(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0013164
(43) 공개일자 2022년02월04일

(51) 국제특허분류(Int. Cl.)
G06N 3/08 (2006.01) *G06F 16/33* (2019.01)
G06F 16/732 (2019.01) *G06F 16/738* (2019.01)
G06F 16/783 (2019.01) *G06N 3/04* (2006.01)
(52) CPC특허분류
G06N 3/08 (2013.01)
G06F 16/3347 (2019.01)
(21) 출원번호 10-2020-0092340
(22) 출원일자 2020년07월24일
심사청구일자 없음

(71) 출원인
주식회사 엔씨소프트
서울특별시 강남구 테헤란로 509 (삼성동)
이화여자대학교 산학협력단
서울특별시 서대문구 이화여대길 52 (대현동, 이화여자대학교)
(72) 발명자
강제원
서울특별시 서대문구 이화여대길 52 아산공학관 524호
김나영
서울특별시 서대문구 이화여대길 52 아산공학관 524호
하성종
경기도 성남시 분당구 대왕판교로 644번길 12
(74) 대리인
함영욱

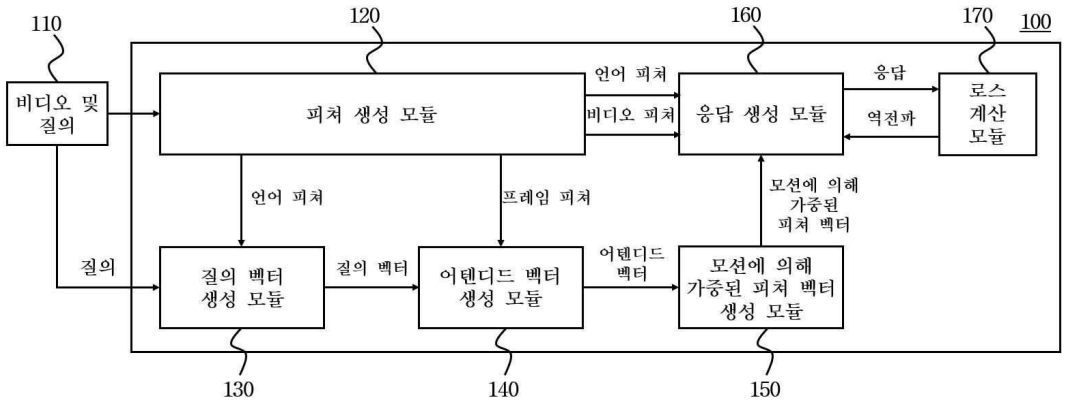
전체 청구항 수 : 총 26 항

(54) 발명의 명칭 **비디오 질의 응답 학습 장치, 비디오 질의 응답 학습 방법, 비디오 질의 응답 장치 및 비디오 질의 응답 방법**

(57) 요 약

본 발명의 일실시예에 따르면, 비디오 질의 응답 학습 장치에 있어서, 적어도 하나의 프로세서를 포함하고, 상기 적어도 하나의 프로세서는, 비디오에 대한 비디오 피쳐(feature) 및 질의에 대한 언어 피쳐(feature)를 생성하고, 인코딩 한 상기 질의를 기초로 질의 벡터를 생성하고, 상기 질의 벡터, 상기 비디오에 포함된 적어도 하나의 프레임에 대한 프레임 피쳐(frame feature) 및 상기 비디오의 비디오 압축 정보를 기초로 모션에 의해 가중된 피쳐 벡터(motion weighted feature vector)를 생성한다.

대 표 도 - 도1



(52) CPC특허분류
G06F 16/732 (2019.01)
G06F 16/738 (2019.01)
G06F 16/786 (2019.01)
G06N 3/04 (2013.01)

명 세 서

청구범위

청구항 1

비디오 질의 응답 학습 장치에 있어서,

적어도 하나의 프로세서를 포함하고,

상기 적어도 하나의 프로세서는,

비디오에 대한 비디오 피쳐(feature) 및 질의에 대한 언어 피쳐(feature)를 생성하고,

인코딩 한 상기 질의를 기초로 질의 벡터를 생성하고,

상기 질의 벡터, 상기 비디오에 포함된 적어도 하나의 프레임에 대한 프레임 피쳐(frame feature) 및 상기 비디오의 비디오 압축 정보를 기초로 모션에 의해 가중된 피쳐 벡터(motion weighted feature vector)를 생성하는 비디오 질의 응답 학습 장치.

청구항 2

제1항에 있어서,

상기 적어도 하나의 프로세서는,

상기 생성한 모션에 의해 가중된 피쳐 벡터(motion weighted feature vector), 상기 비디오 피쳐(feature) 및 상기 언어 피쳐(feature)를 기초로 응답을 생성하도록 뉴럴 네트워크를 학습시키는 비디오 질의 응답 학습 장치.

청구항 3

제1항에 있어서,

상기 적어도 하나의 프로세서는,

제1 인코더를 기초로 상기 질의를 인코딩하여 특징 벡터를 생성하고,

상기 생성한 특징 벡터 및 상기 언어 피쳐(feature)를 기초로 질의 벡터를 생성하는 비디오 질의 응답 학습 장치.

청구항 4

제1항에 있어서,

상기 적어도 하나의 프로세서는,

상기 질의 벡터, 상기 비디오에 포함된 적어도 하나의 프레임에 대한 프레임 피쳐(frame feature) 및 상기 비디오 압축 정보에 포함된 모션 벡터 중 적어도 어느 하나를 기초로 적어도 하나의 어텐디드 벡터(attended vector)를 생성하고,

상기 생성한 적어도 하나의 어텐디드 벡터(attended vector) 및 상기 비디오 압축 정보에 포함된 잔차(residual)를 기초로 모션에 의해 가중된 피쳐 벡터(Motion weighted feature vector)를 생성하는 비디오 질의 응답 학습 장치.

청구항 5

제4항에 있어서,

상기 적어도 하나의 프로세서는,

상기 질의 벡터 및 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피쳐(frame feature)를 기초

로 제1 어텐션 벡터(attention vector)를 생성하고,

상기 생성한 제1 어텐션 벡터(attention vector)와 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature)를 기초로 제1 어텐디드 벡터(attended vector)를 생성하고,

상기 질의 벡터 및 피처 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피처(frame feature)를 기초로 제2 어텐션 벡터(attention vector)를 생성하고,

상기 생성한 제2 어텐션 벡터(attention vector), 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature) 및 피처 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피처(frame feature)를 기초로 제2 어텐디드 벡터(attended vector)를 생성하는 비디오 질의 응답 학습 장치.

청구항 6

제5항에 있어서,

상기 적어도 하나의 프로세서는,

상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature) 및 상기 비디오 압축 정보에 포함된 모션 벡터를 기초로 피처 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피처(frame feature)를 생성하는 비디오 질의 응답 학습 장치.

청구항 7

제5항에 있어서,

상기 적어도 하나의 프로세서는,

상기 비디오 압축 정보에 포함된 잔차(residual)를 기초로 상기 제1 어텐디드 벡터(attended vector) 및 상기 제2 어텐디드 벡터(attended vector)의 가중치 합을 결정하고,

제2 인코더를 기초로 상기 결정한 가중치 합을 인코딩하여 모션에 의해 가중된 피처 벡터(Motion weighted feature vector)를 생성하는 비디오 질의 응답 학습 장치.

청구항 8

비디오에 대한 비디오 피처(feature) 및 질의에 대한 언어 피처(feature)를 생성하는 동작;

인코딩 한 상기 질의를 기초로 질의 벡터를 생성하는 동작; 및

상기 질의 벡터, 상기 비디오에 포함된 적어도 하나의 프레임에 대한 프레임 피처(frame feature) 및 상기 비디오의 비디오 압축 정보를 기초로 모션에 의해 가중된 피처 벡터(motion weighted feature vector)를 생성하는 동작

을 포함하는 비디오 질의 응답 학습 방법.

청구항 9

제8항에 있어서,

상기 비디오 질의 응답 학습 방법은,

상기 생성한 모션에 의해 가중된 피처 벡터(motion weighted feature vector), 상기 비디오 피처(feature) 및 상기 언어 피처(feature)를 기초로 응답을 생성하도록 뉴럴 네트워크를 학습시키는 동작

을 더 포함하는 비디오 질의 응답 학습 방법.

청구항 10

제8항에 있어서,

상기 질의 벡터를 생성하는 동작은,

제1 인코더를 기초로 상기 질의를 인코딩하여 특징 벡터를 생성하는 동작; 및

상기 생성한 특징 벡터 및 상기 언어 피처(feature)를 기초로 질의 벡터를 생성하는 동작

을 포함하는 비디오 질의 응답 학습 방법.

청구항 11

제8항에 있어서,

상기 모션에 의해 가중된 피처 벡터(motion weighted feature vector)를 생성하는 동작은,

상기 질의 벡터, 상기 비디오에 포함된 적어도 하나의 프레임에 대한 프레임 피처(frame feature) 및 상기 비디오 압축 정보에 포함된 모션 벡터 중 적어도 어느 하나를 기초로 적어도 하나의 어텐디드 벡터(attended vector)를 생성하는 동작; 및

상기 생성한 적어도 하나의 어텐디드 벡터(attended vector) 및 상기 비디오 압축 정보에 포함된 잔차(residual)를 기초로 모션에 의해 가중된 피처 벡터(Motion weighted feature vector)를 생성하는 동작

을 포함하는 비디오 질의 응답 학습 방법.

청구항 12

제11항에 있어서,

상기 적어도 하나의 어텐디드 벡터(attended vector)를 생성하는 동작은,

상기 질의 벡터 및 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature)를 기초로 제1 어텐션 벡터(attention vector)를 생성하는 동작;

상기 생성한 제1 어텐션 벡터(attention vector)와 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature)를 기초로 제1 어텐디드 벡터(attended vector)를 생성하는 동작;

상기 질의 벡터 및 피처 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피처(frame feature)를 기초로 제2 어텐션 벡터(attention vector)를 생성하는 동작; 및

상기 생성한 제2 어텐션 벡터(attention vector), 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature) 및 피처 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피처(frame feature)를 기초로 제2 어텐디드 벡터(attended vector)를 생성하는 동작

을 포함하는 비디오 질의 응답 학습 방법.

청구항 13

제12항에 있어서,

상기 제2 어텐션 벡터(attention vector)를 생성하는 동작은,

상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature) 및 상기 비디오 압축 정보에 포함된 모션 벡터를 기초로 피처 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피처(frame feature)를 생성하는 동작

을 포함하는 비디오 질의 응답 학습 방법.

청구항 14

제12항에 있어서,

상기 모션에 의해 가중된 피처 벡터(Motion weighted feature vector)를 생성하는 동작은,

상기 비디오 압축 정보에 포함된 잔차(residual)를 기초로 상기 제1 어텐디드 벡터(attended vector) 및 상기 제2 어텐디드 벡터(attended vector)의 가중치 합을 결정하는 동작; 및

제2 인코더를 기초로 상기 결정한 가중치 합을 인코딩하여 모션에 의해 가중된 피처 벡터(Motion weighted

feature vector)를 생성하는 동작

을 포함하는 비디오 질의 응답 학습 방법.

청구항 15

비디오 질의 응답 장치에 있어서,

적어도 하나의 프로세서를 포함하고,

상기 적어도 하나의 프로세서는,

비디오에 대한 비디오 피처(feature) 및 질의에 대한 언어 피처(feature)를 생성하고,

인코딩 한 상기 질의를 기초로 질의 벡터를 생성하고,

상기 질의 벡터, 상기 비디오에 포함된 적어도 하나의 프레임에 대한 프레임 피처(frame feature) 및 상기 비디오의 비디오 압축 정보를 기초로 모션에 의해 가중된 피처 벡터(motion weighted feature vector)를 생성하고,

상기 생성한 모션에 의해 가중된 피처 벡터(motion weighted feature vector), 상기 비디오 피처(feature) 및 상기 언어 피처(feature)를 기초로 학습이 완료된 뉴럴 네트워크를 이용하여 응답을 생성하는 비디오 질의 응답 장치.

청구항 16

제15항에 있어서,

상기 적어도 하나의 프로세서는,

학습이 완료된 제1 인코더를 기초로 상기 질의를 인코딩하여 특징 벡터를 생성하고,

상기 생성한 특징 벡터 및 상기 언어 피처(feature)를 기초로 질의 벡터를 생성하는 비디오 질의 응답 장치.

청구항 17

제15항에 있어서,

상기 적어도 하나의 프로세서는,

상기 질의 벡터, 상기 비디오에 포함된 적어도 하나의 프레임에 대한 프레임 피처(frame feature) 및 상기 비디오 압축 정보에 포함된 모션 벡터 중 적어도 어느 하나를 기초로 적어도 하나의 어텐디드 벡터(attended vector)를 생성하고,

상기 생성한 적어도 하나의 어텐디드 벡터(attended vector) 및 상기 비디오 압축 정보에 포함된 잔차(residual)를 기초로 모션에 의해 가중된 피처 벡터(Motion weighted feature vector)를 생성하는 비디오 질의 응답 장치.

