



(19) 대한민국특허청(KR)  
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0006544  
(43) 공개일자 2025년01월13일

(51) 국제특허분류(Int. Cl.)  
G06T 13/20 (2011.01) G06N 3/0455 (2023.01)  
G06T 3/40 (2024.01)  
(52) CPC특허분류  
G06T 13/20 (2013.01)  
G06N 3/0455 (2023.01)  
(21) 출원번호 10-2023-0086466  
(22) 출원일자 2023년07월04일  
심사청구일자 2023년07월04일

(71) 출원인  
계명대학교 산학협력단  
대구광역시 달서구 달구벌대로 1095, 계명대학교  
산학협력관 201호(신당동)  
(72) 발명자  
고병철  
대구광역시 수성구 범어천로 107  
정찬혁  
대구광역시 달서구 월곡로 320  
(74) 대리인  
김건우

전체 청구항 수 : 총 16 항

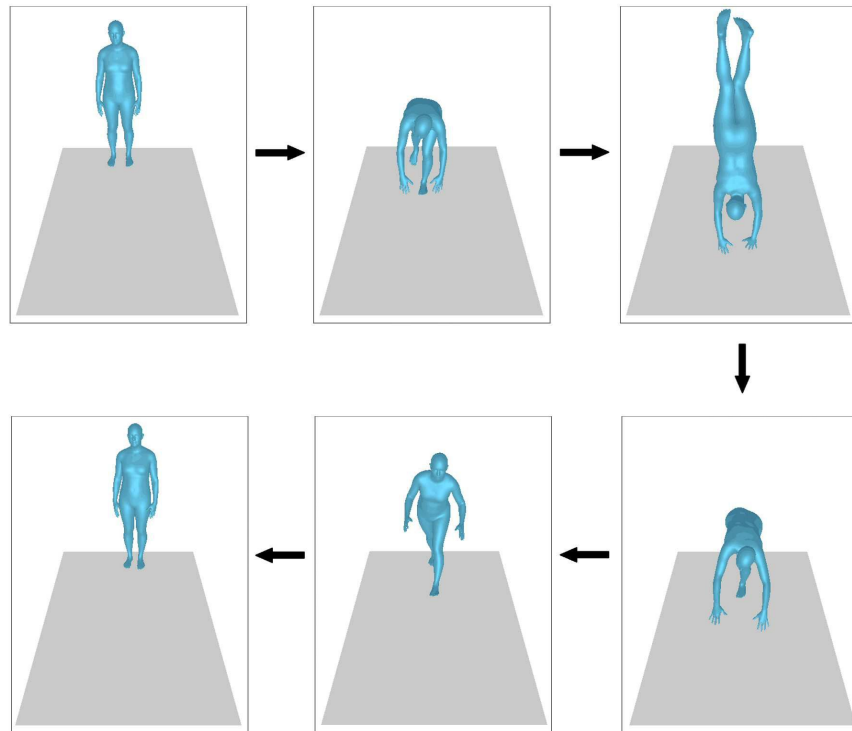
(54) 발명의 명칭 이산화된 잠재 공간을 활용한 모션 생성 장치 및 방법

(57) 요약

본 발명은 이산화된 잠재 공간을 활용한 모션 생성 장치에 관한 것으로서, 보다 구체적으로는 모션 생성 장치로서, 인코더가 만들어진 임베딩을 벡터 양자화하여 이산 토큰으로 임베딩을 만드는 코드북(Codebook)을 포함하여, 상기 코드북을 이용해 연속된 데이터인 모션 시퀀스를 이산화된 잠재 공간으로 학습하는 벡터 양자화-분산 오토

(뒷면에 계속)

대표도 - 도1



인코더(Vector Quantized Variational Autoencoder: VQ-VAE)를 학습하는 제1 학습부; 상기 제1 학습부에서 학습한 상기 VQ-VAE의 인코더로 상기 모션 시퀀스를 토큰화하여 자기회귀 생성 모델을 학습하는 제2 학습부; 및 상기 제2 학습부에서 학습한 상기 자기회귀 생성 모델로 생성한 모션 토큰을 상기 제1 학습부에서 학습한 상기 VQ-VAE의 디코더를 사용해 모션으로 복원하는 모션 생성부를 포함하는 것을 그 구성상의 특징으로 한다.

또한, 본 발명은 이산화된 잠재 공간을 활용한 모션 생성 방법에 관한 것으로서, 보다 구체적으로는 컴퓨터에서 각 단계가 수행되는 모션 생성 방법으로서, (1) 인코더가 만들어낸 임베딩을 벡터 양자화하여 이산 토큰으로 임베딩을 만드는 코드북(Codebook)을 포함하여, 상기 코드북을 이용해 연속된 데이터인 모션 시퀀스를 이산화된 잠재 공간으로 학습하는 벡터 양자화-분산 오토인코더(Vector Quantized Variational Autoencoder: VQ-VAE)를 학습하는 제1 학습 단계; (2) 상기 제1 학습 단계에서 학습한 상기 VQ-VAE의 인코더로 상기 모션 시퀀스를 토큰화하여 자기회귀 생성 모델을 학습하는 제2 학습 단계; 및 (3) 상기 제2 학습 단계에서 학습한 상기 자기회귀 생성 모델로 생성한 모션 토큰을 상기 제1 학습 단계에서 학습한 상기 VQ-VAE의 디코더를 사용해 모션으로 복원하는 모션 생성 단계를 포함하는 것을 그 구성상의 특징으로 한다.

본 발명에서 제안하고 있는 이산화된 잠재 공간을 활용한 모션 생성 장치 및 방법에 따르면, 인코더가 만들어낸 임베딩을 벡터 양자화하여 이산 토큰으로 임베딩을 만드는 코드북(Codebook)을 포함하여, 코드북을 이용해 연속된 데이터인 모션 시퀀스를 이산화된 잠재 공간으로 학습하는 벡터 양자화(Vector Quantized)-분산 오토인코더(VQ-VAE)를 이용해, 모션 시퀀스 데이터에 자기회귀 생성 모델을 적용할 수 있도록 구성함으로써, 빠른 속도로 더 정확하고 다양한 모션을 생성할 수 있다.

(52) CPC특허분류

**G06T 3/40** (2024.01)

G06T 2207/20081 (2013.01)

G06T 2207/20084 (2013.01)

이 발명을 지원한 국가연구개발사업

|             |                                |
|-------------|--------------------------------|
| 과제고유번호      | 1345363225                     |
| 과제번호        | 2022R1I1A3058128               |
| 부처명         | 교육부                            |
| 과제관리(전문)기관명 | 한국연구재단                         |
| 연구사업명       | 지역대학우수과학자지원사업                  |
| 연구과제명       | 자기-지도학습 기반의 경량 시공간 그래프 트랜스포머연구 |
| 기 여 율       | 1/1                            |
| 과제수행기관명     | 계명대학교                          |
| 연구기간        | 2023.03.01 ~ 2024.02.29        |

## 명세서

### 청구범위

#### 청구항 1

모션 생성 장치(100)로서,

인코더(10)가 만들어낸 임베딩을 벡터 양자화하여 이산 토큰으로 임베딩을 만드는 코드북(Codebook)(30)을 포함하여, 상기 코드북(30)을 이용해 연속된 데이터인 모션 시퀀스를 이산화된 잠재 공간으로 학습하는 벡터 양자화-분산 오토인코더(Vector Quantized Variational Autoencoder: VQ-VAE)를 학습하는 제1 학습부(110);

상기 제1 학습부(110)에서 학습한 상기 VQ-VAE의 인코더(10)로 상기 모션 시퀀스를 토큰화하여 자기회귀 생성 모델을 학습하는 제2 학습부(120); 및

상기 제2 학습부(120)에서 학습한 상기 자기회귀 생성 모델로 생성한 모션 토큰을 상기 제1 학습부(110)에서 학습한 상기 VQ-VAE의 디코더(20)를 사용해 모션으로 복원하는 모션 생성부(130)를 포함하는 것을 특징으로 하는, 이산화된 잠재 공간을 활용한 모션 생성 장치(100).

#### 청구항 2

제1항에 있어서, 상기 제1 학습부(110)는,

상기 VQ-VAE를 구성하는 인코더(10), 디코더(20) 및 코드북(30)을 학습하며,

상기 VQ-VAE는, 상기 인코더(10)에서 만든 임베딩을 벡터 양자화를 통해 상기 코드북(30)의 임베딩으로 토큰화하는 것을 특징으로 하는, 이산화된 잠재 공간을 활용한 모션 생성 장치(100).

