

(12) 특허협력조약에 의하여 공개된 국제출원

(19) 세계지식재산권기구
국제사무국

(43) 국제공개일

2020년 6월 11일 (11.06.2020)



(10) 국제공개번호

WO 2020/117028 A1

(51) 국제특허분류:

G06F 16/783 (2019.01)

G06F 16/432 (2019.01)

G06F 16/73 (2019.01)

G06N 3/08 (2006.01)

(21) 국제출원번호:

PCT/KR2019/017311

(22) 국제출원일:

2019년 12월 9일 (09.12.2019)

(25) 출원언어:

한국어

(26) 공개언어:

한국어

(30) 우선권정보:

10-2018-0157356 2018년 12월 7일 (07.12.2018) KR

10-2019-0162788 2019년 12월 9일 (09.12.2019) KR

(71) 출원인: 서울대학교 산학협력단 (SEOUL NATIONAL UNIVERSITY R&DB FOUNDATION) [KR/KR]; 08826 서울시 관악구 관악로 1 (신림동), Seoul (KR).

(72) 발명자: 장병탁 (ZHANG, Byoung-Tak); 06603 서울시 서초구 서초중앙로24길 33, 103동 501호, Seoul (KR). 최

성호 (CHOI, Seong-Ho); 08748 서울시 관악구 봉천로 41길 18-4, 406호, Seoul (KR).

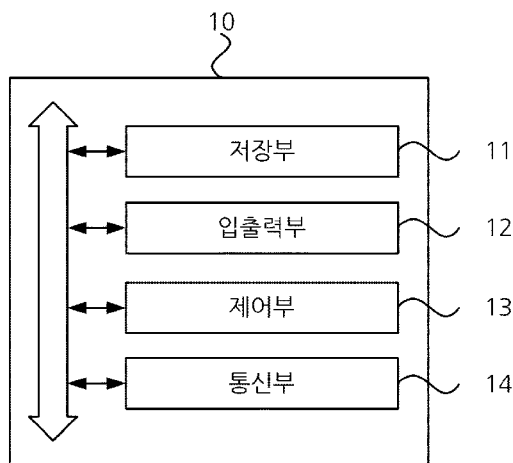
(74) 대리인: 특허법인 아이스퀘어 (ISQUARE PATENT & LAW FIRM); 06174 서울시 강남구 영동대로82길 11, 5층(대치동, 도원빌딩), Seoul (KR).

(81) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 국내 권리의 보호를 위하여): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) 지정국 (별도의 표시가 없는 한, 가능한 모든 종류의 국내 권리의 보호를 위하여): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM,

(54) Title: QUERY RESPONSE DEVICE AND METHOD

(54) 발명의 명칭: 질의 응답 장치 및 방법



11 ... Storage unit

12 ... Input/output unit

13 ... Control unit

14 ... Communication unit

(57) Abstract: A query response method performed by a query response device comprises the steps of: classifying an image frame, audio data, and subtitle data that are included in video data of a data set in one shot unit, on the basis of one same subtitle; extracting a shot feature vector by calculating a feature vector of each of an image frame, audio data, and subtitle data that are included in a shot; extracting, from a pair of query responses included in the data set, a feature vector of question data and a feature vector of each of a plurality of pieces of option data corresponding to the question data; calculating a video feature vector by inputting the shot feature vector to a neural network including a plurality of layers, assigning an attention weight calculated on the basis of the question data to an output vector of each of the layers, and then summing results; and selecting a final answer from among the plurality of pieces of option data, on the basis of a similarity between the video feature vector and an option feature vector.

(57) 요약서: 질의 응답 장치에 의해 수행되는 질의 응답 방법은 데이터 세트의 비디오 데이터에 포함된 이미지 프레임, 오디오 데이터, 자막 데이터를, 동일한 하나의 자막을 기준으로 한 샷 단위로 구분하는 단계; 샷에 포함된 이미지 프레임, 오디오 데이터, 자막 데이터 각각의 특징벡터를 연산하여 샷 특징벡터를 추출하는 단계; 상기 데이터 세트에 포함된 질의 응답 쌍에서 질문 데이터와 상기 질문 데이터에 대응하는 복수의 선택지 데이터 각각의 특징벡터를 추출하는 단계; 복수의 계층의 신경망에 상기 샷 특징벡터를 입력한 후, 각 계층의 출력벡터에 상기 질문 데이터에 기초하여 연산된 주의 가중치를 부여하여 합산함으로써 비디오 특징벡터를 산출하는 단계; 그리고 상기 비디오 특징벡터와 선택지 특징벡터 사이의 유사도에 기초하여 상기 복수의 선택지 데이터 중 최종 답안을 선택하는 단계를 포함한다.

ZW), 유라시아 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 유럽 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

공개:

- 국제조사보고서와 함께 (조약 제21조(3))
- 청구범위 보정 기한 만료 전의 공개이며, 보정서를 접수하는 경우 그에 관하여 별도 공개함 (규칙 48.2(h))

명세서

발명의 명칭: 질의 응답 장치 및 방법

기술분야

- [1] 본 명세서에 개시되는 실시예들은 계층적 비디오 스토리 모델링과 이를 이용하여 비디오를 대상으로 하는 질의에 응답하는 질의 응답 장치 및 방법에 대한 것이다.
- [2] 2019년 과제정보 및 사사표기
- [3] 1-1. 과제고유번호 : 1711081135
- [4] 1-2. 사사표기 : 본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 혁신성장동력프로젝트사업의 연구결과로 수행되었음(IITP-2017-0-01772-003).
- [5] 2-1. 과제고유번호 : 1711081008
- [6] 2-2. 사사표기 : 본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 SW컴퓨팅산업원천기술개발(R&D, 정보화)의 연구결과로 수행되었음(IITP-2015-0-00310-005).

배경기술

- [7] 최근 비디오 스토리 모델링 분야는 시각 및 자연어 처리 연구를 융합하는 고난도의 분야로 새롭게 떠오르고 있다. 이러한 시각 및 자연어 처리의 융합된 영역을 다루는 대표적인 기술로서 Visual Question Answering(VQA) 기술이 있다. VQA기술은 비디오에 내포된 의미나 줄거리에 대한 질의에 응답할 수 있도록 비디오에 포함된 데이터를 분석하고 그에 따라 질문에 대한 답변을 추론한다.
- [8] 그러나 비디오 스토리 질의응답 문제는 단순한 이미지와는 차원이 다르게 많은 정보를 내포하는 비디오 자체의 특성과, 그 특성에서 오는 질의의 다양한 종류로 인한 문제 난이도의 증가로 인해, 해결하기 어려운 분야로 평가 받고 있다.
- [9] 그에 따라 데이터 세트의 수집이 어려울 뿐 아니라, 현재 선행되어 있는 비디오 스토리 질의응답 데이터 세트는 여러 한계점을 가지고 있다.
- [10] 또한 종래의 비디오 질의응답 모델들은 자막과 영상의 이미지 프레임 정보만을 사용했기 때문에, 자막이나 영상에서 추론되지 않는 등장인물의 목소리에 담긴 화자나 감정에 대한 정보를 이용하기 어렵다는 문제가 있었다.
- [11] 따라서 비디오 질의응답의 난이도를 개선함과 동시에, 화자나 감정을 구분하여 질의에 대한 적절한 응답을 추론을 수행할 수 있는 방법이 요구된다.
- [12] 한편, 전술한 배경기술은 발명자가 본 발명의 도출을 위해 보유하고 있었거나, 본 발명의 도출 과정에서 습득한 기술 정보로서, 반드시 본 발명의 출원 전에 일반 공중에게 공개된 공지기술이라 할 수는 없다.

발명의 상세한 설명

기술적 과제

- [13] 본 명세서에서 개시되는 실시예들은 질의 응답 장치 및 방법을 제시하는 데에

그 목적이 있다.

