(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0096149
(43) 공개일자 2024년06월26일

(51) 국제특허분류(Int. Cl.)
G10L 21/10 (2013.01) *G06F 40/279* (2020.01)
G06N 3/04 (2023.01) *G10L 15/08* (2006.01)
G10L 15/26 (2006.01)
(52) CPC특허분류
G10L 21/10 (2013.01)
G06F 40/279 (2020.01)
(21) 출원번호 10-2022-0178525
(22) 출원일자 2022년12월19일
심사청구일자 2022년12월19일

(71) 출원인
울산과학기술원
울산광역시 울주군 언양읍 유니스트길 50
(72) 발명자
김태환
울산광역시 울주군 언양읍 유니스트길 50
이태경
울산광역시 울주군 언양읍 유니스트길 50
(74) 대리인
리앤목특허법인

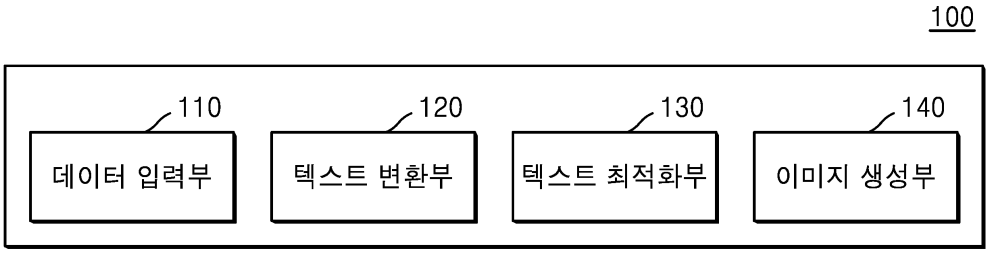
전체 청구항 수 : 총 13 항

(54) 발명의 명칭 **이미지 생성 장치, 오디오 데이터에 대응하는 이미지를 생성하는 방법 및 컴퓨터 프로그램**

(57) 요 약

본 개시의 실시예들에 따르면, 오디오 데이터를 캡션 데이터로 변환하고 캡션 데이터에서 명사가 강조된 문장 어텐션 데이터를 생성하며, 캡션 데이터와 문장 어텐션 데이터를 결합한 벡터를 입력하여 오디오 데이터에 대응하는 이미지를 생성할 수 있다.

대 표 도 - 도1



(52) CPC특허분류
G06N 3/04 (2023.01)
G10L 15/08 (2013.01)
G10L 15/26 (2013.01)
G10L 2015/088 (2013.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711160496
과제번호	2022-0-00608-001
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	사람중심인공지능핵심원천기술개발
연구과제명	(1세부) 인간과 교감하는 멀티모달 인터랙션 인공지능 기술
기 여 율	1/1
과제수행기관명	울산과학기술원
연구기간	2022.04.01 ~ 2022.12.31

명 세 서

청구범위

청구항 1

이미지 생성 장치가, 오디오 데이터를 읽어들여 오디오 변환기에 입력하고, 상기 오디오 변환기를 통해 상기 오디오 데이터에 대응하는 캡션 데이터 및 오디오 어텐션 데이터를 출력하는 단계;

상기 이미지 생성 장치가, 상기 캡션 데이터를 명사 추출기에 입력하여 문장 어텐션 데이터를 출력하는 단계;

상기 이미지 생성 장치가, 상기 캡션 데이터와 상기 오디오 어텐션 데이터를 인코딩하여 인코딩된 데이터를 생성하고, 상기 인코딩된 데이터와 상기 문장 어텐션 데이터를 결합시켜 특징 벡터를 생성하는 단계;

상기 이미지 생성 장치가, 상기 캡션 데이터를 텍스트 임베딩하여 임베딩 데이터를 생성하고, 상기 임베딩 데이터와 상기 특징 벡터를 곱한 입력 벡터를 산출하는 단계; 및

상기 이미지 생성 장치가, 상기 입력 벡터를 이미지 생성 모델에 입력하여 상기 오디오 데이터에 대응하는 이미지를 생성하고, 생성한 이미지를 출력하는 단계;를 포함하는, 오디오 데이터에 대응하는 이미지를 생성하는 방법.

청구항 2

제1항에 있어서,

상기 이미지를 출력하는 단계는,

상기 이미지와 상기 캡션 데이터를 비교하여 비교값에 대응하는 제1 결과값을 산출하고, 상기 이미지와 상기 오디오 데이터를 비교하여 비교값에 대응하는 제2 결과값을 산출하며, 상기 제1 결과값과 상기 제2 결과값에 기초한 최종 결과값을 고려하여 상기 입력 벡터를 다시 생성하고 다시 생성한 입력 벡터에 대응하는 이미지를 다시 생성하는, 오디오 데이터에 대응하는 이미지를 생성하는 방법.

청구항 3

제1항에 있어서,

상기 오디오 변환기는,

미리 저장된 오디오 데이터들과, 상기 오디오 데이터들에 대한 캡션 데이터들의 학습 데이터로 학습된 모델로, 오디오 데이터를 입력으로 하고 캡션 데이터를 출력으로 하는 것인, 오디오 데이터에 대응하는 이미지를 생성하는 방법.

청구항 4

제1항에 있어서,

상기 명사 추출기는,

캡션 데이터에서 명사에 해당하는 단어를 출력하도록 학습된 모델로,

캡션 데이터를 입력으로 하고 상기 캡션 데이터에 포함된 각 단어에 대한 명사될 확률값을 출력하는 것인, 오디오 데이터에 대응하는 이미지를 생성하는 방법.

청구항 5

제1항에 있어서,

상기 인코딩된 데이터는,

상기 캡션 데이터에 포함된 각 단어에 대해서 각 단어의 위치 정보를 추가하여 인코딩 처리한 데이터인, 오디오 데이터에 대응하는 이미지를 생성하는 방법.

청구항 6

제1항에 있어서,

상기 이미지 생성 모델은,

미리 저장된 텍스트 데이터들과, 상기 텍스트 데이터들에 대한 이미지들의 학습 데이터로 학습된 모델로, 텍스트 데이터를 입력으로 하고 텍스트 데이터에 대응하는 이미지를 출력으로 하는 것인, 오디오 데이터에 대응하는 이미지를 생성하는 방법.

청구항 7

프로세서와 컴퓨터 판독 가능한 메모리를 포함하고,

상기 프로세서가 상기 메모리에 저장된 명령어들을 판독하여

오디오 데이터를 읽어들여 오디오 변환기에 입력하고, 상기 오디오 변환기를 통해 상기 오디오 데이터에 대응하는 캡션 데이터 및 오디오 어텐션 데이터를 출력하고,

상기 캡션 데이터를 명사 추출기에 입력하여 문장 어텐션 데이터를 출력하며,

상기 캡션 데이터와 상기 오디오 어텐션 데이터를 인코딩하여 인코딩된 데이터를 생성하고, 상기 인코딩된 데이터와 상기 문장 어텐션 데이터를 결합시켜 특징 벡터를 생성하고,

상기 캡션 데이터를 텍스트 임베딩하여 임베딩 데이터를 생성하고, 상기 임베딩 데이터와 상기 특징 벡터를 곱한 입력 벡터를 산출하며,

상기 입력 벡터를 이미지 생성 모델에 입력하여 상기 오디오 데이터에 대응하는 이미지를 생성하고, 생성한 이미지를 출력하는, 이미지 생성 장치.

