

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)(11) 공개번호 10-2025-0070982
(43) 공개일자 2025년05월21일

(51) 국제특허분류(Int. Cl.)

G06F 16/732 (2019.01) G06F 16/432 (2019.01)
G06F 16/738 (2019.01) G06F 16/783 (2019.01)
G06N 3/08 (2023.01) G06V 10/70 (2022.01)
G10L 25/30 (2013.01) G10L 25/48 (2013.01)

(52) CPC특허분류

G06F 16/7328 (2019.01)
G06F 16/433 (2019.01)

(21) 출원번호 10-2023-0157475

(22) 출원일자 2023년11월14일

심사청구일자 2023년11월14일

(71) 출원인

경희대학교 산학협력단

경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)

(72) 발명자

김정욱

경기도 안양시 동안구 귀인로 258, 106동 802호

박규리

서울특별시 강남구 언주로 123, 4동 511호

오영민

서울특별시 마포구 대흥로4길 28, 301호

(74) 대리인

두호특허법인

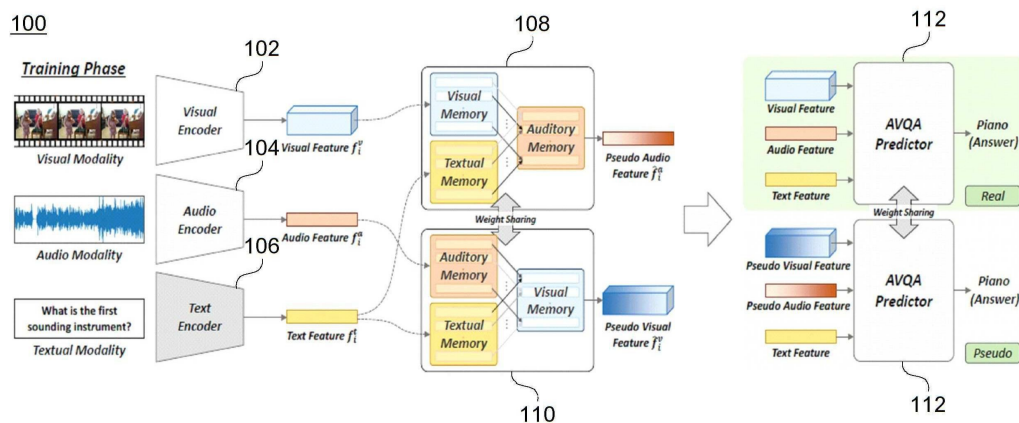
전체 청구항 수 : 총 19 항

(54) 발명의 명칭 딥러닝 기반의 오디오 비주얼 질문 응답의 학습 방법 및 이를 수행하기 위한 컴퓨팅 장치

(57) 요약

오디오 비주얼 질문 응답의 학습 방법 및 이를 수행하기 위한 컴퓨팅 장치가 개시된다. 개시되는 일 실시예에 따른 오디오 비주얼 질문 응답의 학습 방법은, 하나 이상의 프로세서들, 및 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨팅 장치에서 수행되는 방법으로서, 비주얼 모달리티 정보를 비주얼 인코더로 입력하여 비주얼 특징을 추출하는 단계, 오디오 모달리티 정보를 오디오 인코더로 입력하여 오디오 특징을 추출하는 단계, 텍스트 모달리티 정보를 텍스트 인코더로 입력하여 텍스트 특징을 추출하는 단계, 비주얼 특징 및 텍스트 특징을 제1 TMA(Trans-Modal Associative) 메모리로 입력하여 의사 오디오 특징을 생성하는 단계, 및 오디오 특징 및 텍스트 특징을 제2 TMA 메모리로 입력하여 의사 비주얼 특징을 생성하는 단계를 포함한다.

대표도



(52) CPC특허분류

G06F 16/434 (2019.01)
G06F 16/738 (2019.01)
G06F 16/7834 (2019.01)
G06F 16/7844 (2019.01)
G06F 16/7847 (2019.01)
G06N 3/08 (2023.01)
G06V 10/70 (2023.08)
G10L 25/30 (2013.01)
G10L 25/48 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711197142
과제번호	00252391
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)
연구과제명	멀티모달 데이터가 온전히 주어지지 않는 자율주행 환경에서 강인한 객체 검출을 위
한 인간의 뇌를 모방한 메모리 구조에 관한 연구	
기 여 율	1/1
과제수행기관명	경희대학교산학협력단
연구기간	2023.06.01 ~ 2024.02.29

명세서

청구범위

청구항 1

하나 이상의 프로세서들, 및

상기 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨팅 장치에서 수행되는 방법으로서,

비주얼 모달리티 정보를 비주얼 인코더로 입력하여 비주얼 특징을 추출하는 단계;

오디오 모달리티 정보를 오디오 인코더로 입력하여 오디오 특징을 추출하는 단계;

텍스트 모달리티 정보를 텍스트 인코더로 입력하여 텍스트 특징을 추출하는 단계;

상기 비주얼 특징 및 상기 텍스트 특징을 제1 TMA(Trans-Modal Associative) 메모리로 입력하여 의사 오디오 특징을 생성하는 단계; 및

상기 오디오 특징 및 상기 텍스트 특징을 제2 TMA 메모리로 입력하여 의사 비주얼 특징을 생성하는 단계를 포함하는, 오디오 비주얼 질문 응답의 학습 방법.

청구항 2

청구항 1에 있어서,

상기 제1 TMA 메모리는, 오디오 모달리티 정보가 누락된 상태에서 비주얼 특징 및 텍스트 특징 간의 연관 관계에 따라 오디오 모달리티 정보를 연상하여 의사 오디오 특징을 생성하도록 마련되고,

상기 제2 TMA 메모리는, 비주얼 모달리티 정보가 누락된 상태에서 오디오 특징 및 텍스트 특징 간의 연관 관계에 따라 비주얼 모달리티 정보를 연상하여 의사 비주얼 특징을 생성하도록 마련되는, 오디오 비주얼 질문 응답의 학습 방법.

청구항 3

청구항 2에 있어서,

상기 제1 TMA 메모리는,

각각 L개의 슬롯을 가지는 제1 비주얼 메모리, 제1 텍스트 메모리, 및 제1 오디오 메모리를 포함하고,

상기 제1 비주얼 메모리, 제1 텍스트 메모리, 및 제1 오디오 메모리의 L개의 슬롯은 학습 가능한 가중치를 갖도록 마련되는, 오디오 비주얼 질문 응답의 학습 방법.

청구항 4

청구항 3에 있어서,

상기 의사 오디오 특징을 생성하는 단계는,

상기 비주얼 특징을 상기 제1 비주얼 메모리의 각 슬롯에 할당하고, 상기 비주얼 특징과 상기 제1 비주얼 메모리의 각 슬롯 간의 연관 관계에 기초하여 비주얼 주소 벡터를 산출하는 단계;

상기 텍스트 특징을 상기 제1 텍스트 메모리의 각 슬롯에 할당하고, 상기 텍스트 특징과 상기 제1 텍스트 메모리의 각 슬롯 간의 연관 관계에 기초하여 텍스트 주소 벡터를 산출하는 단계;

상기 비주얼 주소 벡터와 상기 텍스트 주소 벡터에 기초하여 비주얼 텍스트 주소 벡터를 산출하는 단계; 및
상기 비주얼 텍스트 주소 벡터와 상기 제1 오디오 메모리의 각 슬롯에 저장된 값에 기초하여 상기 의사 오디오 특징을 산출하는 단계를 포함하는, 오디오 비주얼 질문 응답의 학습 방법.

청구항 5

청구항 4에 있어서,

상기 비주얼 주소 벡터의 각 요소는 하기 수학식에 의해 산출되고,

(수학식)

$$a_{ij}^v = \frac{\exp(s_{ij}^v)}{\sum_{m=1}^L \exp(s_{im}^v)}, s_{ij}^v = \frac{f_i^v \cdot M_j^v}{\sqrt{c}}$$

L : 슬롯의 개수

s_{ij}^v : 비주얼 특징(f_i^v)과 j번째 제1 비주얼 메모리의 슬롯(M_j^v) 간의 유사도

c : 비주얼 특징의 채널 수

상기 텍스트 주소 벡터의 각 요소는 하기 수학식에 의해 산출되는, 오디오 비주얼 질문 응답의 학습 방법.

(수학식)

$$a_{ij}^t = \frac{\exp(s_{ij}^t)}{\sum_{m=1}^L \exp(s_{im}^t)}, s_{ij}^t = \frac{f_i^t \cdot M_j^t}{\sqrt{c}}$$

s_{ij}^t : 텍스트 특징(f_i^t)과 j번째 제1 텍스트 메모리의 슬롯(M_j^t) 간의 유사도

c : 텍스트 특징의 채널 수

청구항 6

청구항 5에 있어서,

상기 의사 오디오 특징은 하기 수학식에 의해 산출되는, 오디오 비주얼 질문 응답의 학습 방법.