#### 청구항 3

제1항에 있어서, 상기 인코더(10)는,

상기 모션 시퀀스를 다운샘플링하는 다운샘플링 모듈(11);

상기 다운샘플링된 모션 시퀀스에서 키(key), 쿼리(query) 및 값(value)을 추출하고, 셀프-어텐션(Self-Attention)으로 상기 모션 시퀀스의 프레임 간 특징을 학습하는 어텐션 모듈(12); 및

복수의 신경망 블록을 사용해 모션의 특징을 학습하는 특징 학습 모듈(13)을 포함하는 것을 특징으로 하는, 이산화된 잠재 공간을 활용한 모션 생성 장치(100).

#### 청구항 4

제3항에 있어서, 상기 어텐션 모듈(12)은,

셀프-어텐션을 적용할 때마다 위치 인코딩(position encoding)을 더해주어, 프레임 간 특징을 섞을 때 순서를 고려하는 것을 특징으로 하는, 이산화된 잠재 공간을 활용한 모션 생성 장치(100).

#### 청구항 5

제3항에 있어서, 상기 특징 학습 모듈(13)은,

ResBlock을 사용해 특징을 학습하되, 상기 ResBlock의 컨볼루션 계층(Convolution layer)에 팽창(dilation)을 적용해 필터의 크기를 유지하면서 수용장(receptive field)을 늘려주는 것을 특징으로 하는, 이산화된 잠재 공

간을 활용한 모션 생성 장치(100).

#### 청구항 6

제5항에 있어서, 상기 팽창은,

지수적으로 줄여서 적용하는 것을 특징으로 하는, 이산화된 잠재 공간을 활용한 모션 생성 장치(100).

#### 청구항 7

제1항에 있어서, 상기 제2 학습부(120)는,

텍스트 임베딩을 조건으로 조건부 자기회귀 생성 모델을 학습하여, 새로운 모션을 생성할 수 있도록 학습하는 것을 특징으로 하는, 이산화된 잠재 공간을 활용한 모션 생성 장치(100).

#### 청구항 8

제1항에 있어서, 상기 자기회귀 생성 모델은,

복수의 트랜스포머 디코더 레이어로 구성되는 것을 특징으로 하는, 이산화된 잠재 공간을 활용한 모션 생성 장치(100).

#### 청구항 9

컴퓨터에서 각 단계가 수행되는 모션 생성 방법으로서,

(1) 인코더(10)가 만들어낸 임베딩을 벡터 양자화하여 이산 토큰으로 임베딩을 만드는 코드북(Codebook)(30)을 포함하여, 상기 코드북(30)을 이용해 연속된 데이터인 모션 시퀀스를 이산화된 잠재 공간으로 학습하는 벡터 양자화-분산 오토인코더(Vector Quantized Variational Autoencoder: VQ-VAE)를 학습하는 제1 학습 단계;

(2) 상기 제1 학습 단계에서 학습한 상기 VQ-VAE의 인코더(10)로 상기 모션 시퀀스를 토큰화하여 자기회귀 생성 모델을 학습하는 제2 학습 단계; 및

(3) 상기 제2 학습 단계에서 학습한 상기 자기회귀 생성 모델로 생성한 모션 토큰을 상기 제1 학습 단계에서 학습한 상기 VQ-VAE의 디코더(20)를 사용해 모션으로 복원하는 모션 생성 단계를 포함하는 것을 특징으로 하는, 이산화된 잠재 공간을 활용한 모션 생성 방법.

#### 청구항 10

제9항에 있어서, 상기 제1 학습부(110)는,

상기 VQ-VAE를 구성하는 인코더(10), 디코더(20) 및 코드북(30)을 학습하며,

상기 VQ-VAE는, 상기 인코더(10)에서 만든 임베딩을 벡터 양자화를 통해 상기 코드북(30)의 임베딩으로 토큰화하는 것을 특징으로 하는, 이산화된 잠재 공간을 활용한 모션 생성 방법.

#### 청구항 11

제9항에 있어서, 상기 인코더(10)는,

상기 모션 시퀀스를 다운샘플링하는 다운샘플링 모듈(11);

상기 다운샘플링 된 모션 시퀀스에서 키(key), 쿼리(query) 및 값(value)을 추출하고, 셀프-어텐션(Self-

Attention)으로 상기 모션 시퀀스의 프레임 간 특징을 학습하는 어텐션 모듈(12); 및

복수의 신경망 블록을 사용해 모션의 특징을 학습하는 특징 학습 모듈(13)을 포함하는 것을 특징으로 하는, 이산화된 잠재 공간을 활용한 모션 생성 방법.

#### 청구항 12

제11항에 있어서, 상기 어텐션 모듈(12)은,

셀프-어텐션을 적용할 때마다 위치 인코딩(position encoding)을 더해주어, 프레임 간 특징을 섞을 때 순서를 고려하는 것을 특징으로 하는, 이산화된 잠재 공간을 활용한 모션 생성 방법.

#### 청구항 13

제11항에 있어서, 상기 특징 학습 모듈(13)은,

ResBlock을 사용해 특징을 학습하되, 상기 ResBlock의 컨볼루션 계층(Convolution layer)에 팽창(dilation)을 적용해 필터의 크기를 유지하면서 수용장(receptive field)을 늘려주는 것을 특징으로 하는, 이산화된 잠재 공간을 활용한 모션 생성 방법.

#### 청구항 14

제13항에 있어서, 상기 팽창은,

지수적으로 줄여서 적용하는 것을 특징으로 하는, 이산화된 잠재 공간을 활용한 모션 생성 방법.

#### 청구항 15

제9항에 있어서, 상기 제2 학습 단계에서는,

텍스트 임베딩을 조건으로 조건부 자기회귀 생성 모델을 학습하여, 새로운 모션을 생성할 수 있도록 학습하는 것을 특징으로 하는, 이산화된 잠재 공간을 활용한 모션 생성 방법.

#### 청구항 16

제9항에 있어서, 상기 자기회귀 생성 모델은,

복수의 트랜스포머 디코더 레이어로 구성되는 것을 특징으로 하는, 이산화된 잠재 공간을 활용한 모션 생성 방법.

### 발명의 설명

#### 기술 분야

[0001] 본 발명은 모션 생성 장치 및 방법에 관한 것으로서, 보다 구체적으로는 이산화된 잠재 공간을 활용한 모션 생성 장치 및 방법에 관한 것이다.

#### 배경 기술

[0002] 이 부분에 기술된 내용은 단순히 본 발명의 일 실시예에 대한 배경 정보를 제공할 뿐 종래기술을 구성하는 것은 아니다.

[0004] 도 1은 모션 시퀀스를 예를 들어 도시한 도면이다. 도 1에 도시된 바와 같은 사람 모션 시퀀스를 만들기 위해서는 모션 촬영을 위한 배우 또는 캡처 장비가 필요하다는 단점이 있다. 최근 이러한 문제점을 해결하기 위해 텍스트 및 오디오의 멀티모달 특징을 이용한 조건부 모션 생성을 할 수 있는 기술이 연구되고 있다. 모션 생성 모델은 사용자가 원하는 다양한 모션을 만들 수 있어야 하며, 이러한 모션을 만들기 위해서는 많은 샘플을 생성해야 한다. 그렇기 때문에 모션 생성에서 속도와 다양성이 특히 중요하다.

[0006] 도 2는 기존의 확산 모델을 사용한 모션 생성을 설명하기 위해 도시한 도면이다. 도 2에 도시된 바와 같이, 기존의 모션 생성 모델들은 확산 모델(비특허문헌 0001, 0002)을 사용하여 무작위 노이즈에서 조금씩 노이즈를 제거하여 모션을 생성한다. 확산 모델은 가우시안 분포를 사용하기 때문에 모델이 신체 구조상 나타낼 수 없는 자세를 피하면서 노이즈를 제거해야 한다. 따라서, 모션을 생성하기 위해서는 많은 단계와 시간적 비용이 필요하다.

[0008] 한편, 자기회귀 생성 모델은 다음 토큰을 예측하여 데이터를 생성하기 때문에 확산 모델보다 비교적 적은 비용으로 모션을 생성할 수 있다. 도 3은 기존의 자기회귀 생성 모델의 구조를 도시한 도면이다. 이러한 자기회귀 생성 모델은 카테고리 분포를 예측하기 때문에 이산 잠재 공간이 필요하다. 하지만 모션에서 잠재 공간을 학습하는 기존의 모델들은 일반화를 위해서 연속 잠재 공간을 가지는 분산 오토인코더(Variational Autoencoder: VAE)(비특허문헌 0003)를 주로 사용한다.

[0010] 전술한 배경 기술은 발명자가 본 발명의 도출을 위해 보유하고 있었거나, 본 발명의 도출 과정에서 습득한 기술 정보로서, 반드시 본 발명의 출원 전에 일반 공중에게 공개된 공지 기술이라 할 수는 없다.

