Sensor Data Representation with Transformer-Based Contrastive Learning for Human Action Recognition and Detection

Lei Yang
Dept. of Computer Science
Tokyo Institute of Technology
Tokyo, Japan
yanglei@ks.c.titech.ac.jp

Yuzhe Hao
Dept. of Computer Science
Tokyo Institute of Technology
Tokyo, Japan
yuzhe@ks.c.titech.ac.jp

Koichi Shinoda
Dept. of Computer Science
Tokyo Institute of Technology
Tokyo, Japan
shinoda@c.titech.ac.jp

***Abstract*—Feature extraction is an important process in human activity recognition (HAR) with wearable sensors. Recent studies have shown that learned features are more effective than manually engineered features in related fields. However, the scarcity and expensiveness of labeled data are limiting the development of sensor data representation learning. Our work focuses on this issue and introduces a self-supervised learning method that uses unlabeled data to improve the quality of learned sensor representations. We hypothesize that unlabeled wearable sensor data in human activities have long-term and short-term temporal contextual correlations and exploit such correlations with Transformer and Contrastive Predictive Coding (CPC) framework. The learned representation is evaluated on human activity recognition and detection tasks in real-life scenarios. The experiments show that our method outperforms previous state-of-the-art methods on MotionSense and MobiAct datasets on the HAR task and gets a remarkable performance on the EVARS dataset on the action detection task.**

***Index Terms*—IMU sensor, self-supervised learning, representation learning, human activity recognition, temporal action localization**

I. Introduction

With the development of ubiquitous and mobile computing technologies, human-centric wearable equipment with inertial measurement units (IMU) has become popular. The growing prevalence of wearable devices makes large amounts of sensor data accessible and promotes the blossoming of human activity-related applications, including but not limited to action recognition, posture prediction, and health assessment. Generally, the workflow of these applications can be conducted in five steps—data collection, pre-processing, segmentation (data split), feature extraction (i.e., representation learning), and classification (or regression), which is called activity recognition chains (ARC) [8]. While the omnipresent IMU sensors grant the availability of unlabeled data, the collection of annotated sensor data remains to be a challenge due to privacy and costliness. Driven by this significant insufficiency of labeling data, self-supervised learning that utilizes unlabeled sensor data have been proposed to improve the performance of related applications.

Self-supervised learning aims to extract high-quality vectorized representations from raw IMU sensor data with abundant unlabeled data to alleviate the impact caused by the lack of annotations. These methods consist of two parts. The architecture used for the representation encoder and the training strategy to pre-train the encoder. Previous studies have shown that deep learning methods, including Convolutional Neural Networks (CNNs) [9] and Recurrent Neural Networks (RNNs) [3], can effectively learn representations of IMU sensor data. Then, Transformer [10], which has excelled in representation learning of data for other modalities, was also introduced for encoding sensor learning. The validity of these encoders indicates that there is intrinsic information in the unlabeled sensor data that can be explored.

For the training strategy, one major method is masked reconstruction. It involves masking a portion of the input data and training the model to reconstruct the missing portion. The main idea of masked reconstruction is that there is a bidirectional correlation in sequential signals that can be learned by the model. Another widely used method is contrastive learning which trains the model to differentiate between similar and dissimilar samples. This method assumes that there is a commonality in some given similar or continuous inputs, and learns this commonality by making the model discriminate among randomly acquired false samples.

In this paper, we introduce a Transformer-based contrastive learning framework with sensor data specialized front-process module. This is the first work to combine contrastive learning with Transformer for sensor data representation learning. This combination of the methods allows the inherent nature of sensor data to be fully utilized by self-supervised pre-training, hence generating informative representations that can be easily adapted to different downstream tasks. Firstly, unlabeled sensor data is used for pre-training. Self-supervised learning is applied here to enable the encoder to learn the latent structure and temporal relationship inside sensor data. After the pre-trained is finished, the encoder is transferred to a downstream task with limited supervised training.