(수학식)

$$\hat{f}_i^a = \sum_{j=1}^L a_{ij}^{vt} \cdot M_j^a$$

$\hat{f}_i^a$: 의사 오디오 특징

M_j^a : j번째 제1 오디오 메모리의 슬롯

a_{ij}^{vt} : 비주얼 텍스트 주소 벡터

청구항 7

청구항 4에 있어서,

상기 제2 TMA 메모리는,

각각 L개의 슬롯을 가지는 제2 비주얼 메모리, 제2 텍스트 메모리, 및 제2 오디오 메모리를 포함하고,

상기 제2 비주얼 메모리, 제2 텍스트 메모리, 및 제2 오디오 메모리의 L개의 슬롯은 학습 가능한 가중치를 갖도록 마련되는, 오디오 비주얼 질문 응답의 학습 방법.

청구항 8

청구항 7에 있어서,

상기 의사 비주얼 특징을 생성하는 단계는,

상기 오디오 특징을 상기 제2 오디오 메모리의 각 슬롯에 할당하고, 상기 오디오 특징과 상기 제2 오디오 메모리의 각 슬롯 간의 연관 관계에 기초하여 오디오 주소 벡터를 산출하는 단계;

상기 텍스트 특징을 상기 제2 텍스트 메모리의 각 슬롯에 할당하고, 상기 텍스트 특징과 상기 제2 텍스트 메모리의 각 슬롯 간의 연관 관계에 기초하여 텍스트 주소 벡터를 산출하는 단계;

상기 오디오 주소 벡터와 상기 텍스트 주소 벡터에 기초하여 오디오 텍스트 주소 벡터를 산출하는 단계; 및

상기 오디오 텍스트 주소 벡터와 상기 제2 비주얼 메모리의 각 슬롯에 저장된 값에 기초하여 상기 의사 비주얼 특징을 산출하는 단계를 포함하는, 오디오 비주얼 질문 응답의 학습 방법.

청구항 9

청구항 8에 있어서,

상기 제1 TMA 메모리 및 상기 제2 TMA 메모리는, 하기의 수학적식에 의한 손실 함수(L_{tmr})(Trans-Modal Recalling Loss)를 가지는, 오디오 비주얼 질문 응답의 학습 방법.

(수학적식)

$$L_{tmr} = \lambda_1 L_a + \lambda_2 L_v$$

L_a : 의사 오디오 특징과 실제 오디오 특징 간의 차이를 최소화 하기 위한 손실 함수

L_v : 의사 비주얼 특징과 실제 비주얼 특징 간의 차이를 최소화 하기 위한 손실 함수

λ_1, λ_2 : 기 설정되는 하이퍼 파라미터

청구항 10

청구항 9에 있어서,

상기 오디오 비주얼 질문 응답의 학습 방법은,

상기 비주얼 특징, 상기 오디오 특징, 및 상기 텍스트 특징을 예측기에 입력하여 상기 텍스트 모달리티 정보에 따른 답변을 예측하도록 하는 단계; 및

상기 의사 비주얼 특징, 상기 의사 오디오 특징, 및 상기 텍스트 특징을 예측기에 입력하여 상기 텍스트 모달리티 정보에 따른 답변을 예측하도록 하는 단계를 더 포함하는, 오디오 비주얼 질문 응답의 학습 방법.

청구항 11

청구항 10에 있어서,

상기 오디오 비주얼 질문 응답의 학습 방법은, 하기 수학식에 의한 최종 손실 함수(L_{Total})를 가지는, 오디오 비주얼 질문 응답의 학습 방법.

(수학식)

$$L_{Total} = L_{avqa}^r + L_{avqa}^p + L_{tmr}$$

L_{avqa}^r : 비주얼 특징, 오디오 특징, 및 텍스트 특징이 모두 입력된 경우 예측기의 손실 함수

L_{avqa}^p : 의사 비주얼 특징, 의사 오디오 특징, 및 텍스트 특징이 입력된 경우 예측기의 손실 함수

청구항 12

하나 이상의 프로세서들;

메모리; 및

하나 이상의 프로그램들을 포함하고,

상기 하나 이상의 프로그램들은 상기 메모리에 저장되고, 상기 하나 이상의 프로세서들에 의해 실행되도록 구성되며,

상기 하나 이상의 프로그램들은,

비주얼 모달리티 정보를 비주얼 인코더로 입력하여 비주얼 특징을 추출하기 위한 명령;

오디오 모달리티 정보를 오디오 인코더로 입력하여 오디오 특징을 추출하기 위한 명령;

텍스트 모달리티 정보를 텍스트 인코더로 입력하여 텍스트 특징을 추출하기 위한 명령;

상기 비주얼 특징 및 상기 텍스트 특징을 제1 TMA(Trans-Modal Associative) 메모리로 입력하여 의사 오디오 특징을 생성하기 위한 명령; 및

상기 오디오 특징 및 상기 텍스트 특징을 제2 TMA 메모리로 입력하여 의사 비주얼 특징을 생성하기 위한 명령을 포함하는, 컴퓨팅 장치.

청구항 13

청구항 12에 있어서,

상기 제1 TMA 메모리는, 오디오 모달리티 정보가 누락된 상태에서 비주얼 특징 및 텍스트 특징 간의 연관 관계에 따라 오디오 모달리티 정보를 연상하여 의사 오디오 특징을 생성하도록 마련되고,

상기 제2 TMA 메모리는, 비주얼 모달리티 정보가 누락된 상태에서 오디오 특징 및 텍스트 특징 간의 연관 관계에 따라 비주얼 모달리티 정보를 연상하여 의사 비주얼 특징을 생성하도록 마련되는, 컴퓨팅 장치.

청구항 14

청구항 13에 있어서,

상기 제1 TMA 메모리는,

각각 L개의 슬롯을 가지는 제1 비주얼 메모리, 제1 텍스트 메모리, 및 제1 오디오 메모리를 포함하고,
상기 제1 비주얼 메모리, 제1 텍스트 메모리, 및 제1 오디오 메모리의 L개의 슬롯은 학습 가능한 가중치를 갖도록 마련되는, 컴퓨팅 장치.

청구항 15

청구항 14에 있어서,

상기 의사 오디오 특징을 생성하기 위한 명령은,

상기 비주얼 특징을 상기 제1 비주얼 메모리의 각 슬롯에 할당하고, 상기 비주얼 특징과 상기 제1 비주얼 메모리의 각 슬롯 간의 연관 관계에 기초하여 비주얼 주소 벡터를 산출하기 위한 명령;

상기 텍스트 특징을 상기 제1 텍스트 메모리의 각 슬롯에 할당하고, 상기 텍스트 특징과 상기 제1 텍스트 메모리의 각 슬롯 간의 연관 관계에 기초하여 텍스트 주소 벡터를 산출하기 위한 명령;

상기 비주얼 주소 벡터와 상기 텍스트 주소 벡터에 기초하여 비주얼 텍스트 주소 벡터를 산출하기 위한 명령; 및

상기 비주얼 텍스트 주소 벡터와 상기 제1 오디오 메모리의 각 슬롯에 저장된 값에 기초하여 상기 의사 오디오 특징을 산출하기 위한 명령을 포함하는, 컴퓨팅 장치.

청구항 16

청구항 15에 있어서,

상기 제2 TMA 메모리는,

각각 L개의 슬롯을 가지는 제2 비주얼 메모리, 제2 텍스트 메모리, 및 제2 오디오 메모리를 포함하고,

상기 제2 비주얼 메모리, 제2 텍스트 메모리, 및 제2 오디오 메모리의 L개의 슬롯은 학습 가능한 가중치를 갖도록 마련되는, 컴퓨팅 장치.

청구항 17

청구항 16에 있어서,

상기 의사 비주얼 특징을 생성하기 위한 명령은,

상기 오디오 특징을 상기 제2 오디오 메모리의 각 슬롯에 할당하고, 상기 오디오 특징과 상기 제2 오디오 메모리의 각 슬롯 간의 연관 관계에 기초하여 오디오 주소 벡터를 산출하기 위한 명령;

상기 텍스트 특징을 상기 제2 텍스트 메모리의 각 슬롯에 할당하고, 상기 텍스트 특징과 상기 제2 텍스트 메모리의 각 슬롯 간의 연관 관계에 기초하여 텍스트 주소 벡터를 산출하기 위한 명령;

상기 오디오 주소 벡터와 상기 텍스트 주소 벡터에 기초하여 오디오 텍스트 주소 벡터를 산출하기 위한 명령; 및

상기 오디오 텍스트 주소 벡터와 상기 제2 비주얼 메모리의 각 슬롯에 저장된 값에 기초하여 상기 의사 비주얼 특징을 산출하기 위한 명령을 포함하는, 컴퓨팅 장치.