## 선행기술문헌

### 비특허문헌

- [0011] (비특허문헌 0001) Tevet, Guy, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model, arXiv preprint arXiv:2209.14916, 2022.
- (비특허문헌 0002) Chen, Xin, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your Commands via Motion Diffusion in Latent Space, arXiv preprint arXiv:2212.04048, 2022.
- (비특허문헌 0003) Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In International Conference on Learning Representations, 2014.
- (비특허문헌 0004) Van Den Oord, Aaron, and Oriol Vinyals. Neural discrete representation learning, Advances in Neural Information Processing Systems 30, 2017.
- (비특허문헌 0005) Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, Advances in Neural Information Processing Systems 30, 2017.
- (비특허문헌 0006) Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In British Machine Vision Conference, 2016.
- (비특허문헌 0007) Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio, arXiv preprint arXiv:1609.03499, 2016.
- (비특허문헌 0008) Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, In Advances in Neural Information Processing Systems, pp. 12449-12460, 2020.
- (비특허문헌 0009) Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini

Agarwal, Girish Sastry et al. Learning transferable visual models from natural language supervision, In International Conference on Machine Learning, pp. 8748-8763, 2021.

(비특허문헌 0010) Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018.

(비특허문헌 0011) Mahmood, Naureen, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes, In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5442-5451, 2019.

(비특허문헌 0012) Guo, Chuan, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions, In Proceedings of the 28th ACM International Conference on Multimedia, pp. 2021-2029, 2020.

(비특허문헌 0013) Guo, Chuan, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5152-5161, 2022.

(비특허문헌 0014) Plappert, Matthias, Christian Mandery, and Tamim Asfour. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks, Robotics and Autonomous Systems 109, pp. 13-26, 2018.

(비특허문헌 0015) Ahuja, Chaitanya, and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting, In 2019 International Conference on 3D Vision, pp. 719-728, IEEE, 2019.

(비특허문헌 0016) Bhattacharya, U., Rewkowski, N., Banerjee, A., Guhan, P., Bera, A., & Manocha, D. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents, In 2021 IEEE Virtual Reality and 3D User Interfaces, pp. 1-10, IEEE, Mar. 2021.

(비특허문헌 0017) Ghosh, Anindita, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions, In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1396-1406, 2021.

(비특허문헌 0018) Petrovich, Mathis, Michael J. Black, and G. J. Varol. TEMOS: Generating diverse human motions from textual descriptions, In Proceedings of the European Conference on Computer Vision, 2022.

(비특허문헌 0019) Guo, Chuan, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5152-5161, 2022.

## 발명의 내용

### 해결하려는 과제

[0012] 본 발명은 기존에 제안된 방법들의 상기와 같은 문제점들을 해결하기 위해 제안된 것으로서, 인코더가 만들어진 임베딩을 벡터 양자화하여 이산 토큰으로 임베딩을 만드는 코드북(Codebook)을 포함하여, 코드북을 이용해 연속된 데이터인 모션 시퀀스를 이산화된 잠재 공간으로 학습하는 벡터 양자화(Vector Quantized)-분산 오토인코더(VQ-VAE)를 이용해, 모션 시퀀스 데이터에 자기회귀 생성 모델을 적용할 수 있도록 구성함으로써, 빠른 속도로 더 정확하고 다양한 모션을 생성할 수 있는, 이산화된 잠재 공간을 활용한 모션 생성 장치 및 방법을 제공하는 것을 그 목적으로 한다.

[0014] 다만, 본 발명이 이루고자 하는 기술적 과제는 상기한 바와 같은 기술적 과제로 한정되지 않으며, 또 다른 기술적 과제들이 존재할 수 있고, 명시적으로 언급하지 않더라도 과제의 해결수단이나 실시 형태로부터 파악될 수

있는 목적이나 효과도 이에 포함됨은 물론이다.

### 과제의 해결 수단

- [0015] 상기한 목적을 달성하기 위한 본 발명의 특징에 따른 이산화된 잠재 공간을 활용한 모션 생성 장치는,
- [0016] 모션 생성 장치로서,
- [0017] 인코더가 만들어낸 임베딩을 벡터 양자화하여 이산 토큰으로 임베딩을 만드는 코드북을 포함하여, 상기 코드북을 이용해 연속된 데이터인 모션 시퀀스를 이산화된 잠재 공간으로 학습하는 벡터 양자화-분산 오토인코더(Vector Quantized Variational Autoencoder: VQ-VAE)를 학습하는 제1 학습부;
- [0018] 상기 제1 학습부에서 학습한 상기 VQ-VAE의 인코더로 상기 모션 시퀀스를 토큰화하여 자기회귀 생성 모델을 학습하는 제2 학습부; 및
- [0019] 상기 제2 학습부에서 학습한 상기 자기회귀 생성 모델로 생성한 모션 토큰을 상기 제1 학습부에서 학습한 상기 VQ-VAE의 디코더를 사용해 모션으로 복원하는 모션 생성부를 포함하는 것을 그 구성상의 특징으로 한다.
- [0021] 바람직하게는, 상기 제1 학습부는,
- [0022] 상기 VQ-VAE를 구성하는 인코더, 디코더 및 코드북을 학습하며,
- [0023] 상기 VQ-VAE는, 상기 인코더에서 만든 임베딩을 벡터 양자화를 통해 상기 코드북의 임베딩으로 토큰화할 수 있다.
- [0025] 바람직하게는, 상기 인코더는,
- [0026] 상기 모션 시퀀스를 다운샘플링하는 다운샘플링 모듈;
- [0027] 상기 다운샘플링 된 모션 시퀀스에서 키(key), 쿼리(query) 및 값(value)을 추출하고, 셀프-어텐션(Self-Attention)으로 상기 모션 시퀀스의 프레임 간 특징을 학습하는 어텐션 모듈; 및
- [0028] 복수의 신경망 블록을 사용해 모션의 특징을 학습하는 특징 학습 모듈을 포함할 수 있다.
- [0030] 더욱 바람직하게는, 상기 어텐션 모듈은,
- [0031] 셀프-어텐션을 적용할 때마다 위치 인코딩(position encoding)을 더해주어, 프레임 간 특징을 섞을 때 순서를 고려할 수 있다.
- [0033] 더욱 바람직하게는, 상기 특징 학습 모듈은,
- [0034] ResBlock을 사용해 특징을 학습하되, 상기 ResBlock의 컨볼루션 계층(Convolution layer)에 팽창(dilation)을 적용해 필터의 크기를 유지하면서 수용장(receptive field)을 늘려줄 수 있다.
- [0036] 더더욱 바람직하게는, 상기 팽창은, 지수적으로 줄여서 적용할 수 있다.
- [0038] 바람직하게는, 상기 제2 학습부는,
- [0039] 텍스트 임베딩을 조건으로 조건부 자기회귀 생성 모델을 학습하여, 새로운 모션을 생성할 수 있도록 학습할 수 있다.
- [0041] 바람직하게는, 상기 자기회귀 생성 모델은,

- [0042] 복수의 트랜스포머 디코더 레이어로 구성될 수 있다.
- [0044] 상기한 목적을 달성하기 위한 본 발명의 특징에 따른 이산화된 잠재 공간을 활용한 모션 생성 방법은,
- [0045] 컴퓨터에서 각 단계가 수행되는 모션 생성 방법으로서,
- [0046] (1) 인코더가 만들어낸 임베딩을 벡터 양자화하여 이산 토큰으로 임베딩을 만드는 코드북을 포함하여, 상기 코드북을 이용해 연속된 데이터인 모션 시퀀스를 이산화된 잠재 공간으로 학습하는 벡터 양자화-분산 오토인코더(VQ-VAE)를 학습하는 제1 학습 단계;
- [0047] (2) 상기 제1 학습 단계에서 학습한 상기 VQ-VAE의 인코더로 상기 모션 시퀀스를 토큰화하여 자기회귀 생성 모델을 학습하는 제2 학습 단계; 및
- [0048] (3) 상기 제2 학습 단계에서 학습한 상기 자기회귀 생성 모델로 생성한 모션 토큰을 상기 제1 학습 단계에서 학습한 상기 VQ-VAE의 디코더를 사용해 모션으로 복원하는 모션 생성 단계를 포함하는 것을 그 구성상의 특징으로 한다.
- [0050] 바람직하게는, 상기 제1 학습부는,
- [0051] 상기 VQ-VAE를 구성하는 인코더, 디코더 및 코드북을 학습하며,
- [0052] 상기 VQ-VAE는, 상기 인코더에서 만든 임베딩을 벡터 양자화를 통해 상기 코드북의 임베딩으로 토큰화할 수 있다.
- [0054] 바람직하게는, 상기 인코더는,
- [0055] 상기 모션 시퀀스를 다운샘플링하는 다운샘플링 모듈;
- [0056] 상기 다운샘플링 된 모션 시퀀스에서 키, 쿼리 및 값을 추출하고, 셀프-어텐션으로 상기 모션 시퀀스의 프레임 간 특징을 학습하는 어텐션 모듈; 및
- [0057] 복수의 신경망 블록을 사용해 모션의 특징을 학습하는 특징 학습 모듈을 포함할 수 있다.
- [0059] 더욱 바람직하게는, 상기 어텐션 모듈은,
- [0060] 셀프-어텐션을 적용할 때마다 위치 인코딩을 더해주어, 프레임 간 특징을 섞을 때 순서를 고려할 수 있다.
- [0062] 더욱 바람직하게는, 상기 특징 학습 모듈은,
- [0063] ResBlock을 사용해 특징을 학습하되, 상기 ResBlock의 컨볼루션 계층에 팽창을 적용해 필터의 크기를 유지하면서 수용장을 늘려줄 수 있다.
- [0065] 더더욱 바람직하게는, 상기 팽창은, 지수적으로 줄여서 적용할 수 있다.
- [0067] 바람직하게는, 상기 제2 학습 단계에서는,
- [0068] 텍스트 임베딩을 조건으로 조건부 자기회귀 생성 모델을 학습하여, 새로운 모션을 생성할 수 있도록 학습할 수 있다.
- [0070] 바람직하게는, 상기 자기회귀 생성 모델은,

