GAN 기반 데이터 증강기법을 통한 가속도 데이터 생성에 대한 연구

강성환[O], 조위덕[*]
[O]아주대학교 지식정보공학과,
[*]아주대학교 전자공학과
e-mail: kumi1688@ajou.ac.kr[O], chowd@ajou.ac.kr[*]

A Study of GAN-based data augmentation technique on Acceleration Data Gereration

Sung-Hwan Kang[O], We-Duke Chow[*]
[O]Dept. of Knowelege Information Engineering, Ajou University,
[*]Dept. of Electric Engineering, Ajou University

● 요 약 ●

본 데이터 GAN 기법 데이터 증강기법을 적용하여 가속도 데이터를 증강하는 방법에 대해 연구한다. 가속도 데이터는 사람의 활동패턴을 인지하는데 있어 가장 기본적인 데이터로 활용된다. 가속도 데이터를 증강한 뒤, 활동패턴을 인지하는 머신러닝 모델 훈련에 사용한 결과 생성한 데이터가 육안으로 확인하였을 때 실제 데이터와 유사한 패턴을 형성하였고, 실제 활동패턴인지 모델 훈련에 사용한 결과 정확도(Accuracy)는 기존 데이터로만 훈련한 경우 74%인데 비해 증강된 데이터를 혼합하여 훈련하였을 때 약 88%로 개선된 것을 확인하였다.

키워드: 데이터 증강(Data Augmentation), 활동인지(Human Activity Recogniotion), 생성적 적대 신경망(GAN)

I. Introduction

가속도 데이터는 3D 공간에서 물체의 운동성을 파악하는 근거가 되는 기초 데이터로 가속도 데이터만으로 해당 물체의 운동 방향과 이동속도를 설명할 수 있다. 가속도 데이터 수집 센서는 비교적 소형 IoT 센서이고 간단하기 때문에 스마트폰 등에 탑재되어 기기의 움직임을 감지한 뒤 사용자의 편의성을 증대하는 방식으로 기기의 상태를 제어하는데 사용되며 사람의 활동 (Activity of Daily Living)을 판단하는데도 사용된다.

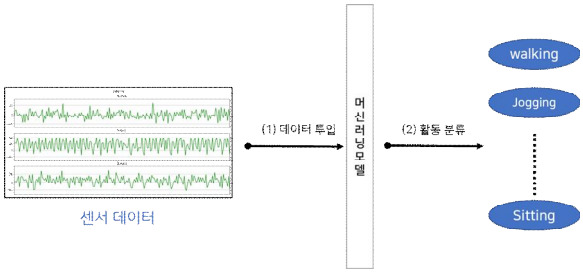

Fig. 1. 활동인지(Human Activity Recognition, HAR)

위의 그림 1처럼 일정 시간동안 가속도 데이터를 확보한 뒤 SVM, 랜덤포레스트 같은 머신러닝 모델이나 딥러닝 모델을 훈련시키면 사람의 활동을 분류할 수 있게 된다.

활동인지(HAR, Human Activity Recognition)용 머신러닝 훈련에 사용되는 가속도 데이터는 많을수록 성능이 개선된다[1]. 하지만 가속도 데이터는 실험자 모집, 주위환경 통제, 수집 기기 제한 등으로 수집이 어려운 면이 있다.

본 논문에서는 GAN 기반 데이터증강기법을 사용하여 HAR 머신러닝 모델을 훈련하기에 적합한 가속도 데이터를 합성하는 방법에 대해 연구해보고자 한다.

II. Preliminaries

2.1 관련연구

현재 GAN을 바탕으로 사람의 행동패턴을 파악하고자하는 연구가 다양하게 진행되고 있다. WiFi 신호를 분석해 실내 활동을 분석하는 머신러닝 모델을 훈련하는데 GAN으로 증강한 데이터를 사용하는

연구가 있다[2]. 또한 사용자의 움직임을 분석하는 머신러닝 모델 훈련에 GAN으로 증강한 데이터가 실제 데이터를 대체하거나 보완할 수 있다는 연구도 있다.[3]

III. The Proposed Scheme

위에서 제시한 것처럼 최근 연구에서는 정확한 활동패턴분류(HAR)를 위해 머신러닝모델에 GAN이 생성한 데이터를 훈련데이터로 써서 모델의 성능을 개선하는 과정을 보이고 있다.