The quality of the representation learning framework is

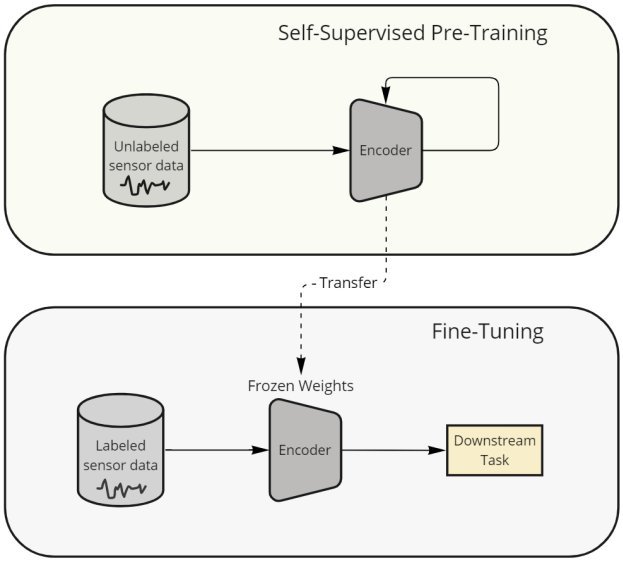

Fig. 1: "Pretrain-then-Finetune" Workflow: The encoder (i.e. representation extractor) is pre-trained with unlabeled data, and then fine-tuned on a downstream task with limited labeled data

evaluated based on the performance of downstream tasks, human activity recognition and temporal action localization. Experiment results show that our model has a deep understanding of sensor data after pre-training on unlabeled data in the proposed framework, outperforming state-of-the-art supervised and self-supervised methods. For the evaluation of temporal action localization (TAL), which is the first attempt to do action detection related tasks using sensor data, we also get competitive results. Through this task, we demonstrate that the representations learned by our model are versatile and have the potential to be applied to different tasks in various scenarios.

II. Previous Studies

Massive research has been conducted for sensor data representation learning. Saeed, *et al.* [9] introduced the "pretrain-then-finetune" workflow into self-supervised sensor representation learning. They used convolutional neural networks as the feature encoder for local information extraction. The encoder is pre-trained with raw unlabeled sensor data on several signal transformation prediction tasks such as noised, scaled, rotated, etc. Then, the model is fine-tuned to human action recognition to evaluate the performance of the pre-trained encoder. The paper shows the potential of such workflow in sensor data representation learning, but the general signal transformation prediction task can not capture the character of sensor data in depth.

Haresamudram *et al.* [4] choose to reconstruct masked input sensor data as their pre-training task for better sensor representations, which also follows the "pretrain-then-finetune" workflow as in [9]. However, the masked reconstruction task may be to difficult for sensor data since it do not contain strong bidirectional relationship.

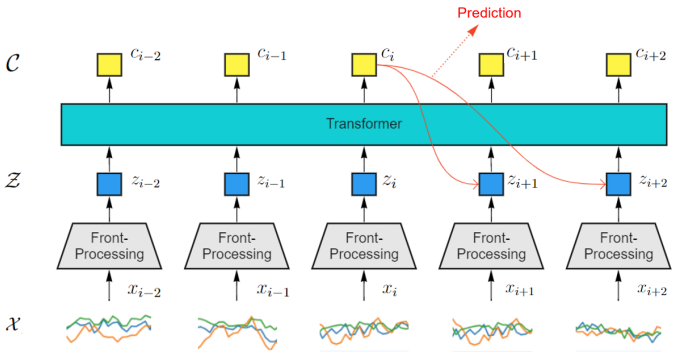

Fig. 2: Overview of the Framework

Temporal information of sensor data is utilized through contrastive learning in [3], proving the importance of contextual relationships inside sensor data. The paper uses recurrent neural networks as encoder to encode the input embeddings into informative representations. The model is pre-trained by predicting future embeddings of data with information provided by the current representation. After pre-training, the encoder is frozen and applied to a human activity recognition task for evaluation. The CPC framework has been proven to be effective in several sequential signal learning tasks [6] and it also shows competitive performance on sensor representation learning. The problem of this paper is also the lack of attention to the inherent characteristic of sensor data and it just follows a successful framework.

III. Contrastive Learning of Transformer

A. Overview

In this section, we describe the design of the proposed framework and how it is applied to human activity related applications.