청구항 18

청구항 17에 있어서,

상기 하나 이상의 프로그램들은,

상기 비주얼 특징, 상기 오디오 특징, 및 상기 텍스트 특징을 예측기에 입력하여 상기 텍스트 모달리티 정보에 따른 답변을 예측하도록 하기 위한 명령; 및

상기 의사 비주얼 특징, 상기 의사 오디오 특징, 및 상기 텍스트 특징을 예측기에 입력하여 상기 텍스트 모달리티 정보에 따른 답변을 예측하도록 하기 위한 명령을 더 포함하는, 컴퓨팅 장치.

청구항 19

비일시적 컴퓨터 판독 가능한 저장 매체(non-transitory computer readable storage medium)에 저장된 컴퓨터 프로그램으로서,

상기 컴퓨터 프로그램은 하나 이상의 명령어들을 포함하고, 상기 명령어들은 하나 이상의 프로세서들을 갖는 컴퓨팅 장치에 의해 실행될 때, 상기 컴퓨팅 장치로 하여금,

비주얼 모달리티 정보를 비주얼 인코더로 입력하여 비주얼 특징을 추출하는 단계;

오디오 모달리티 정보를 오디오 인코더로 입력하여 오디오 특징을 추출하는 단계;

텍스트 모달리티 정보를 텍스트 인코더로 입력하여 텍스트 특징을 추출하는 단계;

상기 비주얼 특징 및 상기 텍스트 특징을 제1 TMA(Trans-Modal Associative) 메모리로 입력하여 의사 오디오 특징을 생성하는 단계; 및

상기 오디오 특징 및 상기 텍스트 특징을 제2 TMA 메모리로 입력하여 의사 비주얼 특징을 생성하는 단계를 수행하도록 하는, 비일시적 컴퓨터 판독 가능한 저장 매체에 저장된 컴퓨터 프로그램.

발명의 설명

기술 분야

[0001] 본 발명의 실시예는 딥러닝 기반의 오디오 비주얼 질문 응답 기술과 관련된다.

배경 기술

[0003] 영상의 시각 정보와 청각 정보를 통해 상황을 이해하고 질문에 대한 답변을 제시하는 기술(Audio-Visual Question Answering: AVQA)은 멀티 모달 딥러닝(Multi-Modal Deep Learning) 분야에서 활발히 연구가 되고 있는 기술 중 하나이다. 사람은 어떤 질문에 답을 할 때 시각적인 정보와 청각적인 정보를 단서로 사용한다. 이러한 인간의 인식 능력을 질문 응답 시스템에서 모방하기 위해 오디오-비주얼 질문 응답(AVQA) 기술이 등장했다. 이 기술은 기계가 질문 텍스트와 관련된 오디오-비주얼 정보를 결합하여 이해하고 응답하는 것을 돕는다. 이러한 특성 때문에 AVQA 작업은 자율 주행과 상호작용이 필요한 교육 분야 등 다양한 실용적인 실제 세계 응용 분야와 밀접하게 연결되어 있다.

[0004] 오디오-비주얼 모달 정보는 일반적으로 효과적인 AVQA 작업에 사용되지만, 실제 상황에서는 기계의 고장 등으로 한 모달리티(오디오 또는 비주얼)의 입력이 누락되는 경우가 자주 발생하여 질문에 답하는 것이 어려워지며 그로 인해 AVQA의 성능 저하가 발생하게 된다. 따라서 다양한 응용 분야에서 AVQA 작업을 수행하기 위해서는 실제 세계 시나리오에서 누락된 모달정보를 처리할 수 있는 AVQA 접근법 개발이 중요하다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 한국등록특허공보 제10-2122918호(2020.06.26)

발명의 내용

해결하려는 과제

[0007] 본 발명의 실시예는 모달리티 정보가 누락된 경우에도 질문에 정확한 답변을 할 수 있도록 하는 오디오 비주얼 질문 응답 장치의 학습 방법 및 이를 수행하기 위한 컴퓨팅 장치를 제공하기 위한 것이다.

과제의 해결 수단

[0009] 개시되는 일 실시예에 따른 오디오 비주얼 질문 응답의 학습 방법은, 하나 이상의 프로세서들, 및 상기 하나 이상의 프로세서들에 의해 실행되는 하나 이상의 프로그램들을 저장하는 메모리를 구비한 컴퓨팅 장치에서 수행되는 방법으로서, 비주얼 모달리티 정보를 비주얼 인코더로 입력하여 비주얼 특징을 추출하는 단계; 오디오 모달리티 정보를 오디오 인코더로 입력하여 오디오 특징을 추출하는 단계; 텍스트 모달리티 정보를 텍스트 인코더로 입력하여 텍스트 특징을 추출하는 단계; 상기 비주얼 특징 및 상기 텍스트 특징을 제1 TMA(Trans-Modal Associative) 메모리로 입력하여 의사 오디오 특징을 생성하는 단계; 및 상기 오디오 특징 및 상기 텍스트 특징을 제2 TMA 메모리로 입력하여 의사 비주얼 특징을 생성하는 단계를 포함한다.

[0010] 상기 제1 TMA 메모리는, 오디오 모달리티 정보가 누락된 상태에서 비주얼 특징 및 텍스트 특징 간의 연관 관계에 따라 오디오 모달리티 정보를 연상하여 의사 오디오 특징을 생성하도록 마련되고, 상기 제2 TMA 메모리는, 비주얼 모달리티 정보가 누락된 상태에서 오디오 특징 및 텍스트 특징 간의 연관 관계에 따라 비주얼 모달리티 정보를 연상하여 의사 비주얼 특징을 생성하도록 마련될 수 있다.

[0011] 상기 제1 TMA 메모리는, 각각 L개의 슬롯을 가지는 제1 비주얼 메모리, 제1 텍스트 메모리, 및 제1 오디오 메모리를 포함하고, 상기 제1 비주얼 메모리, 제1 텍스트 메모리, 및 제1 오디오 메모리의 L개의 슬롯은 학습 가능한 가중치를 갖도록 마련될 수 있다.

[0012] 상기 의사 오디오 특징을 생성하는 단계는, 상기 비주얼 특징을 상기 제1 비주얼 메모리의 각 슬롯에 할당하고, 상기 비주얼 특징과 상기 제1 비주얼 메모리의 각 슬롯 간의 연관 관계에 기초하여 비주얼 주소 벡터를 산출하는 단계; 상기 텍스트 특징을 상기 제1 텍스트 메모리의 각 슬롯에 할당하고, 상기 텍스트 특징과 상기 제1 텍스트 메모리의 각 슬롯 간의 연관 관계에 기초하여 텍스트 주소 벡터를 산출하는 단계; 상기 비주얼 주소 벡터와 상기 텍스트 주소 벡터에 기초하여 비주얼 텍스트 주소 벡터를 산출하는 단계; 및 상기 비주얼 텍스트 주소 벡터와 상기 제1 오디오 메모리의 각 슬롯에 저장된 값에 기초하여 상기 의사 오디오 특징을 산출하는 단계를 포함할 수 있다.

[0013] 상기 비주얼 주소 벡터의 각 요소는 하기 수학식에 의해 산출되고,

[0014] (수학식)

$$a_{ij}^v = \frac{\exp(s_{ij}^v)}{\sum_{m=1}^L \exp(s_{im}^v)}, s_{ij}^v = \frac{f_i^v \cdot M_j^v}{\sqrt{c}}$$

[0015]

[0016] L : 슬롯의 개수

[0017] s_{ij}^v : 비주얼 특징(f_i^v)과 j번째 제1 비주얼 메모리의 슬롯(M_j^v) 간의 유사도

[0018] c : 비주얼 특징의 채널 수

[0019] 상기 텍스트 주소 벡터의 각 요소는 하기 수학식에 의해 산출될 수 있다.

[0020] (수학식)

$$a_{ij}^t = \frac{\exp(s_{ij}^t)}{\sum_{m=1}^L \exp(s_{im}^t)}, s_{ij}^t = \frac{f_i^t \cdot M_j^t}{\sqrt{c}}$$

[0021]

[0022] s_{ij}^t : 텍스트 특징(f_i^t)과 j번째 제1 텍스트 메모리의 슬롯(M_j^t) 간의 유사도

[0023] c : 텍스트 특징의 채널 수

[0024] 상기 의사 오디오 특징은 하기 수학식에 의해 산출될 수 있다.