- [14] 본 명세서에서 개시되는 실시예들은 질의에 따른 주의집중을 이용하여 비디오 스토리에 대한 질의응답을 수행하는 질의 응답 장치 및 방법을 제시하는 데에 그 목적이 있다.
- [15] 본 명세서에 개시되는 실시예들은 계층적으로 비디오 스토리를 모델링하고 그에 기초하여 질의응답을 수행하는 질의 응답 장치 및 방법을 제시하는 데에 그 목적이 있다.
- [16] 나아가 본 명세서에 개시되는 실시예들은 비디오 데이터 세트에 포함된 오디오 정보를 이미지 프레임 및 자막 정보와 함께 이용하여 화자나 감정을 구분하여 질의에 대한 적절한 응답을 수행하는 질의 응답 장치 및 방법을 제시하는 데에 그 목적이 있다.

과제 해결 수단

- [17] 질의 응답 장치에 의해 수행되는 질의 응답 방법은, 데이터 세트의 비디오 데이터에 포함된 이미지 프레임, 오디오 데이터, 자막 데이터를, 동일한 하나의 자막을 기준으로 한 샷 단위로 구분하는 단계; 샷에 포함된 이미지 프레임, 오디오 데이터, 자막 데이터 각각의 특징벡터를 연산하여 샷 특징벡터를 추출하는 단계; 상기 데이터 세트에 포함된 질의응답 쌍에서 질문 데이터와 상기 질문 데이터에 대응하는 복수의 선택지 데이터 각각의 특징벡터를 추출하는 단계; 복수의 계층의 신경망에 상기 샷 특징벡터를 입력한 후, 각 계층의 출력벡터에 상기 질문 데이터에 기초하여 연산된 주의 가중치를 부여하여 합산함으로써 비디오 특징벡터를 산출하는 단계; 그리고 상기 비디오 특징벡터와 선택지 특징벡터 사이의 유사도에 기초하여 상기 복수의 선택지 데이터 중 최종 답안을 선택하는 단계를 포함한다.

발명의 효과

- [18] 전술한 본 발명의 과제 해결 수단 중 어느 하나에 의하면, 질의 응답 장치 및 방법을 제시할 수 있다.
- [19] 본 발명의 과제 해결 수단 중 어느 하나에 의하면, 질의에 따른 주의집중을 이용하여 비디오 스토리에 대한 질의응답을 수행하는 질의 응답 장치 및 방법을 제시할 수 있다.
- [20] 본 발명의 과제 해결 수단 중 어느 하나에 의하면, 계층적으로 비디오 스토리를 모델링하고 그에 기초하여 질의응답을 수행하는 질의 응답 장치 및 방법을 제시할 수 있다.
- [21] 본 발명의 과제 해결 수단 중 어느 하나에 의하면, 비디오 데이터 세트에 포함된 오디오 정보를 이미지 프레임 및 자막 정보와 함께 이용하여 화자나 감정을 구분하여 질의에 대한 적절한 응답을 수행하는 질의 응답 장치 및 방법을 제시할 수 있다.
- [22] 본 발명에서 얻을 수 있는 효과는 이상에서 언급한 효과들로 제한되지 않으며,

언급하지 않은 또 다른 효과들은 아래의 기재로부터 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

- [23] 도 1은 일 실시예에 따른 질의 응답 장치의 구성을 도시한 도면이다.
- [24] 도 2는 일 실시예에 따른 질의 응답 장치에서 사용하는 데이터 세트의 일례를 도시한 도면이다.
- [25] 도 3은 일 실시예에 따른 질의 응답 장치에 의해 수행되는 질의 응답 방법을 단계적으로 도시한 도면이다.
- [26] 도 4는 일 실시예에 따른 질의 응답 방법에서 샷 특징을 추출하는 과정을 개념적으로 도시한 도면이다.
- [27] 도 5는 일 실시예에 따른 질의 응답 방법에서 질문에 따른 최종 답안을 선택하는 과정을 개념적으로 도시한 도면이다.

발명의 실시를 위한 최선의 형태

- [28] 상술한 기술적 과제를 달성하기 위한 기술적 수단으로서, 일 실시예에 따르면, 질의 응답 장치에 의해 수행되는 질의 응답 방법은, 데이터 세트의 비디오 데이터에 포함된 이미지 프레임, 오디오 데이터, 자막 데이터를, 동일한 하나의 자막을 기준으로 한 샷 단위로 구분하는 단계; 샷에 포함된 이미지 프레임, 오디오 데이터, 자막 데이터 각각의 특징벡터를 연산하여 샷 특징벡터를 추출하는 단계; 상기 데이터 세트에 포함된 질의응답 쌍에서 질문 데이터와 상기 질문 데이터에 대응하는 복수의 선택지 데이터 각각의 특징벡터를 추출하는 단계; 복수의 계층의 신경망에 상기 샷 특징벡터를 입력한 후, 각 계층의 출력벡터에 상기 질문 데이터에 기초하여 연산된 주의 가중치를 부여하여 합산함으로써 비디오 특징벡터를 산출하는 단계; 그리고 상기 비디오 특징벡터와 선택지 특징벡터 사이의 유사도에 기초하여 상기 복수의 선택지 데이터 중 최종 답안을 선택하는 단계를 포함한다.
- [29] 다른 실시예에 따르면, 질의 응답 장치는 비디오 데이터와 질의응답 쌍을 포함하는 데이터 세트를 저장하는 저장부; 그리고 상기 비디오 데이터에 포함된 이미지 프레임, 오디오 데이터, 자막 데이터를, 동일한 하나의 자막을 기준으로 한 샷 단위로 구분한 후, 각각의 샷에 포함된 이미지 프레임, 오디오 데이터, 자막 데이터를 분석함으로써 샷 특징벡터를 추출하고, 복수의 계층의 신경망에 상기 샷 특징벡터를 입력하며, 상기 질의응답 쌍에서 질문 데이터와 상기 질문 데이터에 대응하는 복수의 선택지 데이터 각각의 특징벡터를 추출하여, 상기 복수의 계층의 신경망의 각 계층의 출력벡터를 이용하여 비디오 특징벡터를 산출하고, 상기 비디오 특징벡터와 선택지 특징벡터 사이의 유사도에 기초하여 상기 복수의 선택지 데이터 중 최종 답안을 선택하는 제어부를 포함할 수 있다.
- [30] 또 다른 실시예에 따르면, 질의 응답 방법을 수행하는 프로그램이 기록된 컴퓨터 판독 가능한 기록 매체로서, 상기 질의 응답 방법은 데이터 세트의

비디오 데이터에 포함된 이미지 프레임, 오디오 데이터, 자막 데이터를, 동일한 하나의 자막을 기준으로 한 샷 단위로 구분하는 단계; 샷에 포함된 이미지 프레임, 오디오 데이터, 자막 데이터 각각의 특징벡터를 연산하여 샷 특징벡터를 추출하는 단계; 상기 데이터 세트에 포함된 질의응답 쌍에서 질문 데이터와 상기 질문 데이터에 대응하는 복수의 선택지 데이터 각각의 특징벡터를 추출하는 단계; 복수의 계층의 신경망에 상기 샷 특징벡터를 입력한 후, 각 계층의 출력벡터에 상기 질문 데이터에 기초하여 연산된 주의 가중치를 부여하여 합산함으로써 비디오 특징벡터를 산출하는 단계; 그리고 상기 비디오 특징벡터와 선택지 특징벡터 사이의 유사도에 기초하여 상기 복수의 선택지 데이터 중 최종 답안을 선택하는 단계를 포함한다.

- [31] 또 다른 실시예에 따르면, 질의 응답 장치에 의해 수행되며, 질의 응답 방법을 수행하기 위해 매체에 저장된 컴퓨터 프로그램으로서, 상기 질의 응답 방법은, 데이터 세트의 비디오 데이터에 포함된 이미지 프레임, 오디오 데이터, 자막 데이터를, 동일한 하나의 자막을 기준으로 한 샷 단위로 구분하는 단계; 샷에 포함된 이미지 프레임, 오디오 데이터, 자막 데이터 각각의 특징벡터를 연산하여 샷 특징벡터를 추출하는 단계; 상기 데이터 세트에 포함된 질의응답 쌍에서 질문 데이터와 상기 질문 데이터에 대응하는 복수의 선택지 데이터 각각의 특징벡터를 추출하는 단계; 복수의 계층의 신경망에 상기 샷 특징벡터를 입력한 후, 각 계층의 출력벡터에 상기 질문 데이터에 기초하여 연산된 주의 가중치를 부여하여 합산함으로써 비디오 특징벡터를 산출하는 단계; 그리고 상기 비디오 특징벡터와 선택지 특징벡터 사이의 유사도에 기초하여 상기 복수의 선택지 데이터 중 최종 답안을 선택하는 단계를 포함한다.