[0071] 복수의 트랜스포머 디코더 레이어로 구성될 수 있다.

### 발명의 효과

[0072] 본 발명에서 제안하고 있는 이산화된 잠재 공간을 활용한 모션 생성 장치 및 방법에 따르면, 인코더가 만들어낸 임베딩을 벡터 양자화하여 이산 토큰으로 임베딩을 만드는 코드북을 포함하여, 코드북을 이용해 연속된 데이터인 모션 시퀀스를 이산화된 잠재 공간으로 학습하는 벡터 양자화-분산 오토인코더(VQ-VAE)를 이용해, 모션 시퀀스 데이터에 자기회귀 생성 모델을 적용할 수 있도록 구성함으로써, 빠른 속도로 더 정확하고 다양한 모션을 생성할 수 있다.

[0074] 더불어, 본 발명의 다양하면서도 유익한 장점과 효과는 상술한 내용에 한정되지 않으며, 본 발명의 구체적인 실시 형태를 설명하는 과정에서 보다 쉽게 이해될 수 있을 것이다.

### 도면의 간단한 설명

[0075] 도 1은 모션 시퀀스를 예를 들어 도시한 도면.

도 2는 기존의 확산 모델을 사용한 모션 생성을 설명하기 위해 도시한 도면.

도 3은 기존의 자기회귀 생성 모델의 구조를 도시한 도면.

도 4는 본 발명의 일실시예에 따른 이산화된 잠재 공간을 활용한 모션 생성 장치의 구성을 도시한 도면.

도 5는 본 발명의 일실시예에 따른 이산화된 잠재 공간을 활용한 모션 생성 장치에서 사용하는 VQ-VAE의 인코더 구성을 도시한 도면.

도 6은 본 발명의 일실시예에 따른 이산화된 잠재 공간을 활용한 모션 생성 장치에서 사용하는 VQ-VAE의 디코더 구성을 도시한 도면.

도 7은 본 발명의 일실시예에 따른 이산화된 잠재 공간을 활용한 모션 생성 장치에서 사용하는 VQ-VAE의 인코더-디코더 구조를 도시한 도면.

도 8은 본 발명의 일실시예에 따른 이산화된 잠재 공간을 활용한 모션 생성 장치에서 사용하는 자기회귀 생성 모델의 구조를 도시한 도면.

도 9는 본 발명의 일실시예에 따른 이산화된 잠재 공간을 활용한 모션 생성 방법의 흐름을 도시한 도면.

도 10은 본 발명의 일실시예에 따른 이산화된 잠재 공간을 활용한 모션 생성 장치 및 방법의 성능을 HumanML3D 데이터셋에서 기존 방법들과 비교하여 나타낸 도면.

### 발명을 실시하기 위한 구체적인 내용

[0076] 아래에서는 첨부한 도면을 참조하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 본 발명의 실시예를 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.

[0078] 명세서 전체에서, 어떤 부분이 다른 부분과 "연결"되어 있다고 할 때, 이는 "직접적으로 연결"되어 있는 경우뿐 아니라, 그 중간에 다른 소자를 사이에 두고 "간접적으로 연결"되어 있는 경우도 포함한다. 또한 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 의미하며, 하나 또는 그 이상의 다른 특징이나 숫자, 단계, 동작, 구성요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0080] 이하의 실시예는 본 발명의 이해를 돕기 위한 상세한 설명이며, 본 발명의 권리 범위를 제한하는 것이 아니다.

따라서 본 발명과 동일한 기능을 수행하는 동일 범위의 발명 역시 본 발명의 권리 범위에 속할 것이다.

- [0082] 또한, 본 발명의 각 실시예에 포함된 각 구성, 과정, 공정 또는 방법 등은 기술적으로 상호간 모순되지 않는 범위 내에서 공유될 수 있다.
- [0084] 또한, 본 발명에 있어서 단말, 장치 또는 디바이스가 수행하는 것으로 기술된 동작이나 기능 중 일부는 해당 단말, 장치 또는 디바이스와 연결된 서버에서 대신 수행될 수 있다. 마찬가지로, 서버가 수행하는 것으로 기술된 동작이나 기능 중 일부도 해당 서버와 연결된 단말, 장치 또는 디바이스에서 수행될 수도 있다.
- [0086] 특히, 본 발명의 각 실시예에 따른 시스템을 실행시키기 위한 수단으로는 애플리케이션(Application), 또는 웹 서버일 수 있으며, 이 애플리케이션, 또는 웹 서버를 기록한 기록매체를 읽을 수 있는 수단인 단말로는, 일반적인 데스크톱이나 노트북 등의 일반 PC뿐만 아니라, 스마트 폰, 태블릿 PC, 등의 모바일 단말기를 포함할 수 있다.
- [0088] 이하, 첨부된 도면을 참고하여 본 발명의 실시예들을 상세히 설명하도록 한다.
- [0090] 도 4는 본 발명의 일실시예에 따른 이산화된 잠재 공간을 활용한 모션 생성 장치(100)의 구성을 도시한 도면이다. 도 4에 도시된 바와 같이, 본 발명의 일실시예에 따른 이산화된 잠재 공간을 활용한 모션 생성 장치(100)는, 모션 생성 장치(100)로서, 인코더(10)가 만들어진 임베딩을 벡터 양자화하여 이산 토큰으로 임베딩을 만드는 코드북(30)을 포함하여, 코드북(30)을 이용해 연속된 데이터인 모션 시퀀스를 이산화된 잠재 공간으로 학습하는 벡터 양자화-분산 오토인코더(VQ-VAE)를 학습하는 제1 학습부(110); 제1 학습부(110)에서 학습한 VQ-VAE의 인코더(10)로 모션 시퀀스를 토큰화하여 자기회귀 생성 모델을 학습하는 제2 학습부(120); 및 제2 학습부(120)에서 학습한 자기회귀 생성 모델로 생성한 모션 토큰을 제1 학습부(110)에서 학습한 VQ-VAE의 디코더(20)를 사용해 모션으로 복원하는 모션 생성부(130)를 포함하여 구성될 수 있다.
- [0092] 즉, 본 발명에서는 벡터 양자화를 사용하여 모션 임베딩을 토큰화한 후 모션 토큰을 생성하는 자기회귀 생성 모델을 제안한다. 학습 과정은 크게 두 단계로 나눌 수 있는데, 첫 번째 단계는 제1 학습부(110), 두 번째 단계는 제2 학습부(120)에서 각각 처리될 수 있다. 먼저, 제1 학습부(110)는, 모션 시퀀스를 벡터 양자화로 토큰화할 수 있는 인코더, 디코더와 코드북을 학습할 수 있다. 코드북  $E=\{e_1, \dots, e_K\}$ ,  $e_k \in \mathbb{R}^D$ 은 차원이 D인 K개의 벡터를 가지고 있으며 VAE의 인코더가 만들어진 임베딩을 벡터 양자화하여 이산 토큰으로 임베딩을 만들 수 있다. 그 다음, 제2 학습부(120)는, 학습한 인코더(10)로 모션 시퀀스를 토큰화하여 자기회귀 생성 모델을 학습할 수 있다. 이때, 텍스트 임베딩을 조건으로 조건부 자기회귀 생성 모델을 학습하면 새로운 토큰에 대한 모션도 생성할 수 있다. 제1 학습부(110) 및 제2 학습부(120)에 의한 두 단계의 학습이 완료되면, 모션 생성부(130)는 자기회귀 생성 모델로 생성한 모션 토큰들을 디코더를 사용하여 다시 모션으로 복원할 수 있다.
- [0094] 이하에서는, 도 4를 참조하여 본 발명의 일실시예에 따른 이산화된 잠재 공간을 활용한 모션 생성 장치(100)의 각 구성에 대해 상세히 설명하도록 한다.
- [0096] 제1 학습부(110)는, 인코더가 만들어진 임베딩을 벡터 양자화하여 이산 토큰으로 임베딩을 만드는 코드북을 포함하여, 코드북을 이용해 연속된 데이터인 모션 시퀀스를 이산화된 잠재 공간으로 학습하는 벡터 양자화-분산 오토인코더(VQ-VAE)를 학습할 수 있다. 보다 구체적으로, 제1 학습부(110)는, VQ-VAE를 구성하는 인코더, 디코더 및 코드북을 학습하며, VQ-VAE는 인코더에서 만든 임베딩을 벡터 양자화를 통해 코드북의 임베딩으로 토큰화할 수 있다.

