헬스 트레이닝 데이터를 활용한 Multi-Encoder VAE 기반 Real-to-Sim 전이 기법

남현원, 한민호*, 고영배
아주대학교 소프트웨어학과, 아주대학교 AI 융합네트워크학과*

hwnam1129@ajou.ac.kr, vosej2414@ajou.ac.kr, youngko@ajou.ac.kr

Multi-Encoder VAE based Real-to-Sim Framework by Utilizing Health Training Data

Nam Hyun Won, Min-Ho Han*, Young-Bae Ko

Department of Software, Ajou University
Department of Artificial Intelligence Convergence Network, Ajou University*

요 약

본 논문에서는 운동 동작을 통해 얻는 대량의 실제 센서 데이터의 수집의 어려움을 해결하는 것을 목표로 하고 있다. 실제 데이터를 물리 기반 시뮬레이션 환경에서 얻은 가상 데이터 도메인 특성에 변환하는 Real2Sim 프레임 워크를 두 개의 인코더를 사용한 (Variational autoencoder)VAE 기반 구조로 설계하여 적용한다. 추가로, 두 인코더 간의 정보 독립성을 강화하기 위해 MINE(Mutual Information Neural Estimation) 기법을 도입하였다. 제안된 방법의 유효성을 검증하기 위해, 가상 데이터로만 학습된 스쿼트 동작 단계 인식 모델에 일반 VAE 모델과 제안 기법을 통해 변환된 데이터를 적용하여 성능을 비교하였다. 이를 통해 제안된 기법을 통해 변환된 데이터의 활용 가능성을 검증하였다.

Ⅰ. 서 론

최근에 MuJoCo 와 같은 물리 기반 시뮬레이션 환경을 활용하여 대량의 실제 데이터의 수집의 어려움을 해결하고자 하는 연구가 진행되고 있다. 지금까지는 실제 데이터 간의 간극을 줄이고 시뮬레이터의 특성에 과도하게 종속되는 경향을 줄이기 위해 Sim2Real 프레임 워크 개발에 많은 노력이 집중되어 왔다[1]. 특히, 최근에 CycleGAN 기반 모델을 활용한 Sim2Real 기법이 많이 제안되고 있다[2]. 하지만, 실제 데이터의 양이 제한적인 경우 일반화 성능이 저하되는 한계가 있으며[3], 대부분 이미지 기반 프레임워크에 초점이 맞춰져 있거나, 도메인 일반화가 부족한 시계열 기반 GAN 모델을 사용하는 경향이 있다[4].

따라서, 본논문에서는 데이터 부족 문제를 해결하기 위해 Sim2Real 프레임 워크와 반대되는 시계열 데이터 기반 Real2Sim 프레임 워크를 VAE 기반 구조로 설계하여 적용한다. 이를 통해 실제 데이터를 가상 데이터 기반 모델에 적용 가능하도록 변환한다. 추가로, Real2Sim 변환 시 시뮬레이터 특성에 과도하게 종속되는 경향을 막기 위해 두 개의 인코더를 사용하고 두 인코더 사이의 정보 분리를 위해 MINE 기반 손실 함수를 사용한다[5].

Ⅱ. 제안 방식

본논문에서는 시계열 센서 데이터 기반 Real2Sim 기법을 위해 RNN 기반 VAE 생성 모델을 사용한다. 이를 통해 실제 데이터의 주요 특성은 유지하되 가상 데이터와의 간극을 좁히는 것을 보여준다. 본논문에서 사용하는 구조는 일반적인 단일 인코더 VAE 가 아닌 그림 1 과 같이 두 개의 인코더를 사용하는 VAE 를 채택하였다. 기존의 여러 인코더를 동시에 학습시키는