이번 연구에서 진행할 부분은 활동분류(HAR) 관련 머신러닝 모델을 훈련하되, 적은 데이터가 있는 경우를 상정한다. 딥러닝 기반 HAR 머신러닝 모델은 데이터가 적은 경우 저조한 성능을 보이게 된다. 이 때 GAN 모델을 훈련시킨 뒤, GAN으로 합성한 데이터를 HAR 머신러닝 모델에 넣고 다시 훈련시켜 성능 향상 수준을 측정해보고, 생성된 데이터의 품질을 확인해보는 연구를 진행하고자 한다.

3.1 사용한 데이터 소개 및 전처리

GAN 기반 데이터 증강은 데이터 표본이 필요하다. 표본으로 사용할 가속도 데이터는 UCI Machine Learning Repository에서 제공하는 HAR 데이터셋을 사용하였다.

해당 데이터는 19-48세의 실험자 30명이 가속도센서가 탑재된 스마트폰을 허리에 착용한 채로 수집을 진행하였다. 걷기, 계단 올라가기, 계단 내려가기, 앉기, 서기, 눕기 등의 총 6가지의 활동을 하였으며, 각 활동마다 3축 가속도 센서, 자이로스코프 센서 데이터를 수집하였다.

가속도 데이터는 신호 데이터로 신호처리가 필요하다. [4] 파이선 신호처리 모듈인 Scipy의 고속 푸리에 변환 및 필터를 사용하여 일부 잡음을 제거하였다. 또한 기존 데이터는 50Hz로 샘플링 된 데이터가 128개(2.56초)가 1단위로 묶여 있었는데, 2.56초는 하나의 활동을 판단하기에는 너무 짧은 시간이라고 판단되어 이를 512개(10.24초)를 1단위로 재조정하여 보다 긴 시간동안 활동을 판단할 수 있도록 하였다.

3.2 데이터 증강 모델 설계

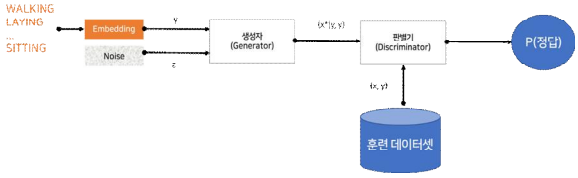

Fig. 2. 데이터 증강 모델 구조도

기호	의미
y	걷기, 뛰기, 정지 상태 등을 나타내는 Label
z	Random noise vector
x*	생성기가 생성한 데이터
x	실제 데이터
x*|y	어떤 레이블 y에 맞도록 생성기가 생성한 데이터
(x*|y, y)	어떤 레이블 y에 맞도록 생성기가 생성한 데이터와 레이블 y의 쌍
(x,y)	실제 데이터 x와 그에 맞는 레이블쌍 y

Fig. 3. 데이터 증강 모델 용어 소개

위의 그림2, 3는 데이터를 증강하기 위해 설계한 모델의 구조와 설명을 나타낸다. GAN 기반 모델이기 때문에 생성자-판별기가 쌍으로 구성하였으며, 특정 레이블에 해당되는 데이터를 생성하기 위해 고안된 CGAN(Conditional GAN)을 참고하여 생성자에 들어가는 Noise에 레이블 embedding을 덧씌우는 과정을 추가하였다. 생성자는 레이블 embedding을 기반으로 데이터를 생성하게 된다.

생성자와 판별기가 고품질의 데이터를 생성하도록 훈련하는 과정은 다음과 같다. 생성자는 랜덤 생성된 노이즈와 주어진 레이블을 바탕으로 데이터를 생성하여 판별기로 넘긴다. 판별기는 생성자가 생성한 가짜 데이터와 훈련 데이터셋에서 가져온 실제 데이터 셋을 비교하여 어떤 레이블(y)에 대해 가짜 데이터(x*|y), 진짜 데이터(x) 여부를 판별한다. 예를 들어 걷기 레이블(y)에 대해 이것이 진짜 데이터(x)인지, 가짜 데이터(x*|y)인지를 판별한다.

생성자와 판별기는 번갈아 가며 훈련을 진행하며, 판별기는 분류 손실을 최소화하는 방향으로 판별기 파라미터를 업데이트하고, 생성자는 판별기의 분류 손실을 최대화하도록 생성자 파라미터를 업데이트 한다. 훈련과정이 반복되면 생성자가 생성한 데이터는 고품질 데이터가 된다.

한 번의 iteration동안 데이터 증강 모델은 판별자 파라미터 업데이트, 생성자 파라미터 업데이트를 각각 1번씩 진행하며, 총 30000번 정도의 Iteration을 진행하였다.