발명의 실시를 위한 형태

- [32] 아래에서는 첨부한 도면을 참조하여 다양한 실시예들을 상세히 설명한다. 아래에서 설명되는 실시예들은 여러 가지 상이한 형태로 변형되어 실시될 수도 있다. 실시예들의 특징을 보다 명확히 설명하기 위하여, 이하의 실시예들이 속하는 기술분야에서 통상의 지식을 가진 자에게 널리 알려져 있는 사항들에 관해서 자세한 설명은 생략하였다. 그리고, 도면에서 실시예들의 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.
- [33] 명세서 전체에서, 어떤 구성이 다른 구성과 "연결"되어 있다고 할 때, 이는 '직접적으로 연결'되어 있는 경우뿐 아니라, '그 중간에 다른 구성을 사이에 두고 연결'되어 있는 경우도 포함한다. 또한, 어떤 구성이 어떤 구성을 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한, 그 외 다른 구성을 제외하는 것이 아니라 다른 구성들을 더 포함할 수도 있음을 의미한다.
- [34] 이하 첨부된 도면을 참고하여 실시예들을 상세히 설명하기로 한다.
- [35] 도 1은 일 실시예에 따른 질의 응답 장치의 구성을 도시한 도면이다. 질의 응답

장치(10)는 비디오를 분석하여 비디오에 대한 질의에 대하여 최적의 응답을 산출하는 정보 처리 장치로 구성될 수 있다.

- [36] 구체적으로 질의 응답 장치(10)는 저장부(11), 입출력부(12), 제어부(13), 그리고 통신부(14)를 포함할 수 있다.
- [37] 저장부(11)는 질의응답에 필요한 각종 데이터나 프로그램들을 저장할 수 있다. 저장부(11)는 비디오에 대한 질의 응답의 학습 과정에 필요한 데이터 세트를 저장할 수 있다. 이때 데이터 세트는 비디오 데이터에 포함되는 이미지 프레임, 오디오 데이터, 자막 데이터와 해당 비디오 데이터에 대응하는 질의 데이터와 응답이 될 선택지 데이터, 정답 데이터 등에 대한 정보를 포함할 수 있다. 특히 데이터 세트는 비디오에 대한 학습(training) 세트, 확인(Validation) 세트, 테스트(Test) 세트를 모두 포함할 수 있다.
- [38] 또한 저장부(11)는 비디오에 대한 질의응답을 학습하고, 그에 따라 질의에 대한 응답을 수행할 수 있도록 하는 프로그램과, 학습 및 질의응답에 이용되는 각각의 모델이나 신경망에 연관된 데이터들을 적어도 일시적으로 보유하거나 갱신할 수 있다.
- [39] 한편 입출력부(12)는 데이터나 사용자 명령을 수신하고, 사용자의 입력에 따라 데이터를 연산하고 처리한 결과를 출력하는 구성이다. 일 실시예에 따르면 입출력부(12)는 키보드나 마우스, 터치패널 등의 사용자 입력수단과, 모니터나 스피커 등의 출력수단을 포함할 수 있다.
- [40] 그리고 제어부(13)는 CPU 등과 같은 적어도 하나의 프로세서를 포함하는 구성으로서, 질의 응답 장치(10)의 전반적인 동작과 구성을 제어할 수 있다. 특히 제어부(13)는 저장부(11)에 저장된 프로그램을 실행하거나 데이터를 읽어 비디오 스토리 모델링을 수행하고 그에 따라 질의에 대한 적절한 응답을 선택할 수 있다.
- [41] 특히 제어부(13)는 비디오의 이미지 프레임과, 오디오, 자막을 각각 분석하여 비디오 줄거리를 모델링한다. 제어부(13)가 비디오 스토리를 모델링하고 그에 따라 질의 응답을 수행하는 구체적인 과정은 추후 더욱 구체적으로 설명한다.
- [42] 한편 통신부(14)는 질의 응답 장치(10)가 다른 장치와 데이터를 교환하도록 하는 구성이다. 통신부(14)는 제어부(13)가 분석할 비디오 데이터 세트를 수신하거나, 제어부(13)가 비디오 데이터 세트를 분석하거나 이를 이용하여 비디오 스토리를 모델링하는 과정에서 필요한 신경망과 관련된 데이터를 수신하여 제공할 수도 있다. 나아가 통신부(14)는 타 단말과 통신하여 제어부(13)로 타 단말에서 수신한 질의를 전달하거나 제어부(13)가 연산한 응답을 타 단말로 제공할 수도 있다.
- [43] 이하에서는 질의 응답 장치(10)가 구체적으로 수행하는 질의 응답 과정을 설명한다. 다만 이에 앞서 우선 비디오 스토리 모델링에 이용하는 데이터 세트에 대해 설명한다. 도 2는 일 실시예에 따른 질의 응답 장치에서 사용하는 데이터 세트의 일례를 도시한 도면이다.

- [44] 우선 비디오 데이터 세트는 도 2에 도시된 바와 같이 특정 비디오 콘텐츠를 구성하는 이미지 프레임, 오디오, 자막 데이터를 포함한다. 그리고 데이터 세트에는 질의응답 쌍이 포함되는데, 이때 질의응답 쌍에는 질문 데이터와 질문 대한 복수의 선택지 데이터가 포함될 수 있다. 또한 이때 선택지 중 정답 데이터가 다른 선택지와 구분 가능하게 포함될 수 있다.
- [45] 도 2에 도시된 실시예는 미국 드라마 ‘프렌즈’의 영상에 대한 질의응답 쌍을 포함하는 FriendsQA 데이터 세트를 예시한 것이다. 도시된 것처럼 데이터 세트에는 특정 에피소드를 구성하는 장면의 이미지 프레임들과 오디오 데이터, 자막 데이터가 포함되며, 관련 영상에 대한 질문 데이터와 대응하는 정답을 포함하는 선택지 데이터가 하나 이상 포함된다.
- [46] 일 실시예에서 이미지 프레임과 오디오 데이터, 그리고 자막 데이터는, 자막 데이터를 기준으로 샷(Shot) 단위로 구분될 수 있다. 즉 하나의 자막이 표시되는 시간 범위 내의 이미지 프레임과 오디오 데이터, 그리고 해당 자막 데이터의 집합이 하나의 샷을 구성하도록 구분될 수 있다.
- [47] 질의 응답 장치(10)가 수행하는 질의 응답 방법은 도 3 내지 도 5를 참조하여 설명한다.
- [48] 도 3은 일 실시예에 따른 질의 응답 장치에 의해 수행되는 질의 응답 방법을 단계적으로 도시한 도면이고, 도 4는 일 실시예에 따른 질의 응답 방법에서 샷 특징을 추출하는 과정을 개념적으로 도시한 도면이며, 도 5는 일 실시예에 따른 질의 응답 방법에서 질문에 따른 최종 답안을 선택하는 과정을 개념적으로 도시한 도면이다.
- [49] 도면에 도시된 바와 같이 질의 응답 장치(10)는 비디오 스토리 모델링을 수행할 비디오 데이터 세트를 준비한다(S310).
- [50] 이때 데이터 세트는 상술한 바와 같이 비디오를 구성하는 이미지 프레임, 오디오 데이터 및 자막 데이터와, 질의응답 쌍을 구성하는 질문 데이터 및 선택지 데이터를 포함할 수 있다.
- [51] 그리고 질의 응답 장치(10)는 이미지 프레임, 오디오 데이터, 그리고 자막 데이터를 샷 단위로 구분할 수 있다. 즉 질의 응답 장치(10)는 하나의 자막의 시작부터 끝까지의 한 세그먼트(Segment)를 각각의 샷으로 구분한다(S320).
- [52] 이때 샷으로 나누어진 비디오는 $X = \{v_i, a_i, s_i\}_{i=1, \dots, |X|}$ 이라고 정의할 수 있다. 이때 v_i 는 이미지 시퀀스(image sequence)의 특징벡터이며, a_i 는 추출된 오디오의 특징벡터, s_i 는 자막을 워드 임베딩(word embedding)하여 산출한 특징벡터이다. 그리고 각 샷의 특징은 이들의 집합으로 표현할 수 있다.
- [53] 그리고 질의 응답 장치(10)는 이어서 각각의 샷의 특징을 추출한다(S330).
- [54] 이때 S330단계에서 샷의 특징을 추출하는 과정은 다시 S331단계 내지 S334단계로 구분될 수 있다.
- [55] 우선 질의 응답 장치(10)는 오디오 특징벡터(a_i)를 추출할 수 있다(S331). S331단계에서는 오디오 데이터에 포함된 소리의 특징을 추출하는 과정으로