[0025] (수학식)

$$\hat{f}_i^a = \sum_{j=1}^L a_{ij}^{vt} \cdot M_j^a$$

[0026]

[0027] $\hat{f}_i^a$: 의사 오디오 특징

[0028] M_j^a : j번째 제1 오디오 메모리의 슬롯

[0029] a_{ij}^{vt} : 비주얼 텍스트 주소 벡터

[0030] 상기 제2 TMA 메모리는, 각각 L개의 슬롯을 가지는 제2 비주얼 메모리, 제2 텍스트 메모리, 및 제2 오디오 메모리를 포함하고, 상기 제2 비주얼 메모리, 제2 텍스트 메모리, 및 제2 오디오 메모리의 L개의 슬롯은 학습 가능한 가중치를 갖도록 마련될 수 있다.

[0031] 상기 의사 비주얼 특징을 생성하는 단계는, 상기 오디오 특징을 상기 제2 오디오 메모리의 각 슬롯에 할당하고, 상기 오디오 특징과 상기 제2 오디오 메모리의 각 슬롯 간의 연관 관계에 기초하여 오디오 주소 벡터를 산출하는 단계; 상기 텍스트 특징을 상기 제2 텍스트 메모리의 각 슬롯에 할당하고, 상기 텍스트 특징과 상기 제2 텍스트 메모리의 각 슬롯 간의 연관 관계에 기초하여 텍스트 주소 벡터를 산출하는 단계; 상기 오디오 주소 벡터와 상기 텍스트 주소 벡터에 기초하여 오디오 텍스트 주소 벡터를 산출하는 단계; 및 상기 오디오 텍스트 주소 벡터와 상기 제2 비주얼 메모리의 각 슬롯에 저장된 값에 기초하여 상기 의사 비주얼 특징을 산출하는 단계를 포함할 수 있다.

[0032] 상기 제1 TMA 메모리 및 상기 제2 TMA 메모리는, 하기의 수학식에 의한 손실 함수(L_{tmr})(Trans-Modal Recalling Loss)를 가질 수 있다.

[0033] (수학식)

$$L_{tmr} = \lambda_1 L_a + \lambda_2 L_v$$

[0034]

[0035] L_a : 의사 오디오 특징과 실제 오디오 특징 간의 차이를 최소화 하기 위한 손실 함수

[0036] L_v : 의사 비주얼 특징과 실제 비주얼 특징 간의 차이를 최소화 하기 위한 손실 함수

[0037] λ_1, λ_2 : 기 설정되는 하이퍼 파라미터

[0038] 상기 오디오 비주얼 질문 응답의 학습 방법은, 상기 비주얼 특징, 상기 오디오 특징, 및 상기 텍스트 특징을 예측기에 입력하여 상기 텍스트 모달리티 정보에 따른 답변을 예측하도록 하는 단계; 및 상기 의사 비주얼 특징, 상기 의사 오디오 특징, 및 상기 텍스트 특징을 예측기에 입력하여 상기 텍스트 모달리티 정보에 따른 답변을 예측하도록 하는 단계를 더 포함할 수 있다.

[0039] 상기 오디오 비주얼 질문 응답의 학습 방법은, 하기 수학식에 의한 최종 손실 함수(L_{Total})를 가질 수 있다.

[0040] (수학식)

$$L_{Total} = L_{avqa}^r + L_{avqa}^p + L_{tmr}$$

[0042] L_{avqa}^r : 비주얼 특징, 오디오 특징, 및 텍스트 특징이 모두 입력된 경우 예측기의 손실 함수

[0043] L_{avqa}^p : 의사 비주얼 특징, 의사 오디오 특징, 및 텍스트 특징이 입력된 경우 예측기의 손실 함수

[0044] 개시되는 일 실시예에 따른 컴퓨팅 장치는, 하나 이상의 프로세서들; 메모리; 및 하나 이상의 프로그램들을 포함하고, 상기 하나 이상의 프로그램들은 상기 메모리에 저장되고, 상기 하나 이상의 프로세서들에 의해 실행되도록 구성되며, 상기 하나 이상의 프로그램들은, 비주얼 모달리티 정보를 비주얼 인코더로 입력하여 비주얼 특징을 추출하기 위한 명령; 오디오 모달리티 정보를 오디오 인코더로 입력하여 오디오 특징을 추출하기 위한 명령; 텍스트 모달리티 정보를 텍스트 인코더로 입력하여 텍스트 특징을 추출하기 위한 명령; 상기 비주얼 특징 및 상기 텍스트 특징을 제1 TMA(Trans-Modal Associative) 메모리로 입력하여 의사 오디오 특징을 생성하기 위한 명령; 및 상기 오디오 특징 및 상기 텍스트 특징을 제2 TMA 메모리로 입력하여 의사 비주얼 특징을 생성하기 위한 명령을 포함한다.

발명의 효과

[0046] 개시되는 실시예에 의하면, 비주얼 모달리티 및 오디오 모달리티 중 어느 하나의 모달리티 정보가 누락된 경우, 나머지 모달리티 정보와 텍스트 간의 연관 관계에 기초하여 누락된 모달리티 정보를 연상하도록 함으로써, 특정 모달리티 정보가 누락된 상태에서도 질문에 정확한 답변을 제공할 수 있게 된다.

도면의 간단한 설명

[0048] 도 1은 본 발명의 일 실시예에서 모달리티 정보가 누락된 경우 누락된 모달리티 정보를 연상하도록 하여 질문에 답하는 상태를 개략적으로 나타낸 도면

도 2는 본 발명의 일 실시예에 따른 오디오 비주얼 질문 응답 장치의 학습 방법을 개략적으로 나타낸 도면

도 3은 본 발명의 일 실시예에서 제1 TMA 메모리가 의사 오디오 특징을 생성하는 과정을 나타낸 도면

도 4는 본 발명의 일 실시예에서 제2 TMA 메모리가 의사 비주얼 특징을 생성하는 과정을 나타낸 도면

도 5는 본 발명의 일 실시예에서 오디오 비주얼 질문 응답 장치의 추론 단계를 개략적으로 나타낸 도면

도 6은 예시적인 실시예들에서 사용되기에 적합한 컴퓨팅 장치를 포함하는 컴퓨팅 환경을 예시하여 설명하기 위한 블록도

발명을 실시하기 위한 구체적인 내용

[0049] 이하, 도면을 참조하여 본 발명의 구체적인 실시형태를 설명하기로 한다. 이하의 상세한 설명은 본 명세서에서 기술된 방법, 장치 및/또는 시스템에 대한 포괄적인 이해를 돕기 위해 제공된다. 그러나 이는 예시에 불과하며 본 발명은 이에 제한되지 않는다.

[0050] 본 발명의 실시예들을 설명함에 있어서, 본 발명과 관련된 공지기술에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략하기로 한다. 그리고, 후술되는 용어들은 본 발명에서의 기능을 고려하여 정의된 용어들로서 이는 사용자, 운용자의 의도 또는 관례 등에 따라 달라질 수 있다. 그러므로 그 정의는 본 명세서 전반에 걸친 내용을 토대로 내려져야 할 것이다. 상세한 설명에서 사용되는 용어는 단지 본 발명의 실시예들을 기술하기 위한 것이며, 결코 제한적이어서는 안 된다. 명확하게 달리 사용되지 않는 한, 단수 형태의 표현은 복수 형태의 의미를 포함한다. 본 설명에서, "포함" 또는 "구비"와 같은 표현은 어떤 특성들, 숫자들, 단계들, 동작들, 요소들, 이들의 일부 또는 조합을 가리키기 위한 것이며, 기술된 것 이외에 하나 또는 그 이상의 다른 특성, 숫자, 단계, 동작, 요소, 이들의 일부 또는 조합의 존재 또는 가능

성을 배제하도록 해석되어서는 안 된다.

