

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)(11) 공개번호 10-2025-0085281
(43) 공개일자 2025년06월12일

(51) 국제특허분류(Int. Cl.)

G06F 16/732 (2019.01) G06F 16/738 (2019.01)
G06F 16/783 (2019.01) G06F 40/284 (2020.01)
G06N 3/0455 (2023.01) G06N 3/0464 (2023.01)
G06N 3/096 (2023.01)

(52) CPC특허분류

G06F 16/732 (2019.01)
G06F 16/738 (2019.01)

(21) 출원번호 10-2023-0174402

(22) 출원일자 2023년12월05일

심사청구일자 2023년12월05일

(71) 출원인

경희대학교 산학협력단

경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)

(72) 발명자

박성배

서울특별시 송파구 오금로 396, 1402호 (가락동, 에스케이허브파크)

김봉민

경기도 용인시 기흥구 서그대로53번길 3-17, 304호 (서천동, 서천빌리지)

(뒷면에 계속)

(74) 대리인

유미특허법인

전체 청구항 수 : 총 14 항

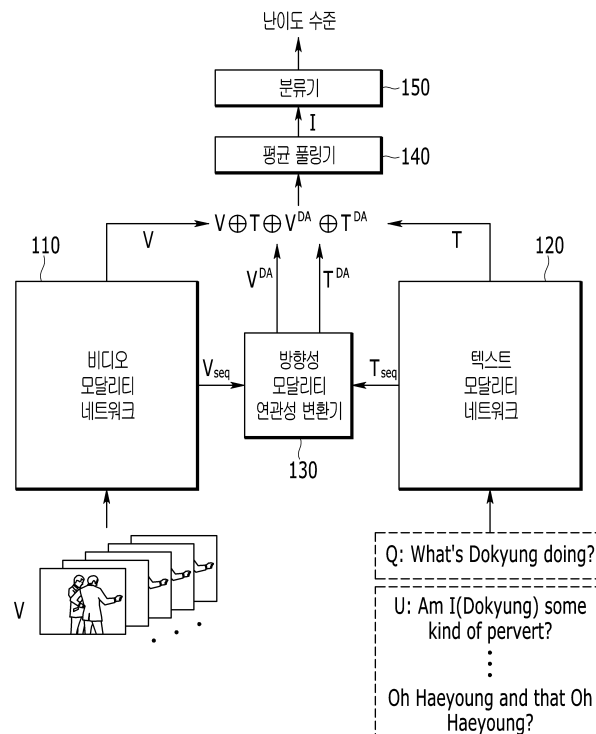
(54) 발명의 명칭 비디오 질의 응답을 위한 질의 난이도 추정 장치 및 그 방법

(57) 요약

비디오 질의 응답을 위한 질의 난이도 추정 장치 및 그 방법이 개시된다. 이 장치는 비디오 질의 응답 (Question-Answering, QA)을 위한 질의 난이도를 추정하는 장치로서, 입력되는 비디오를 대응하는 비디오 표현으로 인코딩하고, 인코딩 중에 비디오의 순차적 표현을 생성하는 비디오 모달리티 네트워크 및 비디오에 대응하는

(뒷면에 계속)

대표도 - 도1



질의와 자막에 대응하는 텍스트 표현을 생성하고, 텍스트 표현의 생성 중에 텍스트의 순차적 표현을 생성하는 텍스트 모달리티 네트워크를 포함한다. 이 장치는 비디오의 순차적 표현과 텍스트의 순차적 표현을 사용하여 텍스트 모달리티에 의해 수반된 비디오 표현과 비디오 모달리티에 의해 수반된 텍스트 표현을 생성하는 방향성 모달리티 연관성 변환기를 더 포함한다. 이 장치는 비디오 표현, 텍스트 표현, 텍스트 모달리티에 의해 수반된 비디오 표현 및 비디오 모달리티에 의해 수반된 텍스트 표현의 연결을 평균 풀링하여 최종 표현을 생성하는 평균 풀링기, 그리고 최종 표현에 기초하여 질의에 대한 난이도 수준을 결정하는 분류기를 더 포함한다.

(52) CPC특허분류

G06F 16/7844 (2019.01)
G06F 40/284 (2020.01)
G06N 3/0455 (2023.01)
G06N 3/0464 (2023.01)
G06N 3/096 (2023.01)

이준범

경기도 화성시 동탄대로24가길 30, 134동 303호 (영천동, 동탄 파크 한양수자인)

(72) 발명자

박규민

경기도 용인시 기흥구 서천동로43번길 11-11, 204호 (서천동)

김소연

경기도 화성시 동탄대로24가길 30, 134동 303호 (영천동, 동탄 파크 한양수자인)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193916
과제번호	2022-0-00951
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	사람중심인공지능핵심원천기술개발
연구과제명	스스로 불확실성을 자각하며 질문하면서 성장하는 에이전트 기술 개발
기 여 율	1/2
과제수행기관명	한국전자통신연구원
연구기간	2023.01.01 ~ 2023.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호	1711194179
과제번호	RS-2023-00207816
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	기초연구실 후속연구
연구과제명	Meta Federated Learning 기반 Satellite-Air-Ground 통합 네트워킹 시스템 핵심 구
조 개발	
기 여 율	1/2
과제수행기관명	경희대학교산학협력단
연구기간	2023.03.01 ~ 2024.02.29

명세서

청구범위

청구항 1

비디오 질의 응답(Question-Answering, QA)을 위한 질의 난이도를 추정하는 장치로서,
 입력되는 비디오를 대응하는 비디오 표현으로 인코딩하고, 상기 인코딩 중에 상기 비디오의 순차적 표현을 생성하는 비디오 모달리티 네트워크,
 상기 비디오에 대응하는 질의와 자막에 대응하는 텍스트 표현을 생성하고, 상기 텍스트 표현의 생성 중에 텍스트의 순차적 표현을 생성하는 텍스트 모달리티 네트워크,
 상기 비디오의 순차적 표현과 상기 텍스트의 순차적 표현을 사용하여 텍스트 모달리티(modality)에 의해 수반된 비디오 표현과 비디오 모달리티에 의해 수반된 텍스트 표현을 생성하는 방향성 모달리티 연관성 변환기,
 상기 비디오 표현, 상기 텍스트 표현, 상기 텍스트 모달리티에 의해 수반된 비디오 표현 및 상기 비디오 모달리티에 의해 수반된 텍스트 표현의 연결을 평균 풀링(average pooling)하여 최종 표현을 생성하는 평균 풀링(average pooling)기, 그리고
 상기 최종 표현에 기초하여 질의에 대한 난이도 수준을 결정하는 분류기를 포함하는 질의 난이도 추정 장치.

청구항 2

제1항에 있어서,
 상기 비디오 모달리티 네트워크는,
 전이 학습의 딥러닝 모델을 사용하여 상기 비디오로부터 공간적 특징을 추출하는 CNN(Convolutional Neural Network) 백본,
 상기 공간적 특징에 시간적 정보를 할당하여 상기 비디오의 모든 순차적 정보를 결합하는 비디오 벡터와 상기 비디오의 순차적 표현을 생성하는 변환기, 그리고
 상기 비디오 벡터에 대해 풀링을 수행하여 상기 비디오 표현을 생성하는 제1 풀링기를 포함하는, 질의 난이도 추정 장치.

청구항 3

제2항에 있어서,
 상기 텍스트 모달리티 네트워크는,
 상기 질의와 자막의 연결을 사용하여 상기 텍스트의 순차적 표현을 생성하고, 상기 텍스트의 순차적 표현을 토큰의 벡터로 요약하는 언어 표현 모델, 그리고
 상기 토큰의 벡터에 대해 풀링을 수행하여 상기 텍스트 표현을 생성하는 제2 풀링기를 포함하는, 질의 난이도 추정 장치.

청구항 4

제3항에 있어서,
 상기 방향성 모달리티 연관성 변환기는,
 상기 비디오의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관을 생성하는 방향성 어텐션 계층,

상기 비디오의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관에 대해 정규화 처리를 수행하는 제1 Add & Norm,

정규화 처리된 상기 비디오의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관에 대해 피드 포워드 처리를 수행하는 피드 포워드 네트워크 계층, 그리고

상기 피드 포워드 처리가 수행된 상기 비디오의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관에 대해 정규화 처리를 수행하여 상기 텍스트 모달리티에 의해 수반된 비디오 표현과 상기 비디오 모달리티에 의해 수반된 텍스트 표현을 생성하는 제2 Add & Norm

을 포함하는, 질의 난이도 추정 장치.

청구항 5

제3항에 있어서,

상기 언어 표현 모델은 RoBERTa(Robustly Optimized BERT Pretraining Approach) 모델인,

질의 난이도 추정 장치.

청구항 6

제4항에 있어서,

상기 난이도 수준은 상기 분류기의 훈련 가능한 파라미터, 상기 난이도 수준의 개수 및 상기 평균 풀링기에 의해 생성된 최종 표현에 기초하여 산출되는,

질의 난이도 추정 장치.