수행된다. 도 4에 도시된 실시예에 의하면 S331단계는 엠에프씨씨(MFCC: Mel-Frequency Cepstral Coefficients)기반의 콘볼루션신경망(CNN: Convolution Neural Network) 모델을 이용하여 수행될 수 있다. 하나의 샷의 오디오 데이터에 대하여 40ms의 해밍 윈도우즈(Hamming windows)를 25ms 만큼 중복되도록 구간을 나누어 적용하여, 가장 낮은 39차원의 특징을 추출한다. 추출된 (N, 39) 차원의 특징에 대하여 필터(filter)의 개수가 다수인 1-D CNN을 두 번 적용하고, 글로벌 맵 풀링(global max pooling)을 이용하여 고정된 차원의 오디오 특징을 추출할 수 있다. 그리고 추출된 오디오 특징에 한 층의 전체 연결 레이어(Fully connected layer)를 적용하여 300차원의 오디오 특징벡터(a_i)를 추출할 수 있다.

- [56] 한편 질의 응답 장치(10)는 자막 특징벡터(s_i)를 추출할 수 있다(S332). 질의 응답 장치(10)는 S332단계에서 자연어 처리 모델을 이용하여 워드 임베딩(word embedding)을 수행하고 워드 단위의 특징을 신경망 모델에 입력하여 고정된 차원의 특징벡터를 산출한다. 도 4에 도시된 실시예에 의하면, 질의 응답 장치(10)는 S332단계에서 자연어 처리를 위한 프리-트레인드(pre-trained) 모델을 사용하여 워드 임베딩을 수행하고, 이 워드 단위의 특징을 양방향 장단기메모리(LSTM: Long Short-Term Memory) 모델의 입력으로 넣어 고정된 차원의 특징벡터를 구함으로써, 자막 특징벡터(s_i)를 산출할 수 있다.

- [57] 한편 도 3에는 도시되어 있지 않으나, 도 4에 도시된 실시예에서와 같이 오디오 특징벡터(a_i)와 자막 특징벡터(s_i)를 멀티모드 조인트 임베딩(Multimodal Joint Embedding)을 통해서 이미지 특징벡터(v_i)의 주의집중 특징 벡터로 사용할 수 있다. 구체적으로는 오디오 특징벡터(a_i)와 자막 특징벡터(s_i)를 후술할 이미지 특징벡터(v_i)와 동일한 차원으로 임베딩한 후 더하여 조인트 임베딩(joint embedding) 벡터(e_i)를 만들 수 있다. 그리고 질의 응답 장치(10)는 조인트 임베딩 벡터(e_i)를 이미지 프레임 특징벡터(v_i)를 산출하는 과정에서 주의집중 특징벡터로 이용할 수 있다.

- [58] 한편 질의 응답 장치(10)는 각각의 샷에 대한 이미지 프레임 특징벡터(v_i)를 추출할 수 있다(S333). 이 단계에서 질의 응답 장치(10)는 하나의 샷에 대해서 초당 3개의 이미지 프레임을 샘플링할 수 있다. 그리고 질의 응답 장치(10)는 샘플된 이미지를 신경망에 입력하여 각각의 이미지에 대한 특징벡터를 산출할 수 있다.

- [59] 예를 들어, 도 4에 도시된 실시예에서 질의 응답 장치(10)는 샘플된 이미지마다 ImageNet pre-trained ResNet-152를 적용하여 (7, 7, 2048) 차원의 텐서를 추출할 수 있다. 그리고 추출된 텐서와 앞서 구한 조인트 임베딩 벡터(e_i)를 내적하여 (7, 7) 차원의 주의맵을 얻을 수 있다. 그리고 질의 응답 장치(10)는 주의맵에 포함된 각각의 주의 가중치를 추출된 텐서에 곱한 후 49개의 2,048차원의 특징 벡터를 모두 더하면 특징벡터를 구할 수 있다. 이렇게 이미지 프레임에 대한 특징벡터를 구하면 하나의 샷에 대해서 가변 길이의 특징벡터를 얻을 수 있는데, 이를 평균 내어 하나의 샷에 대한 이미지 프레임 특징벡터(v_i)로 나타낼 수 있다.

- [60] 그리고 질의 응답 장치(10)는 하나의 샷에 대하여 구한 세 가지 특징벡터를 연결(Concatenation)하여 하나의 샷 특징벡터를 구할 수 있다.
- [61] 한편 질의 응답 장치(10)는 각 질문에 대하여 질문 특징벡터(q)를 산출할 수 있고(S340), 각각의 선택지에 대해서도 선택지 특징벡터(α_i)를 추출할 수 있다(S350). 이때, 각각 자연어로 구성되는 질문과 선택지의 특징벡터를 추출하기 위해 질의 응답 장치(10)는 앞서 자막 특징벡터(s_i)를 추출할 때와 마찬가지로 자연어 처리 모델을 이용하여 워드 임베딩(word embedding)을 수행하고 워드 단위의 특징을 신경망 모델에 입력하여 고정된 차원의 특징벡터를 산출할 수 있다. 예를 들어, 도 4에 도시된 바와 같이 질의 응답 장치(10)는 자연어 처리를 위한 프리-트레인드(pre-trained) 모델을 사용하여 워드 임베딩을 수행하고, 이 워드 단위의 특징을 양방향 장단기메모리(LSTM: Long Short-Term Memory) 모델의 입력으로 넣어 고정된 차원의 특징벡터를 구함으로써 질문 특징벡터(q)와 선택지 특징벡터(α_i)를 추출할 수 있다.
- [62] 그리고 한편 질의 응답 장치(10)는 질문 특징벡터(q)를 선형 변환하여 후술할 S380단계에서 이용될 주의집중 특징벡터를 산출할 수 있다(S360).
- [63] 한편 질의 응답 장치(10)는 S330단계를 거쳐 산출된 각각의 샷 특징벡터를 여러 계층의 컨볼루션(convolution)을 통해 계층적으로 나타낼 수 있다. 일 실시예에서 제안하는 모델은 복수의 계층의 신경망, 예를 들어 세 개의 층의 1-D CNN을 이용한 방식이다. 질의 응답 장치(10)는 세 개의 1-D CNN의 각 계층에 샷 특징벡터를 임베딩(S370)할 수 있다. 그리고 질의 응답 장치(10)는 글로벌 맥스-풀링(global max-pooling)하여 여러 수준의 출력 벡터를 산출할 수 있다.
- [64] 이어서 질의 응답 장치(10)는 S360단계에서 산출한 주의집중 특징벡터에 포함된 각 주의 가중치를 각각의 계층의 출력 벡터에 부여한 후 합산하여 가중합을 산출할 수 있다(S380). 이러한 가중합은 질문에 대해 주의집중 되었기 때문에, 질문에 가장 적절한 '비디오 특징벡터'가 된다.
- [65] 그리고 질의 응답 장치(10)는 비디오 특징벡터와 선택지 특징벡터(α_i)간의 내적 유사도(dot-product similarity)에 기초하여, 복수의 선택지 중 하나의 최종 답안을 선택한다.
- [66] 한편 질의 응답 장치(10)는 복수의 질의응답 쌍에 대하여 상술한 과정을 반복 수행하면서 최종 답안이 미리 설정된 정답이 되도록 복수의 계층의 신경망을 학습시킬 수 있다.
- [67] 그리고 나아가 질의 응답 장치(10)는 위와 같이 데이터 세트를 통해 학습된 복수의 계층의 신경망을 이용하여 사용자의 질의에 대응하는 적절한 응답을 사용자에게 제공할 수 있다.
- [68] 실시예에서 사용되는 '~부'라는 용어는 소프트웨어 또는 FPGA(field programmable gate array) 또는 ASIC 와 같은 하드웨어 구성요소를 의미하며, '~부'는 어떤 역할들을 수행한다. 그렇지만 '~부'는 소프트웨어 또는 하드웨어에 한정되는 의미는 아니다. '~부'는 어드레싱할 수 있는 저장 매체에 있도록 구성될

수도 있고 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다. 따라서, 일 예로서 '~부'는 소프트웨어 구성요소들, 객체지향 소프트웨어 구성요소들, 클래스 구성요소들 및 태스크 구성요소들과 같은 구성요소들과, 프로세스들, 함수들, 속성들, 프로시저들, 서브루틴들, 프로그램특허 코드의 세그먼트들, 드라이버들, 펌웨어, 마이크로코드, 회로, 데이터, 데이터베이스, 데이터 구조들, 테이블들, 어레이들, 및 변수들을 포함한다.