청구항 8

제7항에 있어서,

상기 프로세서는,

상기 이미지와 상기 캡션 데이터를 비교하여 비교값에 대응하는 제1 결과값을 산출하고, 상기 이미지와 상기 오디오 데이터를 비교하여 비교값에 대응하는 제2 결과값을 산출하며, 상기 제1 결과값과 상기 제2 결과값에 기초한 최종 결과값을 고려하여 상기 입력 벡터를 다시 생성하고 다시 생성한 입력 벡터에 대응하는 이미지를 다시 생성하는, 이미지 생성 장치.

청구항 9

제7항에 있어서,

상기 오디오 변환기는,

미리 저장된 오디오 데이터들과, 상기 오디오 데이터들에 대한 캡션 데이터들의 학습 데이터로 학습된 모델로, 오디오 데이터를 입력으로 하고 캡션 데이터를 출력으로 하는 것인, 이미지 생성 장치.

청구항 10

제7항에 있어서,

상기 명사 추출기는,

캡션 데이터에서 명사에 해당하는 단어를 출력하도록 학습된 모델로,

캡션 데이터를 입력으로 하고 상기 캡션 데이터에 포함된 각 단어에 대한 명사될 확률값을 출력하는 것인, 이미지 생성 장치.

청구항 11

제7항에 있어서,

상기 인코딩된 데이터는,

상기 캡션 데이터에 포함된 각 단어에 대해서 각 단어의 위치 정보를 추가하여 인코딩 처리한 데이터인, 이미지 생성 장치.

청구항 12

제7항에 있어서,

상기 이미지 생성 모델은,

미리 저장된 텍스트 데이터들과, 상기 텍스트 데이터들에 대한 이미지들의 학습 데이터로 학습된 모델로, 텍스트 데이터를 입력으로 하고 텍스트 데이터에 대응하는 이미지를 출력으로 하는 것인, 이미지 생성 장치.

청구항 13

컴퓨터를 이용하여 제1항 내지 6항의 방법을 실행시키기 위하여 컴퓨터 판독 가능한 저장 매체에 저장된 컴퓨터 프로그램.

발명의 설명

기 술 분 야

[0001] 본 개시의 실시예들은, 이미지 생성 장치, 오디오 데이터에 대응하는 오디오 데이터에 대응하는 이미지를 생성하는 방법 및 컴퓨터 프로그램에 관한 것으로, 보다 구체적으로는, 오디오 데이터를 캡션 데이터로 변환하고 캡션 데이터에서 명사가 강조된 문장 어텐션 데이터를 생성하며, 캡션 데이터와 문장 어텐션 데이터를 결합한 벡터를 입력하여 오디오 데이터에 대응하는 이미지를 생성하는 점을 특징으로 한다.

배 경 기 술

[0002] 인간의 두뇌를 모델링하는 뉴럴 네트워크 연구는 1940년대 신경 세포의 모델링부터 시작하여 현재까지 다양한 기술이 축적되어 왔다. 또한, 2006년에는, 성능이 향상될 수 있는 DBN, RBM 개념을 시작으로 다계층 구조에서도 훈련이 가능한 방법들이 소개되면서 다시 뉴럴 네트워크가 주목받기 시작했다.

[0003] 특히, DBN 구조의 딥러닝 기술이 음성 인식에 활용하면서 기존의 GMM-HMM 프레임 워크에서 가지고 있던 성능의 한계를 넘어서는 음성 인식 기술이 출현했다. 음성 인식 기술은, 음성 데이터를 그에 대응하는 텍스트로 변환할 수 있게 하였다.

[0004] 그러나, 천둥이 치는 소리, 강아지가 짖는 소리, 아이가 우는 소리, 사람이 소리치는 소리 등의 하나 이상의 소리를 포함하는 오디오 데이터에 대응하여 소리와 대응하는 이미지를 생성하지 못하는 문제점이 있었다.

발명의 내용

해결하려는 과제

[0005] 본 명세서에서 개시되는 실시예들은, 오디오 데이터를 캡션 데이터로 변환하고 캡션 데이터에서 명사가 강조된 문장 어텐션 데이터를 생성하며, 캡션 데이터와 문장 어텐션 데이터를 결합한 벡터를 입력하여 오디오 데이터에 대응하는 이미지를 생성하는 이미지 생성 장치, 오디오 데이터에 대응하는 이미지를 생성하는 방법 및 컴퓨터 프로그램을 제시하는데 목적이 있다.

과제의 해결 수단

[0006] 본 개시의 실시 예들에 따른 방법은, 이미지 생성 장치가, 오디오 데이터를 읽어들여 오디오 변환기에 입력하고, 상기 오디오 변환기를 통해 상기 오디오 데이터에 대응하는 캡션 데이터 및 오디오 어텐션 데이터를 출력하는 단계; 상기 이미지 생성 장치가, 상기 캡션 데이터를 명사 추출기에 입력하여 문장 어텐션 데이터를 출력하는 단계; 상기 이미지 생성 장치가, 상기 캡션 데이터와 상기 오디오 어텐션 데이터를 인코딩하여 인코딩된 데이터를 생성하고, 상기 인코딩된 데이터와 상기 문장 어텐션 데이터를 결합시켜 특징 벡터를 생성하는 단

계; 상기 이미지 생성 장치가, 상기 캡션 데이터를 텍스트 임베딩하여 임베딩 데이터를 생성하고, 상기 임베딩 데이터와 상기 특징 벡터를 곱한 입력 벡터를 산출하는 단계; 및 상기 이미지 생성 장치가, 상기 입력 벡터를 이미지 생성 모델에 입력하여 상기 오디오 데이터에 대응하는 이미지를 생성하고, 생성한 이미지를 출력하는 단계;를 포함할 수 있다.

[0007] 상기 이미지를 출력하는 단계는, 상기 이미지와 상기 캡션 데이터를 비교하여 비교값에 대응하는 제1 결과값을 산출하고, 상기 이미지와 상기 오디오 데이터를 비교하여 비교값에 대응하는 제2 결과값을 산출하며, 상기 제1 결과값과 상기 제2 결과값에 기초한 최종 결과값을 고려하여 상기 입력 벡터를 다시 생성하고 다시 생성한 입력 벡터에 대응하는 이미지를 다시 생성할 수 있다.

[0008] 상기 오디오 변환기는, 미리 저장된 오디오 데이터들과, 상기 오디오 데이터들에 대한 캡션 데이터들의 학습 데이터로 학습된 모델로, 오디오 데이터를 입력으로 하고 캡션 데이터를 출력으로 하는 것일 수 있다.

[0009] 상기 명사 추출기는, 캡션 데이터에서 명사에 해당하는 단어를 출력하도록 학습된 모델로, 캡션 데이터를 입력으로 하고 상기 캡션 데이터에 포함된 각 단어에 대한 명사될 확률값을 출력하는 것 일 수 있다.

[0010] 상기 인코딩된 데이터는, 상기 캡션 데이터에 포함된 각 단어에 대해서 각 단어의 위치 정보를 추가하여 인코딩 처리한 데이터일 수 있다.

[0011] 상기 이미지 생성 모델은, 미리 저장된 텍스트 데이터들과, 상기 텍스트 데이터들에 대한 이미지들의 학습 데이터로 학습된 모델로, 텍스트 데이터를 입력으로 하고 텍스트 데이터에 대응하는 이미지를 출력으로 하는 것일 수 있다.