[0098] 여기서는, 먼저 토큰 임베딩에 대해 설명하기로 한다.

[0100] 일반적으로 사용되는 모션 VAE에서는 입력  $x$ 를 받아 이산 표현 벡터 (또는 잠재 공간, latent space)  $z$ 를 출력한다. 잠재 공간  $z \sim q(z)$ 는 가우시안 분포로 표현되며 데이터  $x$ 를 조건으로 하는 조건부 확률분포  $q(z|x)$ 로 데이터를 확률적으로 인코딩한다. 가우시안 분포는 평균과 분산으로 나타낼 수 있기 때문에 VAE의 인코더는  $q(z|x)$ 의 평균과 분산을 예측하며 디코더는  $q(z|x)$ 에서 샘플링한  $z$ 로  $x$ 를 복원한다.

[0102] 본 발명의 VQ-VAE에서는 샘플링 과정 없이 인코더에서 만든 임베딩을 벡터 양자화를 통해 코드북에 있는 임베딩으로 토큰화할 수 있다. VQ-VAE의 사후 확률  $q(z|x)$ 는 양자화되기 때문에 카테고리 분포로 다음 수학적 식 1과 같이 정의할 수 있다.

### 수학적 식 1

$$q(z = k|x) = \begin{cases} 1, & \text{for } k = \operatorname{argmin}_j \|z - e_j\|_2 \\ 0, & \text{otherwise} \end{cases}$$

[0106] 여기서, 확률분포가 직접 정의되었기 때문에 샘플링 과정 없이 계산이 가능하다. 인코더  $f_e$ 를 양자화되지 않은 인코더라 하면 양자화된 인코더  $f_q$ 는 다음 수학적 식 2와 같이 정의할 수 있다.

### 수학적 식 2

$$f_q(x) = e_k, \quad k = \operatorname{argmin}_j \|f_e(x) - e_j\|_2$$

[0110] 수학적 식 2와 같이, 인코더  $f_e$ 로 인코딩된 임베딩  $f_e(x)$ 는 코드북의 임베딩  $\{e_j | 1 \leq j \leq K\}$  중 임베딩  $f_e(x)$ 와 가장 가까운  $k$ 번째 임베딩  $e_k$ 로 양자화될 수 있다. 양자화된 임베딩은 코드북에 있는  $K$ 개의 벡터밖에 표현하지 못하기 때문에 일반적인 가우시안 분포에 비해서 표현력이 매우 제한적이다. 그러나  $K$ 개의 벡터를 중복을 허용하여 조합하면 비교적 다양한 모션을 표현할 수 있다. 또한, 시퀀스마다 임베딩을 공유하므로 학습 데이터보다 긴 모션도 표현할 수 있으며 시간과 독립적으로 모션을 표현할 수 있다.

[0112] 도 5는 본 발명의 일실시예에 따른 이산화된 잠재 공간을 활용한 모션 생성 장치(100)에서 사용하는 VQ-VAE의 인코더(10) 구성을 도시한 도면이고, 도 6은 본 발명의 일실시예에 따른 이산화된 잠재 공간을 활용한 모션 생성 장치(100)에서 사용하는 VQ-VAE의 디코더(20) 구성을 도시한 도면이고, 도 7은 본 발명의 일실시예에 따른 이산화된 잠재 공간을 활용한 모션 생성 장치(100)에서 사용하는 VQ-VAE의 인코더(10)-디코더(20) 구조를 도시한 도면이다.

[0114] 도 5에 도시된 바와 같이, 본 발명의 일실시예에 따른 이산화된 잠재 공간을 활용한 모션 생성 장치(100)에서 사용하는 VQ-VAE의 인코더(10)는, 모션 시퀀스를 다운샘플링하는 다운샘플링 모듈(11); 다운샘플링된 모션 시

퀀스에서 키, 쿼리 및 값 추출하고, 셀프-어텐션으로 모션 시퀀스의 프레임 간 특징을 학습하는 어텐션 모듈(12); 및 복수의 신경망 블록을 사용해 모션의 특징을 학습하는 특징 학습 모듈(13)을 포함하여 구성될 수 있다. 도 6 및 도 7에 도시된 바와 같이, 디코더(20) 모듈은 인코더(10)의 다운샘플링 모듈(11) 대신 업샘플링 모듈(21)을 포함하며, 어텐션 모듈(22) 및 특징 학습 모듈(23)을 포함해 인코더(10)와 반대 방향으로 구성될 수 있다.

[0116] 인코더(10)와 디코더(20)의 전체적인 구조는 도 7에 도시된 바와 같다. 인코더(10)와 디코더(20)는 T개의 모션 벡터를 T/M개의 임베딩으로 인코딩하고 디코딩할 수 있다. 여기서, T는 모션 벡터의 수, M은 다운샘플링 스테이지 수이다. 각 임베딩은 벡터 양자화를 사용하여 코드북(30)에 있는 임베딩 또는 토큰들로 양자화될 수 있다.

[0118] 인코더(10)는 T개의 모션 벡터를 T/M개의 임베딩으로 압축하기 위해 M개의 다운샘플링과 업샘플링 스테이지를 사용한다. 인코더(10)의 다운샘플링 모듈(11)은, 이전 층의 길이가 t일 때 t/2로 다운샘플링할 수 있다.

[0120] 인코더(10)의 어텐션 모듈(12)은, 다운샘플링 된 시퀀스의 키, 쿼리, 값을 컨볼루션 계층을 사용하여 만들고, 셀프-어텐션(비특허문헌 0005)으로 시퀀스의 프레임 간 특징을 학습할 수 있다. 다음 층의 컨볼루션 계층들은 필터 범위 밖에 있는 특징들도 고려할 수 있게 된다. 어텐션 모듈(12)은, 셀프-어텐션을 적용할 때마다 위치 인코딩을 더해주어, 프레임 간 특징을 섞을 때 순서를 고려할 수 있다. 위치 인코딩을 더해줌으로써 프레임간 특징을 섞을 때 순서를 유지해 줄 수 있게 된다.

[0122] 셀프 어텐션을 적용한 다음에는 N개의 ResBlock(비특허문헌 0006)을 사용하여 모션의 특징을 학습할 수 있다. 보다 구체적으로, 특징 학습 모듈(13)은, ResBlock을 사용해 특징을 학습하되, ResBlock의 컨볼루션 계층에 팽창을 적용해 필터의 크기를 유지하면서 수용장을 늘려줄 수 있다. 모션은 컨볼루션 계층의 필터 크기에 비해 훨씬 길기 때문에 오디오 모델에 자주 사용되는 CNN(비특허문헌 0007, 0008)과 같이 ResBlock의 컨볼루션 계층에 팽창을 사용하여 필터의 크기를 유지하면서 수용장을 늘려줄 수 있다. 오디오 데이터에서는 빠르게 변하는 주파수를 학습하기 위해서 팽창을 {1, 3, 9, ...}로 지수적으로 늘려주지만, 사람의 움직임은 비교적 느리게 변하기 때문에 {..., 9, 3, 1}로 지수적으로 줄여서 사용할 수 있다.

[0124] 도 6 및 도 7에 도시된 바와 같이, 디코더(20)의 모델 구조는 인코더(10)의 다운샘플링 모듈(11) 대신 업샘플링 모듈(21)을 사용할 수 있다.

[0126] 제2 학습부(120)는, 제1 학습부(110)에서 학습한 VQ-VAE의 인코더(10)로 모션 시퀀스를 토큰화하여 자기회귀 생성 모델을 학습할 수 있다. 이때, 제2 학습부(120)는, 텍스트 임베딩을 조건으로 조건부 자기회귀 생성 모델을 학습하여, 새로운 모션을 생성할 수 있도록 학습할 수 있다. 즉, 제2 학습부(120)는, 학습된 잠재 시퀀스를 사용하여 자기회귀 모션 생성 모델을 학습시켜 모션을 생성할 수 있다. 학습 과정에서 초기 토큰을 CLIP(비특허문헌 0009)으로 만든 조건 임베딩으로 사용하면 조건부 모션을 생성할 수 있다.