청구항 7

질의 난이도 추정 장치가 비디오 질의 응답을 위한 질의 난이도를 추정하는 방법으로서,

입력되는 비디오를 대응하는 비디오 표현으로 인코딩하고, 상기 인코딩 중에 상기 비디오의 순차적 표현을 생성하며, 이와 동시에, 상기 비디오에 대응하는 질의와 자막에 대응하는 텍스트 표현을 생성하고, 상기 텍스트 표현의 생성 중에 텍스트의 순차적 표현을 생성하는 단계,

상기 비디오의 순차적 표현과 상기 텍스트의 순차적 표현을 사용하여 텍스트 모달리티에 의해 수반된 비디오 표현과 비디오 모달리티에 의해 수반된 텍스트 표현을 생성하는 단계,

상기 비디오 표현, 상기 텍스트 표현, 상기 텍스트 모달리티에 의해 수반된 비디오 표현 및 상기 비디오 모달리티에 의해 수반된 텍스트 표현의 연결을 평균 풀링하여 최종 표현을 생성하는 단계, 그리고

상기 최종 표현에 기초하여 질의에 대한 난이도 수준을 결정하는 단계

를 포함하는 질의 난이도 추정 방법.

청구항 8

제7항에 있어서,

상기 텍스트의 순차적 표현을 생성하는 단계는,

전이 학습의 딥러닝 모델을 사용하여 상기 비디오로부터 공간적 특징을 추출하는 단계,

상기 공간적 특징에 시간적 정보를 할당하여 상기 비디오의 모든 순차적 정보를 결합하는 비디오 벡터와 상기 비디오의 순차적 표현을 생성하는 단계, 그리고

상기 비디오 벡터에 대해 풀링을 수행하여 상기 비디오 표현을 생성하는 단계

를 포함하는, 질의 난이도 추정 방법.

청구항 9

제8항에 있어서,

상기 텍스트의 순차적 표현을 생성하는 단계는,

상기 질의와 자막의 연결을 사용하여 상기 텍스트의 순차적 표현을 생성하는 단계,

언더 표현 모델을 사용하여 상기 텍스트의 순차적 표현을 토큰의 벡터로 요약하는 단계, 그리고

상기 토큰의 벡터에 대해 풀링을 수행하여 상기 텍스트 표현을 생성하는 단계

를 더 포함하는, 질의 난이도 추정 방법.

청구항 10

제9항에 있어서,

상기 비디오의 순차적 표현과 상기 텍스트의 순차적 표현을 사용하여 텍스트 모달리티에 의해 수반된 비디오 표현과 비디오 모달리티에 의해 수반된 텍스트 표현을 생성하는 단계는,

상기 비디오의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관을 생성하는 단계,

상기 비디오의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관에 대해 정규화 처리를 수행하는 단계,

정규화 처리된 상기 비디오의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관에 대해 피드 포워드 처리를 수행하는 단계, 그리고

상기 피드 포워드 처리가 수행된 상기 비디오의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관에 대해 정규화 처리를 수행하여 상기 텍스트 모달리티에 의해 수반된 비디오 표현과 상기 비디오 모달리티에 의해 수반된 텍스트 표현을 생성하는 단계

를 포함하는, 질의 난이도 추정 방법.

청구항 11

비디오 질의 응답을 위한 질의 난이도를 추정하는 장치로서,

프로세서 및 메모리를 포함하며,

상기 메모리는 코드의 집합을 저장하도록 구성되고,

상기 코드는,

입력되는 비디오를 대응하는 비디오 표현으로 인코딩하고, 상기 인코딩 중에 상기 비디오의 순차적 표현을 생성하며, 이와 동시에, 상기 비디오에 대응하는 질의와 자막에 대응하는 텍스트 표현을 생성하고, 상기 텍스트 표현의 생성 중에 텍스트의 순차적 표현을 생성하는 프로세스,

상기 비디오의 순차적 표현과 상기 텍스트의 순차적 표현을 사용하여 텍스트 모달리티에 의해 수반된 비디오 표현과 비디오 모달리티에 의해 수반된 텍스트 표현을 생성하는 프로세스,

상기 비디오 표현, 상기 텍스트 표현, 상기 텍스트 모달리티에 의해 수반된 비디오 표현 및 상기 비디오 모달리티에 의해 수반된 텍스트 표현의 연결을 평균 풀링하여 최종 표현을 생성하는 프로세스, 그리고

상기 최종 표현에 기초하여 질의에 대한 난이도 수준을 결정하는 프로세스

를 실행하도록 상기 프로세서를 제어하는 데 사용되는,

질의 난이도 추정 장치.

청구항 12

제11항에 있어서,

상기 텍스트의 순차적 표현을 생성하는 프로세스는,
 전이 학습의 딥러닝 모델을 사용하여 상기 비디오로부터 공간적 특징을 추출하는 프로세스,
 상기 공간적 특징에 시간적 정보를 할당하여 상기 비디오의 모든 순차적 정보를 결합하는 비디오 벡터와 상기 비디오의 순차적 표현을 생성하는 프로세스, 그리고
 상기 비디오 벡터에 대해 풀링을 수행하여 상기 비디오 표현을 생성하는 프로세스
 를 포함하는, 질의 난이도 추정 장치.

청구항 13

제12항에 있어서,
 상기 텍스트의 순차적 표현을 생성하는 프로세스는,
 상기 질의와 자막의 연결을 사용하여 상기 텍스트의 순차적 표현을 생성하는 프로세스,
 언더 표현 모델을 사용하여 상기 텍스트의 순차적 표현을 토큰의 벡터로 요약하는 프로세스, 그리고
 상기 토큰의 벡터에 대해 풀링을 수행하여 상기 텍스트 표현을 생성하는 프로세스
 를 더 포함하는, 질의 난이도 추정 장치.

청구항 14

제13항에 있어서,
 상기 비디오의 순차적 표현과 상기 텍스트의 순차적 표현을 사용하여 텍스트 모달리티에 의해 수반된 비디오 표현과 비디오 모달리티에 의해 수반된 텍스트 표현을 생성하는 프로세스는,
 상기 비디오의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관을 생성하는 프로세스,
 상기 비디오의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관에 대해 정규화 처리를 수행하는 프로세스,
 정규화 처리된 상기 비디오의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관에 대해 피드 포워드 처리를 수행하는 프로세스, 그리고
 상기 피드 포워드 처리가 수행된 상기 비디오의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관에 대해 정규화 처리를 수행하여 상기 텍스트 모달리티에 의해 수반된 비디오 표현과 상기 비디오 모달리티에 의해 수반된 텍스트 표현을 생성하는 프로세스
 를 포함하는, 질의 난이도 추정 장치.

발명의 설명

기술 분야

[0001] 본 발명은 비디오 질의 응답을 위한 질의 난이도 추정 장치 및 그 방법에 관한 것이다.

배경 기술

[0002] 질의 응답(Question-Answering, QA)은 질의에 대한 참고 정보를 이용하여 주어진 질문에 대한 적절한 답변을 제공하는 작업이다. 답변을 추론하는 데 있어 필수적인 단계 중 하나는 질의의 난이도 수준을 추정하는 것이며 이를 질의 난이도 추정(question difficulty estimation, QDE)이라고 한다. 추정된 질의 난이도는 QA 시스템의 성능을 향상시키는 데 사용될 수 있다.

[0003] 최근, 비디오 기반의 QA가 개발되고 있다. 그러나, 비디오 QA에 대한 집중적인 연구에도 불구하고, 비디오 QA에 있어서 QDE에 대한 연구는 부족한 편이다.

- [0004] 비디오 QA에서의 질의는 일반적으로 텍스트 쿼리와 참조 비디오로 구성된다. 따라서, 비디오 QA에서 QDE의 두 가지 중요한 인자는 (i) 비디오 표현과 (ii) 텍스트와 비디오 정보의 연관 표현이다.
- [0005] 특히, 비디오 표현은 이미지의 시퀀스로 보여질 수 있기 때문에 비디오의 시간적 특성을 보존하는 것이 비디오 표현에 있어서 중요한 의미를 갖는다.
- [0006] 비디오 QA에 대한 초기 연구에서는 LSTM(Long Short-Term Memory)을 사용하여 비디오 표현에서 시간 정보를 반영하였다. 그러나, 최근에는 LSTM 대신에 변환기를 채택하고 있다. 왜냐하면, 변환기는 긴 비디오의 롱텀(long-term) 종속성을 처리하는 데 LSTM보다 훨씬 효율적이기 때문이다.
- [0007] 또한, 쿼리는 텍스트이고, 레퍼런스는 텍스트와 다른 모달리티의 비디오라는 사실에 기인하여, 쿼리와 참고 비디오 사이의 의미 있는 연관성을 포착하기 어렵다. 다중 모달리티에 대한 일반적인 접근 방식은 각 모달리티의 특징을 독립적으로 추출한 다음 이를 단일 표현으로 통합하는 것이다. 그러나, 이러한 모든 기술은 비디오 QA에서 질의에 대한 답변에만 중점을 두었고 비디오 QA에서의 QDE와는 아무런 관련이 없다.
- [0008] 또한, 비디오 QA는 본질적으로 텍스트와 비디오 모두를 포함하는 다중 모달리티 작업이지만 이전의 질의 난이도 예측 모델은 여러 가지 모달리티 간의 연관성을 고려하지 않고 있다.
- [0009] 따라서, 비디오 QA에서의 질의 난이도 예측시 여러 가지 모달리티 간의 연관성을 고려한 모델이 요구된다.