청구항 18

제17항에 있어서,

상기 적어도 하나의 프로세서는,

상기 질의 벡터 및 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature)를 기초로 제1 어텐션 벡터(attention vector)를 생성하고,

상기 생성한 제1 어텐션 벡터(attention vector)와 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature)를 기초로 제1 어텐디드 벡터(attended vector)를 생성하고,

상기 질의 벡터 및 피처 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피처(frame feature)를 기초로 제2 어텐션 벡터(attention vector)를 생성하고,

상기 생성한 제2 어텐션 벡터(attention vector), 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature) 및 피처 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피처(frame feature)를 기초로 제2 어텐디드 벡터(attended vector)를 생성하는 비디오 질의 응답

장치.

청구항 19

제18항에 있어서,

상기 적어도 하나의 프로세서는,

상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature) 및 상기 비디오 압축 정보에 포함된 모션 벡터를 기초로 피처 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피처(frame feature)를 생성하는 비디오 질의 응답 장치.

청구항 20

제18항에 있어서,

상기 적어도 하나의 프로세서는,

상기 비디오 압축 정보에 포함된 잔차(residual)를 기초로 상기 제1 어텐디드 벡터(attended vector) 및 상기 제2 어텐디드 벡터(attended vector)의 가중치 합을 결정하고,

학습이 완료된 제2 인코더를 기초로 상기 결정한 가중치 합을 인코딩하여 모션에 의해 가중된 피처 벡터(Motion weighted feature vector)를 생성하는 비디오 질의 응답 장치.

청구항 21

비디오에 대한 비디오 피처(feature) 및 질의에 대한 언어 피처(feature)를 생성하는 동작;

인코딩 한 상기 질의를 기초로 질의 벡터를 생성하는 동작;

상기 질의 벡터, 상기 비디오에 포함된 적어도 하나의 프레임에 대한 프레임 피처(frame feature) 및 상기 비디오의 비디오 압축 정보를 기초로 모션에 의해 가중된 피처 벡터(motion weighted feature vector)를 생성하는 동작; 및

상기 생성한 모션에 의해 가중된 피처 벡터(motion weighted feature vector), 상기 비디오 피처(feature) 및 상기 언어 피처(feature)를 기초로 학습이 완료된 뉴럴 네트워크를 이용하여 응답을 생성하는 비디오 질의 응답 방법.

청구항 22

제21항에 있어서,

상기 질의 벡터를 생성하는 동작은,

학습이 완료된 제1 인코더를 기초로 상기 질의를 인코딩하여 특징 벡터를 생성하는 동작; 및

상기 생성한 특징 벡터 및 상기 언어 피처(feature)를 기초로 질의 벡터를 생성하는 동작

을 포함하는 비디오 질의 응답 방법.

청구항 23

제21항에 있어서,

상기 모션에 의해 가중된 피처 벡터(motion weighted feature vector)를 생성하는 동작은,

상기 질의 벡터, 상기 비디오에 포함된 적어도 하나의 프레임에 대한 프레임 피처(frame feature) 및 상기 비디오 압축 정보에 포함된 모션 벡터 중 적어도 어느 하나를 기초로 적어도 하나의 어텐디드 벡터(attended vector)를 생성하는 동작; 및

상기 생성한 적어도 하나의 어텐디드 벡터(attended vector) 및 상기 비디오 압축 정보에 포함된 잔차(residual)를 기초로 모션에 의해 가중된 피처 벡터(Motion weighted feature vector)를 생성하는 동작

을 포함하는 비디오 질의 응답 방법.

청구항 24

제23항에 있어서,

상기 적어도 하나의 어텐디드 벡터(attended vector)를 생성하는 동작은,

상기 질의 벡터 및 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature)를 기초로 제1 어텐션 벡터(attention vector)를 생성하는 동작;

상기 생성한 제1 어텐션 벡터(attention vector)와 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature)를 기초로 제1 어텐디드 벡터(attended vector)를 생성하는 동작;

상기 질의 벡터 및 피쳐 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피처(frame feature)를 기초로 제2 어텐션 벡터(attention vector)를 생성하는 동작; 및

상기 생성한 제2 어텐션 벡터(attention vector), 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature) 및 피쳐 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피처(frame feature)를 기초로 제2 어텐디드 벡터(attended vector)를 생성하는 동작

을 포함하는 비디오 질의 응답 방법.

청구항 25

제24항에 있어서,

상기 제2 어텐션 벡터(attention vector)를 생성하는 동작은,

상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature) 및 상기 비디오 압축 정보에 포함된 모션 벡터를 기초로 피쳐 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피처(frame feature)를 생성하는 동작

을 포함하는 비디오 질의 응답 방법.

청구항 26

제24항에 있어서,

상기 모션에 의해 가중된 피쳐 벡터(Motion weighted feature vector)를 생성하는 동작은,

상기 비디오 압축 정보에 포함된 잔차(residual)를 기초로 상기 제1 어텐디드 벡터(attended vector) 및 상기 제2 어텐디드 벡터(attended vector)의 가중치 합을 결정하는 동작; 및

학습이 완료된 제2 인코더를 기초로 상기 결정한 가중치 합을 인코딩하여 모션에 의해 가중된 피쳐 벡터(Motion weighted feature vector)를 생성하는 동작

을 포함하는 비디오 질의 응답 방법.

발명의 설명

기 술 분 야

[0001] 아래의 실시예들은 비디오 질의 응답 학습 장치, 비디오 질의 응답 학습 방법, 비디오 질의 응답 장치 및 비디오 질의 응답 방법에 관한 것이다.

배 경 기 술

[0002] 머신 러닝(machine learning)은 인공 지능의 한 분야로, 패턴인식과 컴퓨터 학습 이론의 연구로부터 진화한 분야이며, 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야를 말한다. 머신 러닝의 핵심은 표현(representation)과 일반화(generalization)에 있다. 표현이란 데이터의 평가이며, 일반화란 아직 알 수 없는 데이터에 대한 처리이다. 이는 전산 학습 이론 분야이기도 하다.

[0003] 딥 러닝(deep learning)은 여러 비선형 변환기법의 조합을 통해 높은 수준의 추상화를 시도하는 기계학습(machine learning) 알고리즘의 집합으로 정의되며, 큰 틀에서 사람의 사고방식을 컴퓨터에게 가르치는 기계학

습의 한 분야라고 이야기할 수 있다.

[0004] 시각적 질의응답(Visual Question Answering)은 사진을 보고 다양한 질문에 대한 답을 할 수 있는 시각언어 멀티 모달 체계이자 다양한 종류의 질문에 대해 답을 하는 멀티태스킹(Multitasking) 체계이다.

[0005] 비디오 질의응답(Video Question Answering)은 비디오에 대해 다양한 질문에 대한 답을 할 수 있는 시각언어 멀티 모달 체계이자 다양한 종류의 질문에 대해 답을 하는 멀티태스킹(Multitasking) 체계이다.

발명의 내용

해결하려는 과제

[0006] 본 발명의 실시예에 따르면, 모션에 의해 가중된 피처 벡터(motion weighted feature vector), 언어 피처(feature) 및 비디오 피처(feature)를 기초로 응답을 생성하도록 뉴럴 네트워크를 학습시킬 수 있는 비디오 질의 응답 학습 장치 및 비디오 질의 응답 학습 방법을 제공할 수 있다.

[0007] 또한, 본 발명의 다른 실시예에 따르면, 질의를 인코딩한 특징 벡터 및 언어 피처를 기초로 질의 벡터를 생성할 수 있는 비디오 질의 응답 학습 장치 및 비디오 질의 응답 학습 방법을 제공할 수 있다.

[0008] 또한, 본 발명의 또 다른 실시예에 따르면, 질의 벡터, 프레임 피처(frame feature) 및 비디오 압축 정보에 포함된 모션 벡터를 기초로 적어도 하나의 어텐디드 벡터(attended vector)를 생성할 수 있는 비디오 질의 응답 학습 장치 및 비디오 질의 응답 학습 방법을 제공할 수 있다.

[0009] 또한, 본 발명의 또 다른 실시예에 따르면, 적어도 하나의 어텐디드 벡터(attended vector) 및 비디오 압축 정보에 포함된 잔차(residual)를 기초로 모션에 의해 가중된 피처 벡터(Motion weighted feature vector)를 생성할 수 있는 비디오 질의 응답 학습 장치 및 비디오 질의 응답 학습 방법을 제공할 수 있다.

[0010] 또한, 본 발명의 또 다른 실시예에 따르면, 모션에 의해 가중된 피처 벡터(motion weighted feature vector), 언어 피처(feature) 및 비디오 피처(feature)를 기초로 응답을 생성할 수 있는 비디오 질의 응답 장치 및 비디오 질의 응답 방법을 제공할 수 있다.

[0011] 또한, 본 발명의 또 다른 실시예에 따르면, 질의를 인코딩한 특징 벡터 및 언어 피처를 기초로 질의 벡터를 생성할 수 있는 비디오 질의 응답 장치 및 비디오 질의 응답 방법을 제공할 수 있다.

[0012] 또한, 본 발명의 또 다른 실시예에 따르면, 질의 벡터, 프레임 피처(frame feature) 및 비디오 압축 정보에 포함된 모션 벡터를 기초로 적어도 하나의 어텐디드 벡터(attended vector)를 생성할 수 있는 비디오 질의 응답 장치 및 비디오 질의 응답 방법을 제공할 수 있다.

[0013] 또한, 본 발명의 또 다른 실시예에 따르면, 적어도 하나의 어텐디드 벡터(attended vector) 및 비디오 압축 정보에 포함된 잔차(residual)를 기초로 모션에 의해 가중된 피처 벡터(Motion weighted feature vector)를 생성할 수 있는 비디오 질의 응답 장치 및 비디오 질의 응답 방법을 제공할 수 있다.

과제의 해결 수단

[0014] 본 발명의 일실시예에 따르면, 비디오 질의 응답 학습 장치에 있어서, 적어도 하나의 프로세서를 포함하고, 상기 적어도 하나의 프로세서는, 비디오에 대한 비디오 피처(feature) 및 질의에 대한 언어 피처(feature)를 생성하고, 인코딩 한 상기 질의를 기초로 질의 벡터를 생성하고, 상기 질의 벡터, 상기 비디오에 포함된 적어도 하나의 프레임에 대한 프레임 피처(frame feature) 및 상기 비디오의 비디오 압축 정보를 기초로 모션에 의해 가중된 피처 벡터(motion weighted feature vector)를 생성한다.

[0015] 또한, 상기 적어도 하나의 프로세서는, 상기 생성한 모션에 의해 가중된 피처 벡터(motion weighted feature vector), 상기 비디오 피처(feature) 및 상기 언어 피처(feature)를 기초로 응답을 생성하도록 뉴럴 네트워크를 학습시킬 수 있다.

[0016] 또한, 상기 적어도 하나의 프로세서는, 제1 인코더를 기초로 상기 질의를 인코딩하여 특징 벡터를 생성하고, 상기 생성한 특징 벡터 및 상기 언어 피처(feature)를 기초로 질의 벡터를 생성할 수 있다.

[0017] 또한, 상기 적어도 하나의 프로세서는, 상기 질의 벡터, 상기 비디오에 포함된 적어도 하나의 프레임에 대한 프레임 피처(frame feature) 및 상기 비디오 압축 정보에 포함된 모션 벡터 중 적어도 어느 하나를 기초로 적어도 하나의 어텐디드 벡터(attended vector)를 생성하고, 상기 생성한 적어도 하나의 어텐디드 벡터(attended

vector) 및 상기 비디오 압축 정보에 포함된 잔차(residual)를 기초로 모션에 의해 가중된 피쳐 벡터(Motion weighted feature vector)를 생성할 수 있다.

[0018] 또한, 상기 적어도 하나의 프로세서는, 상기 질의 벡터 및 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피쳐(frame feature)를 기초로 제1 어텐션 벡터(attention vector)를 생성하고, 상기 생성한 제1 어텐션 벡터(attention vector)와 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피쳐(frame feature)를 기초로 제1 어텐디드 벡터(attended vector)를 생성하고, 상기 질의 벡터 및 피쳐 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피쳐(frame feature)를 기초로 제2 어텐션 벡터(attention vector)를 생성하고, 상기 생성한 제2 어텐션 벡터(attention vector), 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피쳐(frame feature) 및 피쳐 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피쳐(frame feature)를 기초로 제2 어텐디드 벡터(attended vector)를 생성할 수 있다.

[0019] 또한, 상기 적어도 하나의 프로세서는, 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피쳐(frame feature) 및 상기 비디오 압축 정보에 포함된 모션 벡터를 기초로 피쳐 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피쳐(frame feature)를 생성할 수 있다.