[0128] 도 8은 본 발명의 일실시예에 따른 이산화된 잠재 공간을 활용한 모션 생성 장치(100)에서 사용하는 자기회귀 생성 모델의 구조를 도시한 도면이다. 도 8에 도시된 바와 같이, 본 발명의 일실시예에 따른 이산화된 잠재 공간을 활용한 모션 생성 장치(100)에서 사용하는 자기회귀 생성 모델은, 복수의 트랜스포머 디코더 레이어로 구성될 수 있다. 본 발명의 자기회귀 생성 모델의 구조는 GPT(비특허문헌 0010) 모델과 유사하게 복수의 트랜스포머(비특허문헌 0005) 디코더 레이어를 사용할 수 있다. 보다 구체적으로는, 18개의 트랜스포머 디코더 레이어를 사용하고, 각 레이어에서는 12개의 헤드를 사용하며 레이어의 차원수는 512를 사용할 수 있다.

- [0130] 모션 생성부(130)는, 제2 학습부(120)에서 학습한 자기회귀 생성 모델로 생성한 모션 토큰을 제1 학습부(110)에서 학습한 VQ-VAE의 디코더(20)를 사용해 모션으로 복원할 수 있다.
- [0132] 도 9는 본 발명의 일실시예에 따른 이산화된 잠재 공간을 활용한 모션 생성 방법의 흐름을 도시한 도면이다. 도 9에 도시된 바와 같이, 본 발명의 일실시예에 따른 이산화된 잠재 공간을 활용한 모션 생성 방법은, 인코더(10)가 만들어낸 임베딩을 벡터 양자화하여 이산 토큰으로 임베딩을 만드는 코드북(30)을 포함하여, 코드북(30)을 이용해 연속된 데이터인 모션 시퀀스를 이산화된 잠재 공간으로 학습하는 벡터 양자화-분산 오토인코더(VQ-VAE)를 학습하는 제1 학습 단계(S110), 제1 학습 단계에서 학습한 VQ-VAE의 인코더(10)로 모션 시퀀스를 토큰화하여 자기회귀 생성 모델을 학습하는 제2 학습 단계(S120) 및 제2 학습 단계에서 학습한 자기회귀 생성 모델로 생성한 모션 토큰을 제1 학습 단계에서 학습한 VQ-VAE의 디코더(20)를 사용해 모션으로 복원하는 모션 생성 단계(S130)를 포함하여 구현될 수 있다.
- [0134] 각각의 단계들과 관련된 상세한 내용들은, 앞서 본 발명의 일실시예에 따른 이산화된 잠재 공간을 활용한 모션 생성 장치(100)와 관련하여 충분히 설명되었으므로, 상세한 설명은 생략하기로 한다.
- [0136] 실험
- [0137] 데이터셋
- [0138] 실험은 AMASS 데이터셋(비특허문헌 0011)과 HumanAct12 데이터셋(비특허문헌 0012)을 결합하고 텍스트 레이블을 추가한 HumanML3D(비특허문헌 0013)를 활용하여 수행한다. HumanML3D 데이터셋은 총14,616개의 모션 샘플과 44,970개의 텍스트 레이블을 포함한다. 하나의 모션 시퀀스는 여러 개의 텍스트 레이블을 포함하며, 각 레이블은 모션의 다른 부분에 대한 설명을 포함한다.
- [0140] 모션 시퀀스에서 64프레임을 자르고 무작위로 샘플링하여 옮기고 모션 시퀀스의 왼쪽과 오른쪽을 뒤집는 등의 데이터 증강 기법을 적용했다.
- [0142] 비교 실험
- [0143] 실험은 HumanML3D(비특허문헌 0013)에서 제안한 방법으로 진행한다. 즉, 사전 학습된 모델을 사용하여 추출한 임베딩으로 생성한 데이터의 가능도(likelihood)와 다양성을 계산하고 다른 모델과 비교 및 분석한다. 본 실험에서는 하나의 텍스트에 대해 여러 모션을 생성하여 텍스트 조건부 다양성을 실험했다.
- [0145] 도 10은 본 발명의 일실시예에 따른 이산화된 잠재 공간을 활용한 모션 생성 장치(100) 및 방법의 성능을 HumanML3D 데이터셋에서 기존 방법들과 비교하여 나타낸 도면이다. 도면에서 방법 옆에 기재된 “[숫자]”는 비특허문헌의 번호를 나타낸다. 제안하는 모델은 모션을 압축하고 다시 원래의 모션을 복원하는 재복원 단계와 텍스트 레이블을 조건으로 주고 모션을 생성하는 두 가지 단계로 나뉘어 학습된다. 두 단계를 구분하기 위해 재복원 실험의 경우 Ours(recons)로 표현한다. 테스트 데이터를 재복원하는 실험에서는 테스트 데이터와 유사하게 복원하는 것을 확인할 수 있다. 특히 R-precision와 MM-Dist에서 다른 기존 모델들과 비교하였을 때 제안하는 모델(Ours(recons))이 실제 데이터와 가장 유사한 값을 보이는 것을 확인할 수 있다. 또한, 텍스트 레이블을 조건으로 주고 모션을 생성했을 때도 우수한 성능을 보인다(Ours). R-precision 과 MM-Dist 그리고 FID에서 가장 우수한 성능을 보임을 입증했다. 따라서, 제안하는 모델은 실제 모션과 유사하며 다양한 데이터를 생성할 수 있다.
- [0147] 본 발명에서는 모션 임베딩을 이산적으로 학습하는 새로운 인코더(10)와 디코더(20) 모델과 자기회귀 생성 모델을 제안하였다. 인코딩과 디코딩 과정에서 팽창 컨볼루션(Dilated Convolution)을 사용하여 큰 움직임을 학습

하고 멀티-헤드(Multi-Head) 셀프-어텐션으로 미세한 움직임 학습할 수 있었다. 학습한 인코더(10)와 디코더(20)를 통해 모션 시퀀스를 토큰화하여 텍스트 임베딩을 조건으로 하는 자기회귀 모션 생성 모델을 학습할 수 있었다. 실험을 통해 제안하는 모델이 기존의 최신 모델들과 비교했을 때 더 정확하게 모션을 생성하면서도 다양성을 유지할 수 있음을 확인했다. 또한, 재복원에서는 실제 데이터를 잘 복원할 수 있음을 실험을 통해 입증하였으며, 텍스트 조건이 주어졌을 때 설명에 맞는 모션 임베딩을 생성할 수 있음을 증명했다.

[0149] 전술한 바와 같이, 본 발명에서 제안하고 있는 이산화된 잠재 공간을 활용한 모션 생성 장치(100) 및 방법에 따르면, 인코더(10)가 만들어낸 임베딩을 벡터 양자화하여 이산 토큰으로 임베딩을 만드는 코드북(30)(Codebook)을 포함하여, 코드북(30)을 이용해 연속된 데이터인 모션 시퀀스를 이산화된 잠재 공간으로 학습하는 벡터 양자화(Vector Quantized)-분산 오토인코더(VQ-VAE)를 이용해, 모션 시퀀스 데이터에 자기회귀 생성 모델을 적용할 수 있도록 구성함으로써, 빠른 속도로 더 정확하고 다양한 모션을 생성할 수 있다.

[0151] 한편, 본 발명은 다양한 통신 단말기로 구현되는 동작을 수행하기 위한 프로그램 명령을 포함하는 컴퓨터에서 판독 가능한 매체를 포함할 수 있다. 예를 들어, 컴퓨터에서 판독 가능한 매체는, 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 장치를 포함할 수 있다.

[0153] 이와 같은 컴퓨터에서 판독 가능한 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 이때, 컴퓨터에서 판독 가능한 매체에 기록되는 프로그램 명령은 본 발명을 구현하기 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 예를 들어, 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해 실행될 수 있는 고급 언어 코드를 포함할 수 있다.

[0155] 전술한 본 발명의 설명은 예시를 위한 것이며, 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 본 발명의 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 설명된 각 구성요소는 분산되어 실시될 수도 있으며, 마찬가지로 분산된 것으로 설명된 구성 요소들도 결합된 형태로 실시될 수 있다.

[0157] 본 발명의 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.