발명의 내용

해결하려는 과제

- [0010] 본 발명이 해결하고자 하는 과제는 비디오 모달리티와 텍스트 모달리티 간의 방향 연관성 모델링시 텍스트와 비디오의 순차적 정보를 사용하여 방향 연관성을 산출함으로써 보다 향상된 질의 난이도 추정이 가능한 질의 난이도 추정 장치 및 그 방법을 제공하는 것이다.

과제의 해결 수단

- [0011] 상기한 바와 같은 본 발명의 과제를 달성하고, 후술하는 본 발명의 특징적인 효과를 실현하기 위한, 본 발명의 특징적인 구성은 하기와 같다.
- [0012] 본 발명의 일 측면에 따르면, 질의 난이도 추정 장치가 제공되며, 이 장치는,
- [0013] 비디오 질의 응답(Question-Answering, QA)을 위한 질의 난이도를 추정하는 장치로서, 입력되는 비디오를 대응하는 비디오 표현으로 인코딩하고, 상기 인코딩 중에 상기 비디오의 순차적 표현을 생성하는 비디오 모달리티 네트워크, 상기 비디오에 대응하는 질의와 자막에 대응하는 텍스트 표현을 생성하고, 상기 텍스트 표현의 생성 중에 텍스트의 순차적 표현을 생성하는 텍스트 모달리티 네트워크, 상기 비디오의 순차적 표현과 상기 텍스트의 순차적 표현을 사용하여 텍스트 모달리티(modality)에 의해 수반된 비디오 표현과 비디오 모달리티에 의해 수반된 텍스트 표현을 생성하는 방향성 모달리티 연관성 변환기, 상기 비디오 표현, 상기 텍스트 표현, 상기 텍스트 모달리티에 의해 수반된 비디오 표현 및 상기 비디오 모달리티에 의해 수반된 텍스트 표현의 연결을 평균 풀링(average pooling)하여 최종 표현을 생성하는 평균 풀링(average pooling)기, 그리고 상기 최종 표현에 기초하여 질의에 대한 난이도 수준을 결정하는 분류기를 포함한다.
- [0014] 여기서, 상기 비디오 모달리티 네트워크는, 전이 학습의 딥러닝 모델을 사용하여 상기 비디오로부터 공간적 특징을 추출하는 CNN(Convolutional Neural Network) 백본, 상기 공간적 특징에 시간적 정보를 할당하여 상기 비디오의 모든 순차적 정보를 결합하는 비디오 벡터와 상기 비디오의 순차적 표현을 생성하는 변환기, 그리고 상기 비디오 벡터에 대해 풀링을 수행하여 상기 비디오 표현을 생성하는 제1 풀링기를 포함한다.
- [0015] 또한, 상기 텍스트 모달리티 네트워크는, 상기 질의와 자막의 연결을 사용하여 상기 텍스트의 순차적 표현을 생성하고, 상기 텍스트의 순차적 표현을 토큰의 벡터로 요약하는 언어 표현 모델, 그리고 상기 토큰의 벡터에 대해 풀링을 수행하여 상기 텍스트 표현을 생성하는 제2 풀링기를 포함한다.
- [0016] 또한, 상기 방향성 모달리티 연관성 변환기는, 상기 비디오의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관을 생성하는 방향성 어텐션 계층, 상기 비디오의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관에 대해 정규화 처리를 수행하는 제1 Add & Norm, 정규화 처리된 상기 비디오의

순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관에 대해 피드 포워드 처리를 수행하는 피드 포워드 네트워크 계층, 그리고 상기 피드 포워드 처리가 수행된 상기 비디오의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관에 대해 정규화 처리를 수행하여 상기 텍스트 모듈리티에 의해 수반된 비디오 표현과 상기 비디오 모듈리티에 의해 수반된 텍스트 표현을 생성하는 제2 Add & Norm을 포함한다.

[0017] 또한, 상기 언어 표현 모델은 RoBERTa(Robustly Optimized BERT Pretraining Approach) 모델이다.

[0018] 또한, 상기 난이도 수준은 상기 분류기의 훈련 가능한 파라미터, 상기 난이도 수준의 개수 및 상기 평균 풀링기에 의해 생성된 최종 표현에 기초하여 산출된다.

[0019] 본 발명의 다른 측면에 따르면, 질의 난이도 추정 방법이 제공되며, 이 방법은,

[0020] 질의 난이도 추정 장치가 비디오 질의 응답을 위한 질의 난이도를 추정하는 방법으로서, 입력되는 비디오를 대응하는 비디오 표현으로 인코딩하고, 상기 인코딩 중에 상기 비디오의 순차적 표현을 생성하며, 이와 동시에, 상기 비디오에 대응하는 질의와 자막에 대응하는 텍스트 표현을 생성하고, 상기 텍스트 표현의 생성 중에 텍스트의 순차적 표현을 생성하는 단계, 상기 비디오의 순차적 표현과 상기 텍스트의 순차적 표현을 사용하여 텍스트 모듈리티에 의해 수반된 비디오 표현과 비디오 모듈리티에 의해 수반된 텍스트 표현을 생성하는 단계, 상기 비디오 표현, 상기 텍스트 표현, 상기 텍스트 모듈리티에 의해 수반된 비디오 표현 및 상기 비디오 모듈리티에 의해 수반된 텍스트 표현의 연결을 평균 풀링하여 최종 표현을 생성하는 단계, 그리고 상기 최종 표현에 기초하여 질의에 대한 난이도 수준을 결정하는 단계를 포함한다.

[0021] 여기서, 상기 텍스트의 순차적 표현을 생성하는 단계는, 전이 학습의 딥러닝 모델을 사용하여 상기 비디오로부터 공간적 특징을 추출하는 단계, 상기 공간적 특징에 시간적 정보를 할당하여 상기 비디오의 모든 순차적 정보를 결합하는 비디오 벡터와 상기 비디오의 순차적 표현을 생성하는 단계, 그리고 상기 비디오 벡터에 대해 풀링을 수행하여 상기 비디오 표현을 생성하는 단계를 포함한다.

[0022] 또한, 상기 텍스트의 순차적 표현을 생성하는 단계는, 상기 질의와 자막의 연결을 사용하여 상기 텍스트의 순차적 표현을 생성하는 단계, 언더 표현 모델을 사용하여 상기 텍스트의 순차적 표현을 토큰의 벡터로 요약하는 단계, 그리고 상기 토큰의 벡터에 대해 풀링을 수행하여 상기 텍스트 표현을 생성하는 단계를 더 포함한다.

[0023] 또한, 상기 비디오의 순차적 표현과 상기 텍스트의 순차적 표현을 사용하여 텍스트 모듈리티에 의해 수반된 비디오 표현과 비디오 모듈리티에 의해 수반된 텍스트 표현을 생성하는 단계는, 상기 비디오의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관을 생성하는 단계, 상기 비디오의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관에 대해 정규화 처리를 수행하는 단계, 정규화 처리된 상기 비디오의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관에 대해 피드 포워드 처리를 수행하는 단계, 그리고 상기 피드 포워드 처리가 수행된 상기 비디오의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관에 대해 정규화 처리를 수행하여 상기 텍스트 모듈리티에 의해 수반된 비디오 표현과 상기 비디오 모듈리티에 의해 수반된 텍스트 표현을 생성하는 단계를 포함한다.

[0024] 본 발명의 또 다른 측면에 따르면, 질의 난이도 추정 장치가 제공되며, 이 장치는,

[0025] 비디오 질의 응답을 위한 질의 난이도를 추정하는 장치로서, 프로세서 및 메모리를 포함하며, 상기 메모리는 코드의 집합을 저장하도록 구성되고, 상기 코드는, 입력되는 비디오를 대응하는 비디오 표현으로 인코딩하고, 상기 인코딩 중에 상기 비디오의 순차적 표현을 생성하며, 이와 동시에, 상기 비디오에 대응하는 질의와 자막에 대응하는 텍스트 표현을 생성하고, 상기 텍스트 표현의 생성 중에 텍스트의 순차적 표현을 생성하는 프로세스, 상기 비디오의 순차적 표현과 상기 텍스트의 순차적 표현을 사용하여 텍스트 모듈리티에 의해 수반된 비디오 표현과 비디오 모듈리티에 의해 수반된 텍스트 표현을 생성하는 프로세스, 상기 비디오 표현, 상기 텍스트 표현, 상기 텍스트 모듈리티에 의해 수반된 비디오 표현 및 상기 비디오 모듈리티에 의해 수반된 텍스트 표현의 연결을 평균 풀링하여 최종 표현을 생성하는 프로세스, 그리고 상기 최종 표현에 기초하여 질의에 대한 난이도 수준을 결정하는 프로세스를 실행하도록 상기 프로세서를 제어하는 데 사용된다.