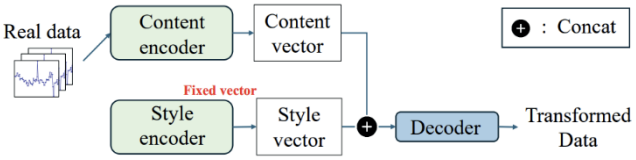

그림 1. 실제 데이터를 가상 데이터 도메인 특성에 적합하게 변환하는 전체 과정

VAE 모델과 다르게 본논문은 개별적으로 학습시키는 구조이다. 이를 통해 두 인코더 사이의 정보 독립성을 강화한다. Content 인코더는 실제 동작 단계(0: 앉는 자세, 1: 정 자세)의 특징을 표현한다. Style 인코더는 가상 데이터의 특징을 표현한다. 인코더에서 추출된 잠재 벡터들은 디코더의 입력으로 사용되어 시퀀스 데이터를 복원하는 구조이다. Content 인코더는 지도 학습 방식으로 학습이 된다. 이후, 수집량이 적은 실제 데이터를 활용하여 파인 튜닝 함으로써, 실제 환경에서도 동작 단계의 특징을 인식할 수 있도록 한다. Style 인코더에서는 가상 데이터의 특징을 의미하는 고정된 잠재 벡터를 추출하여 실제 데이터를 가상 데이터의 특징에 맞게 변환하도록 유도한다. 여기서 고정된 잠재 벡터는 다음과 같다:

$$z_s = \mu + \epsilon \cdot \sigma, \quad \text{where } \epsilon \sim \mathcal{N}(0, I)$$

여기서 μ 와 σ 는 Style 인코더가 추정한 평균 벡터와 표준편차 벡터이다. ϵ 는 정규분포 $\mathcal{N}$ 에서 샘플링된 노이즈 벡터이다.

이 잠재 벡터를 추출하기 위해 가상 데이터만을 활용하여 가상 데이터 특징의 분포를 학습한다. 학습 과정에는 Content 인코더를 고정(frozen)시켜 사용한다. 즉, Content 인코더는 디코더의 입력 구성에는 사용되지만, 학습 중에는 파라미터가 업데이트되지 않도록 한다. 이를 통해 Style 인코더가 디코더와 함께

학습되는 동안 가상 데이터 특징에 집중할 수 있도록 한다. Style 인코더의 학습은 VAE 의 손실 함수 구조를 따른다. 추가로, Style 인코더가 Content 인코더와는 독립적인 표현 학습을 유도하기 위해 MINE 기반 정규화 손실을 도입하였다. 손실 함수는 다음과 같다:

$$\mathcal{L}_{\text{StyleEncoder}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{KLD}} + \lambda \cdot \mathcal{L}_{\text{MINE}}$$

여기서 L_MINE은 다음과 같다:

$$\mathcal{L}_{\text{MINE}} = -\left(E_{p(z_c,z_s)}[f(z_c,z_s)] - \log\left(E_{p(z_c)p(z_s)}[e^{f(z_c,z_s)}]\right)\right)$$

여기서 $f(z_c,z_s)$ 는 Content 인코더와 Style 인코더의 출력을 입력으로 하는 신경망의 출력이며, $p(z_c,z_s)$ 는 실제 샘플 쌍에서 얻은 결합 분포, $p(z_c)p(z_s)$ 는 서로 독립적으로 샘플링한 분포를 의미한다. 이 손실 항은 두 잠재벡터 사이의 상호 정보를 최소화함으로써, Style 인코더가 Content 인코더와 독립적인 정보를 학습하도록 유도한다.

디코더에서는 두 잠재벡터를 입력 받을 때, 학습 가능한 가중치 계수를 도입하여 각 인코더로부터의 정보 비중을 자동으로 조절할 수 있도록 한다. 이를 통해 학습 과정에서 두 정보 간의 균형을 동적으로 조정할 수 있으며, 특정 인코더의 정보가 과도하게 종속되는 현상을 방지한다.

Ⅲ. 데이터 수집

본논문에서는 iPhone 13 mini 와 Apple Watch Series 7 을 통해 실제 데이터를 수집하였다. iOS 의 Nearby Interaction API 를 통해 두 기기 간 UWB 거리 측정 세션을 설정하고, 약 5Hz 주기로 거리 데이터를 수신하였다. 워치의 IMU 센서 데이터는 Core Motion API 를 활용해 가속도 및 각속도 정보를 20Hz 주기로 수집하였다. 서로 다른 주기의 데이터를 동기화하기 위해 UWB 데이터를 선형 보간하여 20Hz 로 업샘플링하였으며, 이를 통해 총 100 회의 스쿼트 동작 데이터를 확보하였다.

가상 데이터는 물리 기반 시뮬레이터인 MuJoCo 를 활용하여 생성하였다. 사람의 신체를 모사한 Humanoid 모델에 PD 제어기를 적용해 스쿼트 동작을 수행하도록 하였으며, 시뮬레이터에 포함된 거리 센서와 IMU 모듈을 통해 동작 중 센서 데이터를 실시간으로 기록하였다. 반복마다 가상 휴대폰과 워치의 위치를 무작위로 설정하여 다양한 조건에서 데이터를 수집하였고, 총 3000 회의 스쿼트 데이터를 확보하였다.