[0012] 본 개시의 실시예들에 따른 이미지 생성 장치는 프로세서와 컴퓨터 판독 가능한 메모리를 포함하고, 상기 프로세서가 상기 메모리에 저장된 명령어들을 판독하여 오디오 데이터를 읽어들여 오디오 변환기에 입력하고, 상기 오디오 변환기를 통해 상기 오디오 데이터에 대응하는 캡션 데이터 및 오디오 어텐션 데이터를 출력하고, 상기 캡션 데이터를 명사 추출기에 입력하여 문장 어텐션 데이터를 출력하며, 상기 캡션 데이터와 상기 오디오 어텐션 데이터를 인코딩하여 인코딩된 데이터를 생성하고, 상기 인코딩된 데이터와 상기 문장 어텐션 데이터를 결합시켜 특징 벡터를 생성하고, 상기 캡션 데이터를 텍스트 임베딩하여 임베딩 데이터를 생성하고, 상기 임베딩 데이터와 상기 특징 벡터를 곱한 입력 벡터를 산출하며, 상기 입력 벡터를 이미지 생성 모델에 입력하여 상기 오디오 데이터에 대응하는 이미지를 생성하고, 생성한 이미지를 출력할 수 있다.

[0013] 본 발명의 실시예에 따른 컴퓨터 프로그램은 컴퓨터를 이용하여 본 발명의 실시예에 따른 방법 중 어느 하나의 방법을 실행시키기 위하여 매체에 저장될 수 있다.

[0014] 이 외에도, 본 발명을 구현하기 위한 다른 방법, 다른 시스템 및 상기 방법을 실행하기 위한 컴퓨터 프로그램을 기록하는 컴퓨터 판독 가능한 기록 매체가 더 제공된다.

[0015] 전술한 것 외의 다른 측면, 특징, 이점이 이하의 도면, 특허청구범위 및 발명의 상세한 설명으로부터 명확해 질 것이다.

발명의 효과

[0016] 전술한 과제 해결 수단 중 어느 하나에 의하면, 오디오 데이터를 캡션 데이터로 변환하고 캡션 데이터에서 명사가 강조된 문장 어텐션 데이터를 생성하며, 캡션 데이터와 문장 어텐션 데이터를 결합한 벡터를 입력하여 오디오 데이터에 대응하는 이미지를 생성할 수 있다.

도면의 간단한 설명

[0017] 도 1은, 본 개시의 실시예들에 따른 소리 데이터에 대응하는 이미지를 생성하는 이미지 생성 장치의 블록도이다.

도 2는 본 개시의 실시예들에 따른 이미지 생성 장치의 구현의 예시 도면이다.

도 3은 본 개시의 실시예들에 따라서 생성된 이미지와 종래의 방법으로 생성된 이미지 사이를 비교한 도면이다.

도 4는 본 개시의 실시예들에 따른 이미지 생성 방법의 흐름도이다.

도 5는, 이미지 생성 단계(S160)의 구체적인 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0018] 이하 첨부된 도면들에 도시된 본 발명에 관한 실시예를 참조하여 본 발명의 구성 및 작용을 상세히 설명한다.

[0019] 본 발명은 다양한 변환을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세한 설명에 상세하게 설명하고자 한다. 본 발명의 효과 및 특징, 그리고 그것들을 달성하는 방법은 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 다양한 형태로 구현될 수 있다.

[0020] 이하, 첨부된 도면을 참조하여 본 발명의 실시예들을 상세히 설명하기로 하며, 도면을 참조하여 설명할 때 동일하거나 대응하는 구성 요소는 동일한 도면부호를 부여하고 이에 대한 중복되는 설명은 생략하기로 한다.

[0021] 본 명세서에서 “학습”, “러닝” 등의 용어는 인간의 교육 활동과 같은 정신적 작용을 지칭하도록 의도된 것이 아닌 절차에 따른 컴퓨팅(computing)을 통하여 기계 학습(machine learning)을 수행함을 일컫는 용어로 해석한다.

[0022] 이하의 실시예에서, 제1, 제2 등의 용어는 한정적인 의미가 아니라 하나의 구성 요소를 다른 구성 요소와 구별하는 목적으로 사용되었다.

[0023] 이하의 실시예에서, 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다.

[0024] 이하의 실시예에서, 포함하다 또는 가지다 등의 용어는 명세서상에 기재된 특징, 또는 구성요소가 존재함을 의미하는 것이고, 하나 이상의 다른 특징들 또는 구성요소가 부가될 가능성을 미리 배제하는 것은 아니다.

[0025] 도면에서는 설명의 편의를 위하여 구성 요소들이 그 크기가 과장 또는 축소될 수 있다. 예컨대, 도면에서 나타난 각 구성의 크기 및 두께는 설명의 편의를 위해 임의로 나타내었으므로, 본 발명이 반드시 도시된 바에 한정되지 않는다.

[0026] 어떤 실시예가 달리 구현 가능한 경우에 특정한 공정 순서는 설명되는 순서와 다르게 수행될 수도 있다. 예를 들어, 연속하여 설명되는 두 공정이 실질적으로 동시에 수행될 수도 있고, 설명되는 순서와 반대의 순서로 진행될 수 있다.

[0027] 본 명세서에서, 오디오 데이터는, 소리를 바이너리 데이터로 추출/변환하여 처리한 데이터를 말한다. 오디오 데이터는, 마이크의 판막이 진동할 때 생성되는 전기 신호를 디지털 데이터(1, 0)으로 변환하여 저장된 것을 말한다. 오디오 데이터의 파일 형식으로는, Mp3, Wav, Aiff, AU, FLAC 등이 있을 수 있으나 이에 한정되지 않고 다양한 형식이 있을 수 있다.

[0028] 본 명세서에서, 캡션 데이터는, 오디오 데이터를 설명하는 텍스트 데이터를 말하며, 복수의 단어들의 조합으로 생성될 수 있다. 캡션 데이터는, 단어, 또는 문장의 형태를 가질 수 있다. 예컨대, 아기가 우는 소리에 대응하여, 아기가 운다와 같은 캡션 데이터가 생성될 수 있다. 강아지가 짖는 소리에 대응하여, 강아지가 짖는다와 같은 캡션 데이터가 생성될 수 있다. 캡션 데이터의 처리를 위해서, 영어와 같은 정해진 언어로 생성될 수 있다. 캡션 데이터는, 텍스트 자체 또는 인코더를 통해서 텍스트가 변환된 데이터를 말할 수 있다.

[0029] 본 명세서에서, 어텐션 데이터는, 딥러닝 모델에서 활용되기 위한 데이터로서, 캡션 데이터에 포함된 단어들 중에서 강조되어야 하는 단어들에 대한 정보, 각 단어에 대한 가중치 값 등을 포함할 수 있다.

[0030] 본 명세서에서, 기계 학습은, 훈련 데이터로부터 하나의 함수를 유추해내기 위한 방법으로, 지도 학습, 비지도 학습, 강화 학습 등의 방법 중 하나를 포함할 수 있다.

[0031] 본 명세서에서, 인공 신경망은, 기계학습과 인지과학에서 생물학의 신경망에서 영감을 얻은 통계학적 학습 알고리즘으로, 스냅스의 결합으로 네트워크를 형성한 인공 뉴런이 학습을 통해 시냅스의 결합 세기를 변화시켜 문제 해결 능력을 가지도록 하는 모델을 말한다.