[0026] 여기서, 상기 텍스트의 순차적 표현을 생성하는 프로세스는, 전이 학습의 딥러닝 모델을 사용하여 상기 비디오로부터 공간적 특징을 추출하는 프로세스, 상기 공간적 특징에 시간적 정보를 할당하여 상기 비디오의 모든 순

차적 정보를 결합하는 비디오 벡터와 상기 비디오의 순차적 표현을 생성하는 프로세스, 그리고 상기 비디오 벡터에 대해 풀링을 수행하여 상기 비디오 표현을 생성하는 프로세스를 포함한다.

[0027] 또한, 상기 텍스트의 순차적 표현을 생성하는 프로세스는, 상기 질의와 자막의 연결을 사용하여 상기 텍스트의 순차적 표현을 생성하는 프로세스, 언더 표현 모델을 사용하여 상기 텍스트의 순차적 표현을 토큰의 벡터로 요약하는 프로세스, 그리고 상기 토큰의 벡터에 대해 풀링을 수행하여 상기 텍스트 표현을 생성하는 프로세스를 더 포함한다.

[0028] 또한, 상기 비디오의 순차적 표현과 상기 텍스트의 순차적 표현을 사용하여 텍스트 모달리티에 의해 수반된 비디오 표현과 비디오 모달리티에 의해 수반된 텍스트 표현을 생성하는 프로세스는, 상기 비디오의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관을 생성하는 프로세스, 상기 비디오의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관에 대해 정규화 처리를 수행하는 프로세스, 정규화 처리된 상기 비디오의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관에 대해 피드 포워드 처리를 수행하는 프로세스, 그리고 상기 피드 포워드 처리가 수행된 상기 비디오의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관과 상기 텍스트의 순차적 표현에 대한 상기 텍스트의 순차적 표현의 연관에 대해 정규화 처리를 수행하여 상기 텍스트 모달리티에 의해 수반된 비디오 표현과 상기 비디오 모달리티에 의해 수반된 텍스트 표현을 생성하는 프로세스를 포함한다.

발명의 효과

[0029] 본 발명에 따르면, 비디오 모달리티와 텍스트 모달리티 간의 방향 연관성을 모델링할 뿐만 아니라 서로 간의 방향 연관성 모델링시 텍스트와 비디오의 순차적 정보를 사용하여 방향 연관성을 산출함으로써 보다 향상된 질의 난이도 추정이 가능하다.

[0030] 이로 인해, 멀티 모달 환경에서 보다 정확하고 향상된 질의 난이도를 분석하여 이에 따른 응답을 제공할 수 있다.

도면의 간단한 설명

[0031] 도 1은 본 발명의 실시예에 따른 질의 난이도 추정 장치의 개략적인 구성 블록도이다.

도 2는 도 1에 도시된 비디오 모달리티 네트워크의 구체적인 구성 블록도이다.

도 3은 도 1에 도시된 텍스트 모달리티 네트워크의 구체적인 구성 블록도이다.

도 4는 도 1에 도시된 방향성 모달리티 연관성 변환기의 구체적인 구성 블록도이다.

도 5는 본 발명의 실시예에 따른 질의 난이도 추정 방법의 개략적인 흐름도이다.

도 6은 본 발명의 다른 실시예에 따른 질의 난이도 추정 장치의 개략적인 구성을 도시한 도면이다.

발명을 실시하기 위한 구체적인 내용

[0032] 아래에서는 첨부한 도면을 참고로 하여 본 발명의 실시예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그러나 본 발명은 여러 가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0033] 명세서 전체에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다. 또한, 명세서에 기재된 "...부", "...기", "모듈" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어 또는 하드웨어 및 소프트웨어의 연관성으로 구현될 수 있다.

[0034] 본 발명에서 설명하는 장치들은 적어도 하나의 프로세서, 메모리 장치, 통신 장치 등을 포함하는 하드웨어로 구성되고, 지정된 장소에 하드웨어와 연관성되어 실행되는 프로그램이 저장된다. 하드웨어는 본 발명의 방법을 실행할 수 있는 구성과 성능을 가진다. 프로그램은 도면들을 참고로 설명한 본 발명의 동작 방법을 구현한 명령어(instructions)를 포함하고, 프로세서와 메모리 장치 등의 하드웨어와 결합하여 본 발명을 실행한다.

- [0035] 이하, 본 발명의 실시예에 따른 질의 난이도 추정 장치 및 그 방법에 대해 설명한다.
- [0036] 본 발명의 실시예에서 제공되는 질의 난이도 추정기는 텍스트와 비디오의 두 가지 상이한 모달리티(modality)를 관리하므로, 두 가지 모달리티의 다중 모드 융합 표현을 통해 비디오 QA의 모든 질의의 난이도 수준을 추정하는 것을 목표로 한다.
- [0037] 이러한 목표를 달성하기 위해, 본 발명의 실시예에서 제공되는 질의 난이도 추정기는 개략적으로 두 단계로 구성된다.
- [0038] 첫 번째 단계에서, 사전 훈련된 CNN(Convolutional Neural Network)을 사용하여 비디오의 특징이 추출된 다음 비디오 특징의 시간 정보가 변환기 인코더에 의해 추가로 인코딩된다. 동시에, 쿼리와 자막같은 모든 텍스트 정보는 사전 훈련된 언어 모델에 의해 인코딩된다.
- [0039] 그 후, 두 번째 단계에서, 향상된 방향성 모달리티 연관성 변환기(Directional Modality Association Transformer ++, DiMAT++)를 사용하여 두 모달리티의 통합된 표현이 생성된다.
- [0040] 쿼리와 그의 참조 비디오는 모두 서로 관련되어 있으므로, 두 모달리티 간의 연관성을 포착하는 것이 중요하다. 따라서, DiMAT++는 실제로 각 모달리티의 방향성 연관성을 생성하여 비디오에서 텍스트로 또는 그 반대로 방향성 영향을 전달한다. 그런 다음, 질의의 난이도는 모달리티와 방향성 연관성의 연결로부터 결정된다.
- [0041] 이하, 본 발명의 실시예에 따른 질의 난이도 추정 장치에 대해 구체적으로 설명한다.
- [0042] 도 1은 본 발명의 실시예에 따른 질의 난이도 추정 장치(100)의 개략적인 구성 블록도이다.
- [0043] 비디오 QA의 QDE는 질의의 난이도 수준이 참조 비디오 $V = \langle v_1, \dots, v_n \rangle$, 그의 자막 $U = \{u_1, \dots, u_m\}$ 및 쿼리 $Q = \{q_1, \dots, q_l\}$ 로부터 결정되는 분류 작업으로 간주된다. 즉, 쿼리(Q)는 참조 비디오와 자막으로 보완된다.
- [0044] V , U 및 Q 는 모두 순차적이다. V 는 n 개의 샘플링된 이미지로 구성되는 반면, U 와 Q 는 각각 m 개 및 l 개의 단어를 갖는다. 대부분의 비디오 QA에서, V 에 나타나는 문자나 객체에 대한 일부 정보를 전달하기 때문에 U 는 V 와 함께 제공된다.
- [0045] 그런 다음, QDE는 모든 Q , U , V 를 포함하는 Q 의 최적 난이도 수준 $d^* \in D$ 를 결정하며, 여기서 $D = \{d_1, \dots, d_k\}$ 는 가능한 난이도 수준의 세트이다.
- [0046] 대부분의 QA에서와 마찬가지로, 비디오 QA도 모든 질의에 대한 답변을 제공한다. 그러나, 본 발명의 실시예에 따른 질의 난이도 추정 장치(100)에서는 질의에 답하기 전에 난이도가 예측되어야 하므로 답변을 정보로서 간주하지 않는다.
- [0047] 도 1에 도시된 바와 같이, 본 발명의 실시예에 따른 질의 난이도 추정 장치(100)는 질의의 난이도 수준을 결정하기 위해 비디오 모달리티 네트워크(110), 텍스트 모달리티 네트워크(120), 방향성 모달리티 연관성 변환기(130), 평균 풀링(average pooling)기(140) 및 분류기(150)를 포함한다.
- [0048] 비디오 모달리티 네트워크(110)는 V 를 비디오 표현 $\mathbf{V} \in \mathbb{R}^h$ 로 인코딩하며, 여기서 h 는 모달리티 벡터의 차원을 나타낸다.
- [0049] 마찬가지로, 텍스트 모달리티 네트워크(120)는 Q 와 U 를 $\mathbf{T} \in \mathbb{R}^h$ 로 나타낸다.
- [0050] 방향성 모달리티 연관성 변환기(130)는 모달리티 간의 방향성 연관성을 모델링한다. 방향성 모달리티 연관성 변환기(130)의 출력은 $\mathbf{V}^{DA}$ 및 $\mathbf{T}^{DA}$ 이며, 여기서 $\mathbf{V}^{DA} \in \mathbb{R}^h$ 는 텍스트 모달리티에 의해 수반된 비디오 표현이고, $\mathbf{T}^{DA} \in \mathbb{R}^h$ 는 비디오 모달리티에 의해 수반된 텍스트 표현이다. $\mathbf{V}^{DA}$ 와 $\mathbf{T}^{DA}$ 는 각각 비디오 모달리티 네트워크(110)의 중간 출력인 $\mathbf{V}_{seq}$ 와 텍스트 모달리티 네트워크(120)의 중간 출력인 $\mathbf{T}_{seq}$ 로부터 생성된다.
- [0051] 평균 풀링기(140)는 비디오 모달리티 네트워크(110)로부터의 출력인 비디오 표현 $\mathbf{V}$, 텍스트 모달리티 네트워크(120)로부터의 출력 텍스트 표현 $\mathbf{T}$, 그리고 방향성 모달리티 연관성 변환기(130)로부터의 출력 $\mathbf{V}^{DA}$ 및 $\mathbf{T}^{DA}$ 의 연결을 평균 풀링함으로써 최종 표현 $\mathbf{I}$ 를 생성한다. 즉, $\mathbf{I}$ 는 다음의 [수학식 1]과 같이 나타낼 수 있다.