- [0051] 또한, 제1, 제2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로 사용될 수 있다. 예를 들어, 본 발명의 권리 범위를 벗어나지 않으면서 제1 구성 요소는 제2 구성 요소로 명명될 수 있고, 유사하게 제2 구성 요소도 제1 구성 요소로 명명될 수 있다.
- [0053] 개시되는 실시예에서는, 오디오 또는 비주얼 정보 중 어느 하나의 모달리티 정보가 누락된 경우, 사용 가능한 모달리티 특징과 텍스트의 단서 간의 연관성을 활용하여 누락된 모달리티 정보를 연상(recall)하도록 하는 TMA(Trans-Modal Associative) 메모리를 제안하고자 한다. 도 1은 본 발명의 일 실시예에서 모달리티 정보가 누락된 경우 누락된 모달리티 정보를 연상하도록 하여 질문에 답하는 상태를 개략적으로 나타낸 도면이다.
- [0054] 도 1의 (a)를 참조하면, 오디오 모달 정보가 없고(누락되고) 오직 비주얼 모달 정보(시각 모달 정보)와 질문만이 사용 가능한 경우, 시각 모달 정보와 질문의 텍스트 간의 연관성을 설정하고 그와 관련된 오디오 단서를 연상(recalled audio)하여 질문에 대한 답을 생성할 수 있다.
- [0055] 도 1의 (b)를 참조하면, 비주얼 모달 정보가 없고 오직 오디오 모달 정보와 질문만이 존재하는 경우, 오디오 모달 정보와 질문의 텍스트 간의 연관성을 설정하고 그와 관련된 비주얼 단서를 연상(recalled visual)하여 질문에 대한 답을 생성할 수 있다.
- [0057] 도 2는 본 발명의 일 실시예에 따른 오디오 비주얼 질문 응답 장치의 학습 방법을 개략적으로 나타낸 도면이다.
- [0058] 도 2를 참조하면, 오디오 비주얼 질문 응답 장치(100)는 비주얼 인코더(102), 오디오 인코더(104), 텍스트 인코더(106), 제1 TMA(Trans-Modal Associative) 메모리(108), 제2 TMA 메모리(110), 및 예측기(Predictor)(112)를 포함할 수 있다.
- [0059] 오디오 비주얼 질문 응답 장치(100)의 학습 과정에서, 비주얼 모달리티 정보(시각 정보), 오디오 모달리티 정보(청각 정보), 및 텍스트 모달리티 정보(질문 정보)가 각각 비주얼 인코더(102), 오디오 인코더(104), 및 텍스트 인코더(106)로 입력될 수 있다.
- [0060] 비주얼 인코더(102)는 입력된 비주얼 모달리티 정보로부터 비주얼 특징(visual feature)(f_i^v)을 추출할 수 있다. 오디오 인코더(104)는 입력된 오디오 모달리티 정보로부터 오디오 특징(audio feature)(f_i^a)을 추출할 수 있다. 텍스트 인코더(106)는 입력된 텍스트 모달리티 정보로부터 텍스트 특징(text feature)(f_i^t)을 추출할 수 있다.
- [0061] 제1 TMA 메모리(108)는 비주얼 특징 및 텍스트 특징을 입력 받아 의사 오디오 특징(pseudo audio feature)을 생성할 수 있다. 즉, 제1 TMA 메모리(108)는 오디오 모달리티 정보가 누락된 상황에서 비주얼 특징 및 텍스트 특징 간의 연관 관계에 기초하여 오디오 모달리티 정보를 연상함으로써 의사 오디오 특징을 생성하도록 마련될 수 있다.
- [0062] 제2 TMA 메모리(110)는 오디오 특징 및 텍스트 특징을 입력 받아 의사 비주얼 특징(pseudo visual feature)을 생성할 수 있다. 즉, 제2 TMA 메모리(110)는 비주얼 모달리티 정보가 누락된 상황에서 오디오 특징 및 텍스트 특징 간의 연관 관계에 기초하여 비주얼 모달리티 정보를 연상함으로써 의사 비주얼 특징을 생성하도록 마련될 수 있다.
- [0063] 학습 과정에서 제1 TMA 메모리(108)와 제2 TMA 메모리(110)는 가중치를 공유할 수 있다. 제1 TMA 메모리(108) 및 제2 TMA 메모리(110)에서 각각 의사 오디오 특징 및 의사 비주얼 특징을 생성하는 구체적인 과정은 도 3 및 도 4를 참조하여 후술하기로 한다.
- [0064] 예측기(Predictor)(112)는 의사 오디오 특징, 의사 비주얼 특징, 및 텍스트 특징을 입력 받을 수 있다. 예측기(112)는 입력된 의사 오디오 특징, 의사 비주얼 특징, 및 텍스트 특징에 기초하여 텍스트 모달리티 정보에 대응한 답변을 예측할 수 있다. 예측기(112)는 예측된 답변과 정답 값을 비교하여 그 차이가 최소화 되도록 학습될 수 있다.

[0065] 또한, 예측기(112)는 모달리티 정보가 누락되지 않고 비주얼 특징, 오디오 특징, 및 텍스트 특징이 모두 입력된 상태에서 예측된 답변과 정답 값 간의 차이가 최소화 되도록 학습될 수 있다. 도 2에서는 설명의 편의 상, 위의 두 경우의 예측 과정을 별도로 도시하였으며, 위의 두 경우의 예측 과정에서 예측기(112)의 가중치는 상호 공유되어 학습될 수 있다.

[0067] 도 3은 본 발명의 일 실시예에서 제1 TMA 메모리가 의사 오디오 특징을 생성하는 과정을 나타낸 도면이다. 도 3을 참조하면, 제1 TMA 메모리(108)는 제1 비주얼 메모리(108a), 제1 텍스트 메모리(108b), 및 제1 오디오 메모리(108c)를 포함할 수 있다. 제1 비주얼 메모리(108a), 제1 텍스트 메모리(108b), 및 제1 오디오 메모리(108c)는 각각 L개의 슬롯으로 구성될 수 있다. 제1 비주얼 메모리(108a), 제1 텍스트 메모리(108b), 및 제1 오디오 메모리(108c)의 각 슬롯은 학습 가능한 가중치를 갖도록 마련될 수 있다. 이 경우, 제1 TMA 메모리(108)는 하기 수학식 1로 나타낼 수 있다.

[0068] (수학식 1)

$$M = \{M_j^v, M_j^a, M_j^t\}_{j=1}^L \quad (M_j^v \in R^{1 \times whc}, M_j^a, M_j^t \in R^{1 \times c})$$

[0070] M^v : 제1 비주얼 메모리

[0071] M^a : 제1 오디오 메모리

[0072] M^t : 제1 텍스트 메모리

[0073] L: 슬롯의 개수

[0074] 수학식 1에서, w, h, c는 특징(feature)에 대한 width, height, channel을 각각 나타낼 수 있다. 메모리의 각 슬롯은 해당하는 모달리티 정보를 기억하는 역할을 한다.

[0075] 여기서, 비주얼 특징(f_i^v) 및 텍스트 특징(f_i^t)이 각각 제1 비주얼 메모리(108a) 및 제1 텍스트 메모리(108b)로 입력되어 각 슬롯에 분할되어 할당될 수 있다. 일 실시예에서, 비주얼 특징은 평탄화(flatten)된 후 제1 비주얼 메모리(108a)로 입력될 수 있다. 즉, 비주얼 특징(f_i^v)은 $w \times h \times c$ 의 3차원 형태이므로 평탄화를 통해 1차원 벡터 형태로 변환한 후 제1 비주얼 메모리(108a)의 각 슬롯에 분할되어 할당될 수 있다.

[0076] 제1 TMA 메모리(108)는 비주얼 특징과 제1 비주얼 메모리(108a)의 각 슬롯 간의 연관 관계를 산출할 수 있다. 이를 위해, 제1 TMA 메모리(108)는 비주얼 특징 및 제1 비주얼 메모리(108a)의 L개 슬롯에 기초하여 비주얼 주소 벡터($a_i^v = \{a_{i1}^v, \dots, a_{iL}^v\} \in R^{1 \times L}$)를 산출할 수 있다. 제1 TMA 메모리(108)는 하기 수학식 2를 통해 비주얼 주소 벡터의 각 요소를 산출할 수 있다.

[0077] (수학식 2)

$$a_{ij}^v = \frac{\exp(s_{ij}^v)}{\sum_{m=1}^L \exp(s_{im}^v)}, s_{ij}^v = \frac{f_i^v \cdot M_j^v}{\sqrt{c}}$$

[0078]

[0079] L: 슬롯의 개수

[0080] s_{ij}^v : 비주얼 특징(f_i^v)과 j번째 제1 비주얼 메모리의 슬롯(M_j^v) 간의 유사도

[0081] c: 비주얼 특징의 채널 수

[0082] 수학식 2에서 비주얼 주소 벡터(a_{ij}^v) 값이 크면, 비주얼 특징(f_i^v)과 j번째 제1 비주얼 메모리의 슬롯(M_j^v) 간에 강한 연관 관계가 있음을 나타내고, 비주얼 주소 벡터(a_{ij}^v) 값이 작으면, 비주얼 특징(f_i^v)과 j번째 제1 비주얼

메모리의 슬롯(M_j^v) 간에 약한 연관 관계가 있음을 나타내게 된다.