## 부호의 설명

[0158] 10: 인코더  
11: 다운샘플링 모듈  
12, 22: 어텐션 모듈  
13, 23: 특징 학습 모듈  
20: 디코더  
21: 업샘플링 모듈  
30: 코드북  
100: 모션 생성 장치

110: 제1 학습부

120: 제2 학습부

130: 모션 생성부

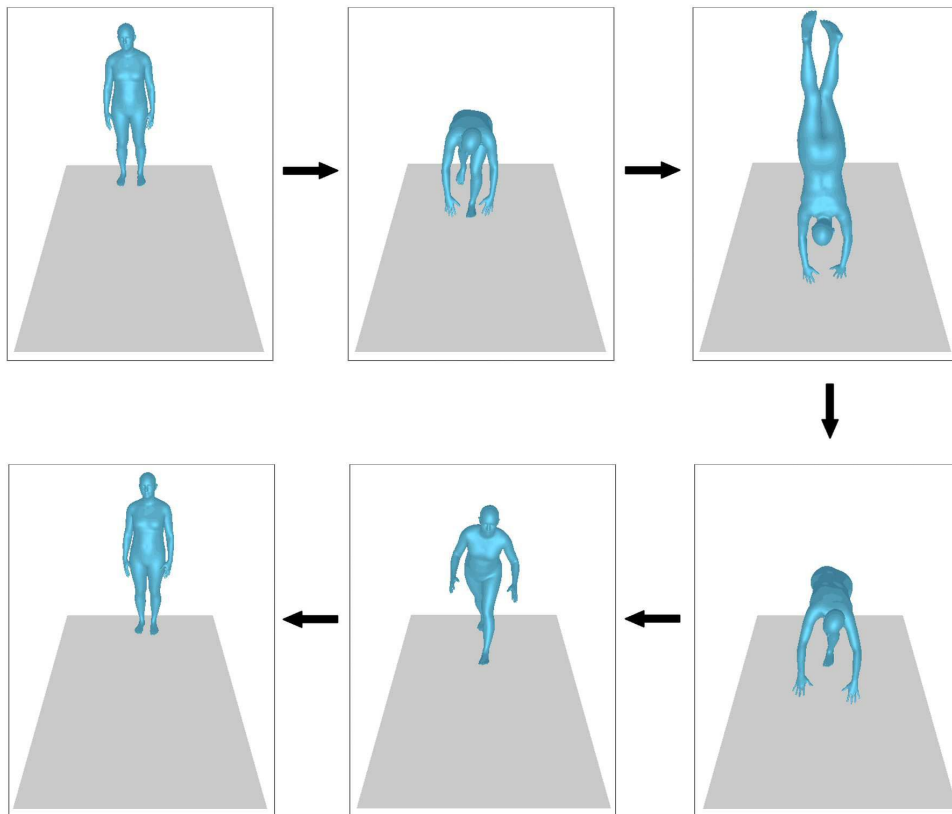
S110: 제1 학습 단계

S120: 제2 학습 단계

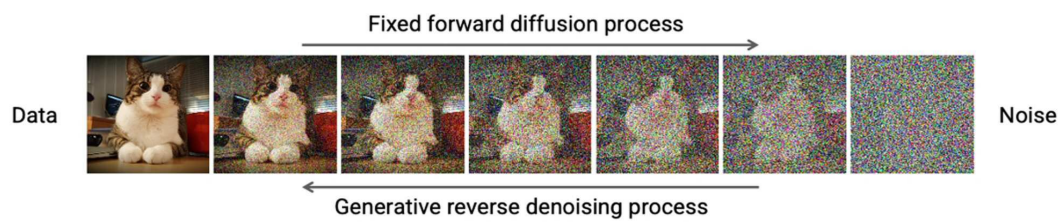
S130: 모션 생성 단계

## 도면

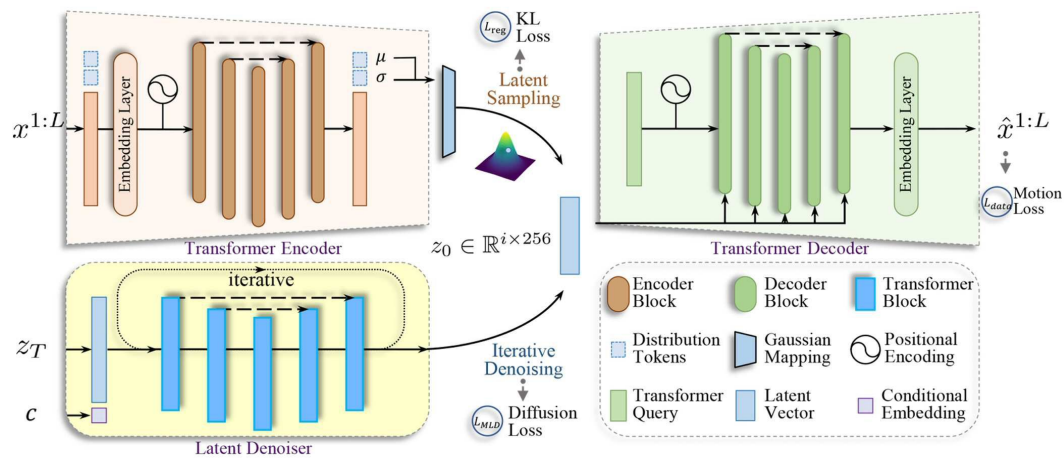
### 도면1



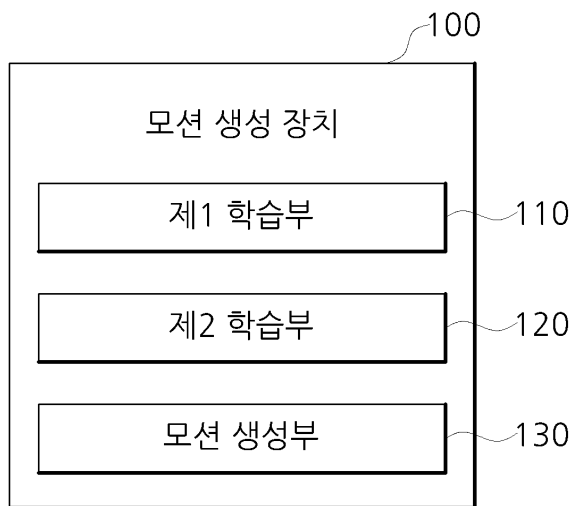
### 도면2



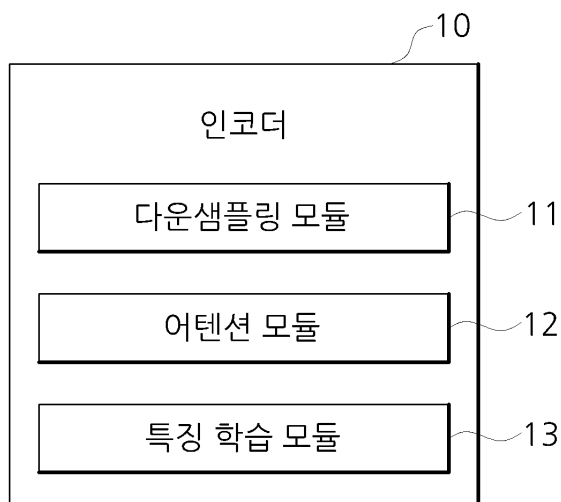
도면3



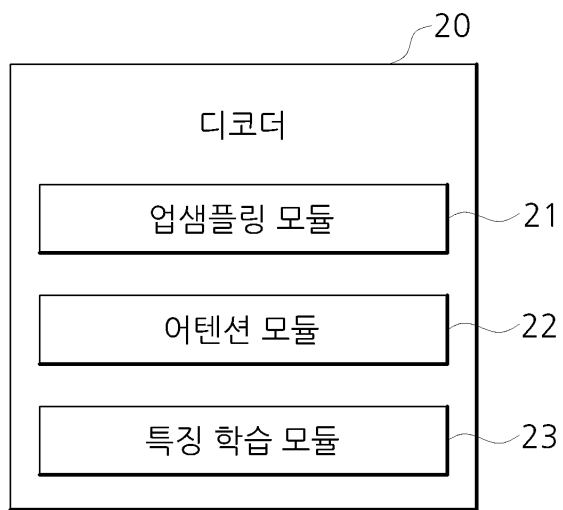
도면4



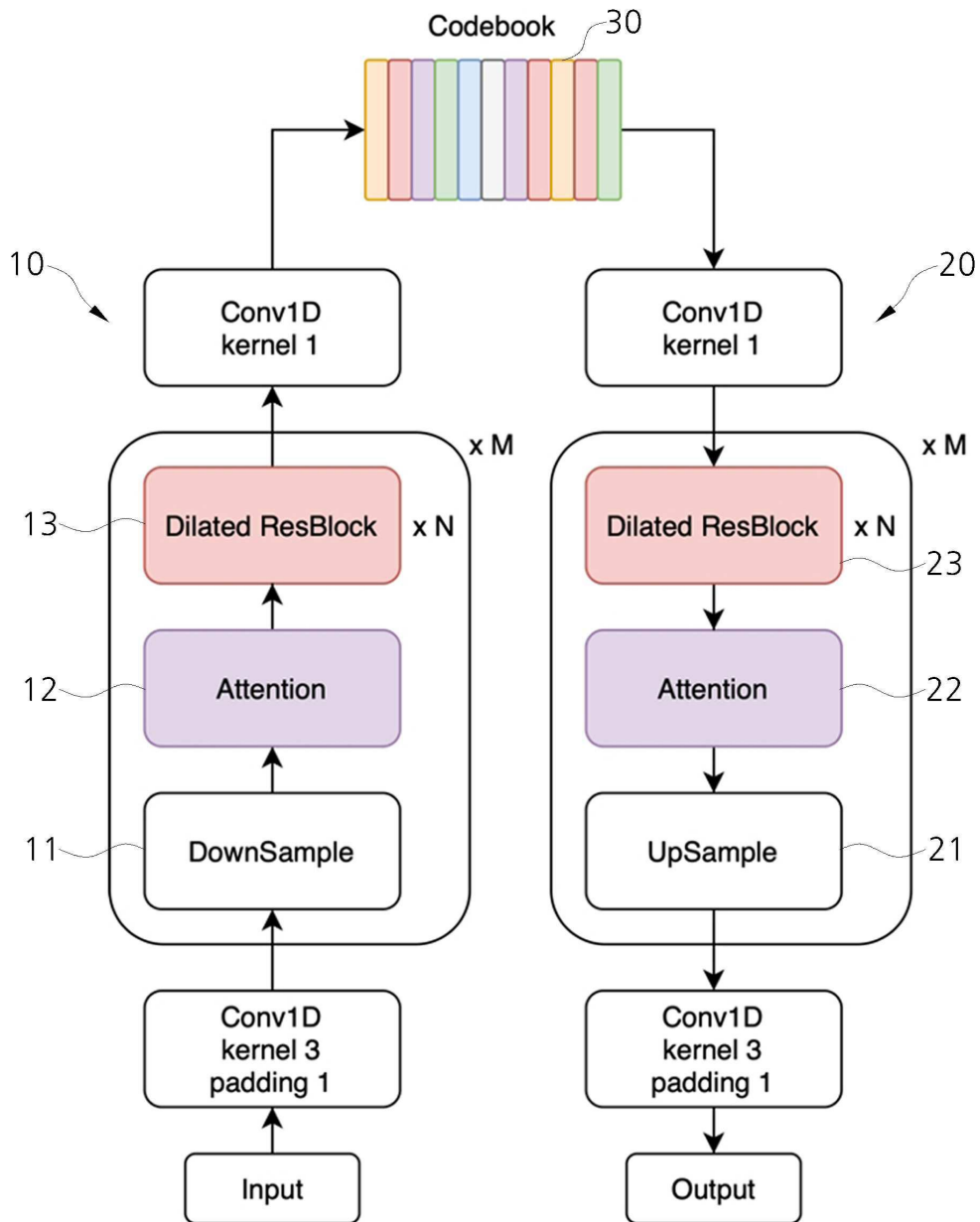
도면5



도면6

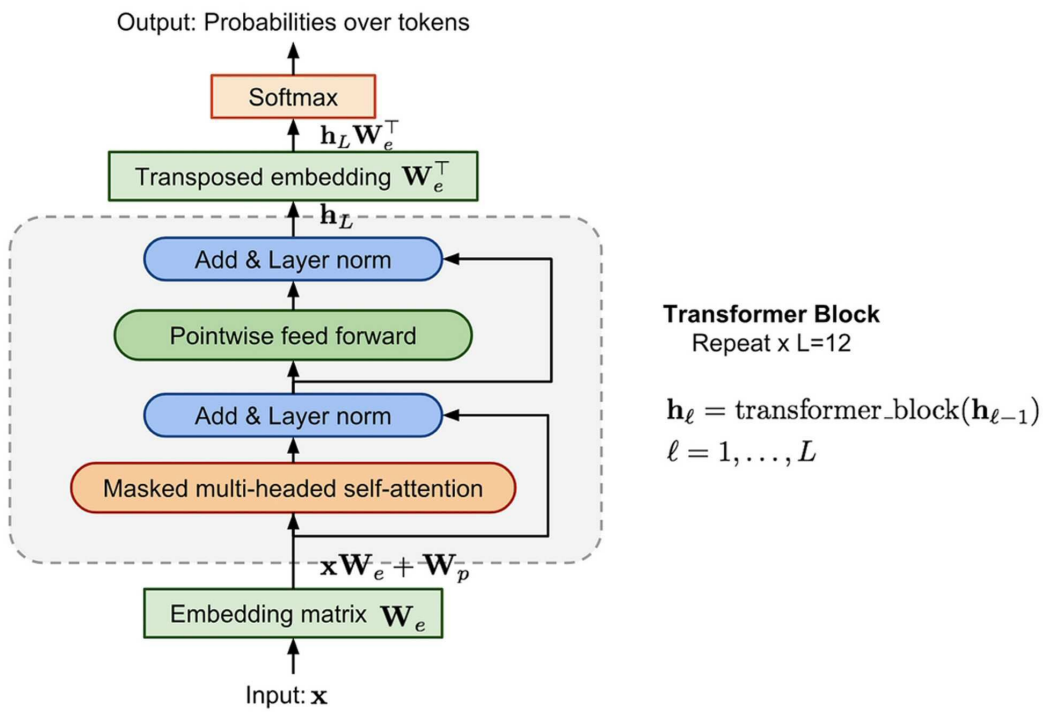


도면7

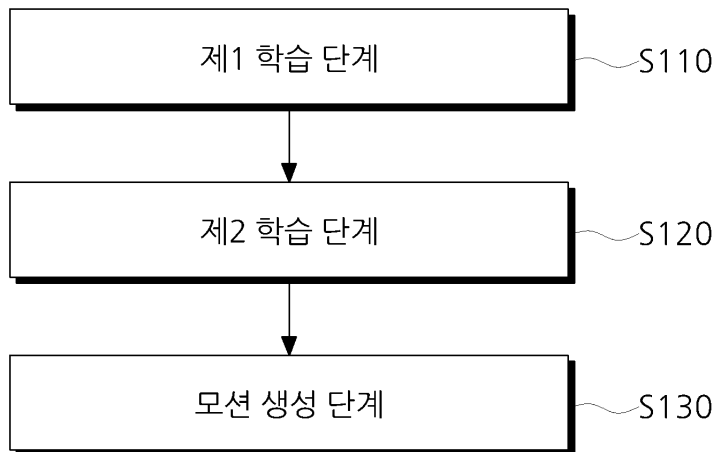


도면8

120



도면9



도면10

|                      | R-Precision $\uparrow$ |              |              | FID $\downarrow$ | MM-Dist $\downarrow$ | Diversity $\rightarrow$ | MModality $\uparrow$ |
|----------------------|------------------------|--------------|--------------|------------------|----------------------|-------------------------|----------------------|
|                      | Top-1                  | Top-2        | Top-3        |                  |                      |                         |                      |
| Real Motion          | 0.511                  | 0.703        | 0.797        | 0.002            | 2.974                | 9.503                   | —                    |
| <b>Ours (recons)</b> | <b>0.493</b>           | <b>0.690</b> | <b>0.78</b>  | <b>0.093</b>     | <b>3.02</b>          | <b>9.715</b>            | —                    |