[0020] 또한, 상기 적어도 하나의 프로세서는, 상기 비디오 압축 정보에 포함된 잔차(residual)를 기초로 상기 제1 어텐디드 벡터(attended vector) 및 상기 제2 어텐디드 벡터(attended vector)의 가중치 합을 결정하고, 제2 인코더를 기초로 상기 결정한 가중치 합을 인코딩하여 모션에 의해 가중된 피쳐 벡터(Motion weighted feature vector)를 생성할 수 있다.

[0021] 본 발명의 다른 실시예에 따르면, 비디오에 대한 비디오 피쳐(feature) 및 질의에 대한 언어 피쳐(feature)를 생성하는 동작, 인코딩 한 상기 질의를 기초로 질의 벡터를 생성하는 동작 및 상기 질의 벡터, 상기 비디오에 포함된 적어도 하나의 프레임에 대한 프레임 피쳐(frame feature) 및 상기 비디오의 비디오 압축 정보를 기초로 모션에 의해 가중된 피쳐 벡터(motion weighted feature vector)를 생성하는 동작을 포함한다.

[0022] 또한, 상기 비디오 질의 응답 학습 방법은, 상기 생성한 모션에 의해 가중된 피쳐 벡터(motion weighted feature vector), 상기 비디오 피쳐(feature) 및 상기 언어 피쳐(feature)를 기초로 응답을 생성하도록 뉴럴 네트워크를 학습시키는 동작을 더 포함할 수 있다.

[0023] 또한, 상기 질의 벡터를 생성하는 동작은, 제1 인코더를 기초로 상기 질의를 인코딩하여 특징 벡터를 생성하는 동작 및 상기 생성한 특징 벡터 및 상기 언어 피쳐(feature)를 기초로 질의 벡터를 생성하는 동작을 포함할 수 있다.

[0024] 또한, 상기 모션에 의해 가중된 피쳐 벡터(motion weighted feature vector)를 생성하는 동작은, 상기 질의 벡터, 상기 비디오에 포함된 적어도 하나의 프레임에 대한 프레임 피쳐(frame feature) 및 상기 비디오 압축 정보에 포함된 모션 벡터 중 적어도 어느 하나를 기초로 적어도 하나의 어텐디드 벡터(attended vector)를 생성하는 동작 및 상기 생성한 적어도 하나의 어텐디드 벡터(attended vector) 및 상기 비디오 압축 정보에 포함된 잔차(residual)를 기초로 모션에 의해 가중된 피쳐 벡터(Motion weighted feature vector)를 생성하는 동작을 포함할 수 있다.

[0025] 또한, 상기 적어도 하나의 어텐디드 벡터(attended vector)를 생성하는 동작은, 상기 질의 벡터 및 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피쳐(frame feature)를 기초로 제1 어텐션 벡터(attention vector)를 생성하는 동작, 상기 생성한 제1 어텐션 벡터(attention vector)와 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피쳐(frame feature)를 기초로 제1 어텐디드 벡터(attended vector)를 생성하는 동작, 상기 질의 벡터 및 피쳐 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피쳐(frame feature)를 기초로 제2 어텐션 벡터(attention vector)를 생성하는 동작 및 상기 생성한 제2 어텐션 벡터(attention vector), 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피쳐(frame feature) 및 피쳐 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피쳐(frame feature)를 기초로 제2 어텐디드 벡터(attended vector)를 생성하는 동작을 포함할 수 있다.

[0026] 또한, 상기 제2 어텐션 벡터(attention vector)를 생성하는 동작은, 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피쳐(frame feature) 및 상기 비디오 압축 정보에 포함된 모션 벡터를 기초로 피쳐 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피쳐(frame feature)를 생성하는 동작을 포함할 수 있다.

[0027] 또한, 상기 모션에 의해 가중된 피쳐 벡터(Motion weighted feature vector)를 생성하는 동작은, 상기 비디오 압축 정보에 포함된 잔차(residual)를 기초로 상기 제1 어텐디드 벡터(attended vector) 및 상기 제2 어텐디드 벡터(attended vector)의 가중치 합을 결정하는 동작 및 제2 인코더를 기초로 상기 결정한 가중치 합을 인코딩하여 모션에 의해 가중된 피쳐 벡터(Motion weighted feature vector)를 생성하는 동작을 포함할 수 있다.

[0028] 본 발명의 또 다른 실시예에 따르면, 비디오 질의 응답 장치에 있어서, 적어도 하나의 프로세서를 포함하고, 상기 적어도 하나의 프로세서는, 비디오에 대한 비디오 피쳐(feature) 및 질의에 대한 언어 피쳐(feature)를 생성하고, 인코딩 한 상기 질의를 기초로 질의 벡터를 생성하고, 상기 질의 벡터, 상기 비디오에 포함된 적어도 하나의 프레임에 대한 프레임 피쳐(frame feature) 및 상기 비디오의 비디오 압축 정보를 기초로 모션에 의해 가중된 피쳐 벡터(motion weighted feature vector)를 생성하고, 상기 생성한 모션에 의해 가중된 피쳐 벡터(motion weighted feature vector), 상기 비디오 피쳐(feature) 및 상기 언어 피쳐(feature)를 기초로 학습이 완료된 뉴럴 네트워크를 이용하여 응답을 생성한다.

[0029] 또한, 상기 적어도 하나의 프로세서는, 학습이 완료된 제1 인코더를 기초로 상기 질의를 인코딩하여 특징 벡터를 생성하고, 상기 생성한 특징 벡터 및 상기 언어 피쳐(feature)를 기초로 질의 벡터를 생성할 수 있다.

[0030] 또한, 상기 적어도 하나의 프로세서는, 상기 질의 벡터, 상기 비디오에 포함된 적어도 하나의 프레임에 대한 프레임 피쳐(frame feature) 및 상기 비디오 압축 정보에 포함된 모션 벡터 중 적어도 어느 하나를 기초로 적어도 하나의 어텐디드 벡터(attended vector)를 생성하고, 상기 생성한 적어도 하나의 어텐디드 벡터(attended vector) 및 상기 비디오 압축 정보에 포함된 잔차(residual)를 기초로 모션에 의해 가중된 피쳐 벡터(Motion weighted feature vector)를 생성할 수 있다.

[0031] 또한, 상기 적어도 하나의 프로세서는, 상기 질의 벡터 및 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피쳐(frame feature)를 기초로 제1 어텐션 벡터(attention vector)를 생성하고, 상기 생성한 제1 어텐션 벡터(attention vector)와 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피쳐(frame feature)를 기초로 제1 어텐디드 벡터(attended vector)를 생성하고, 상기 질의 벡터 및 피쳐 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피쳐(frame feature)를 기초로 제2 어텐션 벡터(attention vector)를 생성하고, 상기 생성한 제2 어텐션 벡터(attention vector), 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피쳐(frame feature) 및 피쳐 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피쳐(frame feature)를 기초로 제2 어텐디드 벡터(attended vector)를 생성할 수 있다.

[0032] 또한, 상기 적어도 하나의 프로세서는, 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피쳐(frame feature) 및 상기 비디오 압축 정보에 포함된 모션 벡터를 기초로 피쳐 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피쳐(frame feature)를 생성할 수 있다.

[0033] 또한, 상기 적어도 하나의 프로세서는, 상기 비디오 압축 정보에 포함된 잔차(residual)를 기초로 상기 제1 어텐디드 벡터(attended vector) 및 상기 제2 어텐디드 벡터(attended vector)의 가중치 합을 결정하고, 학습이 완료된 제2 인코더를 기초로 상기 결정한 가중치 합을 인코딩하여 모션에 의해 가중된 피쳐 벡터(Motion weighted feature vector)를 생성할 수 있다.

[0034] 본 발명의 또 다른 실시예에 따르면, 비디오에 대한 비디오 피쳐(feature) 및 질의에 대한 언어 피쳐(feature)를 생성하는 동작, 인코딩 한 상기 질의를 기초로 질의 벡터를 생성하는 동작, 상기 질의 벡터, 상기 비디오에 포함된 적어도 하나의 프레임에 대한 프레임 피쳐(frame feature) 및 상기 비디오의 비디오 압축 정보를 기초로 모션에 의해 가중된 피쳐 벡터(motion weighted feature vector)를 생성하는 동작 및 상기 생성한 모션에 의해 가중된 피쳐 벡터(motion weighted feature vector), 상기 비디오 피쳐(feature) 및 상기 언어 피쳐(feature)를 기초로 학습이 완료된 뉴럴 네트워크를 이용하여 응답을 생성한다.

[0035] 또한, 상기 질의 벡터를 생성하는 동작은, 학습이 완료된 제1 인코더를 기초로 상기 질의를 인코딩하여 특징 벡터를 생성하는 동작 및 상기 생성한 특징 벡터 및 상기 언어 피쳐(feature)를 기초로 질의 벡터를 생성하는 동작을 포함할 수 있다.

[0036] 또한, 상기 모션에 의해 가중된 피쳐 벡터(motion weighted feature vector)를 생성하는 동작은, 상기 질의 벡터, 상기 비디오에 포함된 적어도 하나의 프레임에 대한 프레임 피쳐(frame feature) 및 상기 비디오 압축 정보에 포함된 모션 벡터 중 적어도 어느 하나를 기초로 적어도 하나의 어텐디드 벡터(attended vector)를 생성하는 동작 및 상기 생성한 적어도 하나의 어텐디드 벡터(attended vector) 및 상기 비디오 압축 정보에 포함된 잔차

(residual)를 기초로 모션에 의해 가중된 피쳐 벡터(Motion weighted feature vector)를 생성하는 동작을 포함할 수 있다.

[0037] 또한, 상기 적어도 하나의 어텐디드 벡터(attended vector)를 생성하는 동작은, 상기 질의 벡터 및 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피쳐(frame feature)를 기초로 제1 어텐션 벡터(attention vector)를 생성하는 동작, 상기 생성한 제1 어텐션 벡터(attention vector)와 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피쳐(frame feature)를 기초로 제1 어텐디드 벡터(attended vector)를 생성하는 동작, 상기 질의 벡터 및 피쳐 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피쳐(frame feature)를 기초로 제2 어텐션 벡터(attention vector)를 생성하는 동작 및 상기 생성한 제2 어텐션 벡터(attention vector), 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피쳐(frame feature) 및 피쳐 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피쳐(frame feature)를 기초로 제2 어텐디드 벡터(attended vector)를 생성할 수 있다.

[0038] 또한, 상기 제2 어텐션 벡터(attention vector)를 생성하는 동작은, 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피쳐(frame feature) 및 상기 비디오 압축 정보에 포함된 모션 벡터를 기초로 피쳐 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피쳐(frame feature)를 생성하는 동작을 포함할 수 있다.

[0039] 또한, 상기 모션에 의해 가중된 피쳐 벡터(Motion weighted feature vector)를 생성하는 동작은, 상기 비디오 압축 정보에 포함된 잔차(residual)를 기초로 상기 제1 어텐디드 벡터(attended vector) 및 상기 제2 어텐디드 벡터(attended vector)의 가중치 합을 결정하는 동작 및 학습이 완료된 제2 인코더를 기초로 상기 결정한 가중치 합을 인코딩하여 모션에 의해 가중된 피쳐 벡터(Motion weighted feature vector)를 생성하는 동작을 포함할 수 있다.

발명의 효과

[0040] 본 발명의 일실시예에 따르면, 모션에 의해 가중된 피쳐 벡터(motion weighted feature vector), 언어 피쳐(feature) 및 비디오 피쳐(feature)를 기초로 응답을 생성하도록 뉴럴 네트워크를 학습시킬 수 있는 효과가 있다.

[0041] 또한, 질의를 인코딩한 특징 벡터 및 언어 피쳐를 기초로 질의 벡터를 생성할 수 있는 효과가 있다.

[0042] 또한, 질의 벡터, 프레임 피쳐(frame feature) 및 비디오 압축 정보에 포함된 모션 벡터를 기초로 적어도 하나의 어텐디드 벡터(attended vector)를 생성할 수 있는 효과가 있다.

[0043] 또한, 적어도 하나의 어텐디드 벡터(attended vector) 및 비디오 압축 정보에 포함된 잔차(residual)를 기초로 모션에 의해 가중된 피쳐 벡터(Motion weighted feature vector)를 생성할 수 있는 효과가 있다.

[0044] 또한, 모션에 의해 가중된 피쳐 벡터(motion weighted feature vector), 언어 피쳐(feature) 및 비디오 피쳐(feature)를 기초로 응답을 생성할 수 있는 효과가 있다.

도면의 간단한 설명

[0045] 도 1은 일실시예에 따른 비디오 질의 응답 학습 장치의 구성을 나타내는 도면이다.

도 2는 일실시예에 따른 비디오 질의 응답 학습 방법을 나타내는 플로우 차트이다.

도 3은 일실시예에 따른 비디오 질의 응답 장치의 구성을 나타내는 도면이다.

도 4는 일실시예에 따른 비디오 질의 응답 방법을 나타내는 플로우 차트이다.