[0052] [수학식 1]

$$[0053] \quad \mathbf{I} = \text{AveragePooling}(\mathbf{V} \oplus \mathbf{T} \oplus \mathbf{V}^{DA} \oplus \mathbf{T}^{DA})$$

[0054] 여기서, $\oplus$ 는 벡터 연결을 나타낸다.

[0055] 분류기(150)는 평균 풀링기(140)의 출력 $\mathbf{I}$ 로부터 Q의 난이도 수준을 결정한다. 일례로, 분류기(150)로서 간단한 다층 퍼셉트론(Multi-Layer Perception, MLP)가 채택되면, 난이도 수준 d^* 는 다음의 [수학식 2]에 따라 산출될 수 있다.

[0056] [수학식 2]

$$[0057] \quad d^* = \text{softmax}(\mathbf{W}_d \mathbf{I}^T + b_d)$$

[0058] 여기서, $\mathbf{W}_d \in \mathbb{R}^{k \times h}$ 및 b_d 는 분류기(150)의 훈련 가능한 파라미터이고, k 는 난이도 수준의 개수이다.

[0059] 이하, 전술한 비디오 모달리티 네트워크(110)에 대해 구체적으로 설명한다.

[0060] 도 2는 도 1에 도시된 비디오 모달리티 네트워크(110)의 구체적인 구성 블록도이다.

[0061] 비디오 모달리티 네트워크(110)의 목적은 참조 비디오 V 를 $\mathbf{V}$ 및 $\mathbf{V}_{seq}$ 로 표현하는 것이며, 여기서 $\mathbf{V}_{seq}$ 는 V 의 순차적 표현이다.

[0062] 일반적으로, 순차적 데이터로부터 시간적 및 공간적 특징을 추출하기 위해 CNN과 LSTM의 조합이 사용되어 왔지만, 순차적 데이터의 긴 종속성을 캡처하는 데 더 효과적이기 때문에 최근에는 변환기(transformer)가 이를 대체하고 있다.

[0063] 이에 따라, 도 2에 도시된 바와 같이, 비디오 모달리티 네트워크(110)는 CNN 백본(111), 변환기(113) 및 풀링기(115)의 조합으로 구성된다.

[0064] CNN 백본(111)은 예를 들어 ImageNet으로 훈련된 ResNet이며, 먼저 입력인 참조 비디오 V 로부터 공간적 특징 V_{emb} 를 추출한다. 여기서, V_{emb} 는 다음의 [수학식 3]에 의해 획득될 수 있다.

[0065] [수학식 3]

$$[0066] \quad V_{emb} = \text{ResNet}(V) \in \mathbb{R}^{n \times h}$$

[0067] 그런 다음, 인코더인 변환기(113)는 V_{emb} 에 시간적 정보를 할당함으로써 V 의 순차적 표현인 $\mathbf{V}_{seq}$ 를 계산할 수 있다. 여기서, $\mathbf{V}_{seq}$ 는 다음의 [수학식 4]를 사용하여 계산될 수 있다.

[0068] [수학식 4]

$$[0069] \quad \mathbf{V}_{seq} = \text{Transformer}(V_{[CLS]} \oplus V_{emb}) \in \mathbb{R}^{(n+1) \times h}$$

[0070] 여기서, $V_{[CLS]}$ 는 모든 순차적 정보를 결합하는 벡터이다.

[0071] 풀링기(115)는 변환기(113)의 출력인 $V_{[CLS]}$ 에 대해 다음의 [수학식 5]와 같이 풀링을 수행하여 최종 비디오 표현 $\mathbf{V}$ 를 출력한다.

[0072] [수학식 5]

$$[0073] \quad \mathbf{V} = \tanh(\mathbf{W}_v (\mathbf{V}_{seq}[0])^T + b_v)$$

[0074] 여기서, $\mathbf{W}_v \in \mathbb{R}^{h \times h}$ 및 b_v 는 훈련 가능한 파라미터이고, $\mathbf{V}_{seq}[0]$ 은 변환기(113)에 의해 생성된 [CLS] 토큰의 특

징 벡터를 지칭한다.

[0075] 다음, 전술한 텍스트 모달리티 네트워크(120)에 대해 구체적으로 설명한다.

[0076] 도 3은 도 1에 도시된 텍스트 모달리티 네트워크(120)의 구체적인 구성 블록도이다.

[0077] 본 발명의 실시예에 따른 텍스트 모달리티 네트워크(120)는 하이퍼파라미터와 학습 데이터의 규모 등을 조절하여 더 길고 오래 학습하는 방식으로 BERT(Bidirectional Encoder Representations from Transformers)의 단점을 보완한 언어 표현 모델인 RoBERTa(Robustly Optimized BERT Pretraining Approach)(121)를 채택하여 Q와 U의 모든 텍스트 정보를 텍스트 모달리티 벡터인 $\mathbf{T}$ 로 표현한다.

[0078] Q와 U의 연결은 RoBERTa(121)의 입력으로 사용되며, 여기서 그들의 분리를 위해 특수 토큰 [SEP]이 사용된다. RoBERTa(121)는 먼저 Q와 U의 연결을 다음의 [수학식 6]과 같이 공식화된 텍스트 순차적 표현 $\mathbf{T}_{seq}$ 로 인코딩한다.

[0079] [수학식 6]

$$\mathbf{T}_{seq} = RoBERTa(Q, U) \in R^{(m+l+2) \times h}$$

[0081] [CLS] 및 [SEP] 토큰으로 인해 $\mathbf{T}_{seq}$ 의 차원에서 2가 필요하다.

[0082] 그런 다음, RoBERTa(121)는 입력 텍스트를 다음의 [수학식 7]을 사용하여 [CLS] 토큰의 벡터인 $h_{[CLS]}$ 로 요약한다.

[0083] [수학식 7]

$$h_{[CLS]} = \mathbf{T}_{seq}[0]$$

[0085] 이와 같이, $h_{[CLS]}$ 가 계산되면, 비디오 모달리티 네트워크(110)에서와 같이 풀링기(123)가 $h_{[CLS]}$ 를 풀링함으로써 최종 텍스트 표현 $\mathbf{T}$ 가 다음의 [수학식 8]에 EK라 획득된다.

[0086] [수학식 8]

$$\mathbf{T} = \tanh(\mathbf{W}_o(h_{[CLS]})^T + b_o)$$

[0088] 여기서, $\mathbf{W}_o \in R^{h \times h}$ 및 b_o 은 훈련 가능한 파라미터 행렬이다.

[0089] 다음, 전술한 방향성 모달리티 연관성 변환기(130)에 대해 구체적으로 설명한다.

[0090] 도 4는 도 1에 도시된 방향성 모달리티 연관성 변환기(130)의 구체적인 구성 블록도이다.

[0091] 방향성 모달리티 연관성 변환기(130)는 하나의 모달리티, 예를 들어 비디오 모달리티가 다른 모달리티, 예를 들어 텍스트 모달리티에 미치는 방향성 영향을 계산한다.

[0092] 도 4에 도시된 바와 같이, 방향성 모달리티 연관성 변환기(130)는 방향성 어텐션(Directional Attention) 계층(131), 피드 포워드(Feed-Forward) 네트워크 계층(133) 및 레이어 정규화와 잔여 연결의 조합인 2개의 Add & Norm(135, 137)을 포함한다. 구체적으로는, 1개의 블록이 방향성 어텐션 계층(131), 피드 포워드 네트워크 계층(133) 및 2개의 Add & Norm(135, 137)을 포함하는 경우, 이러한 블록이 N개가 존재하게 된다.