| Seq2Seq[14]          | 0.180                  | 0.300        | 0.396        | 11.75            | 5.529                | 6.223                   | —                    |
| LJ2P[15]             | 0.246                  | 0.387        | 0.486        | 11.02            | 5.296                | 7.676                   | —                    |
| T2G[16]              | 0.165                  | 0.267        | 0.345        | 7.664            | 6.030                | 6.409                   | —                    |
| Hier[17]             | 0.301                  | 0.425        | 0.552        | 6.532            | 5.012                | 8.332                   | —                    |
| TEMOS[18]            | 0.424                  | 0.612        | 0.722        | 3.734            | 3.703                | 8.973                   | 0.368                |
| T2M[19]              | 0.457                  | 0.639        | 0.740        | 1.067            | 3.340                | 9.188                   | 2.090                |
| MDM[1]               | —                      | —            | 0.611        | 0.544            | 5.566                | 9.559                   | 2.799                |
| MLD[2]               | 0.481                  | 0.673        | 0.782        | 0.473            | 3.196                | 9.724                   | 2.413                |
| <b>Ours</b>          | <b>0.478</b>           | <b>0.675</b> | <b>0.775</b> | <b>0.139</b>     | <b>3.107</b>         | <b>9.711</b>            | <b>1.796</b>         |