[0032] 도 1은, 본 개시의 실시예들에 따른 소리 데이터에 대응하는 이미지를 생성하는 이미지 생성 장치의 블록도이다.

[0033] 데이터 입력부(110)는, 오디오 데이터를 입력한다. 오디오 데이터는 소정의 방법으로 전처리될 수 있다.

[0034] 오디오 데이터는, 사람의 음성, 동물의 울음소리, 천둥 번개와 같은 자연의 소리 등을 말하며, 오디오 데이터는, 하나 이상의 소리를 포함하는 것을 말할 수 있다. 오디오 데이터의 예시로는, 고양이가 우는 소리와

천둥 소리가 결합된 데이터, 아이가 우는 소리와 사람이 소리치는 소리가 결합된 데이터, 강아지가 짓는 소리와 아이가 웃는 소리가 결합된 데이터 등이 있을 수 있다.

[0035] 텍스트 입력부(110)는 오디오 데이터를 시계열의 일정한 시간 구간으로 나누고 각 구간에 대한 스펙트럼을 산출하는 방식으로 전처리할 수 있다. 추가적으로, 텍스트 입력부(110)는, 각 구간에 대한 스펙트럼을 정해진 주파수 단위, 예를 들어 멜 단위(Mel unit)로 변환한 스펙트럼을 생성할 수 있다. 이러한 방식으로 오디오 데이터를 전처리 함으로써, 오디오 데이터의 주파수 특성을 시계열 상에서 확인할 수 있다.

[0036] 텍스트 변환부(120)는, 오디오 데이터에 대응하는 캡션 데이터를 생성할 수 있다. 텍스트 변환부(120)는, 오디오 데이터에 대응하여, 캡션 데이터에 포함된 각각의 단어들의 발생 확률을 포함하는 오디오 어텐션 데이터를 생성할 수 있다. 예를 들어, 캡션 데이터로 'a dog is barking'이 출력된 경우, 오디오 어텐션 데이터는, 각 단어의 확률값인 0.5, 0.9, 0.7, 0.8을 포함하여 생성될 수 있다.

[0037] 텍스트 변환부(120)는, 오디오 데이터를 소정의 방법으로 전처리한 후, 오디오 데이터에 대응하는 캡션 데이터 및 오디오 어텐션 데이터를 생성할 수 있다. 텍스트 변환부(120)는, 기계 학습된 오디오 변환기에 오디오 데이터를 입력으로 하여 캡션 데이터 및 오디오 어텐션 데이터를 출력할 수 있다.

[0038] 오디오 변환기는, 오디오 데이터들과, 오디오 데이터들에 대한 캡션 데이터들의 학습 데이터로 학습된 모델로, 바닐라 변환기 구조를 가질 수 있다. 오디오 변환기의 초기 가중치는 미리 정해진 사전 훈련된 가중치 값으로 초기화될 수 있습니다. 오디오 변환기는, 학습된 인공신경망을 이용하여 오디오 데이터를 입력으로 하면 오디오 데이터에 대한 캡션 데이터를 출력할 수 있다. 오디오 변환기는, 인간의 주의집중을 모방한 인공 신경망을 이용할 수 있다. 인간의 주의집중을 모방한 인공 신경망은, 중요한 입력 부분을 다시 참고하는 기계 학습 기법으로 학습되는 것으로, 점곱 주의집중과 멀티헤드 주의 집중 기법 등으로 학습된 것일 수 있다.

[0039] 오디오 어텐션 데이터는, 캡션 데이터의 각 단어의 확률값을 설정함으로써, 각 단어의 가중치를 조절할 수 있다. 오디오 어텐션 데이터는, 오디오 변환기의 디코더가 오디오에 대응하는 캡션 데이터를 생성할 때의 확률값을 포함할 수 있다. 오디오 어텐션 데이터는, 각 오디오에 대하여 유일한 값(고유한 값, Unique)이며, 오디오의 동적 특성에 따라서 가변적인 값을 포함할 수 있다.

[0040] 텍스트 변환부(120)는, 오디오 데이터들과 오디오 데이터에 대응하는 캡션 데이터들로 학습된 오디오 변환기를 이용하여 입력된 오디오 데이터에 대한 캡션 데이터, 오디오 어텐션 데이터를 출력할 수 있다.

[0041] 텍스트 최적화부(130)는, 오디오 데이터에 대응하는 캡션 데이터를 명사 추출기에 입력하여 문장 어텐션 데이터를 출력할 수 있다.

[0042] 명사 추출기는, 캡션 데이터에서 명사에 해당하는 단어를 추출하는 모델로서, 캡션 데이터들과 명사들의 학습 데이터들로 학습된 것일 수 있다. 명사 추출기를 통해서 캡션 데이터의 명사는 오디오 데이터에서 소리를 발생시키는 대상에 해당하는 것으로, 소리를 발생시키는 대상이 무엇인지 식별하도록 한다.

[0043] 문장 어텐션 데이터는, 소리를 발생시키는 주체에 대한 데이터를 의미할 수 있다. 문장 어텐션 데이터는, 캡션 데이터에서 명사에 해당하는 단어들을 강조하여 표현하는 데이터로서, 예컨대, 명사에 동사보다 더 높은 가중치 값을 부여한 데이터를 말할 수 있다. 문장 어텐션 데이터는, 캡션 데이터의 각 단어가 명사일 확률값을 포함한다.

[0044] 텍스트 최적화부(130)는, 오디오 데이터에 대한, 캡션 데이터와 오디오 어텐션 데이터를 위치적으로 인코딩하고, 인코딩된 데이터에 문장 어텐션 데이터를 결합시켜 오디오에 대한 특징 벡터를 생성할 수 있다. 특징 벡터는, 캡션 데이터에 대해서 오디오 어텐션 데이터와 문장 어텐션 데이터를 고려하여 생성된 것으로, 오디오 데이터에 포함된 소리를 발생시키는 주체에 대한 정보를 포함하도록 변형될 것이다.

[0045] 여기서, 위치적인 인코딩은, 캡션 데이터(D12)에 포함된 각 단어에 대해서 단어의 위치 정보를 포함시켜 인코딩 처리를 하는 것을 말한다. 각 단어에 대응되는 값과 각 단어가 문장 안에서 위치한 위치에 대응하는 값을 포함시켜 인코딩을 수행하는 것을 말한다. 예를 들어, A dog is barking에서, 각 단어의 위치에 따라서, A에는, 1, dog에는, 2, is에는 3, barking에는 4로 설정할 수 있다.

[0046] 이미지 생성부(140)는, 텍스트 최적화부(130)를 통해 출력된 캡션 데이터를 기계 학습된 이미지 생성 모델에 입력하여 오디오 데이터에 대응하는 이미지를 출력할 수 있다. 이미지 생성 모델은, 텍스트 데이터에 대응하는 이미지 데이터를 출력으로 하는 기계 학습 모델을 말한다. 이미지 생성 모델은, 학습된 인공 신경망을 이용하여

입력에 대응하는 이미지 데이터를 출력할 수 있다.