[0083] 제1 TMA 메모리(108)는 텍스트 특징과 제1 텍스트 메모리(108b)의 각 슬롯 간의 연관 관계를 산출할 수 있다. 이를 위해, 제1 TMA 메모리(108)는 텍스트 특징 및 제1 텍스트 메모리(108b)의 L개 슬롯에 기초하여 텍스트 주소 벡터를 산출할 수 있다. 제1 TMA 메모리(108)는 하기 수학적 식 3을 통해 텍스트 주소 벡터의 각 요소를 산출할 수 있다.

[0084] (수학적 식 2)

$$a_{ij}^t = \frac{\exp(s_{ij}^t)}{\sum_{m=1}^L \exp(s_{im}^t)}, s_{ij}^t = \frac{f_i^t \cdot M_j^t}{\sqrt{c}}$$

[0085]

[0086] s_{ij}^t : 텍스트 특징(f_i^t)과 j번째 제1 텍스트 메모리의 슬롯(M_j^t) 간의 유사도

[0087] c : 텍스트 특징의 채널

[0088] 제1 TMA 메모리(108)는 비주얼 주소 벡터(a_i^v)와 텍스트 주소 벡터(a_i^t)를 곱하여 비주얼 텍스트 주소 벡터(a_i^{vt})를 산출할 수 있다. 제1 TMA 메모리(108)는 비주얼 주소 벡터(a_i^v)와 텍스트 주소 벡터(a_i^t)를 element-wise 곱하여 비주얼 텍스트 주소 벡터(a_i^{vt})를 산출할 수 있다. 이는 수학적 식 3으로 나타낼 수 있다.

[0089] (수학적 식 3)

$$a_i^{vt} = \text{softmax}(a_i^v \circ a_i^t)$$

[0090]

[0091] softmax : 소프트 맥스 함수

[0092] 제1 TMA 메모리(108)는 비주얼 텍스트 주소 벡터(a_i^{vt})와 제1 오디오 메모리(108c)의 각 슬롯에 저장된 값에 기초하여 의사 오디오 특징을 산출할 수 있다. 초기에 제1 오디오 메모리(108c)의 각 슬롯에는 랜덤한 값이 저장될 수 있다. 제1 TMA 메모리(108)는 하기 수학적 식 4에 의해 의사 오디오 특징($\hat{f}_i^a$)을 산출할 수 있다.

[0093] (수학적 식 4)

$$\hat{f}_i^a = \sum_{j=1}^L a_{ij}^{vt} \cdot M_j^a$$

[0094]

[0095] M_j^a : j번째 제1 오디오 메모리의 슬롯

[0096] 여기서, 비주얼 텍스트 주소 벡터(a_i^{vt}) 값이 높다면 제1 오디오 메모리의 j번째 슬롯이 오디오 모달리티 정보를 연상하는데 더 많이 고려된다.

[0098] 도 4는 본 발명의 일 실시예에서 제2 TMA 메모리가 의사 비주얼 특징을 생성하는 과정을 나타낸 도면이다. 도 4를 참조하면, 제2 TMA 메모리(110)는 제2 비주얼 메모리(110a), 제2 텍스트 메모리(110b), 및 제2 오디오 메모리(110c)를 포함할 수 있다. 제2 비주얼 메모리(110a), 제2 텍스트 메모리(110b), 및 제2 오디오 메모리(110c)는 각각 L개의 슬롯으로 구성될 수 있다.

[0099] 제2 비주얼 메모리(110a), 제2 텍스트 메모리(110b), 및 제2 오디오 메모리(110c)의 각 슬롯은 제1 비주얼 메모리(108a), 제1 텍스트 메모리(108b), 및 제1 오디오 메모리(108c)의 슬롯과 같이 학습 가능한 가중치를 가질 수 있다. 제1 비주얼 메모리(108a), 제1 텍스트 메모리(108b), 및 제1 오디오 메모리(108c)의 각 슬롯과 제2 비주

열 메모리(110a), 제2 텍스트 메모리(110b), 및 제2 오디오 메모리(110c)의 각 슬롯은 학습 과정에서 가중치를 상호 공유할 수 있다.

여기서, 오디오 특징(f_i^a) 및 텍스트 특징(f_i^t)이 각각 제2 오디오 메모리(110c) 및 제2 텍스트 메모리(110b)로 입력되어 각 슬롯에 분할되어 할당될 수 있다. 제2 TMA 메모리(110)는 오디오 특징과 제2 오디오 메모리(110c)의 각 슬롯 간의 연관 관계에 기초하여 오디오 주소 벡터를 산출할 수 있다. 제2 TMA 메모리(110)는 하기 수학식 5를 통해 오디오 주소 벡터의 각 요소를 산출할 수 있다.

(수학식 5)

$$a_{ij}^a = \frac{\exp(s_{ij}^a)}{\sum_{m=1}^L \exp(s_{im}^a)}, s_{ij}^a = \frac{f_i^a \cdot M_j^a}{\sqrt{c}}$$

L : 슬롯의 개수

s_{ij}^a : 오디오 특징(f_i^a)과 j번째 제2 오디오 메모리의 슬롯(M_j^a) 간의 유사도

c : 오디오 특징의 채널 수

마찬가지로, 제2 TMA 메모리(110)는 텍스트 특징과 제2 텍스트 메모리(110b)의 각 슬롯 간의 연관 관계에 기초하여 텍스트 주소 벡터를 산출할 수 있다. 텍스트 주소 벡터를 구하는 식은 수학식 2와 유사하므로 이에 대한 자세한 설명은 생략하기로 한다.

제2 TMA 메모리(110)는 오디오 주소 벡터(a_i^a)와 텍스트 주소 벡터(a_i^t)를 곱하여 오디오 텍스트 주소 벡터(a_i^{at})를 산출할 수 있다. 제2 TMA 메모리(110)는 오디오 주소 벡터(a_i^a)와 텍스트 주소 벡터(a_i^t)를 element-wise 곱하여 오디오 텍스트 주소 벡터(a_i^{at})를 산출할 수 있다. 이는 수학식 6으로 나타낼 수 있다.

(수학식 6)

$$a_i^{at} = softmax(a_i^a \circ a_i^t)$$

제2 TMA 메모리(110)는 오디오 텍스트 주소 벡터(a_i^{at})와 제2 비주얼 메모리(108a)의 각 슬롯에 저장된 값에 기초하여 의사 비주얼 특징을 산출할 수 있다. 제2 TMA 메모리(110)는 하기 수학식 7에 의해 의사 비주얼 특징($\hat{f}_i^v$)을 산출할 수 있다.

(수학식 7)

$$\hat{f}_i^v = \sum_{j=1}^L a_{ij}^{at} \cdot M_j^v$$

M_j^v : j번째 제2 비주얼 메모리의 슬롯

한편, 제1 TMA 메모리(108) 및 제2 TMA 메모리(110)는 각각 의사 오디오 특징 및 의사 비주얼 특징이 실제 오디오 특징 및 실제 비주얼 특징과 닮아가도록 하는 손실 함수(L_{tmr})(Trans-Modal Recalling Loss)를 가질 수 있다. 손실 함수(L_{tmr})는 하기 수학식 8로 나타낼 수 있다.

[0116] (수학식 8)

$$L_{tmr} = \lambda_1 L_a + \lambda_2 L_v$$

[0117]

[0118] L_a : 의사 오디오 특징과 실제 오디오 특징 간의 차이를 최소화 하기 위한 손실 함수

[0119] L_v : 의사 비주얼 특징과 실제 비주얼 특징 간의 차이를 최소화 하기 위한 손실 함수

[0120] λ_1, λ_2 : 기 설정되는 하이퍼 파라미터

[0121] 또한, 오디오 비주얼 질문 응답 장치(100)의 학습 과정에서 최종 손실 함수는 수학식 9와 같이 나타낼 수 있다.

[0122] (수학식 9)

$$L_{Total} = L_{avqa}^r + L_{avqa}^p + L_{tmr}$$

[0123]

[0124] L_{avqa}^r : 비주얼 특징, 오디오 특징, 및 텍스트 특징이 모두 입력된 경우 예측기의 손실 함수

[0125] L_{avqa}^p : 의사 비주얼 특징, 의사 오디오 특징, 및 텍스트 특징이 입력된 경우 예측기의 손실 함수

[0126] 즉, 손실 함수(L_{avqa}^r)는 모달리티 정보가 누락되지 않고 비주얼 특징, 오디오 특징, 및 텍스트 특징이 모두 입력된 상태에서 예측기(112)가 예측한 답변과 정답 값 간의 차이가 최소화 되도록 하는 손실 함수이다. 또한, 손실 함수(L_{avqa}^p)는 의사 오디오 특징, 의사 비주얼 특징, 및 텍스트 특징이 입력되는 경우 예측기(112)가 예측한 답변과 정답 값 간의 차이가 최소화 되도록 하는 손실 함수이다.

