Wearable Sensor-based Human Activity Recognition Using Supervised Contrastive Learning

Nguyen Thi Hoai Thu, Dong Seog Han*
School of Electronic and Electrical Engineering, Kyungpook National University
thunguyen@knu.ac.kr, *dshan@knu.ac.kr

Abstract

Human activity recognition (HAR) task often needs to handle faces the problems of high inter-class similarity and high intra-class variance. To overcome these challenges, in this paper, we proposed a deep learning-based HAR framework that utilizes the advantages of the supervised contrastive learning method to increase not only the inter-class variability of sensor data from different activities but also the intra-class compactness of sensor data from same activity. Several experiments are carried out on the UCI HAR dataset with different deep learning model and train-test configurations. The results have proven that using supervised contrastive learning can increase the system performance compared to the traditional supervised learning method.

I. Introduction and Related Work

Wearable sensor-based human activity recognition (HAR) is the task of processing and analyzing the data collected from sensors such as acceleration and gyroscopes embedded in wearable devices to detect users' daily-life activity or abnormal activities such as freezing of gait [1]. Recently, with the advantages of automatic feature learning without requiring expert knowledge, several studies have applied deep learning (DL) algorithms such as long short-term memory (LSTM) and convolutional neural network to HAR tasks and achieved better results compared to the conventional machine learning approach.

However, the DL models with the supervised learning method often require a large label dataset to be able to learn deep representations of the data automatically. In addition, each person has a different way of carrying out an activity, while some activities such as jumping and falling have similar patterns in the sensor data. These characteristics of the wearable sensor-based HAR data often lead to the problems of high inter-class similarity and high intra-class variance in deep learning-based classification models. This makes the classification model easily get wrong predictions. Thus, it is necessary to use a representation learning method that can learn an embedding space in which sample pairs of an activity stay close to each other, while sample pairs of two different activities are far apart.

II. Methodology

In this paper, we utilize a representation learning method called supervised contrastive learning [2] to overcome the high intra-class variance and inter-class similarity in HAR. Unlike the well-known self-supervised contrastive representation learning method which often heavily relies on the data augmentation method, we employ a fully-supervised contrastive learning method to effectively leverage the label information. The framework is divided into two stages: supervised contrastive learning and target supervised learning as shown in Fig. 1.

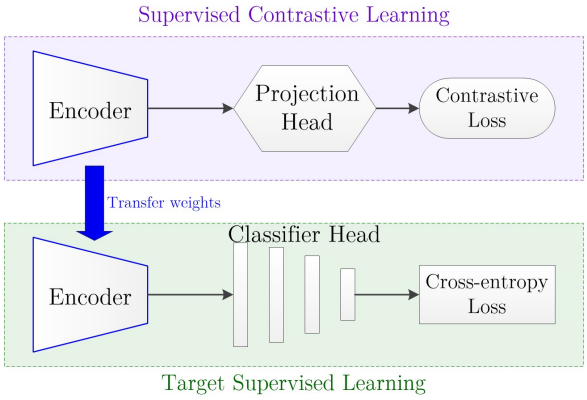

Figure 1. Architecture and training method of the proposed supervised contrastive learning-based HAR model

First, a projection head is attached to the encoder to learn to produce vector representations of input sensor data such that representations of samples from the same activity will be more similar compared to the representations of samples in different activities. During training, the output feature of the encoder is flattened into a 1-D vector and forwarded to the projection head constructed from one fully connected layer with size of a 256. In the contrastive loss, output features of the projection head from all the data samples in a mini batch are then normalized using an L2 norm. The supervised contrastive loss [2] is defined as

$$L = \sum_{i \in I} -\log \left\{ \frac{1}{\left|P(i)\right|} \sum_{p \in P(i)} \frac{\exp\left(z_i \cdot z_p / \tau\right)}{\sum_{a \in A(i)} \exp\left(z_i \cdot z_a / \tau\right)} \right\}$$

, where $P(i)$ is the set of all other samples in the current mini batch that have the same label (a.k.a same activity) with the data sample i; $A(i)$ is all the samples in the current mini batch, exclude the sample i; z_i is the normalized output feature from the projection head of the data sample i; τ is the temperature of the softmax function. This supervised contrastive loss contrasts the set of all samples from the same class as positives against the negatives from the remainder of the batch. After the supervised contrastive learning, the pre-trained encoder is transferred to the target supervised learning. A classifier head consists of three fully connected layers is attached to the output layer of the pre-trained encoder. A cross-entropy loss is used for this training process to obtain the final target model.

To analyze the efficiency of the supervised contrastive learning method in the application of human activity recognition, different experiments are carried out on the UCI HAR dataset [3]. Four different deep learning models including a 2-D convolutional neural network (CNN) and 3 models proposed in [4] (i.e., 1-D CNN, LSTM and hybrid 1-D CNN-LSTM) are used as the encoders in the supervised contrastive learning. The results from the traditional supervised learning method are used as baselines. In the supervised contrastive learning method, there are two cases are considered: 1) the encoder is frozen during the training process of target HAR task, 2) the encoder is also trained together with the classifier head. From the experiment results using accuracy as a performance metric shown in Table I, the supervised contrastive learning method has better performance than the purely supervised learning method in all four deep learning model. The gap between two learning methods can be clearly seen in the case of 1D-CNN and hybrid CNN-LSTM models with the difference of 1.4% and 2.2%, respectively. In terms of the encoder, both trainable encoder and frozen encoder have quite similar performance.

Table I. Comparison results between baseline supervised learning and supervised contrastive learning on different deep learning models

Model	**Baseline**	**Contrastive** (*frozen encoder*)	**Contrastive** (*trainable encoder*)
2D-CNN	91.93%	91.94%	**92.64%**
1D-CNN	92.01%	**93.44%**	93.36%
LSTM	92.32%	92.88%	**92.91%**
1D-CNN-LSTM	91.04%	93.1%	**93.22%**

III. Conclusion

In this study, we proposed a deep learning-based human activity recognition framework that uses the supervised contrastive learning to learn a data representation space in which sample pairs of an activity stay close to each other, while sample pairs of two different activities are far apart. The experiment results on several deep learning models have shown that the supervised contrastive learning outperforms the conventional supervised learning in the human activity recognition task by reducing the intra-class variance and inter-class similarity effectively.

ACKNOWLEDGMENT

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2022-2020-0-01808) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation).

References

[1] Thu, Nguyen Thi Hoai, and Dong Seog Han. "Freezing of Gait Detection Using Discrete Wavelet Transform and Hybrid Deep Learning Architecture." *2021 Twelfth International Conference on Ubiquitous and Future Networks (ICUFN)*, 2021, pp. 448-451.

[2] Khosla, Prannay, et al. "Supervised contrastive learning." *Advances in Neural Information Processing Systems* 33 (2020), pp. 18661-18673.

[3] Anguita, Davide, et al. "A public domain dataset for human activity recognition using smartphones." *Proceedings of the 21th international European symposium on artificial neural networks, computational intelligence and machine learning*, 2013, pp. 437-442.

[4] Thu, Nguyen Thi Hoai, and Dong Seog Han. "An Investigation on Deep Learning-Based Activity Recognition Using IMUs and Stretch Sensors." *2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 2022, pp. 377-382.