[0047] 이미지 생성부(140)는, 입력된 오디오 데이터에 대응하는 캡션 데이터, 오디오 어텐션 데이터, 및 문장 어텐션 데이터를 이미지 생성 모델에 입력하여 오디오 데이터에 대응하는 이미지를 출력할 수 있다.

[0048] 이미지 생성부(140)는, 캡션 데이터를 텍스트 임베딩하여 임베딩 데이터를 생성하고, 임베딩 데이터에 텍스트 최적화부(130)를 통해 생성된 특징 벡터를 적용하여 입력 벡터를 생성할 수 있다. 입력 벡터는, 텍스트 임베딩한 캡션 데이터에 특징 벡터를 곱하여 산출될 수 있다.

[0049] 이미지 생성부(140)는, 상기의 과정을 통해서 생성된 입력 벡터를 이미지 생성 모델에 입력하여 오디오와 대응되는 이미지를 생성할 수 있다.

[0050] 본 개시의 실시예들에 따르면, 이미지 생성 장치(100)는, 오디오를 입력으로 하고 이미지를 출력으로 하도록 학습된 모델이 아닌, 텍스트를 입력으로 하고 이미지를 출력으로 하도록 학습된 모델을 이용할 수 있다.

[0051] 오디오를 입력으로 하고 이미지를 출력으로 하도록 학습된 모델은, 오디오 데이터와 이미지 데이터 사이의 특성의 차이로 인해서, 오디오와 대응되는 이미지를 출력하는데 한계가 있었다. 또한, 오디오를 입력으로 하고 이미지를 출력으로 하도록 학습된 모델은, 오디오와 이미지의 쌍으로 이루어진 훈련 데이터 셋트의 부족으로 인해서 오디오에 대응하는 이미지를 정확하지 않게 출력할 수 있었다.

[0052] 이미지 생성부(140)는, 입력 벡터로 생성된 이미지와 캡션 데이터를 비교하여 제1 비교값을 산출하고, 제1 비교값이 기 설정된 제1 기준 범위를 초과하는 경우, 입력 벡터로 생성된 이미지가 적절하지 않다는 제1 결과값을 산출할 수 있다. 이미지와 캡션 데이터를 비교하여 비교값을 산출하는 것은, 기계 학습된 모델을 이용한 캡션 비교기에 의할 수 있으며, 예시로는, CLIP score를 이용할 수 있다.

[0053] 이미지 생성부(140)는, 입력 벡터로 생성된 이미지와, 오디오 데이터를 비교하여 제2 비교값을 산출하고, 제2 비교값이 기 설정된 제2 기준 범위를 초과하는 경우, 입력 벡터로 생성된 이미지가 적절하지 않다는 제2 결과값을 산출할 수 있다. 오디오 데이터와 이미지를 비교하여 비교값을 산출하는 것은, 기계 학습 모델을 이용한 오디오 비교기에 의할 수 있으며, 예시로는, Audio CLIP score를 이용할 수 있다.

[0054] 이미지 생성부(140)는, 제1 결과값 및/또는 제2 결과값을 기초로 입력 벡터에 대한 최종 결과값을 산출할 수 있다.

[0055] 이미지 생성부(140)는, 최종 결과값이 적절하지 않은 것으로 검출된 경우, 입력 벡터를 다시 생성하고, 다시 생성된 입력 벡터를 이미지 생성 모델에 입력하여 오디오 데이터에 대응하는 이미지를 다시 생성할 수 있다.

[0056] 이미지 생성부(140)는, 최종 결과값에 따라서 대응되는 방법으로 입력 벡터를 다시 생성할 수 있다. 이미지 생성부(140)는, 최종 결과값 중에서, 적절하지 않다는 제1 결과값과 대응하여 설정된 제1 방법으로 입력 벡터를 다시 생성할 수 있다. 이미지 생성부(140)는, 최종 결과값 중에서, 적절하지 않다는 제2 결과값과 대응하여 설정된 제2 방법으로 입력 벡터를 다시 생성할 수 있다.

[0057] 이미지 생성부(140)는, 생성된 이미지를 캡션 데이터 또는 오디오 데이터와 비교하여 이미지의 퀄리티를 평가하는 결과값을 산출하고, 결과값이 기준값에 미치지 못하는 경우, 오디오 데이터에 대응하는 이미지를 다시 생성하는 과정을 반복적으로 수행할 수 있다.

[0058] 도 2는 본 개시의 실시예들에 따른 이미지 생성 장치의 구현의 예시 도면이다.

[0059] 오디오 데이터(Input data)가 오디오 캡션 변환기(Audio Captioning Transformer, M1)에 입력되고, 오디오 어텐션 데이터(D11), 캡션 데이터(D12)가 출력될 수 있다. 캡션 데이터(D12)로부터 문장 어텐션 데이터(D13)가 생성될 수 있다. 문장 어텐션 데이터(D13)는 학습된 명사 추출기를 이용하여 생성될 수 있다.

[0060] 본 개시의 실시예들에 따르면, 오디오 어텐션 데이터(D11)과 캡션 데이터(D12)를 텍스트 임베딩 공간에서 위치적인 인코딩을 수행하여 데이터를 생성하고, 상기의 데이터에, 문장 어텐션 데이터(D13)를 결합시켜, 특징 벡터(D2, Init W)를 출력할 수 있다.

[0061] 본 개시의 실시예들에 따르면, 캡션 데이터를 CLIP Text Encoder(M2)에 입력하여 텍스트 임베딩하여 제1 벡터(Latent z)가 출력될 수 있다. 제1 벡터(Latent z)는 캡션 데이터 만을 포함하는 것으로, 특징 벡터(D2)를 곱하여 제2 벡터(new latent z, D31)로 변환될 수 있다. 제2 벡터(D31)는 오디오에 대한 캡션 데이터에 오디오 어텐션 데이터와 문장 어텐션 데이터의 특징을 포함하는 특징 벡터를 곱한 것일 수 있다.

[0062] 이미지 생성 장치(100)는 제2 벡터(D31)를 이미지 생성 모델(M3)에 입력하여 제1 이미지(Image1)를 생성할 수 있다.

[0063] 이미지 생성 장치(100)는, 캡션 비교기를 이용하여 제1 이미지(Image 1)와 캡션 데이터를 비교하고 오디오 비교기를 이용하여 제1 이미지(Image 1)와 오디오 데이터를 비교할 수 있다.

[0064] 이미지 생성 장치(100)는, 제1 이미지(Image 1)와 캡션 데이터를 비교한 제1 비교값에 대응하는 제1 결과값과, 제1 이미지와 오디오 데이터를 비교한 제2 비교값에 대응하는 제2 결과값 중 적어도 하나를 고려하여 제1 이미지(Image 1)를 다시 생성할 수 있다.

[0065] 이미지 생성 장치(100)는, 제1 결과값과 제2 결과값 중 적어도 하나가 정해진 기준 범위에 미치지 못하는 경우, 이미지를 다시 생성하는 지시 신호를 생성하고, 지시 신호에 대응하여 벡터(D31)를 제1 결과값 또는 제2 결과값에 따라 보완이 필요한 부분에 대해서 변형하여 입력 벡터(D32)를 다시 생성할 수 있다.

[0066] 이미지 생성 장치(100)는, 다시 생성된 입력 벡터(D32)를 이미지 생성 모델에 입력하여 제2 이미지(Iamge2)를 출력할 수 있다.