- [69] 구성요소들과 '~부'들 안에서 제공되는 기능은 더 작은 수의 구성요소들 및 '~부'들로 결합되거나 추가적인 구성요소들과 '~부'들로부터 분리될 수 있다.
- [70] 뿐만 아니라, 구성요소들 및 '~부'들은 디바이스 또는 보안 멀티미디어카드 내의 하나 또는 그 이상의 CPU 들을 재생시키도록 구현될 수도 있다.
- [71] 또한 본 발명의 일 실시예에 따르는 질의 응답 방법은 컴퓨터에 의해 실행 가능한 명령어를 포함하는 컴퓨터 프로그램(또는 컴퓨터 프로그램 제품)으로 구현될 수도 있다. 컴퓨터 프로그램은 프로세서에 의해 처리되는 프로그래밍 가능한 기계 명령어를 포함하고, 고레벨 프로그래밍 언어(High-level Programming Language), 객체 지향 프로그래밍 언어(Object-oriented Programming Language), 어셈블리 언어 또는 기계 언어 등으로 구현될 수 있다. 또한 컴퓨터 프로그램은 유형의 컴퓨터 판독가능 기록매체(예를 들어, 메모리, 하드디스크, 자기/광학 매체 또는 SSD(Solid-State Drive) 등)에 기록될 수 있다.
- [72] 따라서 일 실시예에 따르는 질의 응답 방법은 상술한 바와 같은 컴퓨터 프로그램이 컴퓨팅 장치에 의해 실행됨으로써 구현될 수 있다. 컴퓨팅 장치는 프로세서와, 메모리와, 저장 장치와, 메모리 및 고속 확장포트에 접속하고 있는 고속 인터페이스와, 저속 버스와 저장 장치에 접속하고 있는 저속 인터페이스 중 적어도 일부를 포함할 수 있다. 이러한 성분들 각각은 다양한 버스를 이용하여 서로 접속되어 있으며, 공통 머더보드에 탑재되거나 다른 적절한 방식으로 장착될 수 있다.
- [73] 여기서 프로세서는 컴퓨팅 장치 내에서 명령어를 처리할 수 있는데, 이런 명령어로는, 예컨대 고속 인터페이스에 접속된 디스플레이처럼 외부 입력, 출력 장치상에 GUI(Graphic User Interface)를 제공하기 위한 그래픽 정보를 표시하기 위해 메모리나 저장 장치에 저장된 명령어를 들 수 있다. 다른 실시예로서, 다수의 프로세서 및(또는) 다수의 버스가 적절히 다수의 메모리 및 메모리 형태와 함께 이용될 수 있다. 또한 프로세서는 독립적인 다수의 아날로그 및(또는) 디지털 프로세서를 포함하는 칩들이 이루는 칩셋으로 구현될 수 있다.
- [74] 또한 메모리는 컴퓨팅 장치 내에서 정보를 저장한다. 일례로, 메모리는 휘발성 메모리 유닛 또는 그들의 집합으로 구성될 수 있다. 다른 예로, 메모리는 비휘발성 메모리 유닛 또는 그들의 집합으로 구성될 수 있다. 또한 메모리는 예컨대, 자기 혹은 광 디스크와 같이 다른 형태의 컴퓨터 판독 가능한 매체일 수도 있다.
- [75] 그리고 저장장치는 컴퓨팅 장치에게 대용량의 저장공간을 제공할 수 있다.

저장 장치는 컴퓨터 판독 가능한 매체이거나 이런 매체를 포함하는 구성일 수 있으며, 예를 들어 SAN(Storage Area Network) 내의 장치들이나 다른 구성도 포함할 수 있고, 플로피 디스크 장치, 하드 디스크 장치, 광 디스크 장치, 혹은 테이프 장치, 플래시 메모리, 그와 유사한 다른 반도체 메모리 장치 혹은 장치 어레이일 수 있다.

[76] 전술한 실시예들은 예시를 위한 것이며, 상술한 실시예들이 속하는 기술분야의 통상의 지식을 가진 자는 상술된 실시예들이 갖는 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다. 예를 들어, 단일형으로 설명되어 있는 각 구성 요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명되어 있는 구성 요소들도 결합된 형태로 실시될 수 있다.

[77] 본 명세서를 통해 보호받고자 하는 범위는 상기 상세한 설명보다는 후술하는 특허청구범위에 의하여 나타내어지며, 특허청구범위의 의미 및 범위 그리고 그 균등 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

청구범위

- [청구항 1] 질의 응답 장치에서 수행되는 질의 응답 방법에 있어서,
 데이터 세트의 비디오 데이터에 포함된 이미지 프레임, 오디오 데이터,
 자막 데이터를, 동일한 하나의 자막을 기준으로 한 샷 단위로 구분하는
 단계;
 샷에 포함된 이미지 프레임, 오디오 데이터, 자막 데이터 각각의
 특징벡터를 연산하여 샷 특징벡터를 추출하는 단계;
 상기 데이터 세트에 포함된 질의응답 쌍에서 질문 데이터와 상기 질문
 데이터에 대응하는 복수의 선택지 데이터 각각의 특징벡터를 추출하는
 단계;
 복수의 계층의 신경망에 상기 샷 특징벡터를 입력한 후, 각 계층의
 출력벡터에 상기 질문 데이터의 특징벡터에 기초하여 연산된 주의
 가중치를 부여하여 합산함으로써 비디오 특징벡터를 산출하는 단계;
 그리고
 상기 비디오 특징벡터와 선택지 특징벡터 사이의 유사도에 기초하여
 상기 복수의 선택지 데이터 중 최종 답안을 선택하는 단계를 포함하는
 질의 응답 방법.
- [청구항 2] 제 1 항에 있어서,
 상기 질의 응답 방법은,
 복수의 질의응답 쌍에 대하여 상기 최종 답안이 각각의 질의응답 쌍에
 대한 상기 복수의 선택지 데이터 중 기설정된 정답이 되도록 상기 복수의
 계층의 신경망을 반복적으로 학습시키는 단계를 더 포함하는, 질의 응답
 방법.
- [청구항 3] 제 1 항에 있어서,
 상기 샷 특징벡터를 추출하는 단계는,
 오디오 특징벡터(a_i)를 추출하는 단계;
 자막 특징벡터(s_i)를 추출하는 단계; 그리고
 이미지 프레임 특징벡터(v_i)를 추출하는 단계를 포함하는, 질의 응답
 방법.
- [청구항 4] 제3항에 있어서,
 상기 샷 특징벡터를 추출하는 단계는,
 상기 오디오 특징벡터(a_i), 상기 자막 특징벡터(s_i), 그리고 상기 이미지
 프레임 특징벡터(v_i)를 연결(Concatenation)하여 상기 샷 특징벡터를
 산출하는 단계를 포함하는, 질의 응답 방법.
- [청구항 5] 제 4 항에 있어서,
 상기 이미지 프레임 특징벡터(v_i)를 추출하는 단계는,
 상기 오디오 특징벡터(a_i)와 상기 자막 특징벡터(s_i)를 멀티모드 조인트

임베딩(Multimodal Joint Embedding)하여, 상기 이미지 프레임 특징벡터(v_i)와 동일한 차원의 조인트 임베딩 벡터(e_i)를 구하고, 조인트 임베딩 벡터(e_i)에 포함된 주의 가중치를 이용하여 상기 이미지 프레임 특징벡터(v_i)를 연산하는 단계를 포함하는, 질의 응답 방법.