도 5는 본 발명의 일실시예를 구현하기 위한 예시적인 컴퓨터 시스템의 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0046] 본 명세서에 개시되어 있는 본 발명의 개념에 따른 실시 예들에 대해서 특정한 구조적 또는 기능적 설명들은 단지 본 발명의 개념에 따른 실시 예들을 설명하기 위한 목적으로 예시된 것으로서, 본 발명의 개념에 따른 실시 예들은 다양한 형태들로 실시될 수 있으며 본 명세서에 설명된 실시 예들에 한정되지 않는다.

[0047] 본 발명의 개념에 따른 실시 예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시

예들을 도면에 예시하고 본 명세서에 상세하게 설명하고자 한다. 그러나, 이는 본 발명의 개념에 따른 실시 예들을 특정한 개시 형태들에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물, 또는 대체물을 포함한다.

[0048] 제1 또는 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 목적으로만, 예컨대 본 발명의 개념에 따른 권리 범위로부터 이탈되지 않은 채, 제1구성요소는 제2구성요소로 명명될 수 있고, 유사하게 제2구성요소는 제1구성요소로도 명명될 수 있다.

[0049] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다. 구성요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0050] 본 명세서에서 사용한 용어는 단지 특정한 실시 예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다.

[0051] 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 설명된 특징, 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0052] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미가 있다.

[0053] 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0054] 이하의 설명에서 동일한 식별 기호는 동일한 구성을 의미하며, 불필요한 중복적인 설명 및 공지 기술에 대한 설명은 생략하기로 한다.

[0055] 본 발명의 실시 예에서 '통신', '통신망' 및 '네트워크'는 동일한 의미로 사용될 수 있다. 상기 세 용어들은, 파일을 사용자 단말, 다른 사용자들의 단말 및 다운로드 서버 사이에서 송수신할 수 있는 유무선의 근거리 및 광역 데이터 송수신망을 의미한다.

[0056] 이하, 첨부한 도면을 참조하여 본 발명의 바람직한 실시 예를 설명함으로써, 본 발명을 상세히 설명한다.

[0057] 도 1은 일실시예에 따른 비디오 질의 응답 학습 장치의 구성을 나타내는 도면이다.

[0058] 도 1을 참조하면, 일실시예에 따른 비디오 질의 응답 학습 장치(100)는 피쳐(feature, 이하 '피쳐'라 한다) 생성 모듈(120), 질의 벡터 생성 모듈(130), 어텐디드 벡터(attended vector, 이하 '어텐디드 벡터'라 한다) 생성 모듈(140), 모션에 의해 가중된 피쳐 벡터(Motion weighted feature vector, 이하 '모션에 의해 가중된 피쳐 벡터'라 한다) 생성 모듈(150), 응답 생성 모듈(160) 및 로스(loss, 이하 '로스'라 한다) 계산 모듈(170)을 포함한다.

[0059] 비디오 질의 응답 학습 장치(100)는 비디오(video) 및 상기 비디오에 대한 질의(110)를 획득할 수 있다.

[0060] 피쳐 생성 모듈(120)은 비디오(video) 및 상기 비디오에 대한 질의(110)를 기초로 비디오 피쳐(feature, 이하 '피쳐'라 한다) 및 언어 피쳐(feature)를 생성할 수 있다.

[0061] 피쳐 생성 모듈(120)은 비디오 피쳐를 생성하기 위한 학습이 완료된 적어도 하나의 딥 뉴럴 네트워크(deep neural network)(예컨대, 컨벌루셔널 뉴럴 네트워크)를 포함할 수 있다.

[0062] 피쳐 생성 모듈(120)은 학습이 완료된 컨벌루셔널 뉴럴 네트워크(Convolutional Neural Network)를 이용하여 비디오에 포함된 프레임에 대한 모습(appearance)의 피쳐(프레임 피쳐(frame feature))를 생성할 수 있으나, 피쳐 생성 모듈(120)이 비디오에 포함된 프레임에 대한 모습(appearance)의 피쳐(프레임 피쳐(frame feature))를 생

성하기 위하여 이용하는 학습이 완료된 딥 뉴럴 네트워크(deep neural network)가 이에 한정되는 것은 아니다.

[0063] 피쳐 생성 모듈(120)은 미리 설정된 간격(예컨대, 10 프레임 마다)에 해당하는 프레임만을 선택하여 비디오에 포함된 프레임에 대한 모습(appearance)의 피쳐(프레임 피쳐(frame feature))를 생성할 수 있다.

[0064] 피쳐 생성 모듈(120)은 상기 생성한 비디오에 포함된 프레임에 대한 모습(appearance)의 피쳐(프레임 피쳐(frame feature))를 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 할 수 있다.

[0065] 피쳐 생성 모듈(120)은 학습이 완료된 3D 컨벌루셔널 뉴럴 네트워크(3D Convolutional Neural Network)를 이용하여 비디오에 포함된 움직임(motion)의 피쳐를 생성할 수 있으나, 피쳐 생성 모듈(120)이 비디오에 포함된 움직임(motion)의 피쳐를 생성하기 위하여 이용하는 학습이 완료된 딥 뉴럴 네트워크(deep neural network)가 이에 한정되는 것은 아니다.

[0066] 피쳐 생성 모듈(120)은 상기 생성한 비디오에 포함된 움직임(motion)의 피쳐를 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 할 수 있다.

[0067] 피쳐 생성 모듈(120)은 학습이 완료된 컨벌루셔널 뉴럴 네트워크(Convolutional Neural Network)를 이용하여 비디오에 포함된 연속된 프레임의 피쳐를 생성할 수 있다.

[0068] 피쳐 생성 모듈(120)은 상기 생성한 비디오에 포함된 연속된 프레임의 피쳐를 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 할 수 있다.

[0069] 피쳐 생성 모듈(120)은 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 한 비디오에 포함된 프레임에 대한 모습(appearance)의 피쳐(프레임 피쳐(frame feature)), 비디오에 포함된 움직임(motion)의 피쳐 및 비디오에 포함된 연속된 프레임의 피쳐를 기초로 비디오 피쳐를 생성할 수 있다.

[0070] 피쳐 생성 모듈(120)은 언어 피쳐를 생성하기 위한 학습이 완료된 적어도 하나의 딥 뉴럴 네트워크를 포함할 수 있다.

[0071] 피쳐 생성 모듈(120)은 학습이 완료된 미리 설정된 언어 임베딩 모델(예컨대, word2vec, GloVe, BERT 등)을 사용하여 질의에 대한 적어도 하나의 임베딩 벡터를 생성할 수 있다.

[0072] 피쳐 생성 모듈(120)은 상기 생성한 질의에 대한 적어도 하나의 임베딩 벡터를 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 할 수 있다.

[0073] 피쳐 생성 모듈(120)은 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 한 질의에 대한 적어도 하나의 임베딩 벡터를 기초로 언어 피쳐를 생성할 수 있다.

[0074] 질의 벡터 생성 모듈(130)은 피쳐 생성 모듈(120)이 생성한 언어 피쳐를 획득할 수 있다.

[0075] 질의 벡터 생성 모듈(130)은 질의를 인코딩하기 위한 적어도 하나의 인코더(예컨대, 제1 인코더)를 포함할 수 있다. 이때, 상기 적어도 하나의 인코더는 장단기 메모리(Long Short-Term Memory, LSTM)일 수 있으나, 상기 적어도 하나의 인코더가 이에 한정되는 것은 아니다. 또한, 상기 적어도 하나의 인코더의 내부 파라미터의 초기 값은 피쳐 생성 모듈(120)에서 생성한 비디오에 포함된 프레임에 대한 모습(appearance)의 피쳐(프레임 피쳐(frame feature))를 인코딩 한 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)의 내부 파라미터 값 일 수 있으나, 상기 적어도 하나의 인코더의 내부 파라미터의 초기 값이 이에 한정되는 것은 아니다.

[0076] 질의 벡터 생성 모듈(130)은 상기 질의를 적어도 하나의 인코더(예컨대, 제1 인코더)로 인코딩하여 특징 벡터를 생성할 수 있다.

[0077] 질의 벡터 생성 모듈(130)은 하기 [수학식 1]을 기초로 상기 생성한 특징 벡터와 피쳐 생성 모듈(120)이 생성한 언어 피쳐를 가중치 합하여 질의 벡터를 생성할 수 있다.

수학식 1

$$E_q = W_1 \sum_{j=1}^{N_w} E'_w(j) + W_2 \sum_{j=1}^{N_w} S_w(j)$$

[0078]

[0079] 여기서, N_w는 질의의 단어 개수이고, W_1, W_2는 학습되어지는 파라미터이다.

[0080] 어텐디드 벡터 생성 모듈(140)은 피쳐 생성 모듈(120)이 생성한 비디오에 포함된 프레임에 대한 모습(appearance)의 피쳐(프레임 피쳐(frame feature))를 획득할 수 있다.

[0081] 어텐디드 벡터 생성 모듈(140)은 질의 벡터, 비디오에 포함된 프레임에 대한 모습(appearance)의 피쳐(이하, '프레임 피쳐(frame feature)' 라 한다) 및 비디오 압축 정보에 포함된 모션 벡터 중 적어도 어느 하나를 기초로 적어도 하나의 어텐디드 벡터(attended vector)를 생성할 수 있다.

[0082] 어텐디드 벡터 생성 모듈(140)은 어텐디드 벡터를 생성하기 위한 딥 뉴럴 네트워크(deep neural network)를 포함할 수 있다.

[0083] 어텐디드 벡터 생성 모듈(140)은 하기 [수학식 2]을 기초로 질의 벡터와 비디오에 포함된 적어도 하나의 프레임에 대한 프레임 피쳐(frame feature)(예컨대, 제1 프레임 피쳐)를 가중치 합하여 제1 어텐션 벡터(attention vector)를 생성할 수 있다.

수학식 2

$$Z_s^t = W_5 \tanh\left(W_3 E_q + W_4 F_s^t + b_1\right) + b_2$$

$$A_s^t = \frac{\exp\left(Z_s^t\right)}{\sum_u \exp\left(Z_s^t(u)\right)}$$

[0084]

[0085] 여기서, F_s^t는 제1 프레임 피쳐(frame feature)이고, W_3, W_4, W_5, b_1, b_2는 학습되어지는 파라미터이며, u는 x, y 좌표에서의 위치이다.

[0086] 어텐디드 벡터 생성 모듈(140)이 생성한 제1 어텐션 벡터(attention vector)는 지역 공간적인 모습(appearance) 정보의 세부 집중 영역을 추출한 것 일 수 있다.

[0087] 어텐디드 벡터 생성 모듈(140)은 하기 [수학식 3]을 기초로 상기 생성한 제1 어텐션 벡터(attention vector)와 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피쳐(frame feature)를 결합하여 제1 어텐디드 벡터(attended vector)를 생성할 수 있다.

수학식 3

$$G_s^t = F_s^t \odot A_s^t$$

[0088]

[0089] 여기서, F_s^t 는 제1 프레임 피처(frame feature)이고, A_s^t 제1 어텐션 벡터(attention vector)이다.

[0090] 어텐디드 벡터 생성 모듈(140)이 생성한 제1 어텐디드 벡터(attended vector)는 프레임 피처들에 모습 세부 집중 영역이 반영된 결과일 수 있다.

[0091] 어텐디드 벡터 생성 모듈(140)은 하기 [수학식 4]를 기초로 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature)와 비디오 압축 정보에 포함된 모션 벡터를 결합하여 피처 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피처(frame feature)를 생성할 수 있다.

수학식 4

[0092]
$$F_s^{t+}(u) = F_s^t\left(u + mv^t\left({}^{u}/_{f}\right)\right)$$

[0093] 여기서, F_s^t 는 제1 프레임 피처(frame feature)이고, u 는 x, y 좌표에서의 위치이고, mv는 모션 벡터이고, f 는 프레임이다.

[0094] 어텐디드 벡터 생성 모듈(140)은 하기 [수학식 5]를 기초로 질의 벡터와 피처 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피처(frame feature)를 가중치 합하여 제2 어텐션 벡터(attention vector)를 생성할 수 있다.

수학식 5

[0095]
$$z_s^{t+} = W_5 \tanh\left(W_3 E_q + W_4 F_S^{t+} + b_1\right) + b_2$$

$$A_s^{t+} = \frac{\exp\left(z_s^{t+}\right)}{\Sigma_u \exp\left(z_s^{t+}(u)\right)}$$

[0096] 여기서, F_S^{t+} 는 피처 왑핑된(Feature warping) 제2 프레임 피처(frame feature)이고, W_3, W_4, W_5, b_1, b_2 는 학습되어지는 파라미터이며, u 는 x, y 좌표에서의 위치이다.

[0097] 어텐디드 벡터 생성 모듈(140)은 하기 [수학식 6]을 기초로 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature), 피처 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피처(frame feature) 및 제2 어텐션 벡터(attention vector)를 결합하여 제2 어텐디드 벡터(attended vector)를 생성할 수 있다.

수학식 6

[0098]
$$G_s^{t+} = (F_s^t + F_s^{t+})/2 \odot A_s^{t+}$$

[0099] 여기서, F_s^t 는 제1 프레임 피처(frame feature)이고, F_S^{t+} 는 피처 왑핑된(Feature warping) 제2 프레임 피처(frame feature)이고, A_s^{t+} 는 제2 어텐션 벡터(attention vector)이다.