[0067] 제2 이미지(Iamge2)에 대해서도, 캡션 데이터 및/또는 오디오 데이터와 각각 비교한 결과값을 산출하고, 제2 이미지(Image2)가 기준 범위를 만족하는 결과값을 가지는 경우, 제2 이미지(Iamge2)가 최종 이미지로 출력될 수 있다.

[0068] 도 2에 도시된 바와 같이, 제1 이미지(Image1)에는 아빠의 얼굴이 없는 반면, 제2 이미지(Iamge2)에는 아빠의 얼굴이 표현됨을 알 수 있다. 벡터(D31)에는 오디오 데이터에 포함된 아빠의 소리가 누락됨으로 인한 결과로 볼 수 있다. 본 개시의 실시예들에 따르면, 오디오 데이터에 복수의 주체들의 소리가 포함되는 경우, 복수의 주체들 모두가 누락되지 않고 포함되는 이미지를 생성할 수 있다.

[0069] 도 3은 본 개시의 실시예들에 따라서 생성된 이미지와 종래의 방법으로 생성된 이미지 사이를 비교한 도면이다.

[0070] 제1 오디오 데이터(Input1)가 입력되면 제1 오디오 데이터(Input1)에 대응하는 제1 캡션 데이터(C1)이 출력되고, 제1 캡션 데이터(C1)에 대응하는 벡터를 이미지 생성 모델에 입력하면 제1 비교이미지(O11)이 출력될 수 있다.

[0071] 본 개시의 실시예에 따르면, 이미지 생성 장치(100)는, 제1 오디오 데이터(Input1)에 대응하는 제1 캡션 데이터(C1)를 출력하고, 캡션 데이터(C1)로부터 오디오 어텐션 데이터와 문장 어텐션 데이터를 생성할 수 있다. 이미지 생성 장치(100)는, 캡션 데이터(C1)에 대응하는 벡터와, 제1 캡션 데이터(C1), 오디오 어텐션 데이터와 문장 어텐션 데이터를 결합하여 인코딩한 인코딩 데이터를 곱하여 입력 벡터를 산출할 수 있다. 이미지 생성 장치(100)는, 이렇게 산출된 입력 벡터를 이미지 생성 모델에 입력하여 제1 이미지(O12)를 생성할 수 있다.

[0072] 제2 오디오 데이터(Input2)가 입력되면 제2 오디오 데이터(Input2)에 대응하는 제2 캡션 데이터(C2)이 출력되고, 제2 캡션 데이터(C2)에 대응하는 벡터를 이미지 생성 모델에 입력하면 제2 비교이미지(O21)이 출력될 수 있다.

[0073] 본 개시의 실시예에 따르면, 이미지 생성 장치(100)는, 제2 오디오 데이터(Input2)에 대응하는 제2 캡션 데이터(C2)를 출력하고, 제2 캡션 데이터(C2)로부터 오디오 어텐션 데이터와 문장 어텐션 데이터를 생성할 수 있다. 이미지 생성 장치(100)는, 제2 캡션 데이터(C2)에 대응하는 벡터와, 제2 캡션 데이터(C2), 오디오 어텐션 데이터와 문장 어텐션 데이터를 결합하여 인코딩한 인코딩 데이터를 곱하여 입력 벡터를 산출할 수 있다. 이미지 생성 장치(100)는, 이렇게 산출된 입력 벡터를 이미지 생성 모델에 입력하여 제2 이미지(O22)를 생성할 수 있다.

[0074] 도면에 도시된 바와 같이, 제1 비교이미지(O11)는 두명의 아이가 우는 모습이지만, 제1 이미지(O12)는 아이 대신 엄마의 모습이 나타남을 알 수 있다. 제2 비교이미지(O21)는 고양이만 표현되어 있지만, 제2 이미지(O22)는 천둥과 고양이가 표현됨을 알 수 있다.

[0075] 본 개시의 실시예들에 따르면, 오디오 데이터에 포함된 복수의 객체들 각각이 이미지에 포함되어 표현될 수 있다.

[0076] 도 4는 본 개시의 실시예들에 따른 이미지 생성 방법의 흐름도이다.

[0077] S110에서는, 이미지 생성 장치(100)는, 오디오 데이터를 입력한다. 오디오 데이터는 소정의 방법으로 전처리될 수 있다.

[0078] 이미지 생성 장치(100)는, 오디오 데이터를 시계열의 일정한 시간 구간으로 나누고 각 구간에 대한 스펙트럼을 산출하는 방식으로 전처리할 수 있다. 추가적으로, 텍스트 입력부(110)는, 각 구간에 대한 스펙트럼을 정해진 주파수 단위, 예를 들어 멜 단위(Mel unit)로 변환한 스펙트럼을 생성할 수 있다. 이러한 방식으로 오디오 데이터를 전처리 함으로써, 오디오 데이터의 주파수 특성을 시계열 상에서 확인할 수 있다.

[0079] S120에서는, 이미지 생성 장치(100)는, 오디오 데이터에 대응하는 캡션 데이터를 생성할 수 있다. 이미지 생성 장치(100)는, 오디오 데이터에 대응하여, 캡션 데이터에 포함된 각각의 단어들의 발생 확률을 포함하는 오디오 어텐션 데이터를 생성할 수 있다. 이미지 생성 장치(100)는, 기계 학습된 오디오 변환기에 오디오 데이터를 입력으로 하여 캡션 데이터 및 오디오 어텐션 데이터를 출력할 수 있다.

[0080] 이미지 생성 장치(100)는, 오디오 데이터들과 오디오 데이터에 대응하는 캡션 데이터들로 학습된 오디오 변환기를 이용하여 입력된 오디오 데이터에 대한 캡션 데이터, 오디오 어텐션 데이터를 출력할 수 있다.

[0081] S130에서는, 이미지 생성 장치(100)는, 오디오 데이터에 대응하는 캡션 데이터를 명사 추출기에 입력하여 문장 어텐션 데이터를 출력할 수 있다. 문장 어텐션 데이터는, 소리를 발생시키는 주체에 대한 데이터를 의미할 수 있다. 문장 어텐션 데이터는, 캡션 데이터에서 명사에 해당하는 단어들을 강조하여 표현하는 데이터로서, 예컨대, 명사에 동사보다 더 높은 가중치 값을 부여한 데이터를 말할 수 있다. 문장 어텐션 데이터는, 캡션 데이터의 각 단어가 명사일 확률값을 포함한다.

[0082] S140에서는, 이미지 생성 장치(100)는, 제1 오디오 데이터에 대한, 캡션 데이터와 오디오 어텐션 데이터를 위치적으로 인코딩하고, 인코딩된 데이터에 문장 어텐션 데이터를 결합시켜 오디오에 대한 특징 벡터를 생성할 수 있다.

[0083] S150에서는, 이미지 생성 장치(100)는, 캡션 데이터를 텍스트 임베딩하여 임베딩 데이터를 생성하고, 임베딩 데이터에 텍스트 최적화부(130)를 통해 생성된 특징 벡터를 적용하여 입력 벡터를 생성할 수 있다.

[0084] S160에서는, 이미지 생성 장치(100)는, 상기의 과정을 통해서 생성된 입력 벡터를 이미지 생성 모델에 입력하여 오디오와 대응되는 이미지를 생성할 수 있다.

[0085] 도 5는, 이미지 생성 단계(S160)의 구체적인 흐름도이다.