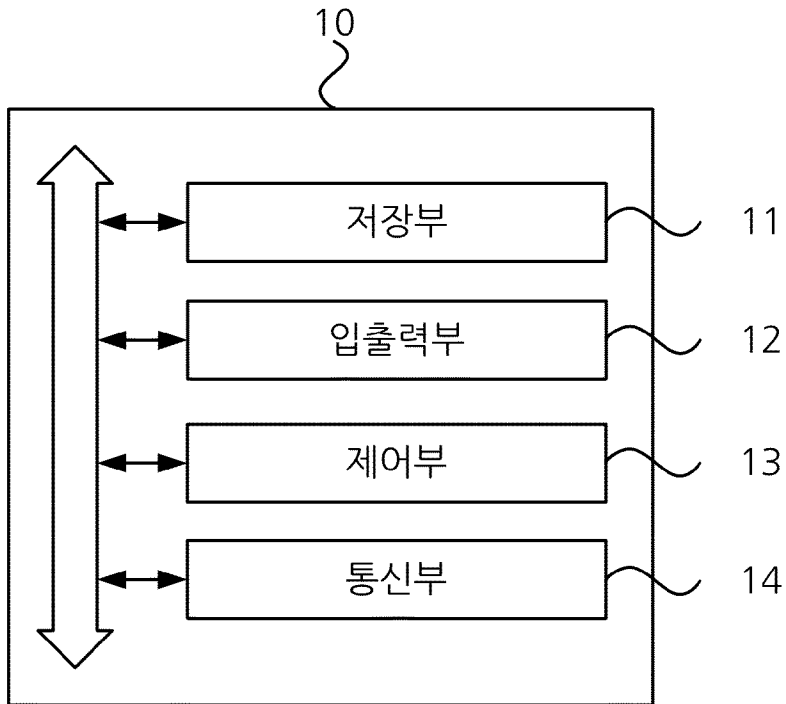
- [청구항 6] 제 1 항에 있어서,
상기 복수의 계층의 신경망은,
세 개의 계층의 1-D CNN으로 구성되는, 질의 응답 방법.
- [청구항 7] 질의 응답 장치에 있어서,
비디오 데이터와 질의응답 쌍을 포함하는 데이터 세트를 저장하는 저장부; 그리고
상기 비디오 데이터에 포함된 이미지 프레임, 오디오 데이터, 자막 데이터를, 동일한 하나의 자막을 기준으로 한 샷 단위로 구분한 후, 각각의 샷에 포함된 이미지 프레임, 오디오 데이터, 자막 데이터를 분석함으로써 샷 특징벡터를 추출하고, 복수의 계층의 신경망에 상기 샷 특징벡터를 입력하며, 상기 질의응답 쌍에서 질문 데이터와 상기 질문 데이터에 대응하는 복수의 선택지 데이터 각각의 특징벡터를 추출하여, 상기 복수의 계층의 신경망의 각 계층의 출력벡터를 이용하여 비디오 특징벡터를 산출하고, 상기 비디오 특징벡터와 선택지 특징벡터 사이의 유사도에 기초하여 상기 복수의 선택지 데이터 중 최종 답안을 선택하는 제어부를 포함하는 질의 응답 장치.
- [청구항 8] 제7항에 있어서,
상기 제어부는,
상기 각 계층의 출력벡터에 상기 질문 데이터의 특징벡터에 기초하여 연산되는 주의 가중치를 부여한 후 합산함으로써 상기 비디오 특징벡터를 산출하는, 질의 응답 장치.
- [청구항 9] 제 7 항에 있어서,
상기 제어부는,
복수의 질의응답 쌍에 대하여 상기 최종 답안이 각각의 질의응답 쌍에 대한 상기 복수의 선택지 데이터 중 기설정된 정답이 되도록 상기 복수의 계층의 신경망을 반복적으로 학습시키는, 질의 응답 장치.
- [청구항 10] 제 7 항에 있어서,
상기 제어부는,
오디오 특징벡터(a_i), 자막 특징벡터(s_i), 그리고 이미지 프레임 특징벡터(v_i)를 각각 추출한 후 상기 오디오 특징벡터(a_i), 자막 특징벡터(s_i), 그리고 이미지 프레임 특징벡터(v_i)를 연결(Concatenation)하여 상기 샷 특징벡터를 산출하는, 질의 응답 장치.
- [청구항 11] 제10항에 있어서,
상기 제어부는,

상기 오디오 특징벡터(a_i)와 상기 자막 특징벡터(s_i)를 멀티모드 조인트 임베딩(Multimodal Joint Embedding)하여, 상기 이미지 프레임 특징벡터(v_i)와 동일한 차원의 조인트 임베딩 벡터(e_i)를 구하고, 조인트 임베딩 벡터(e_i)에 포함된 주의 가중치를 이용하여 상기 이미지 프레임 특징벡터(v_i)를 연산하는, 질의 응답 장치.

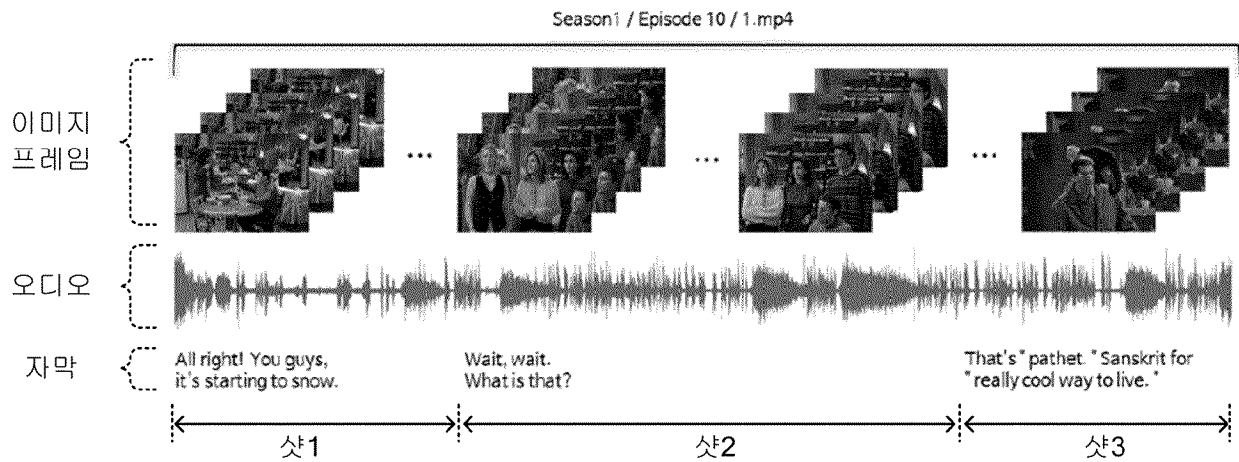
[청구항 12] 제 1 항에 기재된 방법을 수행하는 프로그램이 기록된 컴퓨터 판독 가능한 기록 매체.

[청구항 13] 질의 응답 장치에 의해 수행되며, 제 1 항에 기재된 방법을 수행하기 위해 컴퓨터 판독 가능 매체에 저장된 컴퓨터 프로그램.