[0093] 먼저, 방향성 어텐션 계층(131)의 입력인 M_1 과 M_2 를 두 가지 모달리티의 표현, 즉 $\mathbf{V}_{seq}$ 및 $\mathbf{T}_{seq}$ 로 나타낼 수 있다.

[0094] 방향성 어텐션 계층(131)은 M_1 에 대한 M_2 의 연관을 나타내는, 즉 $\mathbf{V}_{seq}$ 에 대한 $\mathbf{T}_{seq}$ 를 나타내는 M_1^a 을 다음의 [수학식 9]와 같이 산출할 수 있다.

[0095] [수학식 9]

$$M_1^a = \text{softmax}\left(\frac{M_1 M_2^T}{\sqrt{s}}\right) M_2$$

[0096]

[0097] 여기서, s는 스케일링 인자이다.

[0098] 전술한 방향성 어텐션 계층(131)을 제외한 나머지 구성은 일반적인 변환기의 역할과 동일하다. 따라서, 최종 방향성 모달리티 연관성 변환(M_1, M_2)은 M_2 에 의해 수반된 M_1 의 새로운 표현을 생성한다. 실제로, 이러한 새로운 표현은 입력이 M_1^a 인 피드 포워드 네트워크 계층(133)의 출력($DiMAT++(M_1, M_2)$)이며, 다음의 [수학식 10]에 따라 산출될 수 있다.

[0099] [수학식 10]

$$\begin{aligned} DiMAT++(M_1, M_2) &= AN(FFN(AN(M_1^a))) \\ &= AN\left(FFN\left(AN\left(\text{softmax}\left(\frac{M_1 M_2^T}{\sqrt{s}}\right) M_2\right)\right)\right) \end{aligned}$$

[0100]

[0101] 마찬가지로, M_2 에 대한 M_1 의 연관을 나타내는 M_2^a 에 대한 새로운 표현은 피드 포워드 네트워크 계층(133)의 출력($DiMAT++(M_2, M_1)$)이며, 다음의 [수학식 11]에 따라 산출될 수 있다.

[0102] [수학식 11]

$$DiMAT++(M_2, M_1) = AN(FFN(AN(M_2^a)))$$

[0103]

[0104] 여기서, M_2^a 는 방향성 어텐션 계층(131)에 의해 다음의 [수학식 12]와 같이 산출될 수 있다.

[0105] [수학식 12]

$$M_2^a = \text{softmax}\left(\frac{M_2 M_1^T}{\sqrt{s}}\right) M_1$$

[0106]

[0107] 한편, 2개의 Add & Norm(135, 137)은 입력 x에 대해 방향성 어텐션 계층(131) 또는 피드 포워드 네트워크 계층(133)의 출력이 Sublayer(x)인 경우 다음의 [수학식 13]과 같이 출력(AN(x))을 산출할 수 있다.

[0108] [수학식 13]

$$AN(x) = LayerNorm(x + Sublayer(x))$$

[0109]

[0110] 한편, 텍스트와 비디오는 비디오 QA의 두 가지 주요 모달리티이다. 따라서, 방향성 모달리티 연관성 변환기(130)는 텍스트에 의해 영향을 받은 비디오 표현(V^{DA})과 비디오에 의해 영향을 받은 텍스트 표현(T^{DA})을 다음의 [수학식 13]와 같이 산출할 수 있다.

[0111] [수학식 14]

$$\begin{aligned} V^{DA} &= DiMAT++(V_{seq}, T_{seq})[0] \in R^h \\ T^{DA} &= DiMAT++(T_{seq}, V_{seq})[0] \in R^h \end{aligned}$$

[0112]

[0113] 여기서, V_{seq} 와 T_{seq} 는 각각 전술한 [수학식 4]와 [수학식 6]에 의해 산출된 비디오 시퀀스 표현 및 텍스트 시퀀스 표현이다.

[0114] 본 발명의 실시예에 따르면, 비디오 모달리티와 텍스트 모달리티 간의 방향 연관성을 모델링할 뿐만 아니라 서로 간의 방향 연관성 모델링시 텍스트와 비디오의 순차적 정보를 사용하여 방향 연관성을 산출함으로써 보다 향상된 질의 난이도 추정이 가능하다.

[0115] 이로 인해, 멀티 보달 환경에서 보다 정확하고 향상된 질의 난이도를 분석하여 이에 따른 응답을 제공할 수 있다.

[0116] 본 발명의 실시예에 따른 질의 난이도 추정 장치(100)의 효율성은 두 개의 벤치마크 데이터 세트를 통해 검증될 수 있다.

[0117] 하나는, 가장 널리 사용되는 비디오 QA 데이터 세트인 TVQA 데이터 세트이다. 그것은 6개의 TV 프로그램으로 구성되었다. 각 질문의 비디오 클립이 모두 자막이 붙어 있더라도, 질의의 난이도는 제공되지 않는다. 따라서, 각 질의의 난이도 수준은 다음과 같이 할당된다.

[0118] 우선, 모든 질의는 TVQA의 기본 모델을 사용하여 답변된다. 기본 모델의 정확도는 68.48%에 불과하며, 이는 질의에 답변하는 데 약간의 오류가 있음을 의미한다. 질의가 기본 모델에 의해 올바르게 답변되면, 쉬운 질의로 간주된다. 반대로, 기본 모델이 질의에 잘못된 답변을 제공하면 해당 질의는 어려운 것으로 간주된다. TVQA 데이터 세트의 137,292개 질의 중 87,292개 질의가 기본 모델을 훈련하는 데 사용된다. 나머지 50,000개의 질의는 제안된 모델을 평가하는 데 사용된다.

[0119] 다른 데이터 세트는 DramaQA 데이터 세트이다. 이것도 한국 TV쇼 '또 오해영'을 기반으로 한 비디오 QA 데이터 세트이다.

[0120] 이러한 데이터 세트의 각 질의는 비디오 클립, 일부 시각적 메타데이터 및 상호 참조 해결 스크립트로 보완되지만, 본 발명의 실시예에서는 TVQA 데이터 세트와의 일관성을 보장하기 위해 비디오 클립만이 사용된다.

[0121] 이러한 데이터 세트의 주요 특징은 각 질의의 난이도가 메모리 복잡도와 로직 복잡도라는 두 가지 기준으로 태그되어 있다는 것이다. 메모리 복잡도는 질의에 답변하는 데 필요한 메모리 길이를 측정하며 두 가지 난이도 수준으로 구성된다. 반면, 로직 복잡도는 질의에 답변하는 데 필요한 추론 수준과 보조 정보의 양을 측정한다. 네 가지 난이도 수준이 있다.

[0122] 전술한 두 개의 데이터 세트에 대한 간단한 통계가 다음의 [표 1]에 나와 있다.

[0123] [표 1]

데이터 세트	TVQA	DramaQA
훈련 인스턴스 개수	40,000	21,113
검증 인스턴스 개수	5,000	4,385
테스트 인스턴스 개수	5,000	4,463
난이도 수준 개수	2	2 (메모리 복잡도) 4 (로직 복잡도)

[0124]

[0125] 이하, 본 발명의 실시예에 따른 질의 난이도 추정 방법에 대해 설명한다.

[0126] 도 5는 본 발명의 실시예에 따른 질의 난이도 추정 방법의 개략적인 흐름도이다. 도 5에 도시된 본 발명의 실시예에 따른 질의 난이도 추정 방법은 전술한 도 1 내지 도 4를 참조하여 설명한 질의 난이도 추정 장치(100)에 의해 실행될 수 있다.

[0127] 도 5를 참조하면, 먼저, 입력 비디오 V 가 비디오 모달리티 네트워크(110)를 통해 비디오 표현 $\mathbf{V} \in \mathbb{R}^h$ 로 인코딩되고, 인코딩 중에 입력 비디오 V 의 순차적 표현 $\mathbf{V}_{seq}$ 가 생성되고, 이와 동시에, 쿼리 Q 와 자막 U 가 텍스트 모달리티 네트워크(120)에 의해 텍스트 표현 $\mathbf{T} \in \mathbb{R}^h$ 로 변환되고, 변환 중에 텍스트 T 의 순차적 표현 $\mathbf{T}_{seq}$ 가 생성된다(S110). 여기서, 비디오 표현 $\mathbf{V}$, 비디오 V 의 순차적 표현 $\mathbf{V}_{seq}$, 텍스트 표현 $\mathbf{T}$ 및 텍스트 T 의 순차적 표현 $\mathbf{T}_{seq}$ 의 생성에 대해서는 전술한 [수학식 5], [수학식 4], [수학식 8] 및 [수학식 6]을 참조한다.