[0100] 모션에 의해 가중된 피처 벡터 생성 모듈(150)은 제1 어텐디드 벡터(attended vector), 제2 어텐디드 벡터(attended vector) 및 비디오 압축 정보에 포함된 잔차(residual)를 기초로 모션에 의해 가중된 피처 벡터(Motion weighted feature vector)를 생성할 수 있다.

[0101] 모션에 의해 가중된 피처 벡터 생성 모듈(150)은 비디오 압축 정보에 포함된 잔차(residual)를 이용하여 모션의 반영 정도를 제어할 수 있다.

[0102] 모션에 의해 가중된 피처 벡터 생성 모듈(150)은 하기 [수학식 7]을 기초로 제1 어텐디드 벡터(attended vector) 및 제2 어텐디드 벡터(attended vector)의 가중치 합을 결정할 수 있다.

수학식 7

[0103]
$$\alpha = enc_r(R_s^t)$$
$$L_v^t = \alpha W_6 \sum_u G_s^t(u) + (1-\alpha) W_7 \sum_u G_s^{t+}(u)$$

[0104] 여기서, enc_r 은 2개 fully connected layer와 sigmoid 함수로 구성된 네트워크이고, W_6, W_7 은 학습되어지는 파라미터이고, u 는 x, y 좌표에서의 위치이다.

[0105] 모션에 의해 가중된 피처 벡터 생성 모듈(150)은 모션에 의해 가중된 피처 벡터를 생성하기 위한 적어도 하나의 인코더(예컨대, 제2 인코더)를 포함할 수 있다. 이때, 상기 적어도 하나의 인코더는 장단기 메모리(Long Short-Term Memory, LSTM)일 수 있으나, 상기 적어도 하나의 인코더가 이에 한정되는 것은 아니다. 또한, 상기 적어도 하나의 인코더의 내부 파라미터의 초기 값은 피처 생성 모듈(120)에서 생성한 비디오에 포함된 프레임에 대한 모습(appearance)의 피처(프레임 피처(frame feature))를 인코딩 한 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)의 내부 파라미터 값 또는 질의 벡터 생성 모듈(130)에 포함된 학습이 완료된 인코더의 내부 파라미터 값일 수 있으나, 상기 적어도 하나의 인코더의 내부 파라미터의 초기 값이 이에 한정되는 것은 아니다.

[0106] 모션에 의해 가중된 피처 벡터 생성 모듈(150)은 상기 결정한 제1 어텐디드 벡터(attended vector) 및 제2 어텐디드 벡터(attended vector)의 가중치 합을 상기 적어도 하나의 인코더(예컨대, 제2 인코더)로 인코딩하여 모션에 의해 가중된 피처 벡터를 생성할 수 있다.

[0107] 응답 생성 모듈(160)은 피처 생성 모듈(120)이 생성한 언어 피처(feature) 및 비디오 피처(feature)를 획득할 수 있다.

[0108] 응답 생성 모듈(160)은 모션에 의해 가중된 피처 벡터 생성 모듈(150)이 생성한 모션에 의해 가중된 피처 벡터를 획득할 수 있다.

[0109] 응답 생성 모듈(160)은 응답을 생성하기 위한 딥 뉴럴 네트워크(deep neural network)를 포함할 수 있다.

[0110] 응답 생성 모듈(160)은 하기 [수학식 8]을 기초로 모션에 의해 가중된 피처 벡터(motion weighted feature vector), 언어 피처(feature) 및 비디오 피처(feature)를 퓨전(fusion)할 수 있다.

수학식 8

$$T_w = S_w \odot \frac{\exp(W_{10}\tanh(W_8 S_w + W_9 E'_w) + b_3)}{\sum \exp(W_{10}\tanh(W_8 S_w + W_9 E'_w) + b_3)} \qquad T_v = S_v \odot \frac{\exp(W_{13}\tanh(W_{11} S_v + W_{12} E'_w) + b_4)}{\sum \exp(W_{13}\tanh(W_{11} S_v + W_{12} E'_w) + b_4)}$$

$$O_g = W_{14}\sum_{j=1}^{N_w} T_w(j) + W_{15}\sum_{j=1}^{N_v} T_v(j)$$

$$O_s = W_{16}\sum_{t=1}^{N_v} L_v^t + W_{17}\sum_{j=1}^{N_w} E'_w(j)$$

[0111]

[0112] 여기서, N_w는 단어 개수이고, N_v비디오 프레임 개수이고, $W_8, W_9, W_{10}, W_{11}, W_{12}, W_{13}, W_{14}, W_{15}, W_{16}, W_{17}, b_3, b_4$는 학습되어지는 파라미터이다.

[0113] 응답 생성 모듈(160)은 O_g와 O_s를 연결(concat)할 수 있다.

[0114] 응답 생성 모듈(160)은 연결(concat)된 O_g와 O_s를 미리 설정된 개수(예컨대, 2개)의 풀리 커넥티드 레이어를 통과시켜 응답인 O_o를 생성할 수 있다.

[0115] 로스 계산 모듈(170)은 응답 생성 모듈(160)이 생성한 응답과 정답을 비교하여 로스를 계산할 수 있다.

[0116] 비디오 질의 응답 학습 장치(100)는 상기 계산된 로스를 기초로 질의 벡터 생성 모듈(130), 어텐디드 벡터 생성 모듈(140), 모션에 의해 가중된 피쳐 벡터 생성 모듈(150) 및 응답 생성 모듈(160)을 업데이트 할 수 있다.

[0117] 비디오 질의 응답 학습 장치(100)는 질의 벡터 생성 모듈(130), 어텐디드 벡터 생성 모듈(140), 모션에 의해 가중된 피쳐 벡터 생성 모듈(150) 및 응답 생성 모듈(160)을 업데이트 하기 위하여 역전파(backpropagation)를 이용할 수 있다.

[0118] 여기서 사용된 '장치 또는 모듈'이라는 용어는 논리적인 구성 단위를 나타내는 것으로서, 반드시 물리적으로 구분되는 구성 요소가 아니라는 점은 본 발명이 속하는 기술분야의 당업자에게 자명한 사항이다.

[0119] 본 발명의 다른 실시예에 따라, 비디오 질의 응답 학습 장치(100)가 별도의 장치를 통하여 비디오 피쳐(feature), 언어 피쳐(feature) 및 프레임 피쳐(frame feature)를 획득하면, 본 발명의 다른 실시예에 따른 비디오 질의 응답 학습 장치(100)는 질의 벡터 생성 모듈(130), 어텐디드 벡터 생성 모듈(140), 모션에 의해 가중된 피쳐 벡터 생성 모듈(150), 응답 생성 모듈(160) 및 로스 계산 모듈(170)만으로 구현될 수 있다.

[0120] 도 2는 일실시예에 따른 비디오 질의 응답 학습 방법을 나타내는 플로우 차트이다.

[0121] 도 2를 참조하면, 일실시예에 따른 비디오 질의 응답 학습 장치가 질의에 대한 언어 피쳐(feature) 및 비디오에 대한 비디오 피쳐(feature)를 생성한다(200).

[0122] 이때, 상기 비디오 질의 응답 학습 장치는 학습이 완료된 컨벌루셔널 뉴럴 네트워크(Convolutional Neural Network)를 이용하여 비디오에 포함된 프레임에 대한 모습(appearance)의 피쳐(프레임 피쳐(frame feature))를 생성할 수 있다.

[0123] 또한, 상기 비디오 질의 응답 학습 장치는 상기 생성한 비디오에 포함된 프레임에 대한 모습(appearance)의 피쳐(프레임 피쳐(frame feature))를 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 할 수 있다.

[0124] 또한, 상기 비디오 질의 응답 학습 장치는 학습이 완료된 3D 컨벌루셔널 뉴럴 네트워크(3D Convolutional Neural Network)를 이용하여 비디오에 포함된 움직임(motion)의 피쳐를 생성할 수 있다.

[0125] 또한, 상기 비디오 질의 응답 학습 장치는 상기 생성한 비디오에 포함된 움직임(motion)의 피쳐를 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 할 수 있다.

[0126] 또한, 상기 비디오 질의 응답 학습 장치는 학습이 완료된 컨벌루셔널 뉴럴 네트워크(Convolutional Neural Network)를 이용하여 비디오에 포함된 연속된 프레임의 피처를 생성할 수 있다.

[0127] 또한, 상기 비디오 질의 응답 학습 장치는 상기 생성한 비디오에 포함된 연속된 프레임의 피처를 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 할 수 있다.

[0128] 또한, 상기 비디오 질의 응답 학습 장치는 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 한 비디오에 포함된 프레임에 대한 모습(appearance)의 피처(프레임 피처(frame feature)), 비디오에 포함된 움직임(motion)의 피처 및 비디오에 포함된 연속된 프레임의 피처를 기초로 비디오 피처를 생성할 수 있다.

[0129] 또한, 상기 비디오 질의 응답 학습 장치는 학습이 완료된 미리 설정된 언어 임베딩 모델(예컨대, word2vec, GloVe, BERT 등)을 사용하여 질의에 대한 적어도 하나의 임베딩 벡터를 생성할 수 있다.

[0130] 또한, 상기 비디오 질의 응답 학습 장치는 상기 생성한 질의에 대한 적어도 하나의 임베딩 벡터를 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 할 수 있다.

[0131] 또한, 상기 비디오 질의 응답 학습 장치는 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 한 질의에 대한 적어도 하나의 임베딩 벡터를 기초로 언어 피처를 생성할 수 있다.

[0132] 비디오 질의 응답 학습 장치가 인코딩 한 상기 질의를 기초로 질의 벡터를 생성한다(210).

[0133] 이때, 상기 비디오 질의 응답 학습 장치는 상기 질의를 적어도 하나의 인코더(예컨대, 제1 인코더)로 인코딩하여 특징 벡터를 생성할 수 있다.

[0134] 또한, 상기 비디오 질의 응답 학습 장치는 상기 생성한 특징 벡터와 상기 생성한 언어 피처를 가중치 합하여 질의 벡터를 생성할 수 있다.

[0135] 비디오 질의 응답 학습 장치가 상기 질의 벡터, 상기 비디오에 포함된 적어도 하나의 프레임에 대한 프레임 피처(frame feature) 및 상기 비디오 압축 정보에 포함된 모션 벡터 중 적어도 어느 하나를 기초로 적어도 하나의 어텐디드 벡터(attended vector)를 생성한다(220).

[0136] 이때, 상기 비디오 질의 응답 학습 장치는 질의 벡터와 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature)를 가중치 합하여 제1 어텐션 벡터(attention vector)를 생성할 수 있다.

[0137] 또한, 상기 비디오 질의 응답 학습 장치는 상기 생성한 제1 어텐션 벡터(attention vector)와 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature)를 결합하여 제1 어텐디드 벡터(attended vector)를 생성할 수 있다.

[0138] 또한, 상기 비디오 질의 응답 학습 장치는 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature)와 비디오 압축 정보에 포함된 모션 벡터를 결합하여 피처 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피처(frame feature)를 생성할 수 있다.

[0139] 또한, 상기 비디오 질의 응답 학습 장치는 질의 벡터와 피처 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피처(frame feature)를 가중치 합하여 제2 어텐션 벡터(attention vector)를 생성할 수 있다.

[0140] 또한, 상기 비디오 질의 응답 학습 장치는 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature), 피처 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피처(frame feature) 및 제2 어텐션 벡터(attention vector)를 결합하여 제2 어텐디드 벡터(attended vector)를 생성할 수 있다.

[0141] 비디오 질의 응답 학습 장치가 상기 생성한 적어도 하나의 어텐디드 벡터(attended vector) 및 상기 비디오 압축 정보에 포함된 잔차(residual)를 기초로 모션에 의해 가중된 피처 벡터(Motion weighted feature vector)를 생성한다(230).

[0142] 이때, 상기 비디오 질의 응답 학습 장치는 제1 어텐디드 벡터(attended vector) 및 제2 어텐디드 벡터(attended vector)의 가중치 합을 결정할 수 있다.

[0143] 또한, 상기 비디오 질의 응답 학습 장치는 상기 결정한 제1 어텐디드 벡터(attended vector) 및 제2 어텐디드

벡터(attended vector)의 가중치 합을 상기 적어도 하나의 인코더(예컨대, 제2 인코더)로 인코딩하여 모션에 의해 가중된 피처 벡터를 생성할 수 있다.

[0144] 비디오 질의 응답 학습 장치가 상기 생성한 모션에 의해 가중된 피처 벡터(motion weighted feature vector), 상기 비디오 피처(feature) 및 상기 언어 피처(feature)를 기초로 응답을 생성하도록 뉴럴 네트워크를 학습시킨다(240).

[0145] 이때, 상기 비디오 질의 응답 학습 장치는 모션에 의해 가중된 피처 벡터(motion weighted feature vector), 언어 피처(feature) 및 비디오 피처(feature)를 퓨전(fusion)할 수 있다.