[0127] 개시되는 실시예에 의하면, 비주얼 모달리티 및 오디오 모달리티 중 어느 하나의 모달리티 정보가 누락된 경우, 나머지 모달리티 정보와 텍스트 간의 연관 관계에 기초하여 누락된 모달리티 정보를 연상하도록 함으로써, 특정 모달리티 정보가 누락된 상태에서도 질문에 정확한 답변을 제공할 수 있게 된다.

[0129] 도 5는 본 발명의 일 실시예에서 오디오 비주얼 질문 응답 장치의 추론 단계를 개략적으로 나타낸 도면이다. 도 5를 참조하면, 오디오 비주얼 질문 응답 장치(100)의 학습이 완료된 상태에서, 오디오 모달리티 정보가 누락된 경우 추론 단계를 살펴보기로 한다. 비디오 모달리티 정보가 누락된 경우도 이와 동일한 방식으로 추론할 수 있음은 물론이다.

[0130] 비주얼 모달리티 정보는 비주얼 인코더(102)로 입력되고, 비주얼 인코더(102)는 비디오 특징을 추출할 수 있다. 텍스트 모달리티 정보는 텍스트 인코더(106)로 입력되고, 텍스트 인코더(106)는 텍스트 특징을 추출할 수 있다. 비디오 특징 및 텍스트 특징은 제1 TMA 메모리(108)로 입력되고, 제1 TMA 메모리(108)는 의사 오디오 특징을 생성할 수 있다.

[0131] 그러면, 비디오 특징, 의사 오디오 특징, 및 텍스트 특징이 예측기(112)로 입력되고, 예측기(112)는 비디오 특징, 의사 오디오 특징, 및 텍스트 특징으로부터 질문(텍스트)에 대한 답변을 출력할 수 있다.

[0133] 도 6은 예시적인 실시예들에서 사용되기에 적합한 컴퓨팅 장치를 포함하는 컴퓨팅 환경(10)을 예시하여 설명하기 위한 블록도이다. 도시된 실시예에서, 각 컴포넌트들은 이하에 기술된 것 이외에 상이한 기능 및 능력을 가질 수 있고, 이하에 기술된 것 이외에도 추가적인 컴포넌트를 포함할 수 있다.

[0134] 도시된 컴퓨팅 환경(10)은 컴퓨팅 장치(12)를 포함한다. 일 실시예에서, 컴퓨팅 장치(12)는 오디오 비주얼 질문 응답 장치(100)일 수 있다.

[0135] 컴퓨팅 장치(12)는 적어도 하나의 프로세서(14), 컴퓨터 판독 가능 저장 매체(16) 및 통신 버스(18)를 포함한다. 프로세서(14)는 컴퓨팅 장치(12)로 하여금 앞서 언급된 예시적인 실시예에 따라 동작하도록 할 수 있다. 예컨대, 프로세서(14)는 컴퓨터 판독 가능 저장 매체(16)에 저장된 하나 이상의 프로그램들을 실행할 수 있다.

다. 상기 하나 이상의 프로그램들은 하나 이상의 컴퓨터 실행 가능 명령어를 포함할 수 있으며, 상기 컴퓨터 실행 가능 명령어는 프로세서(14)에 의해 실행되는 경우 컴퓨팅 장치(12)로 하여금 예시적인 실시예에 따른 동작들을 수행하도록 구성될 수 있다.

[0136] 컴퓨터 판독 가능 저장 매체(16)는 컴퓨터 실행 가능 명령어 내지 프로그램 코드, 프로그램 데이터 및/또는 다른 적합한 형태의 정보를 저장하도록 구성된다. 컴퓨터 판독 가능 저장 매체(16)에 저장된 프로그램(20)은 프로세서(14)에 의해 실행 가능한 명령어의 집합을 포함한다. 일 실시예에서, 컴퓨터 판독 가능 저장 매체(16)는 메모리(랜덤 액세스 메모리와 같은 휘발성 메모리, 비휘발성 메모리, 또는 이들의 적절한 조합), 하나 이상의 자기 디스크 저장 디바이스들, 광학 디스크 저장 디바이스들, 플래시 메모리 디바이스들, 그 밖에 컴퓨팅 장치(12)에 의해 액세스되고 원하는 정보를 저장할 수 있는 다른 형태의 저장 매체, 또는 이들의 적합한 조합일 수 있다.

[0137] 통신 버스(18)는 프로세서(14), 컴퓨터 판독 가능 저장 매체(16)를 포함하여 컴퓨팅 장치(12)의 다른 다양한 컴포넌트들을 상호 연결한다.

[0138] 컴퓨팅 장치(12)는 또한 하나 이상의 입출력 장치(24)를 위한 인터페이스를 제공하는 하나 이상의 입출력 인터페이스(22) 및 하나 이상의 네트워크 통신 인터페이스(26)를 포함할 수 있다. 입출력 인터페이스(22) 및 네트워크 통신 인터페이스(26)는 통신 버스(18)에 연결된다. 입출력 장치(24)는 입출력 인터페이스(22)를 통해 컴퓨팅 장치(12)의 다른 컴포넌트들에 연결될 수 있다. 예시적인 입출력 장치(24)는 포인팅 장치(마우스 또는 트랙패드 등), 키보드, 터치 입력 장치(터치패드 또는 터치스크린 등), 음성 또는 소리 입력 장치, 다양한 종류의 센서 장치 및/또는 촬영 장치와 같은 입력 장치, 및/또는 디스플레이 장치, 프린터, 스피커 및/또는 네트워크 카드와 같은 출력 장치를 포함할 수 있다. 예시적인 입출력 장치(24)는 컴퓨팅 장치(12)를 구성하는 일 컴포넌트로서 컴퓨팅 장치(12)의 내부에 포함될 수도 있고, 컴퓨팅 장치(12)와는 구별되는 별개의 장치로 컴퓨팅 장치(12)와 연결될 수도 있다.

[0140] 이상에서 본 발명의 대표적인 실시예들을 상세하게 설명하였으나, 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자는 상술한 실시예에 대하여 본 발명의 범주에서 벗어나지 않는 한도 내에서 다양한 변형이 가능함을 이해할 것이다. 그러므로 본 발명의 권리범위는 설명된 실시예에 국한되어 정해져서는 안 되며, 후술하는 특허 청구범위뿐만 아니라 이 특허청구범위와 균등한 것들에 의해 정해져야 한다.

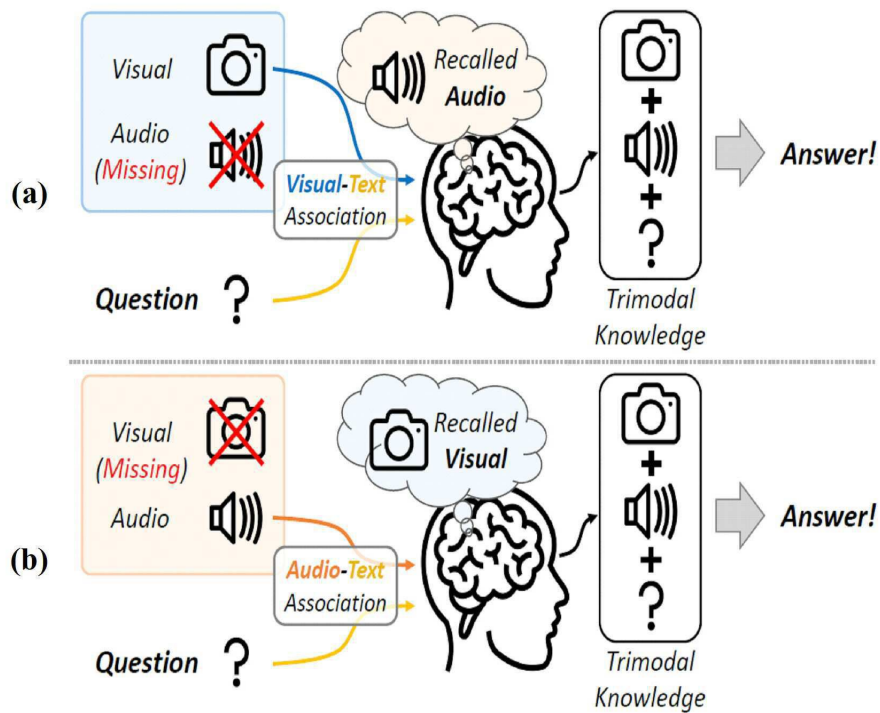
부호의 설명

[0142] 10 : 컴퓨팅 환경
12 : 컴퓨팅 장치
14 : 프로세서
16 : 컴퓨터 판독 가능 저장 매체
18 : 통신 버스
20 : 프로그램
22 : 입출력 인터페이스
24 : 입출력 장치
26 : 네트워크 통신 인터페이스
100 : 오디오 비주얼 질문 응답 장치
102 : 비주얼 인코더
104 : 오디오 인코더
106 : 텍스트 인코더