[0086] S160는 도면에 도시된 바와 같이 복수의 단계들을 포함하여 구현될 수 있다.

[0087] S161에서는, 이미지 생성 장치(100)는, 입력 벡터로 생성된 이미지와 캡션 데이터를 비교하여 제1 비교값을 산출하고, 제1 비교값이 기 설정된 제1 기준 범위를 초과하는 경우, 입력 벡터로 생성된 이미지가 적절하지 않다는 제1 결과값을 산출할 수 있다. 이미지와 캡션 데이터를 비교하여 비교값을 산출하는 것은, 기계 학습된 모델을 이용한 캡션 비교기에 의할 수 있으며, 예시로는, CLIP score를 이용할 수 있다.

[0088] S162에서는, 이미지 생성 장치(100)는, 입력 벡터로 생성된 이미지와, 오디오 데이터를 비교하여 제2 비교값을 산출하고, 제2 비교값이 기 설정된 제2 기준 범위를 초과하는 경우, 입력 벡터로 생성된 이미지가 적절하지 않다는 제2 결과값을 산출할 수 있다. 오디오 데이터와 이미지를 비교하여 비교값을 산출하는 것은, 기계 학습 모델을 이용한 오디오 비교기에 의할 수 있으며, 예시로는, Audio CLIP score를 이용할 수 있다.

[0089] S163에서는, 이미지 생성 장치(100)는, 제1 결과값 및/또는 제2 결과값을 기초로 입력 벡터에 대한 최종 결과값을 산출할 수 있다.

[0090] S164, S166에서는, 이미지 생성 장치(100)는, 최종 결과값이 적절하지 않은 것으로 검출된 경우, 입력 벡터를 다시 생성하고, 다시 생성된 입력 벡터를 이미지 생성 모델에 입력하여 오디오 데이터에 대응하는 이미지를 다시 생성할 수 있다.

[0091] S165에서는, 이미지 생성 장치(100)는, 입력 벡터로 생성된 이미지를 출력할 수 있다. 이미지 생성 장치(100)는, 이미지를 파일로 변환하여 저장할 수 있다. 이미지 생성 장치(100)는, 이미지를 출력 장치를 통해서 디스플레이할 수 있다. 이미지 생성 장치(100)는, 파일로 변환된 이미지를 다른 전자 장치로 전송할 수 있다.

[0092] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령

(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 어플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0093] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(signal wave)에 영구적으로, 또는 일시적으로 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0094] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0095] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