[0146] 또한, 상기 비디오 질의 응답 학습 장치는 퓨전(fusion) 결과를 미리 설정된 개수(예컨대, 2개)의 풀리 커넥티드 레이어를 통과시켜 응답을 생성할 수 있다.

[0147] 또한, 상기 비디오 질의 응답 학습 장치는 생성한 응답과 정답을 비교하여 로스를 계산할 수 있다.

[0148] 또한, 상기 비디오 질의 응답 학습 장치는 역전파(backpropagation)를 이용하여 상기 비디오 질의 응답 장치에 포함된 뉴럴 네트워크를 학습시킬 수 있다.

[0149] 도 3은 일실시예에 따른 비디오 질의 응답 장치의 구성을 나타내는 도면이다.

[0150] 도 3을 참조하면, 일실시예에 따른 비디오 질의 응답 장치(300)는 피처(feature) 생성 모듈(320), 질의 벡터 생성 모듈(330), 어텐디드 벡터(attended vector) 생성 모듈(340), 모션에 의해 가중된 피처 벡터(Motion weighted feature vector) 생성 모듈(350) 및 응답 생성 모듈(360)을 포함한다.

[0151] 비디오 질의 응답 장치(300)는 비디오(video) 및 상기 비디오에 대한 질의(310)를 획득할 수 있다.

[0152] 피처 생성 모듈(320)은 비디오(video) 및 상기 비디오에 대한 질의(310)를 기초로 비디오 피처(feature) 및 언어 피처(feature)를 생성할 수 있다.

[0153] 피처 생성 모듈(320)은 비디오 피처를 생성하기 위한 학습이 완료된 적어도 하나의 딥 뉴럴 네트워크(deep neural network)(예컨대, 컨벌루셔널 뉴럴 네트워크)를 포함할 수 있다.

[0154] 피처 생성 모듈(320)은 학습이 완료된 컨벌루셔널 뉴럴 네트워크(Convolutional Neural Network)를 이용하여 비디오에 포함된 프레임에 대한 모습(appearance)의 피처(프레임 피처(frame feature))를 생성할 수 있으나, 피처 생성 모듈(320)이 비디오에 포함된 프레임에 대한 모습(appearance)의 피처(프레임 피처(frame feature))를 생성하기 위하여 이용하는 학습이 완료된 딥 뉴럴 네트워크(deep neural network)가 이에 한정되는 것은 아니다.

[0155] 피처 생성 모듈(320)은 미리 설정된 간격(예컨대, 10 프레임 마다)에 해당하는 프레임만을 선택하여 비디오에 포함된 프레임에 대한 모습(appearance)의 피처(프레임 피처(frame feature))를 생성할 수 있다.

[0156] 피처 생성 모듈(320)은 상기 생성한 비디오에 포함된 프레임에 대한 모습(appearance)의 피처(프레임 피처(frame feature))를 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 할 수 있다.

[0157] 피처 생성 모듈(320)은 학습이 완료된 3D 컨벌루셔널 뉴럴 네트워크(3D Convolutional Neural Network)를 이용하여 비디오에 포함된 움직임(motion)의 피처를 생성할 수 있으나, 피처 생성 모듈(320)이 비디오에 포함된 움직임(motion)의 피처를 생성하기 위하여 이용하는 학습이 완료된 딥 뉴럴 네트워크(deep neural network)가 이에 한정되는 것은 아니다.

[0158] 피처 생성 모듈(320)은 상기 생성한 비디오에 포함된 움직임(motion)의 피처를 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 할 수 있다.

[0159] 피처 생성 모듈(320)은 학습이 완료된 컨벌루셔널 뉴럴 네트워크(Convolutional Neural Network)를 이용하여 비디오에 포함된 연속된 프레임의 피처를 생성할 수 있다.

[0160] 피처 생성 모듈(320)은 상기 생성한 비디오에 포함된 연속된 프레임의 피처를 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 할 수 있다.

[0161] 피처 생성 모듈(320)은 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 한 비디오에 포함된 프레임에 대한 모습(appearance)의 피처(프레임 피처(frame feature)), 비디오에 포함된 움직임(motion)의 피처 및 비디오에 포함된 연속된 프레임의 피처를 기초로 비디오 피처를 생성할 수 있다.

[0162] 피쳐 생성 모듈(320)은 언어 피쳐를 생성하기 위한 학습이 완료된 적어도 하나의 딥 뉴럴 네트워크를 포함할 수 있다.

[0163] 피쳐 생성 모듈(320)은 학습이 완료된 미리 설정된 언어 임베딩 모델(예컨대, word2vec, GloVe, BERT 등)을 사용하여 질의에 대한 적어도 하나의 임베딩 벡터를 생성할 수 있다.

[0164] 피쳐 생성 모듈(320)은 상기 생성한 질의에 대한 적어도 하나의 임베딩 벡터를 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 할 수 있다.

[0165] 피쳐 생성 모듈(320)은 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 한 질의에 대한 적어도 하나의 임베딩 벡터를 기초로 언어 피쳐를 생성할 수 있다.

[0166] 질의 벡터 생성 모듈(330)은 피쳐 생성 모듈(320)이 생성한 언어 피쳐를 획득할 수 있다.

[0167] 질의 벡터 생성 모듈(330)은 질의를 인코딩하기 위한 학습이 완료된 적어도 하나의 인코더(예컨대, 제1 인코더)를 포함할 수 있다. 이때, 상기 학습이 완료된 적어도 하나의 인코더는 장단기 메모리(Long Short-Term Memory, LSTM)일 수 있으나, 상기 적어도 하나의 인코더가 이에 한정되는 것은 아니다.

[0168] 질의 벡터 생성 모듈(330)은 상기 질의를 학습이 완료된 적어도 하나의 인코더(예컨대, 제1 인코더)로 인코딩하여 특징 벡터를 생성할 수 있다.

[0169] 질의 벡터 생성 모듈(330)은 하기 [수학식 9]를 기초로 상기 생성한 특징 벡터와 피쳐 생성 모듈(320)이 생성한 언어 피쳐를 가중치 합하여 질의 벡터를 생성할 수 있다.

수학식 9

[0170]
$$E_q = W_1 \sum_{j=1}^{N_w} E'_w(j) + W_2 \sum_{j=1}^{N_w} S_w(j)$$

[0171] 여기서, N_w는 질의의 단어 개수이고, W_1, W_2는 학습에 의해 결정된 파라미터이다.

[0172] 어텐디드 벡터 생성 모듈(340)은 피쳐 생성 모듈(320)이 생성한 비디오에 포함된 프레임에 대한 모습(appearance)의 피쳐(프레임 피쳐(frame feature))를 획득할 수 있다.

[0173] 어텐디드 벡터 생성 모듈(340)은 질의 벡터, 비디오에 포함된 프레임에 대한 모습(appearance)의 피쳐(이하, '프레임 피쳐(frame feature)' 라 한다) 및 비디오 압축 정보에 포함된 모션 벡터 중 적어도 어느 하나를 기초로 적어도 하나의 어텐디드 벡터(attended vector)를 생성할 수 있다.

[0174] 어텐디드 벡터 생성 모듈(340)은 어텐디드 벡터를 생성하기 위한 학습이 완료된 딥 뉴럴 네트워크(deep neural network)를 포함할 수 있다.

[0175] 어텐디드 벡터 생성 모듈(340)은 하기 [수학식 10]을 기초로 질의 벡터와 비디오에 포함된 적어도 하나의 프레임에 대한 프레임 피쳐(frame feature)(예컨대, 제1 프레임 피쳐)를 가중치 합하여 제1 어텐션 벡터(attention vector)를 생성할 수 있다.

수학식 10

[0176]
$$Z_s^t = W_5 \tanh\left(W_3 E_q + W_4 F_s^t + b_1\right) + b_2$$
$$A_s^t = \frac{\exp\left(Z_s^t\right)}{\sum_u \exp\left(Z_s^t(u)\right)}$$

[0177] 여기서, F_s^t 는 제1 프레임 피쳐(frame feature)이고, W_3, W_4, W_5, b_1, b_2 는 학습에 의해 결정된 파라미터이며, u 는 x, y 좌표에서의 위치이다.

[0178] 어텐디드 벡터 생성 모듈(340)이 생성한 제1 어텐션 벡터(attention vector)는 지역 공간적인 모습(appearance) 정보의 세부 집중 영역을 추출한 것 일 수 있다.

[0179] 어텐디드 벡터 생성 모듈(340)은 하기 [수학식 11]을 기초로 상기 생성한 제1 어텐션 벡터(attention vector)와 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피쳐(frame feature)를 결합하여 제1 어텐디드 벡터(attended vector)를 생성할 수 있다.

수학식 11

[0180] $$G_s^t = F_s^t \odot A_s^t$$

[0181] 여기서, F_s^t 는 제1 프레임 피쳐(frame feature)이고, A_s^t 제1 어텐션 벡터(attention vector)이다.

[0182] 어텐디드 벡터 생성 모듈(340)이 생성한 제1 어텐디드 벡터(attended vector)는 프레임 피쳐들에 모습 세부 집중 영역이 반영된 결과일 수 있다.

[0183] 어텐디드 벡터 생성 모듈(340)은 하기 [수학식 12]를 기초로 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피쳐(frame feature)와 비디오 압축 정보에 포함된 모션 벡터를 결합하여 피쳐 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피쳐(frame feature)를 생성할 수 있다.

수학식 12

[0184] $$F_s^{t+}(u) = F_s^t\left(u + mv^t\left(u/f\right)\right)$$

[0185] 여기서, F_s^t 는 제1 프레임 피쳐(frame feature)이고, u 는 x, y 좌표에서의 위치이고, mv 는 모션 벡터이고, f 는 프레임이다.

[0186] 어텐디드 벡터 생성 모듈(340)은 하기 [수학식 13]을 기초로 질의 벡터와 피쳐 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피쳐(frame feature)를 가중치 합하여 제2 어텐션 벡터(attention vector)를 생성할 수 있다.

수학식 13

$$z_s^{t+} = W_5 \tanh\left(W_3 E_q + W_4 F_S^{t+} + b_1\right) + b_2$$

$$A_s^{t+} = \frac{\exp\left(z_s^{t+}\right)}{\Sigma_u \exp\left(z_s^{t+}(u)\right)}$$

[0187]

[0188] 여기서, F_S^{t+} 는 피쳐 왑핑된(Feature warping) 제2 프레임 피쳐(frame feature)이고, W_3, W_4, W_5, b_1, b_2 는 학습에 의해 결정된 파라미터이며, u 는 x, y 좌표에서의 위치이다.

[0189] 어텐디드 벡터 생성 모듈(340)은 하기 [수학식 14]를 기초로 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피쳐(frame feature), 피쳐 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피쳐(frame feature) 및 제2 어텐션 벡터(attention vector)를 결합하여 제2 어텐디드 벡터(attended vector)를 생성할 수 있다.

수학식 14

$$G_s^{t+} = (F_s^t + F_s^{t+})/2 \odot A_s^{t+}$$

[0190]

[0191] 여기서, F_s^t 는 제1 프레임 피쳐(frame feature)이고, F_S^{t+} 는 피쳐 왑핑된(Feature warping) 제2 프레임 피쳐(frame feature)이고, A_s^{t+} 는 제2 어텐션 벡터(attention vector)이다.

[0192] 모션에 의해 가중된 피쳐 벡터 생성 모듈(350)은 제1 어텐디드 벡터(attended vector), 제2 어텐디드 벡터(attended vector) 및 비디오 압축 정보에 포함된 잔차(residual)를 기초로 모션에 의해 가중된 피쳐 벡터(Motion weighted feature vector)를 생성할 수 있다.

[0193] 모션에 의해 가중된 피쳐 벡터 생성 모듈(350)은 비디오 압축 정보에 포함된 잔차(residual)를 이용하여 모션의 반영 정도를 제어할 수 있다.

[0194] 모션에 의해 가중된 피쳐 벡터 생성 모듈(350)은 하기 [수학식 15]를 기초로 제1 어텐디드 벡터(attended vector) 및 제2 어텐디드 벡터(attended vector)의 가중치 합을 결정할 수 있다.

수학식 15

$$\alpha = enc_r(R_s^t)$$

$$L_v^t = \alpha W_6 \sum_u G_s^t(u) + (1-\alpha) W_7 \sum_u G_s^{t+}(u)$$

[0195]

[0196] 여기서, enc_r 은 2개 fully connected layer와 sigmoid 함수로 구성된 네트워크이고, W_6, W_7 은 학습에 의해 결정된 파라미터이고, u 는 x, y 좌표에서의 위치이다.

[0197] 모션에 의해 가중된 피쳐 벡터 생성 모듈(350)은 모션에 의해 가중된 피쳐 벡터를 생성하기 위한 학습이 완료된 적어도 하나의 인코더(예컨대, 제2 인코더)를 포함할 수 있다. 이때, 상기 학습이 완료된 적어도 하나의 인코더는 장단기 메모리(Long Short-Term Memory, LSTM)일 수 있으나, 상기 학습이 완료된 적어도 하나의 인코더가 이에 한정되는 것은 아니다.