The framework follows the workflow shown in Fig. 1. Firstly, unlabeled sensor data is used for pre-training. Self-supervised learning is applied at this stage to enable the encoder to learn the latent structure and temporal relationship inside sensor data. At the second stage, the pre-trained encoder is transferred to a downstream task with limited supervised training. The quality of the representation learning framework is evaluated based on its performance on downstream tasks.

Our proposed framework consists of a front-process module and a Transformer encoder as shown in Fig. 2. From the bottom to the top, the first part is a front-process module that shares weight across different timesteps but only processes one timestep at a time. The module focuses on processing the multi-channel sensor data $\mathcal{X}$ into fused intermediate representations $\mathcal{Z}$. The second part is a Transformer encoder that learns the bidirectional temporal correlation inside the clip-level sensor data and outputs the final representation $\mathcal{C}$.

B. Front-Processing Module

In our system, the front-processing module is designed to solve the fusion problem of different types of IMU sensor data. Typically, three different types of data are included in

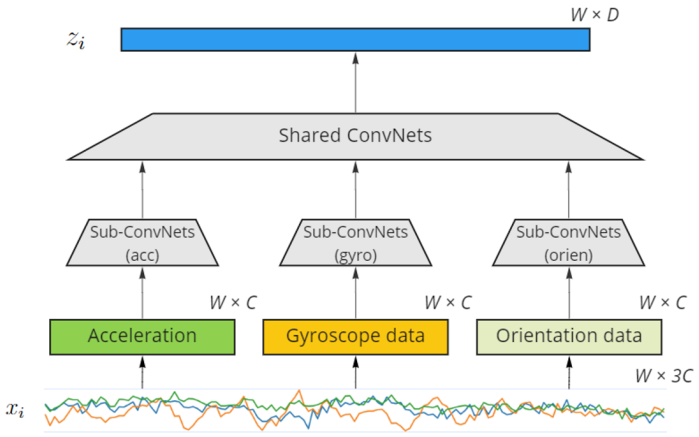

Fig. 3: Front-Processing Module: W refers to window size, i.e., the number of frames in each chip, D refers to the feature dimension, which is a predefined hyperparameter, $C = 3$ is the dimension of each data type

IMU sensor data, which is acceleration, angular velocity, and software-based orientation information. If we fuse different types of data at the beginning stage, data from other types could be some kind of noise for the model to obtain the whole understanding of each type.

To process the raw sensor data into intermediate features that are suitable for deep neural networks, we design a sensor data specialized front-processing module shown in Fig. 3. It firstly transfers the data from different types to a standard latent sub-representation separately, which preserves the inner relationship inside each data type. Then the three orthogonal sensor features are get fused together with shared CNNs, for obtaining the comprehensive sensor features.

C. Transformer Encoder

We use a Transformer-based encoder [10] as the feature encoder shown in Fig. 2. By applying the Transformer-based encoder, we aim to explore the long-term temporal relationship inside input sequences. Such a relationship is an inherent characteristic of wearable IMU sensor data because human activities are persistent and consistent. Unlike other sequential signals such as speech signals in which one utterance can be less than 1 second, human activities always last from tens of seconds to hours (i.e. jogging or walking). And the consistency is cross-activities. For example, if we are running, what we are going to do next is either continue running or slow down, both will show a clear numerical consistency. The Transformer-based encoder can capture such contextual correlation without any extra information. That is also the foundation of our self-supervised learning framework.

Combining the front-process module and Transformer-based encoder, Our encoder can focus on the channel-wise nature of sensor data while also paying attention to the persistence of human activity associated with IMU sensors.