[0128] 다음, 입력 비디오 V 의 순차적 표현 $\mathbf{V}_{seq}$ 와 텍스트 T 의 순차적 표현 $\mathbf{T}_{seq}$ 을 사용하여 텍스트 모달리티에 의해

수반된 비디오 표현 V^{DA} 과 비디오 모달리티에 의해 수반된 텍스트 표현 T^{DA} 가 생성된다(S120). 여기서, 텍스트 모달리티에 의해 수반된 비디오 표현 V^{DA} 과 비디오 모달리티에 의해 수반된 텍스트 표현 T^{DA} 의 생성에 대해서는 전술한 [수학식 14]를 참조한다.

[0129] 그 후, 비디오 표현 V , 텍스트 표현 T , 텍스트 모달리티에 의해 수반된 비디오 표현 V^{DA} 및 비디오 모달리티에 의해 수반된 텍스트 표현 T^{DA} 의 연결을 평균 풀링하여 최종 표현 I 가 생성된다(S130). 여기서, 최종 표현 I 의 생성에 대해서는 전술한 [수학식 1]을 참조한다.

[0130] 마지막으로, 최종 표현 I 에 기초하여 쿼리(질의) Q에 대한 난이도 수준이 결정된다(S140). 여기서, 난이도 수준에 대해서는 전술한 [수학식 2]를 참조한다.

[0131] 전술한 단계들, S110, S120, S130, S140에 대한 구체적인 설명은 전술한 도 1 내지 도 4를 참조하여 설명한 질의 난이도 추정 장치(100)에 대한 설명을 참조하면 쉽게 이해될 것이므로 여기에서는 구체적인 설명을 생략한다.

[0132] 다음, 본 발명의 다른 실시예에 따른 질의 난이도 추정 장치에 대해 설명한다.

[0133] 도 6은 본 발명의 다른 실시예에 따른 질의 난이도 추정 장치(200)의 개략적인 구성을 도시한 도면이다.

[0134] 도 6에 도시된 바와 같이, 본 발명의 실시예에 따른 질의 난이도 추정 장치(200)는 적어도 하나의 프로세서(210), 메모리(220), 통신기(230), 입출력기(240) 및 통신 버스(250)를 포함한다.

[0135] 프로세서(210)는 범용 CPU(Central Processing Unit), 마이크로프로세서, ASIC(Application-Specific Integrated Circuit), 또는 본 출원의 해결수단에서 프로그램 실행을 제어하기 위한 하나 이상의 집적 회로일 수 있다.

[0136] 메모리(220)는 본 발명의 실시예에 따른 질의 난이도 추정을 위한 정보를 저장한다.

[0137] 구체적으로, 메모리(220)는 코드의 집합을 저장하도록 추가로 구성되고, 그 코드는 다음과 같은 프로세스를 실행하기 위해 프로세서(210)를 제어하는 데 사용된다. 이러한 프로세스는, 비디오 모달리티 네트워크(110)를 통해 입력 비디오를 대응하는 비디오 표현으로 인코딩하고 인코딩 중에 입력 비디오의 순차적 표현을 생성하며, 이와 동시에, 텍스트 모달리티 네트워크(120)를 통해 쿼리와 자막에 대응하는 텍스트 표현을 생성하고 변환 중에 텍스트의 순차적 표현을 생성하는 프로세스, 입력 비디오의 순차적 표현과 텍스트의 순차적 표현을 사용하여 텍스트 모달리티에 의해 수반된 비디오 표현과 비디오 모달리티에 의해 수반된 텍스트 표현을 생성하는 프로세스, 비디오 표현, 텍스트 표현, 텍스트 모달리티에 의해 수반된 비디오 표현 및 비디오 모달리티에 의해 수반된 텍스트 표현의 연결을 평균 풀링하여 최종 표현을 생성하는 프로세스, 최종 표현에 기초하여 쿼리(질의)에 대한 난이도 수준을 결정하는 프로세스 등을 포함한다.

[0138] 메모리(220)는 ROM(Read-Only Memory) 또는 명령을 저장할 수 있는 다른 유형의 정적 저장 장치, 또는 RAM(Random Access Memory) 또는 정보 및 명령을 저장할 수 있는 다른 유형의 동적 저장 장치일 수 있거나, 또는 EEPROM(Electrically Erasable Programmable Read-Only Memory), CD-ROM(Compact Disc Read-Only Memory) 또는 다른 콤팩트 디스크 저장 장치 또는 광 디스크 저장 장치(압축 광 디스크, 레이저 디스크, 광 디스크, 디지털 다용도 디스크, 블루레이 디스크 등을 포함함), 자기 디스크 저장 매체 또는 다른 자기 저장 장치, 또는 명령 또는 데이터 구조의 형태로 예상 프로그램 코드를 운반하거나 저장할 수 있으면서 컴퓨터에 의해 액세스될 수 있는 임의의 다른 매체일 수 있으며, 이것은 제한되지 않는다. 메모리(220)는 독립적으로 존재할 수 있으며, 통신 버스(250)에 의해 프로세서(210)에 연결된다.

[0139] 통신기(230)는 외부의 다른 장치와 유무선 통신을 수행하며, 다양한 유선 또는 무선 통신 기술로 구현될 수 있다. 예를 들어, 통신기(230)는 인터넷과 연결되어 서비스를 제공하는 경우 인터넷에서 정보 전송을 위한 표준 프로토콜인 TCP/IP를 따를 수 있다.

[0140] 입출력기(240)는 구체적으로는 입력 장치(241)와 출력 장치(242)로 구성되며, 입력 장치(241)는 프로세서(210)와 통신하고, 복수의 방식으로 사용자의 입력을 수신할 수 있다. 예를 들어, 입력 장치(241)는 마우스, 키보드, 터치 스크린 또는 센싱 장치일 수 있다. 출력 장치(242)는 프로세서(210)와 통신하고, 복수의 방식으로 정보를 디스플레이하거나 음성을 출력할 수 있다. 예를 들어, 출력 장치(242)는 LCD(Liquid Crystal Display),

LED(Light Emitting Diode) 디스플레이, OLED(Organic Light Emitting Diode) 디스플레이, 스피커 등일 수 있다.

[0141] 통신 버스(250)는 단어 의미의 모호성 해소를 위한 의미 유지 문장 혼합 장치(200)의 모든 컴포넌트들, 즉 프로세서(210), 메모리(220), 통신기(230) 및 입출력기(240)를 결합하도록 구성된다.

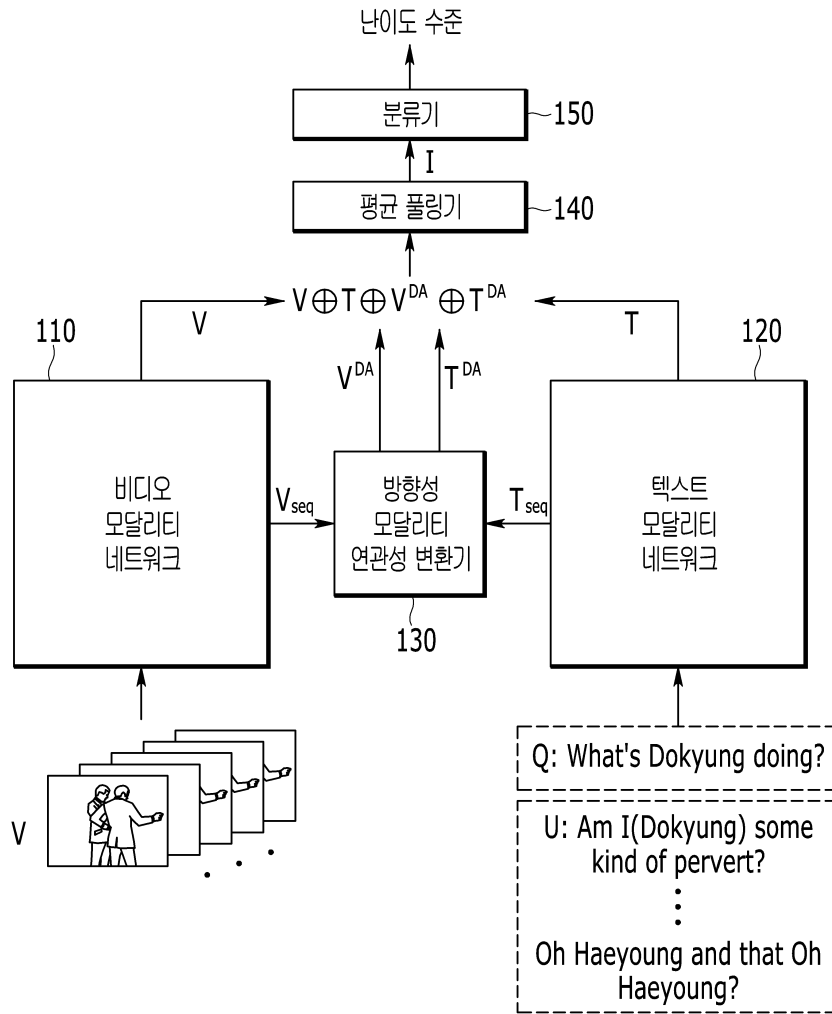
[0142] 본 발명의 실시예에서 사용되는 구성요소 또는 “~부”는 메모리 상의 소정 영역에서 수행되는 태스크, 클래스, 서브 루틴, 프로세스, 오브젝트, 실행 쓰레드, 프로그램과 같은 소프트웨어(software)나, FPGA(field-programmable gate array)나 ASIC(application-specific integrated circuit)과 같은 하드웨어(hardware)로 구현될 수 있으며, 또한 상기 소프트웨어 및 하드웨어의 조합으로 이루어질 수도 있다. 상기 구성요소 또는 ‘~부’ 등은 컴퓨터로 판독 가능한 저장 매체에 포함되어 있을 수도 있고, 복수의 컴퓨터에 그 일부가 분산되어 분포될 수도 있다.