[0198] 모션에 의해 가중된 피쳐 벡터 생성 모듈(350)은 상기 결정한 제1 어텐디드 벡터(attended vector) 및 제2 어텐디드 벡터(attended vector)의 가중치 합을 상기 학습이 완료된 적어도 하나의 인코더(예컨대, 제2 인코더)로 인코딩하여 모션에 의해 가중된 피쳐 벡터를 생성할 수 있다.

[0199] 응답 생성 모듈(360)은 피쳐 생성 모듈(320)이 생성한 언어 피쳐(feature) 및 비디오 피쳐(feature)를 획득할 수 있다.

[0200] 응답 생성 모듈(360)은 모션에 의해 가중된 피쳐 벡터 생성 모듈(350)이 생성한 모션에 의해 가중된 피쳐 벡터를 획득할 수 있다.

[0201] 응답 생성 모듈(160)은 응답을 생성하기 위한 학습이 완료된 딥 뉴럴 네트워크(deep neural network)를 포함할 수 있다.

[0202] 응답 생성 모듈(160)은 하기 [수학식 16]을 기초로 모션에 의해 가중된 피쳐 벡터(motion weighted feature vector), 언어 피쳐(feature) 및 비디오 피쳐(feature)를 퓨전(fusion)할 수 있다.

수학식 16

[0203]

$$T_w = S_w \odot \frac{\exp(W_{10}\tanh(W_8 S_w + W_9 E_w') + b_3)}{\sum \exp(W_{10}\tanh(W_8 S_w + W_9 E_w') + b_3)} \qquad T_v = S_v \odot \frac{\exp(W_{13}\tanh(W_{11} S_v + W_{12} E_w') + b_4)}{\sum \exp(W_{13}\tanh(W_{11} S_v + W_{12} E_w') + b_4)}$$

$$O_g = W_{14}\sum_{j=1}^{N_w} T_w(j) + W_{15}\sum_{j=1}^{N_v} T_v(j)$$

$$O_s = W_{16}\sum_{t=1}^{N_v} L_v^t + W_{17}\sum_{j=1}^{N_w} E_w'(j)$$

[0204] 여기서, N_w는 단어 개수이고, N_v비디오 프레임 개수이고, $W_8, W_9, W_{10}, W_{11}, W_{12}, W_{13}, W_{14}, W_{15}, W_{16}, W_{17}, b_3, b_4$는 학습에 의해 결정된 파라미터이다.

[0205] 응답 생성 모듈(360)은 O_g와 O_s를 연결(concat)할 수 있다.

[0206] 응답 생성 모듈(360)은 연결(concat)된 O_g와 O_s를 미리 설정된 개수(예컨대, 2개)의 폴리 커넥티드 레이어를 통과시켜 응답인 O_o를 생성할 수 있다.

[0207] 비디오 질의 응답 장치(300)는 응답 생성 모듈(360)이 생성한 응답(370)을 출력할 수 있다.

[0208] 여기서 사용된 '장치 또는 모듈'이라는 용어는 논리적인 구성 단위를 나타내는 것으로서, 반드시 물리적으로 구분되는 구성 요소가 아니라는 점은 본 발명이 속하는 기술분야의 당업자에게 자명한 사항이다.

[0209] 본 발명의 다른 실시예에 따라, 비디오 질의 응답 장치(300)가 별도의 장치를 통하여 비디오 피쳐(feature), 언어 피쳐(feature) 및 프레임 피쳐(frame feature)를 획득하면, 본 발명의 다른 실시예에 따른 비디오 질의 응답 장치(300)는 질의 벡터 생성 모듈(330), 어텐디드 벡터 생성 모듈(340), 모션에 의해 가중된 피쳐 벡터 생성 모듈(350) 및 응답 생성 모듈(360)만으로 구현될 수 있다.

[0210] 도 4는 일실시예에 따른 비디오 질의 응답 방법을 나타내는 플로우 차트이다.

[0211] 도 4를 참조하면, 일실시예에 따른 비디오 질의 응답 장치가 질의에 대한 언어 피쳐(feature) 및 비디오에 대한 비디오 피쳐(feature)를 생성한다(400).

[0212] 이때, 상기 비디오 질의 응답 장치는 학습이 완료된 컨벌루셔널 뉴럴 네트워크(Convolutional Neural Network)를 이용하여 비디오에 포함된 프레임에 대한 모습(appearance)의 피처(프레임 피처(frame feature))를 생성할 수 있다.

[0213] 또한, 상기 비디오 질의 응답 장치는 상기 생성한 비디오에 포함된 프레임에 대한 모습(appearance)의 피처(프레임 피처(frame feature))를 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 할 수 있다.

[0214] 또한, 상기 비디오 질의 응답 장치는 학습이 완료된 3D 컨벌루셔널 뉴럴 네트워크(3D Convolutional Neural Network)를 이용하여 비디오에 포함된 움직임(motion)의 피처를 생성할 수 있다.

[0215] 또한, 상기 비디오 질의 응답 장치는 상기 생성한 비디오에 포함된 움직임(motion)의 피처를 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 할 수 있다.

[0216] 또한, 상기 비디오 질의 응답 장치는 학습이 완료된 컨벌루셔널 뉴럴 네트워크(Convolutional Neural Network)를 이용하여 비디오에 포함된 연속된 프레임의 피처를 생성할 수 있다.

[0217] 또한, 상기 비디오 질의 응답 장치는 상기 생성한 비디오에 포함된 연속된 프레임의 피처를 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 할 수 있다.

[0218] 또한, 상기 비디오 질의 응답 장치는 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 한 비디오에 포함된 프레임에 대한 모습(appearance)의 피처(프레임 피처(frame feature)), 비디오에 포함된 움직임(motion)의 피처 및 비디오에 포함된 연속된 프레임의 피처를 기초로 비디오 피처를 생성할 수 있다.

[0219] 또한, 상기 비디오 질의 응답 장치는 학습이 완료된 미리 설정된 언어 임베딩 모델(예컨대, word2vec, GloVe, BERT 등)을 사용하여 질의에 대한 적어도 하나의 임베딩 벡터를 생성할 수 있다.

[0220] 또한, 상기 비디오 질의 응답 장치는 상기 생성한 질의에 대한 적어도 하나의 임베딩 벡터를 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 할 수 있다.

[0221] 또한, 상기 비디오 질의 응답 장치는 학습이 완료된 장단기 메모리(Long Short-Term Memory, LSTM)를 이용하여 인코딩 한 질의에 대한 적어도 하나의 임베딩 벡터를 기초로 언어 피처를 생성할 수 있다.

[0222] 비디오 질의 응답 장치가 인코딩 한 상기 질의를 기초로 질의 벡터를 생성한다(410).

[0223] 이때, 상기 비디오 질의 응답 장치는 상기 질의를 학습이 완료된 적어도 하나의 인코더(예컨대, 제1 인코더)로 인코딩하여 특징 벡터를 생성할 수 있다.

[0224] 또한, 상기 비디오 질의 응답 장치는 상기 생성한 특징 벡터와 상기 생성한 언어 피처를 가중치 합하여 질의 벡터를 생성할 수 있다.

[0225] 비디오 질의 응답 장치가 상기 질의 벡터, 상기 비디오에 포함된 적어도 하나의 프레임에 대한 프레임 피처(frame feature) 및 상기 비디오 압축 정보에 포함된 모션 벡터 중 적어도 어느 하나를 기초로 적어도 하나의 어텐디드 벡터(attended vector)를 생성한다(420).

[0226] 이때, 상기 비디오 질의 응답 장치는 질의 벡터와 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature)를 가중치 합하여 제1 어텐션 벡터(attention vector)를 생성할 수 있다.

[0227] 또한, 상기 비디오 질의 응답 장치는 상기 생성한 제1 어텐션 벡터(attention vector)와 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature)를 결합하여 제1 어텐디드 벡터(attended vector)를 생성할 수 있다.

[0228] 또한, 상기 비디오 질의 응답 장치는 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피처(frame feature)와 비디오 압축 정보에 포함된 모션 벡터를 결합하여 피처 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피처(frame feature)를 생성할 수 있다.

[0229] 또한, 상기 비디오 질의 응답 장치는 질의 벡터와 피처 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피처(frame feature)를 가중치 합하여 제2 어텐션 벡터(attention vector)를 생성할 수 있다.

[0230] 또한, 상기 비디오 질의 응답 장치는 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제1 프레임 피쳐(frame feature), 피쳐 왑핑된(Feature warping) 상기 비디오에 포함된 적어도 하나의 프레임에 대한 제2 프레임 피쳐(frame feature) 및 제2 어텐션 벡터(attention vector)를 결합하여 제2 어텐디드 벡터(attended vector)를 생성할 수 있다.

[0231] 비디오 질의 응답 장치가 상기 생성한 적어도 하나의 어텐디드 벡터(attended vector) 및 상기 비디오 압축 정보에 포함된 잔차(residual)를 기초로 모션에 의해 가중된 피쳐 벡터(Motion weighted feature vector)를 생성한다(430).

[0232] 이때, 상기 비디오 질의 응답 장치는 제1 어텐디드 벡터(attended vector) 및 제2 어텐디드 벡터(attended vector)의 가중치 합을 결정할 수 있다.

[0233] 또한, 상기 비디오 질의 응답 장치는 상기 결정한 제1 어텐디드 벡터(attended vector) 및 제2 어텐디드 벡터(attended vector)의 가중치 합을 상기 학습이 완료된 적어도 하나의 인코더(예컨대, 제2 인코더)로 인코딩하여 모션에 의해 가중된 피쳐 벡터를 생성할 수 있다.

[0234] 비디오 질의 응답 장치가 상기 생성한 모션에 의해 가중된 피쳐 벡터(motion weighted feature vector), 상기 비디오 피쳐(feature) 및 상기 언어 피쳐(feature)를 기초로 응답을 생성한다(440).

[0235] 이때, 상기 비디오 질의 응답 장치는 모션에 의해 가중된 피쳐 벡터(motion weighted feature vector), 언어 피쳐(feature) 및 비디오 피쳐(feature)를 퓨전(fusion)할 수 있다.

[0236] 또한, 상기 비디오 질의 응답 장치는 퓨전(fusion) 결과를 미리 설정된 개수(예컨대, 2개)의 풀리 커넥티드 레이어를 통과시켜 응답을 생성할 수 있다.

[0237] 또한, 상기 비디오 질의 응답 장치는 상기 생성한 응답을 출력할 수 있다.

[0238] 도 5는 본 발명의 일실시예를 구현하기 위한 예시적인 컴퓨터 시스템의 블록도이다.

[0239] 도 5를 참조하면, 본 발명의 일실시예를 구현하기 위한 예시적인 컴퓨터 시스템은 정보를 교환하기 위한 버스 또는 다른 커뮤니케이션 채널(501)을 포함하고, 프로세서(502)는 정보를 처리하기 위하여 버스(501)와 연결된다.

[0240] 컴퓨터 시스템(500)은 정보 및 프로세서(502)에 의해 처리되는 명령들을 저장하기 위하여 버스(501)와 연결된 RAM(Random Access Memory) 또는 다른 동적 저장 장치인 메인 메모리(503)를 포함한다.

[0241] 또한, 메인 메모리(503)는 프로세서(502)에 의한 명령들의 실행동안 임시변수들 또는 다른 중간 정보를 저장하기 위해 사용될 수 있다.

[0242] 컴퓨터 시스템(500)은 프로세서(502)에 대한 정적인 정보 또는 명령들을 저장하기 위하여 버스(501)에 결합된 ROM(Read Only Memory) 및 다른 정적 저장장치(504)를 포함할 수 있다.

[0243] 마그네틱 디스크, 집(zip) 또는 광 디스크 같은 대량 저장장치(505) 및 그것과 대응하는 드라이브 또한 정보 및 명령들을 저장하기 위하여 컴퓨터 시스템(500)에 연결될 수 있다.

[0244] 컴퓨터 시스템(500)은 엔드 유저(end user)에게 정보를 디스플레이 하기 위하여 버스(501)를 통해 음극선관 또는 엘씨디 같은 디스플레이 장치(510)와 연결될 수 있다.

[0245] 키보드(520)와 같은 문자 입력 장치는 프로세서(502)에 정보 및 명령을 전달하기 위하여 버스(501)에 연결될 수 있다.

[0246] 다른 유형의 사용자 입력 장치는 방향 정보 및 명령 선택을 프로세서(502)에 전달하고, 디스플레이(510) 상의 커서의 움직임을 제어하기 위한 마우스, 트랙볼 또는 커서 방향 키들과 같은 커서 컨트롤 장치(530)이다.

[0247] 통신 장치(540) 역시 버스(501)와 연결된다.