[0096] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

부호의 설명

[0097] 100: 이미지 생성 장치

110: 데이터 입력부

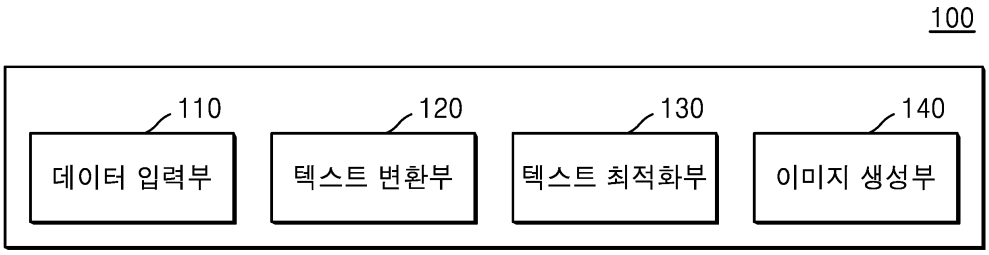
120: 텍스트 변환부

130: 텍스트 최적화부

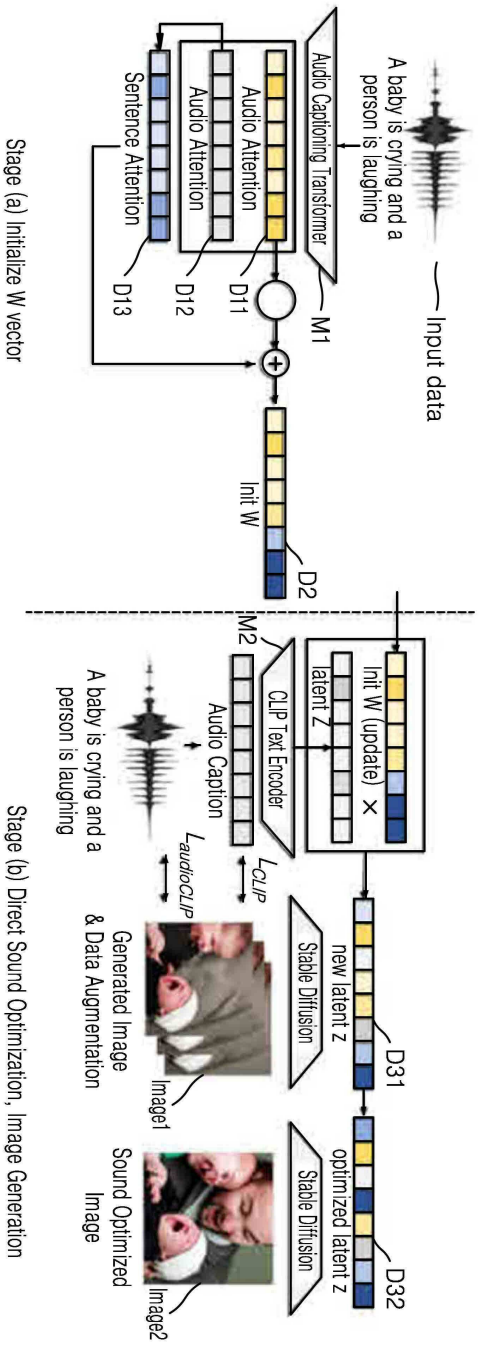
140: 이미지 생성부

도면

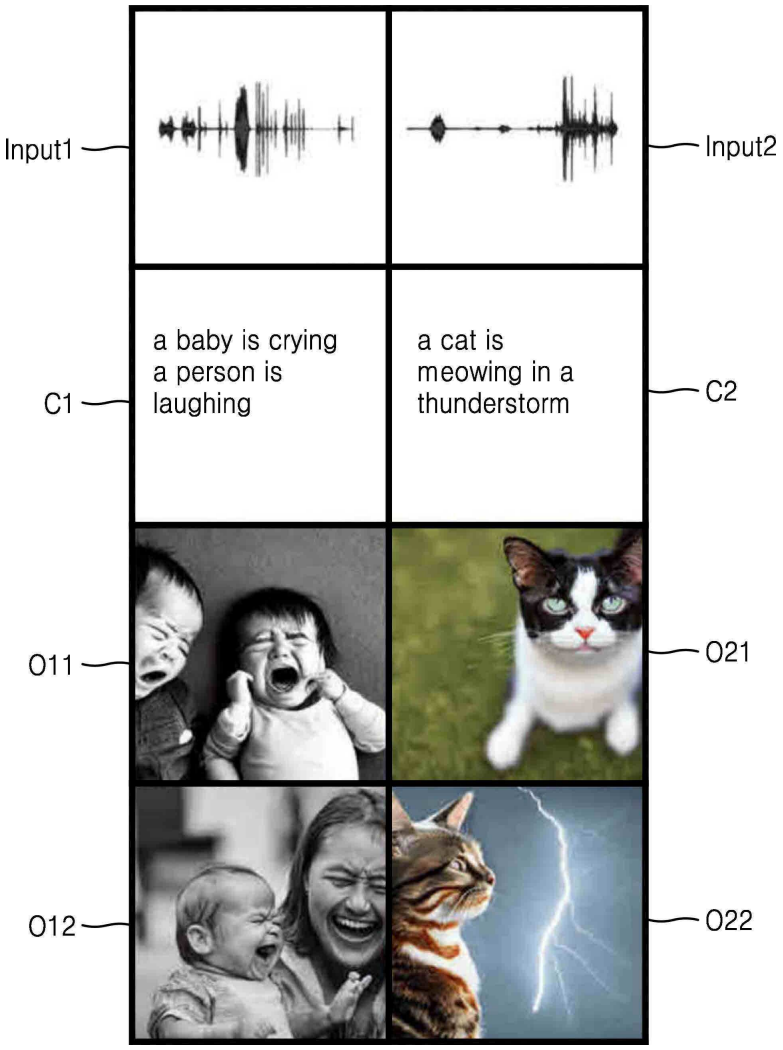
도면1



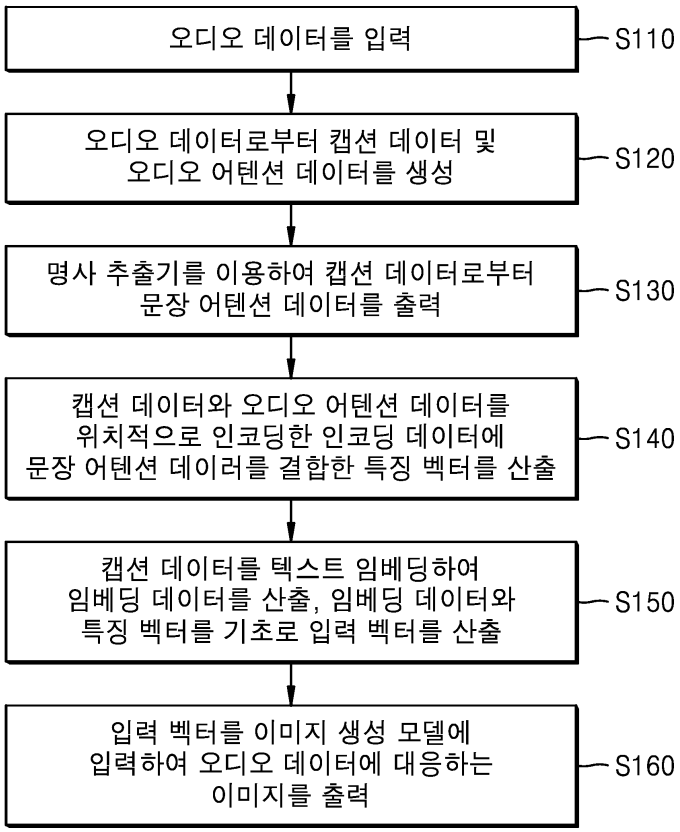
도면2


Input data
A baby is crying and a person is laughing
Audio Captioning Transformer
M1
Audio Attention
Audio Attention
Sentence Attention
D11
D12
D13
Init W
D2
Stage (a) Initialize W vector
M2
CLIP Text Encoder
Audio Caption
A baby is crying and a person is laughing
Init W (update) ×
latent Z
Stable Diffusion
new latent z
D31
Stable Diffusion
optimized latent z
D32
Generated Image & Data Augmentation
Image1
Sound Optimized Image
Image2
Stage (b) Direct Sound Optimization, Image Generation

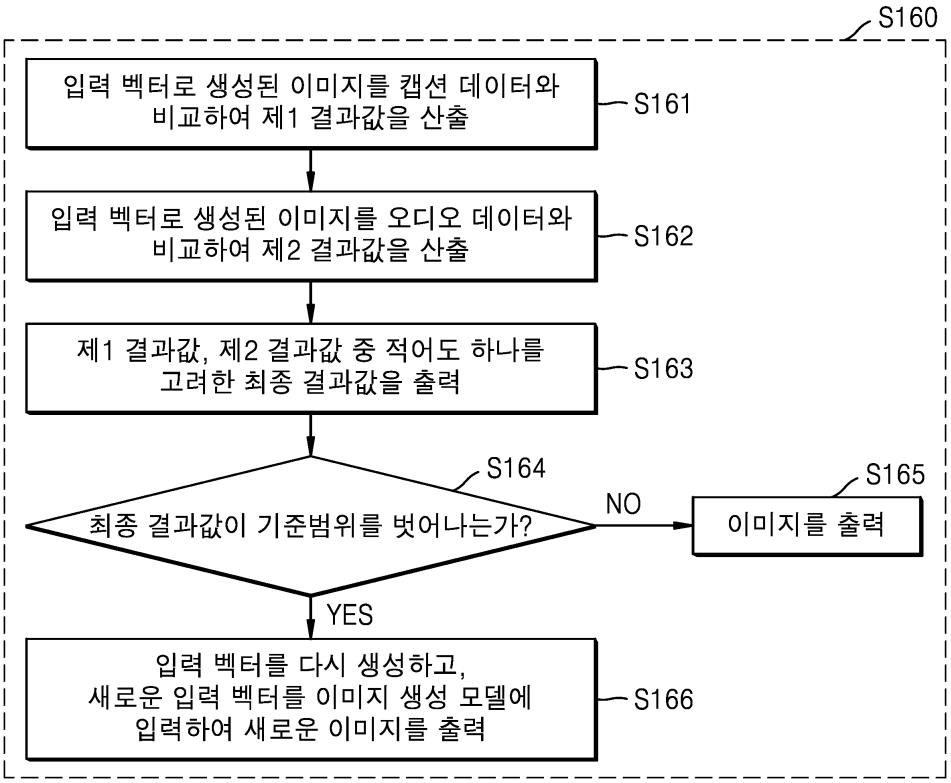
도면3


Input1
Input2
a baby is crying
a person is
laughing
a cat is
meowing in a
thunderstorm
C1
C2
O11
O21
O12
O22

도면4


오디오 데이터를 입력
S110
오디오 데이터로부터 캡션 데이터 및 오디오 어텐션 데이터를 생성
S120
명사 추출기를 이용하여 캡션 데이터로부터 문장 어텐션 데이터를 출력
S130
캡션 데이터와 오디오 어텐션 데이터를 위치적으로 인코딩한 인코딩 데이터에 문장 어텐션 데이터를 결합한 특징 벡터를 산출
S140
캡션 데이터를 텍스트 임베딩하여 임베딩 데이터를 산출, 임베딩 데이터와 특징 벡터를 기초로 입력 벡터를 산출
S150
입력 벡터를 이미지 생성 모델에 입력하여 오디오 데이터에 대응하는 이미지를 출력
S160

도면5


S160
입력 벡터로 생성된 이미지를 캡션 데이터와 비교하여 제1 결과값을 산출
S161
입력 벡터로 생성된 이미지를 오디오 데이터와 비교하여 제2 결과값을 산출
S162
제1 결과값, 제2 결과값 중 적어도 하나를 고려한 최종 결과값을 출력
S163
S164
최종 결과값이 기준범위를 벗어나는가?
NO
S165
이미지를 출력
YES
입력 벡터를 다시 생성하고, 새로운 입력 벡터를 이미지 생성 모델에 입력하여 새로운 이미지를 출력
S166