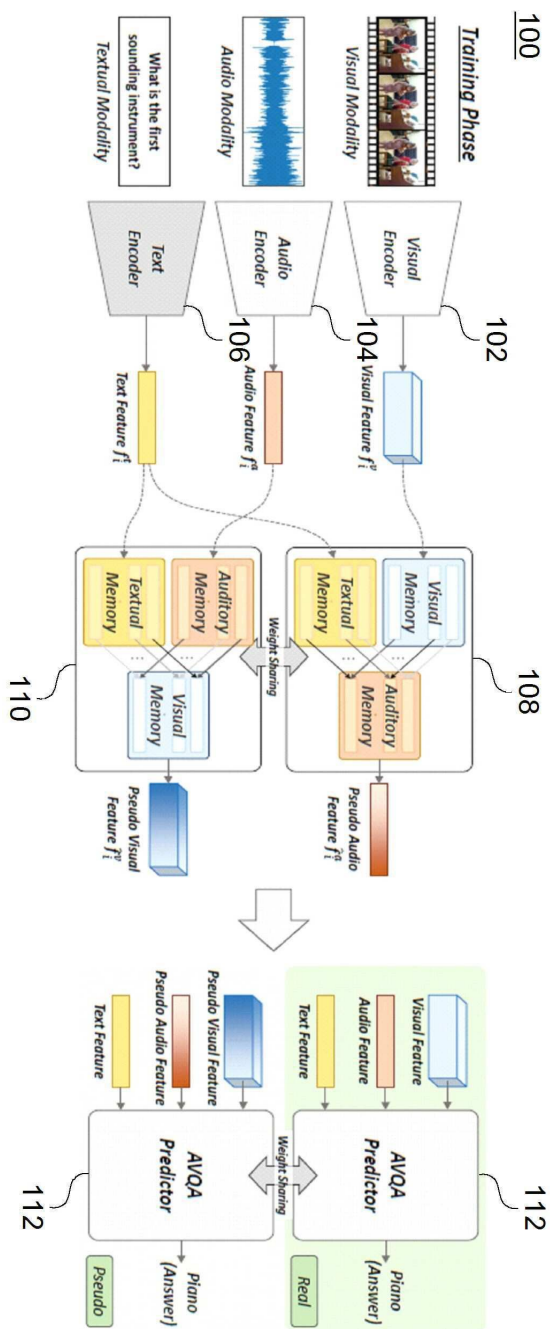
- 108 : 제1 TMA 메모리
- 108a : 제1 비주얼 메모리
- 108b : 제1 텍스트 메모리
- 108c : 제1 오디오 메모리
- 110 : 제2 TMA 메모리
- 110a : 제2 비주얼 메모리
- 110b : 제2 텍스트 메모리
- 110c : 제2 오디오 메모리
- 112 : 예측기

도면

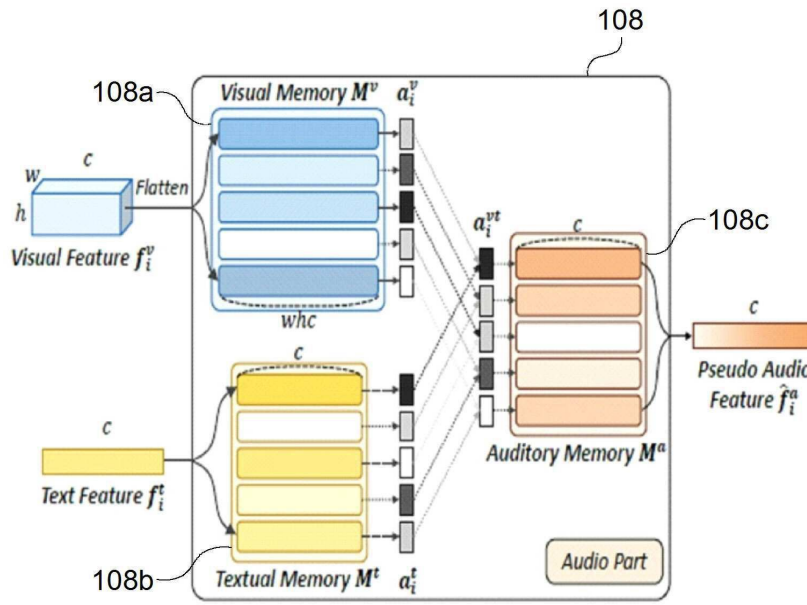
도면1



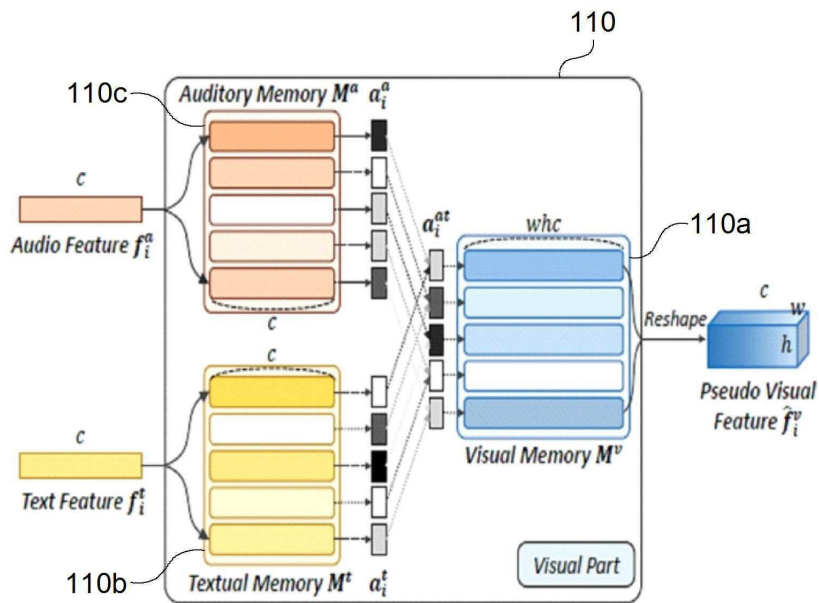
도면2



도면3

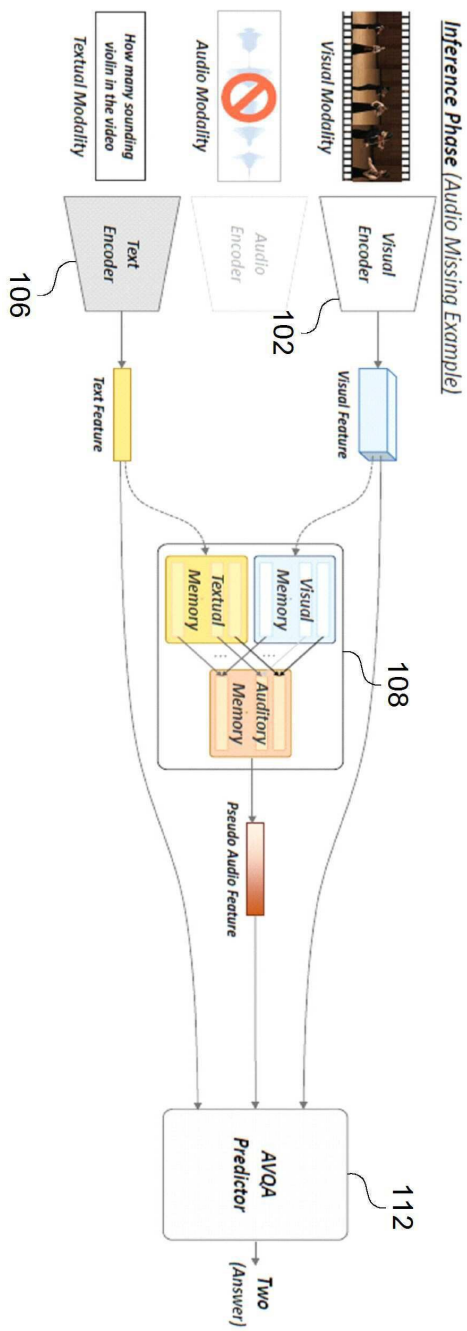


도면4



100

도면5



도면6

10