[0248] 통신 장치(540)는 지역 네트워크 또는 광역망에 접속되는 것을 서포트 하기 위하여 모뎀, 네트워크 인터페이스 카드, 이더넷, 토큰 링 또는 다른 유형의 물리적 결합물과 연결하기 위해 사용되는 인터페이스 장치를 포함할 수 있다. 이러한 방식으로 컴퓨터 시스템(500)은 인터넷 같은 종래의 네트워크 인프라 스트럭처를 통하여 다수의 클라이언트 및 서버와 연결될 수 있다.

[0249] 이상에서, 본 발명의 실시예를 구성하는 모든 구성 요소들이 하나로 결합되거나 결합되어 동작하는 것으로 설명

되었다고 해서, 본 발명이 반드시 이러한 실시예에 한정되는 것은 아니다. 즉, 본 발명의 목적 범위 안에서라면, 그 모든 구성 요소들이 적어도 하나로 선택적으로 결합하여 동작할 수도 있다.

[0250] 또한, 그 모든 구성 요소들이 각각 하나의 독립적인 하드웨어로 구현될 수 있지만, 각 구성 요소들의 그 일부 또는 전부가 선택적으로 조합되어 하나 또는 복수 개의 하드웨어에서 조합된 일부 또는 전부의 기능을 수행하는 프로그램 모듈을 갖는 컴퓨터 프로그램으로서 구현될 수도 있다. 그 컴퓨터 프로그램을 구성하는 코드들 및 코드 세그먼트들은 본 발명의 기술 분야의 당업자에 의해 용이하게 추론될 수 있을 것이다.

[0251] 이러한 컴퓨터 프로그램은 컴퓨터가 읽을 수 있는 저장매체(Computer Readable Media)에 저장되어 컴퓨터에 의하여 읽혀지고 실행됨으로써, 본 발명의 실시예를 구현할 수 있다. 컴퓨터 프로그램의 저장매체로서는 자기 기록매체, 광 기록매체, 등이 포함될 수 있다.

[0252] 또한, 이상에서 기재된 "포함하다", "구성하다" 또는 "가지다" 등의 용어는, 특별히 반대되는 기재가 없는 한, 해당 구성 요소가 내재될 수 있음을 의미하는 것이므로, 다른 구성 요소를 제외하는 것이 아니라 다른 구성 요소를 더 포함할 수 있는 것으로 해석되어야 한다.

[0253] 기술적이거나 과학적인 용어를 포함한 모든 용어들은, 다르게 정의되지 않는 한, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 사전에 정의된 용어와 같이 일반적으로 사용되는 용어들은 관련 기술의 문맥 상의 의미와 일치하는 것으로 해석되어야 하며, 본 발명에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0254] 본 발명에서 개시된 방법들은 상술된 방법을 달성하기 위한 하나 이상의 동작들 또는 단계들을 포함한다. 방법 동작들 및/또는 단계들은 청구항들의 범위를 벗어나지 않으면서 서로 상호 교환될 수도 있다. 다시 말해, 동작들 또는 단계들에 대한 특정 순서가 명시되지 않는 한, 특정 동작들 및/또는 단계들의 순서 및/또는 이용은 청구항들의 범위로부터 벗어남이 없이 수정될 수도 있다.

[0255] 본 발명에서 이용되는 바와 같이, 아이템들의 리스트 중 "그 중 적어도 하나" 를 지칭하는 구절은 단일 멤버들을 포함하여, 이들 아이템들의 임의의 조합을 지칭한다. 일 예로서, "a, b, 또는 c: 중의 적어도 하나" 는 a, b, c, a-b, a-c, b-c, 및 a-b-c 뿐만 아니라 동일한 엘리먼트의 다수의 것들과의 임의의 조합 (예를 들어, a-a, a-a-a, a-a-b, a-a-c, a-b-b, a-c-c, b-b, b-b-b, b-b-c, c-c, 및 c-c-c 또는 a, b, 및 c 의 다른 임의의 순서화한 것) 을 포함하도록 의도된다.

[0256] 본 발명에서 이용되는 바와 같이, 용어 "결정하는"는 매우 다양한 동작들을 망라한다. 예를 들어, "결정하는"는 계산하는, 컴퓨팅, 프로세싱, 도출하는, 조사하는, 룩업하는 (예를 들어, 테이블, 데이터베이스, 또는 다른 데이터 구조에서 룩업하는), 확인하는 등을 포함할 수도 있다. 또한, "결정하는"은 수신하는 (예를 들면, 정보를 수신하는), 액세스하는 (메모리의 데이터에 액세스하는) 등을 포함할 수 있다. 또한, "결정하는"은 해결하는, 선택하는, 고르는, 확립하는 등을 포함할 수 있다.

[0257] 이상의 설명은 본 발명의 기술 사상을 예시적으로 설명한 것에 불과한 것으로서, 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자라면 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 다양한 수정 및 변형이 가능할 것이다.

[0258] 따라서, 본 발명에 개시된 실시예들은 본 발명의 기술 사상을 한정하기 위한 것이 아니라 설명하기 위한 것이고, 이러한 실시예에 의하여 본 발명의 기술 사상의 범위가 한정되는 것은 아니다. 본 발명의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 발명의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

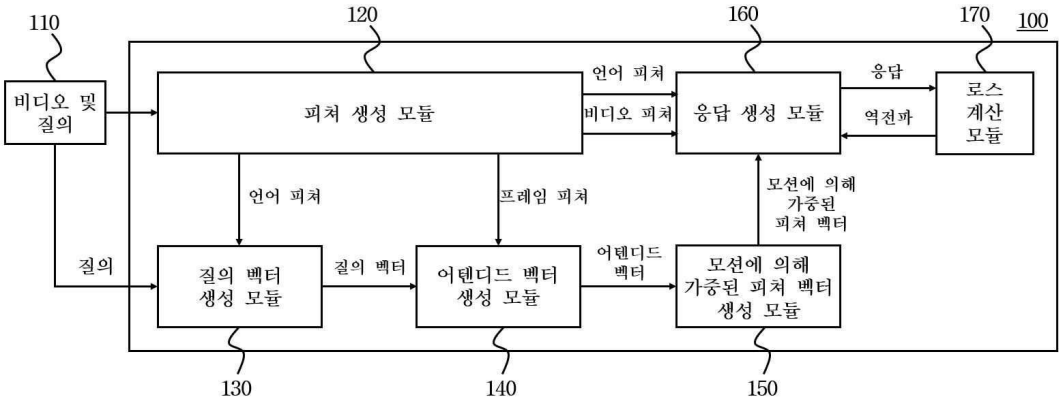
부호의 설명

[0259] 100... 비디오 질의 응답 학습 장치

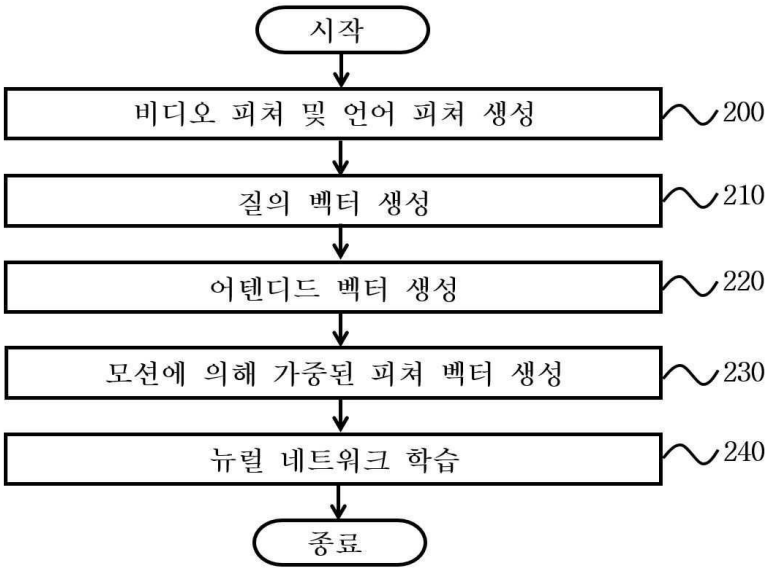
300... 비디오 질의 응답 장치

도면

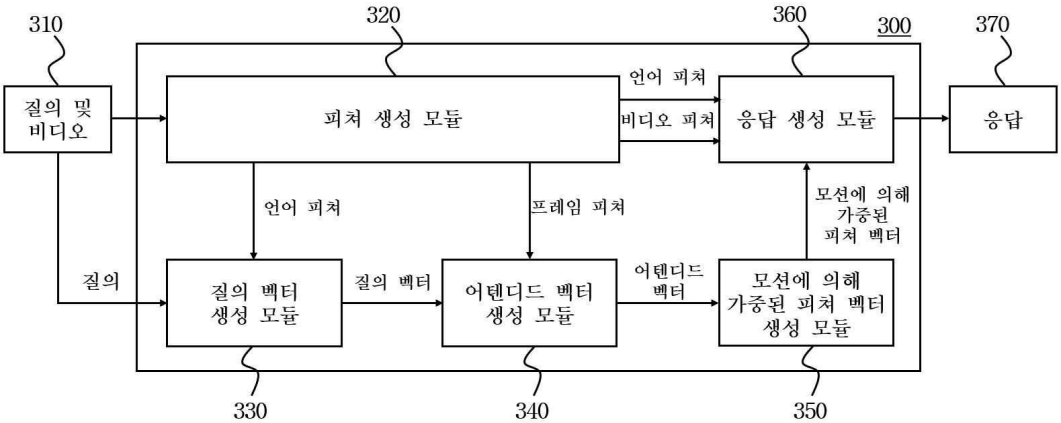
도면1



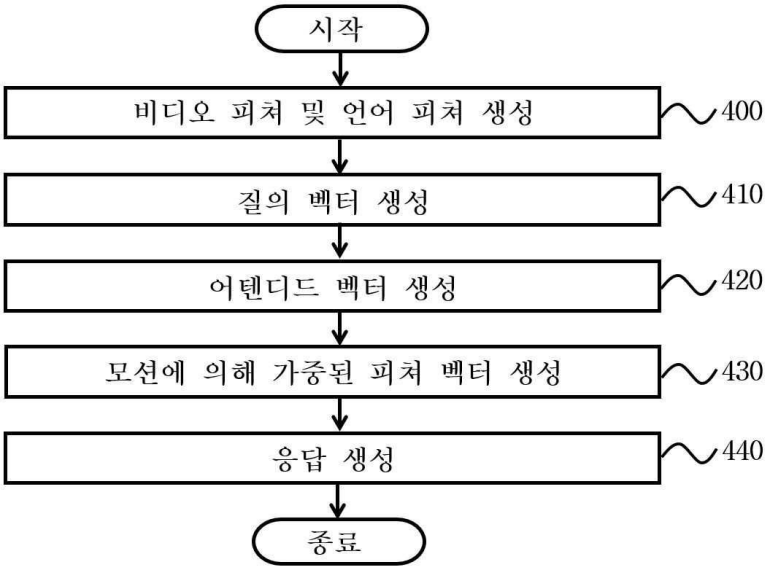
도면2



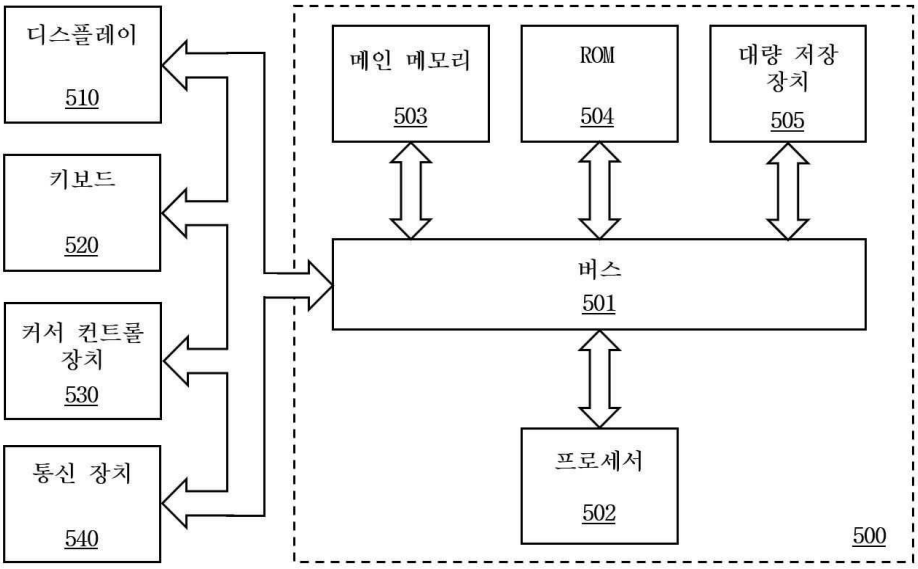
도면3



도면4


시작
비디오 피처 및 언어 피처 생성
400
질의 벡터 생성
410
어텐디드 벡터 생성
420
모션에 의해 가중된 피처 벡터 생성
430
응답 생성
440
종료

도면5


디스플레이
510
키보드
520
커서 컨트롤 장치
530
통신 장치
540
메인 메모리
503
ROM
504
대량 저장 장치
505
버스
501
프로세서
502
500