3.3 데이터 증강 모델 훈련 결과

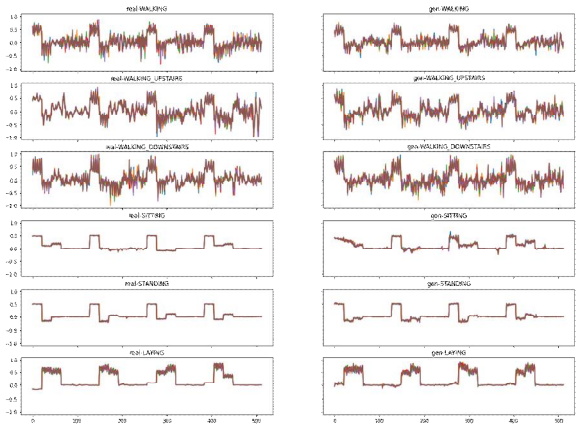

Fig. 4. 실제 데이터 샘플(왼쪽)과 증강모델로 생성한 데이터(오른쪽). 위에서부터 순서대로 걷기, 계단 오르기, 계단 내리기, 앉기, 눕기 순

생성된 데이터와 실제 데이터를 비교한 결과는 위의 그림 4와 같다. 육안으로 비교하였을 때 실제 데이터와 생성 데이터가 유사한 패턴을 띈다는 것을 확인할 수 있다.

3.4 생성 데이터 품질 검증

데이터 증강 모델로 생성한 데이터의 품질을 검증하기 위해 별도의 CNN 기반 HAR 모델을 설계하였다.[5] 해당 모델은 가속도 데이터를 받아 활동을 분류하는 모델이다. 해당 모델의 훈련에 사용한 데이터는 아래의 그림4와 같다. 데이터는 1 unit 당 10.24초간의 가속도 데이터를 묶어서 활용하며, 그림 4의 데이터 단위는 unit이다.

	훈련용 데이터	검증용 데이터	테스트용 데이터
원본 데이터	1,375	457	735
생성 데이터	15,000	5,000	735(원본 데이터)

Fig. 5. 훈련 데이터와 테스트 데이터 셋 정보(단위 unit)

	훈련셋 정확도	테스트셋 정확도
원본 데이터	1.0000	0.7429
생성 데이터	0.9996	0.4204
원본 + 생성 데이터	0.9996	0.8857

Fig. 6. 훈련 결과

위의 그림 6는 그림 5의 데이터로 훈련한 모델의 정확도를 나타낸다. 원본 데이터, 생성 데이터 모두 훈련셋의 정확도는 높지만 테스트셋의 정확도가 낮은 과적합 현상을 보였지만, 원본 데이터와 생성 데이터를 섞어 훈련한 결과 테스트셋 정확도가 상승해 과적합 현상이 해소된 것을 확인할 수 있다.

IV. Conclusions

본 논문에서는 GAN기반 데이터증강모델을 사용하여 활동분류모델(HAR)을 훈련시키기 위한 데이터를 합성하는 방법에 대한 연구를 진행하였다. 데이터 증강모델CNN기반 생성자-판별기 쌍으로 설계하였으며, 두 모델이 경쟁적으로 훈련하여 좋은 품질의 데이터를 생성한다. 훈련 결과 육안으로는 실제 데이터와 유사한 패턴을 보이는 데이터를 생성하였으며, 검증용 HAR 모델을 훈련한 결과 기존의 데이터만으로 훈련한 경우에 비해 정확도(Accuracy) 지표가 74%에서 88%로 개선된 것을 확인할 수 있다.

다만 생성된 데이터셋만으로 훈련하였을 때 낮은 정확도를 보이는 것과 훈련+원본 데이터셋을 혼합하여 훈련하였을 때에도 90% 이상의 정확도를 넘지 못하는 부분은 이번 연구에서 부족한 점으로 들 수 있다.

추후 연구에서는 부족한 점을 개선하여 생성 데이터 셋만으로 높은 정확도를 보이는 데이터 증강 모델을 연구해보고자 한다.

REFERENCES

[1] Deepika singh et al. “Human Activity Recognition Using Recurrent neural Networks”, Springer, 2017

[2] Parisa fard moshiri et al. “Using GAN to Enchance the Accuracy of Indoor Human Activity Recognition”, arxiv, 2020

[3] Junhao Shi et al. “A GAN-based data augmentation method for human activity recognition via the caching ability”, Wiley, 2020

[4] D Anguita et al. “Human Activity recognition on smartphones using a multiclass hardware-friendly support vector machine”, Springer, 2012

[5] Song-Mi Lee et al. “Human Activity Recognition From Accelerometer Data Using Convolutional neural Network”, IEEE, 2017