[0143] 이상에서 설명한 본 발명의 실시예는 장치 및 방법을 통해서만 구현이 되는 것은 아니며, 본 발명의 실시예의 구성에 대응하는 기능을 실현하는 프로그램 또는 그 프로그램이 기록된 기록 매체를 통해 구현될 수도 있다.

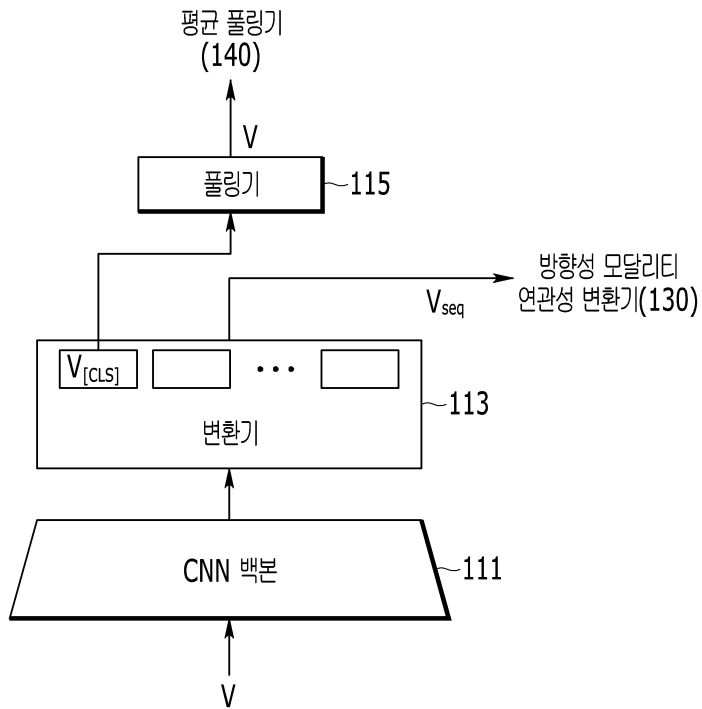
[0144] 이상에서 본 발명의 실시예에 대하여 상세하게 설명하였지만 본 발명의 권리범위는 이에 한정되는 것은 아니고 다음의 청구범위에서 정의하고 있는 본 발명의 기본 개념을 이용한 당업자의 여러 변형 및 개량 형태 또한 본 발명의 권리범위에 속하는 것이다.

도면

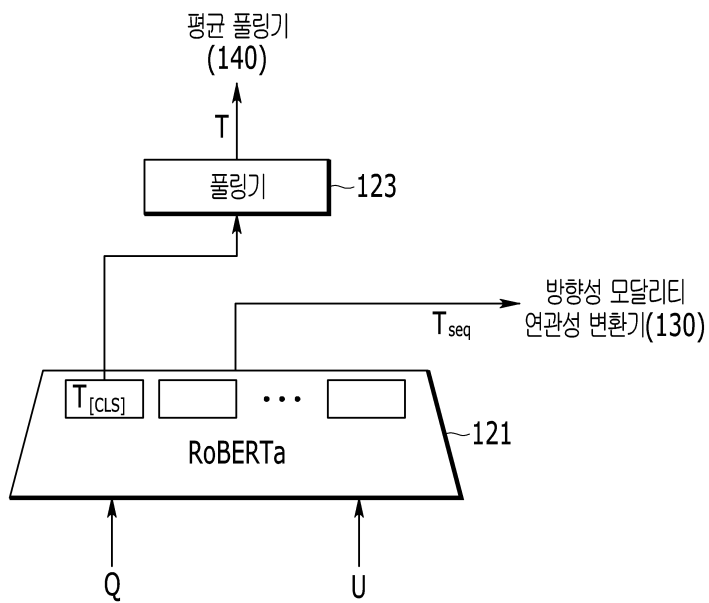
도면1



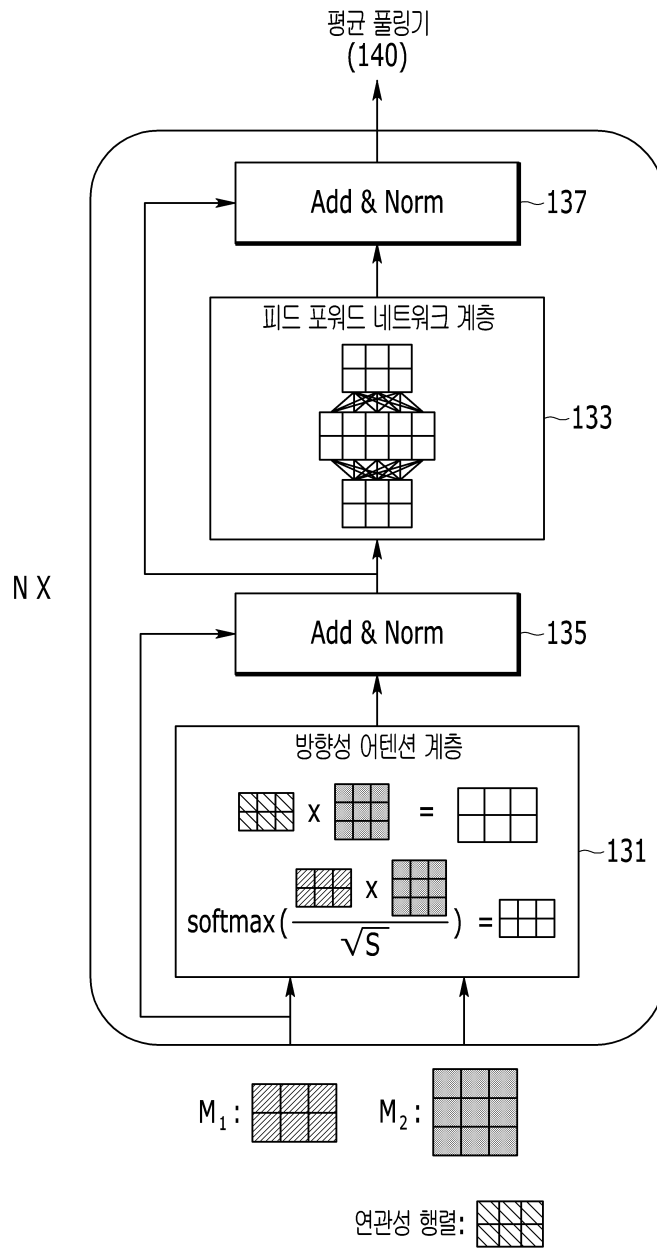
도면2



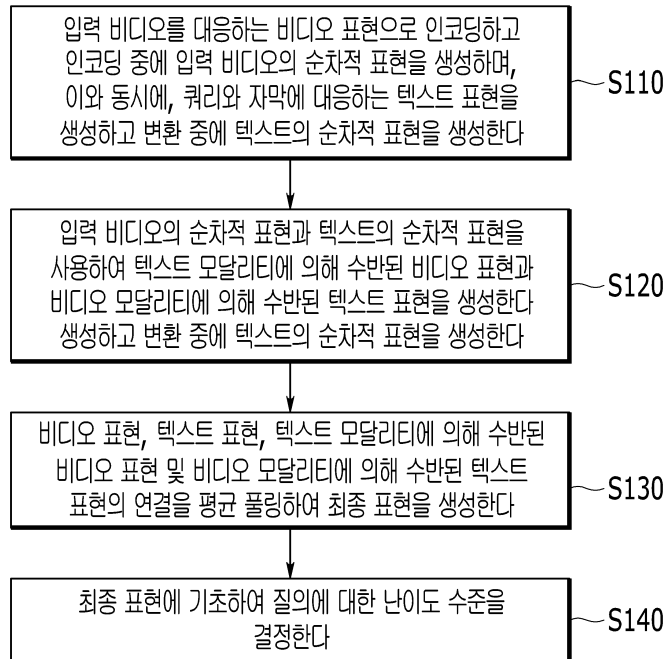
도면3



도면4



도면5



도면6