[도1]



[도2]



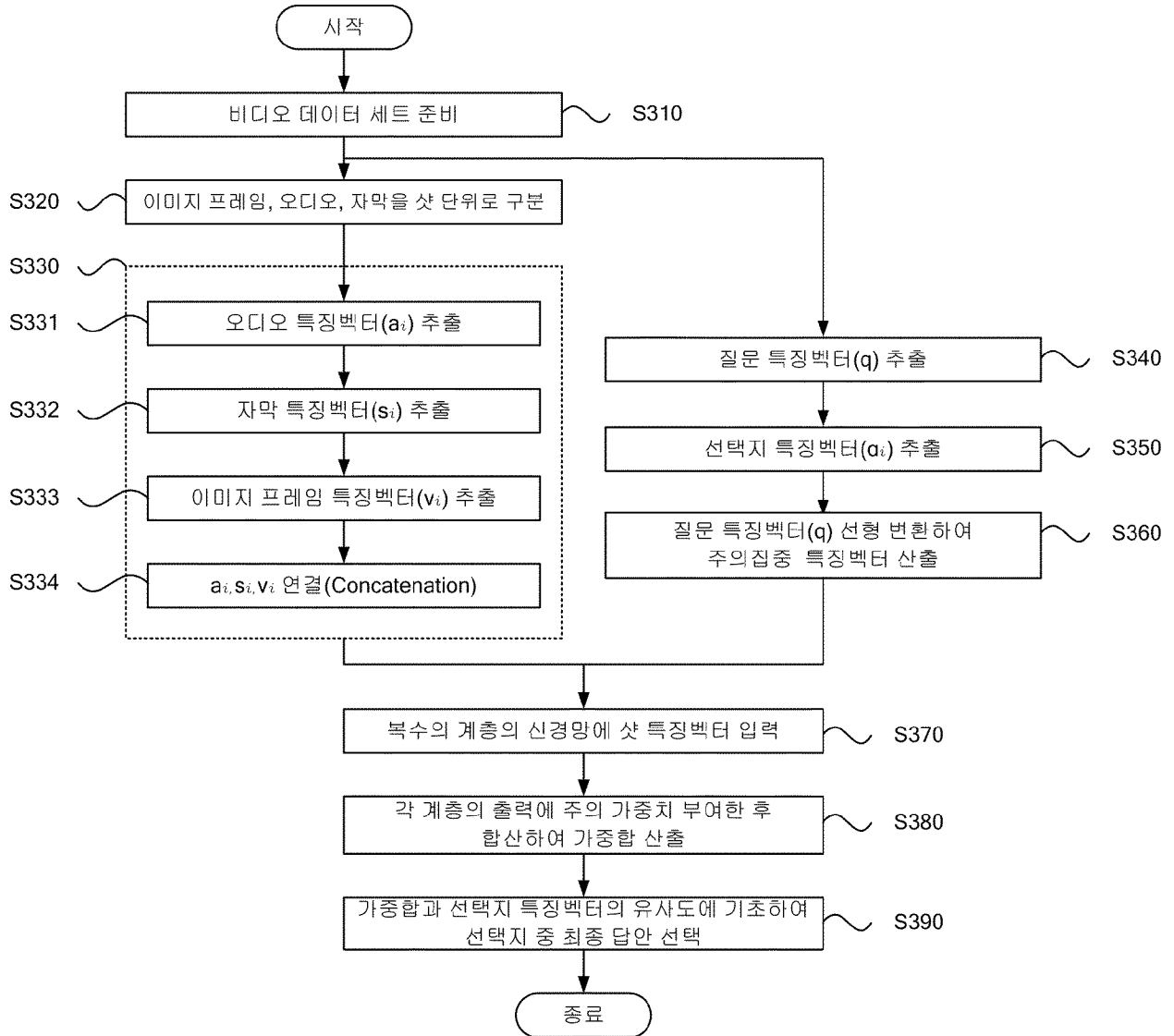
질문

Q: Who does Ross want Joey, Rachel, Monica, Chandler, and Phoebe to meet?

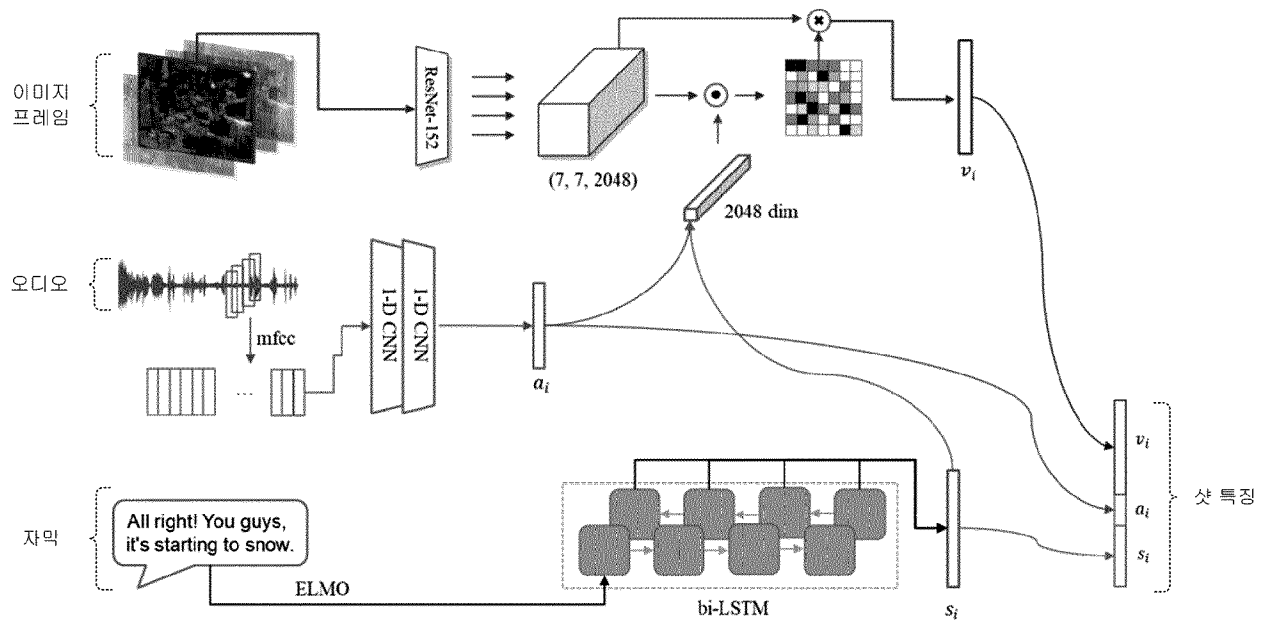
선택지

- A: He wants them to meet his girlfriend's parents.
- A: He wants them to meet Marcel.
- A: He wants them to meet his new girlfriend.
- A: He wants them to meet his son's friend.
- A: He wants them to meet his new boss.

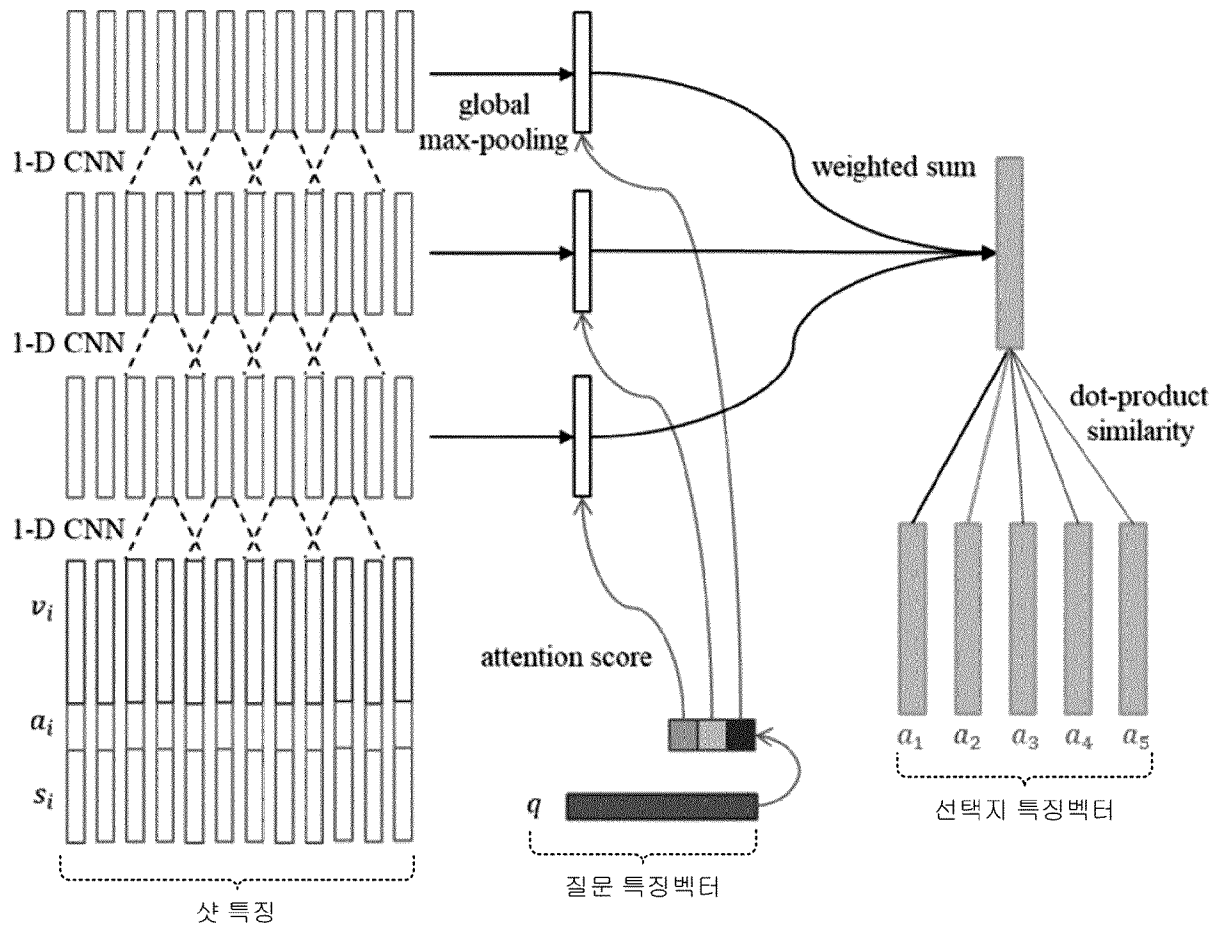
[도3]