Ⅳ. 실험

그림 2 은 본논문에서 제안하는 기법과 일반 VAE 를 통해 가상 데이터를 복원한 결과를 비교하는 그래프이다. 제안하는 기법이 보다 원본 가상 데이터를 더 잘 따라가는 것을 확인할 수 있다. 이를 통해 두 개의 인코더의 정보 간의 균형을 동적으로 조정하고 두 정보 간의 독립을 강화함으로써 보다 가상 데이터 특성을 효과적으로 반영함을 보여준다.

표 1 에는 가상 데이터로만 학습된 Bi-GRU 모델을 기반으로 스쿼트 동작 단계 판별 성능을 평가한 결과를 보여준다. 테스트 데이터셋으로는 Real2Sim 기법을 통해 변환된 데이터를 사용하였다. 이를 통해 Real2Sim 변환이 실제 데이터가 가상 데이터와의 간극을 줄였는지 검증하였다. 일반 VAE 모델을 통해 변환된 데이터보다 제안하는 기법을 통해 변환된 데이터를 통해 테스트했을 시, 정확도가 15.23% 높았다. 인코더를 Content 인코더와 Style 인코더로 분리하여, Style 인코더를 통해 가상 데이터의 특성을 더 잘 반영하도록 하였다. 결과적으로 실제 데이터가 가상 데이터 도메인 특성에 적합하게 변환된 것을 확인할 수 있다.

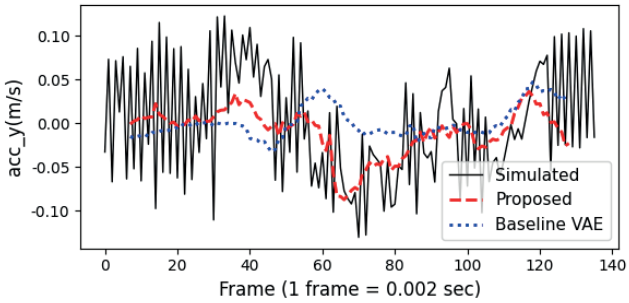

그림 2. 기존 VAE 와 제안한 기법을 이용한 가상 데이터 복원 결과 비교

Train	Test	Accuracy
가상 데이터	실제 데이터	55.80
가상 데이터	변환 데이터 (VAE)	59.77
가상 데이터	**변환 데이터 (Proposed)**	**74.90**

표 1. 가상 데이터로만 학습된 스쿼트 동작 단계 인식 모델의 변환 데이터셋 기반 테스트 성능 비교

Ⅴ. 결론

본 논문에서는 실제 데이터 부족 문제를 완화하고 도메인 일반화가 부족한 시계열 모델을 보완하기 위해 시계열 데이터 기반 Real2Sim 프레임 워크를 VAE 기반 구조로 설계하여 적용하였다. 두 개의 인코더를 사용하였으며, 두 정보의 독립을 강화하고, 두 정보 간의 균형을 동적으로 조절할 수 있게 하였다. 이를 통해 실제 데이터가 가상 데이터와의 간극을 좁히게 할 수 있었다.

ACKNOWLEDGMENT

"본 연구는 2025 년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음"(2022-0-01077)

참 고 문 헌

[1] Chen, Weihang, et al. "General-Purpose Sim2Real Protocol for Learning Contact-Rich Manipulation With Marker-Based Visuotactile Sensors." IEEE Transactions on Robotics 40 (2024): 1509-1526.

[2] Chen, Weihang, et al. "Bidirectional Sim-to-Real Transfer for GelSight Tactile Sensors With CycleGAN." IEEE Robotics and Automation Letters 7.3 (2022): 6187-6194.

[3] Zhou, Zhongchao, et al. "Addressing data imbalance in Sim2Real: ImbalSim2Real scheme and its application in finger joint stiffness self-sensing for soft robot-assisted rehabilitation." Frontiers in Bioengineering and Biotechnology 12 (2024): Article 1334643.

[4] Zhang, Da, et al. "A comprehensive review on GANs for time-series signals." Neural Computing and Applications 34 (2022): 3551-3571.

[5] Belghazi, Mohamed Ishmael, et al. "Mutual Information Neural Estimation." Proceedings of the 35th International Conference on Machine Learning, PMLR 80 (2018): 531-540.