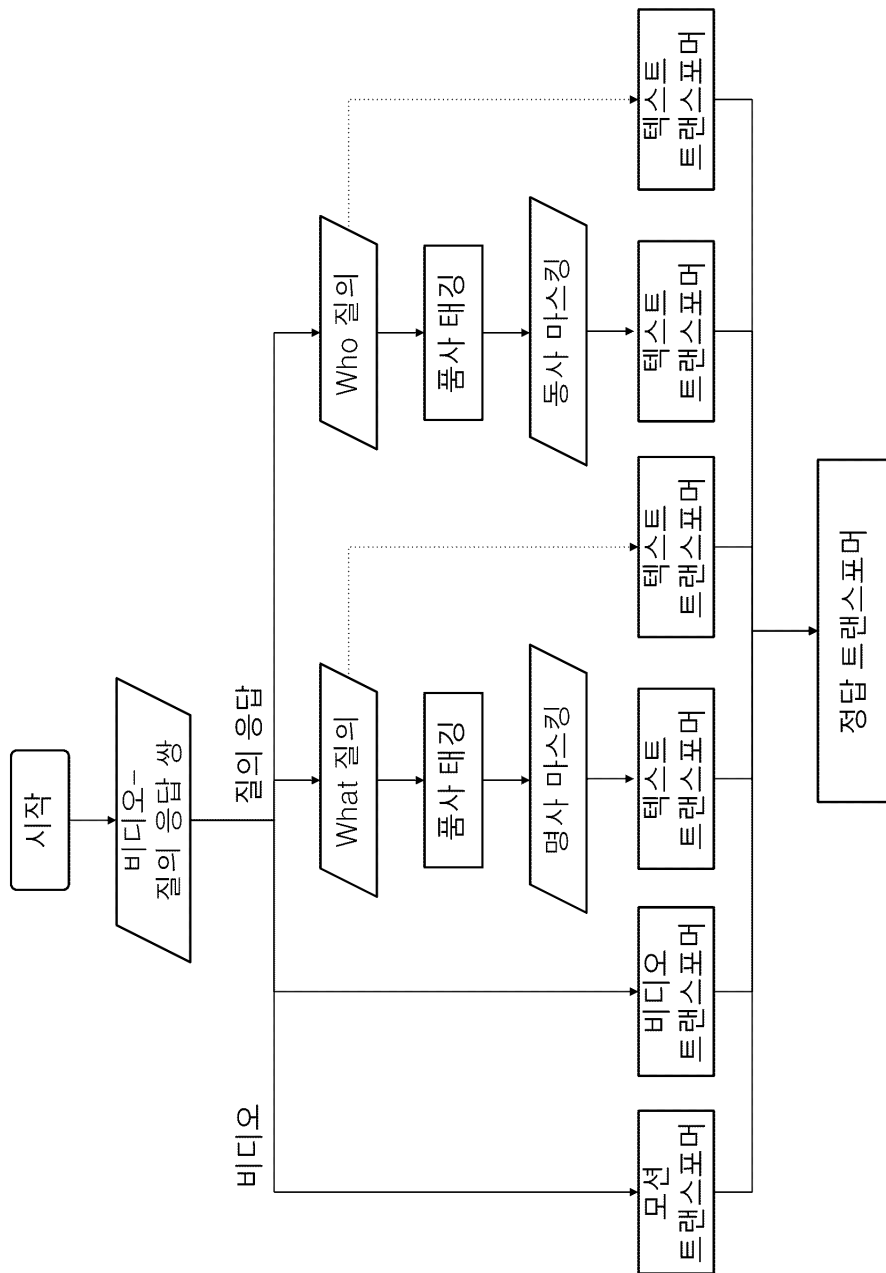
200



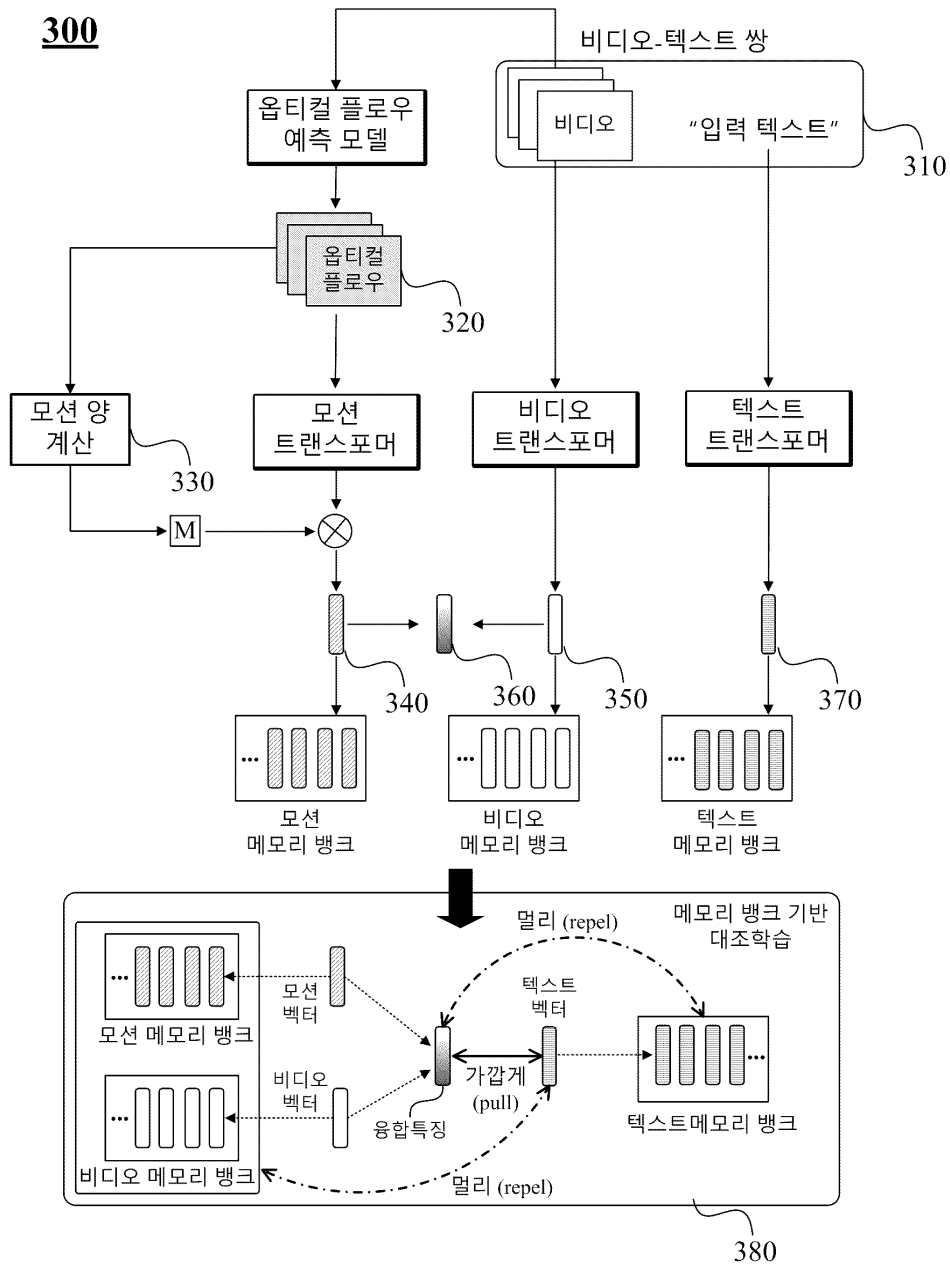
도면3



도면4



도면5



도면6