[도4]



[도5]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/KR2019/017311

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/783(2019.01)i, G06F 16/432(2019.01)i, G06F 16/73(2019.01)i, G06N 3/08(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F 16/783; G06F 16/00; G06F 17/30; G06K 9/00; G06K 9/62; G06N 99/00; G06Q 10/10; H04N 5/92; G06F 16/432; G06F 16/73; G06N 3/08

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models: IPC as above

Japanese utility models and applications for utility models: IPC as above

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS (KIPO internal) & Keywords: VQA(visual question answering), image frame, audio, subtitle, neural network, multimodal joint embedding, similarity

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	KR 10-2018-0101948 A (NAVER CORPORATION) 14 September 2018 See paragraphs [0077]-[0094] and figure 3.	1-13
A	JP 5238339 B2 (NIPPON HOSO KYOKAI) 17 July 2013 See claims 1-3 and figure 1.	1-13
A	US 2017-0154314 A1 (FAMA TECHNOLOGIES, INC.) 01 June 2017 See paragraphs [0021]-[0024] and figure 1.	1-13
A	KR 10-2003-0065860 A (KT CORPORATION) 09 August 2003 See claims 1, 2 and figure 1.	1-13
A	KR 10-2017-0062290 A (INDUSTRY-ACADEMIC COOPERATION FOUNDATION, YONSEI UNIVERSITY) 07 June 2017 See paragraphs [0030]-[0040] and figure 1.	1-13



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

08 APRIL 2020 (08.04.2020)

Date of mailing of the international search report

09 APRIL 2020 (09.04.2020)

Name and mailing address of the ISA/KR

Korean Intellectual Property Office
Government Complex Daejeon Building 4, 189, Cheongsu-ro, Seo-gu,
Daejeon, 35208, Republic of Korea

Facsimile No. +82-42-481-8578

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/KR2019/017311

Patent document cited in search report	Publication date	Patent family member	Publication date
KR 10-2018-0101948 A	14/09/2018	KR 10-1970293 B1	18/04/2019
JP 5238339 B2	17/07/2013	JP 2009-265909 A	12/11/2009
US 2017-0154314 A1	01/06/2017	None	
KR 10-2003-0065860 A	09/08/2003	KR 10-0831550 B1	22/05/2008
KR 10-2017-0062290 A	07/06/2017	WO 2017-090830 A1	01/06/2017

A. 발명이 속하는 기술분류(국제특허분류(IPC))

G06F 16/783(2019.01)i, G06F 16/432(2019.01)i, G06F 16/73(2019.01)i, G06N 3/08(2006.01)i

B. 조사된 분야

조사된 최소문헌(국제특허분류를 기재)

G06F 16/783; G06F 16/00; G06F 17/30; G06K 9/00; G06K 9/62; G06N 99/00; G06Q 10/10; H04N 5/92; G06F 16/432; G06F 16/73; G06N 3/08

조사된 기술분야에 속하는 최소문헌 이외의 문헌

한국등록실용신안공보 및 한국공개실용신안공보: 조사된 최소문헌란에 기재된 IPC
일본등록실용신안공보 및 일본공개실용신안공보: 조사된 최소문헌란에 기재된 IPC

국제조사에 이용된 전산 데이터베이스(데이터베이스의 명칭 및 검색어(해당하는 경우))

eKOMPASS(특허청 내부 검색시스템) & 키워드: VQA(visual question answering), 이미지 프레임(image frame), 오디오(audio), 자막(subtitle), 신경망(neural network), 멀티모드 조인트 임베딩(multimodal joint embedding), 유사도(similarity)

C. 관련 문헌

카테고리*	인용문헌명 및 관련 구절(해당하는 경우)의 기재	관련 청구항
A	KR 10-2018-0101948 A (네이버 주식회사) 2018.09.14 단락 [0077]-[0094] 및 도면 3	1-13
A	JP 5238339 B2 (NIPPON HOSO KYOKAI) 2013.07.17 청구항 1-3 및 도면 1	1-13
A	US 2017-0154314 A1 (FAMA TECHNOLOGIES, INC.) 2017.06.01 단락 [0021]-[0024] 및 도면 1	1-13
A	KR 10-2003-0065860 A (주식회사 케이티) 2003.08.09 청구항 1, 2 및 도면 1	1-13
A	KR 10-2017-0062290 A (연세대학교 산학협력단) 2017.06.07 단락 [0030]-[0040] 및 도면 1	1-13

☐ 추가 문헌이 C(계속)에 기재되어 있습니다.

☒ 대응특허에 관한 별지를 참조하십시오.

* 인용된 문헌의 특별 카테고리:

- “A” 특별히 관련이 없는 것으로 보이는 일반적인 기술수준을 정의한 문헌 “T” 국제출원일 또는 우선일 후에 공개된 문헌으로, 출원과 상충하지 않으며 발명의 기초가 되는 원리나 이론을 이해하기 위해 인용된 문헌
- “D” 본 국제출원에서 출원인이 인용한 문헌
- “E” 국제출원일보다 빠른 출원일 또는 우선일을 가지나 국제출원일 이후 “X” 특별한 관련이 있는 문헌. 해당 문헌 하나만으로 청구된 발명의 신규성 또는 진보성이 없는 것으로 본다.
- “L” 우선권 주장에 의문을 제기하는 문헌 또는 다른 인용문헌의 공개일 또는 다른 특별한 이유(이유를 명시)를 밝히기 위하여 인용된 문헌 “Y” 특별한 관련이 있는 문헌. 해당 문헌이 하나 이상의 다른 문헌과 조합하는 경우로 그 조합이 당업자에게 자명한 경우 청구된 발명은 진보성이 없는 것으로 본다.
- “O” 구두 개시, 사용, 전시 또는 기타 수단을 언급하고 있는 문헌
- “P” 우선일 이후에 공개되었으나 국제출원일 이전에 공개된 문헌 “&” 동일한 대응특허문헌에 속하는 문헌

국제조사의 실제 완료일

2020년 04월 08일 (08.04.2020)

국제조사보고서 발송일

2020년 04월 09일 (09.04.2020)

ISA/KR의 명칭 및 우편주소



대한민국 특허청
(35208) 대전광역시 서구 청사로 189,
4동 (둔산동, 정부대전청사)

팩스 번호 +82-42-481-8578

심사관

변성철

전화번호 +82-42-481-8262



국제조사보고서에서 인용된 특허문헌	공개일	대응특허문헌	공개일
KR 10-2018-0101948 A	2018/09/14	KR 10-1970293 B1	2019/04/18
JP 5238339 B2	2013/07/17	JP 2009-265909 A	2009/11/12
US 2017-0154314 A1	2017/06/01	없음	
KR 10-2003-0065860 A	2003/08/09	KR 10-0831550 B1	2008/05/22
KR 10-2017-0062290 A	2017/06/07	WO 2017-090830 A1	2017/06/01