VAE를 활용한 모션 캡처 데이터 생성

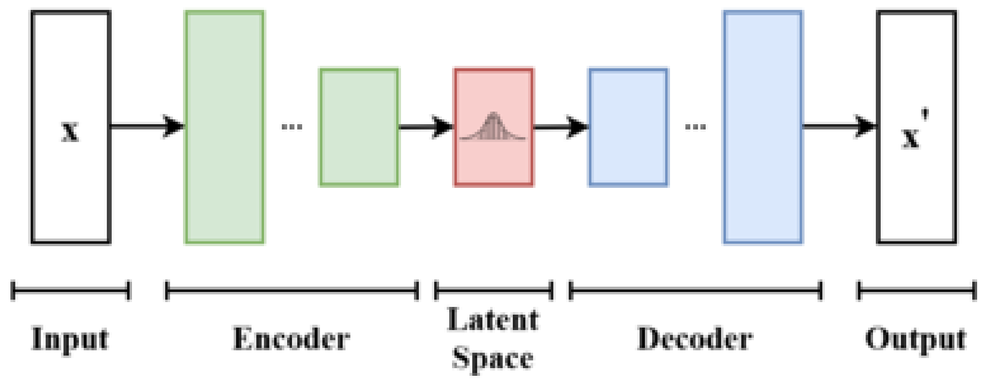
이 름 민경재 **지도교수** 유리

연구 배경

본 연구는 생성형 AI 모델을 활용하여 실제와 같은 모션 캡처 데이터를 생성할 수 있는지 탐구하는 것을 목표로 합니다. 현재 모션 캡처 데이터는 세밀하고 다양한 인간의 움직임을 담을 수 있지만 사람이 직접 여러 센서 장비를 착용하고 녹화해야 한다는 한계점이 존재합니다. 따라서 본 연구에서는 Variational Autoencoder (VAE) 모델을 학습하여 현실적인 모션 데이터를 생성해낼 수 있는지 탐구하였습니다.

관련 선행 연구

Variational Autoencoder (VAE) [1]는 학습 데이터를 Latent Space로 압축시킨 후 다시 복원하는 과정에서 데이터의 통계적 특성을 추출하는 기법입니다. MNIST 등의 이미지 데이터셋을 생성하는 데에 효과적임이 입증되었기에 본 연구의 기준 모델로 활용되었습니다. AMASS 데이터셋 [2] 은 모션 캡쳐 데이터셋 중 제일 방대하며 통일된 모션 형식을 가지기 때문에 본 연구의 학습 데이터셋으로 활용되었습니다.



모션 데이터 구조 분석

AMASS 데이터셋을 구성하는 모션 데이터는 SMPL [3]의 파라미터에 근거합니다. SMPL은 센서를 통해 측정한 모션 데이터를 통일된 구조로 표현하기 위해 Pose, Beta, Theta, Dmpl 등의 학습된 파라미터를 사용합니다. 본 연구에서는 피부와 체형에 관한 정보를 제외하고 모션만을 생성할 수 있도록 Pose와 Theta 파라미터만을 학습에 사용했습니다. Pose는 인체에 존재하는 23개의 관절마다 x, y, z 방향의 회전이 일어난 정도를 기록하며, Theta는 인체의 중심이 공통적으로 x, y, z 방향으로 기울어진 정도를 담고 있습니다. 추가적으로 다양한 모션의 학습을 위해 AMASS에 등록된 11개의 모션 데이터셋을 융합하여 학습 데이터로 활용했습니다.

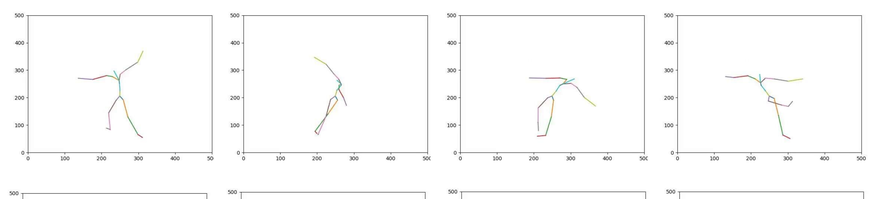
모델 학습 과정

VAE 모델을 학습하기 위해 AMASS 데이터셋을 매 10 프레임마다 30 프레임 길이로 잘랐습니다. 또한 VAE 모델은 Encoder와 Decoder에 Linear Layer가 3단계로 쌓여 있도록 구성하였으며 Latent Vector는 1024의 크기를 가지도록 설계했습니다. Adam optimizer을 이용해 0.0001의 learning rate로 총 5 epoch동안 학습하여 매 epoch마다 학습된 결과를 확인했습니다.

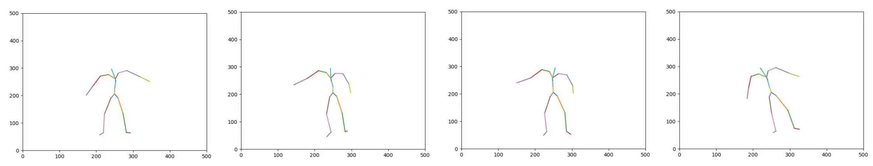
모델 추론 결과 및 분석

학습된 모델의 Decoder에서 Pose와 Theta 파라미터를 복원한 후 오픈소스 라이브러리[4] 를 활용하여 동영상으로 변환하였습니다. 이후 육안으로 결과를 비교하여 학습과정을 정성적으로 평가했습니다.

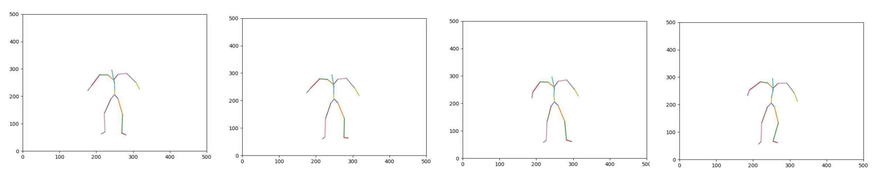
Epoch 0



Epoch 1



Epoch 5



Epoch가 증가함에 따라 무작위적인 모션이 감소하고 균일한 모션이 생성되는 것을 관찰했습니다. 하지만 epoch가 5 이상으로 증가할 때 생성된 영상에서는 미세한 떨림과 단조로운 모션만 생성되는 한계점이 발견되었습니다. 이에 추후에는 더 적은 수의 모션 종류만 학습에 이용하여 현실적이고 의미있는 모션을 생성하는 것이 목표입니다.

참고문헌 및 오픈소스

[1] Kingma, Diederik P., and Max Welling. "An introduction to variational autoencoders." Foundations and Trends® in Machine Learning.

[2] Mahmood, Naureen, et al. "AMASS: Archive of motion capture as surface shapes." Proceedings of the IEEE/CVF international conference on computer vision. 2019.

[3] Loper, Matthew, et al. "SMPL: A skinned multi-person linear model." Seminal Graphics Papers: Pushing the Boundaries.

[4] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2019.