D. Pre-Training with Contrastive Learning

We use a contrastive learning based self-supervised learning method to match the continuity of IMU sensor data. As continuous chips share a significant similarity, contrastive learning methods can learn the relationship between them easily. Our learning method follows the idea of contrastive predictive coding (CPC) [6]. The CPC framework allows a further prediction compared to other methods, making the long-term human activities' information to be obtained. The process of applying CPC to our model is shown in Fig. 2. Firstly, the front-processing module map a clip of sensor data x_i into a sequence of intermediate representations z_i. Then, the Transformer encoder is utilized to process the intermediate representations z_i into output representation c_i. Finally, we use the output representation c_i to predict future timesteps z_{i+k}. As introduced above, CPC does not use a generative model $p_k(x_{i+k}|c_i)$, but the density ratio is applied here as follows [6]:

$$f_k(x_{i+k}, c_i) \propto \frac{p(x_{i+k}|c_i)}{p(x_{i+k})}, \tag{1}$$

here $\propto$ stands for 'proportional to'. And a simple log-bilinear model is used for scoring:

$$f_k(x_{i+k}, c_i) = \exp(z_{i+k}^T W_k c_i). \tag{2}$$

Here, a linear transformation $W_k^T c_i$ is used for the prediction with a different W_k for every timestep k. Using the density ratio $f_k(x_{i+k}, c_i)$ and inferring z_{t+k} with an encoder relieves the model from modeling the high-dimensional distribution x_{t_k}. We can not get $p(x)$ or $p(x|c)$ directly, but samples from these distributions can be used directly for some techniques such as Noise-Contrastive-Estimation. Here, we use the InforNEC loss detailed in [6] to update the network parameters:

$$L_N = -\mathbb{E}_X \left[\log \frac{f_k(x_{i+k}, c_i)}{\Sigma_{x_j \in \mathcal{X}} f_k(x_j, c_i)} \right], \tag{3}$$

where a positive sample that is from future timesteps $p(x_{i+k}|c_i)$ and $N-1$ negative samples from proposal distribution $p(x_{i+k})$ is used for optimizing the objective. The InforNEC loss is a categorical cross-entropy of classifying the positive sample from negative samples, with $\frac{f_k}{\Sigma_{\mathcal{X}} f_k}$ being the prediction of the model.

IV. Experiments

A. Setup

In this section, we introduce the setup for our pre-training and how we use two downstream tasks, human activity recognition (HAR) and temporal action localization (TAL), to evaluate the quality of our learned representation. HAR refers to identifying actions performed by a person based on data collected from the surroundings. Such action involves the main part of common events performed in our daily life, such as walking, jogging, or jumping. TAL is an action detection task, which aims to detect activities in a temporal data stream and output the beginning and ending timestamps. Specifically,

TABLE I: Experiment Results for HAR: The evaluation metrics is mean F1-score, which is the mean of F1-score for each activity class. The results of the best-performing method on both datasets are bold. * indicates this method is a supervised learning method.

Approach	MobiAct	MotionSense
DeepConvLSTM* [7]	82.40	85.15
Multi-task self-supervised [9]	75.41	83.30
Convolutional autoencoder [2]	79.58	82.50
Masked reconstruction [4]	76.81	88.02
Contrastive Learning with RNN [3]	80.97	89.05
Proposed (w/o front-processing)	81.58	88.98
Proposed	**83.84**	**89.23**

we consider the data stream contains action instances and background data (indicating there is no action we care), and the aim of TAL is to find these action instances from the background data.

The pre-training is performed for 100 epochs with the learning rate range from {1e-3, 5e-4} and $k \in \{2,4,8,12,16\}$. For the Transformer encoder, the embedded dimension D is 128, and the number of layers is 8. The network weights are optimized using Adam. The prediction networks, W_k, are linear layers with 128 units. For the experiment on each dataset, the pre-training data is the training set of the same dataset without annotations.

B. Results for Human Activity Recognition

Fine-tuning for human activity recognition is a simple multi-class classification task, which is optimized by cross-entropy loss shown below:

$$L_{CE} = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}), \tag{4}$$

where M is the number of classes, y is the binary indicator (0 or 1) if class label c is the correct classification for observation o. p is the predicted probability observation o is of class c.

We perform HAR on two datasets, MobiAct [11] and MotionSense [5]. The evaluation metric is the mean F1-score defined as:

$$F_m = \frac{2}{|c|} \Sigma_c \frac{\mathrm{prec_c} \times \mathrm{recall_c}}{\mathrm{prec_c} + \mathrm{recall_c}}, \tag{5}$$

where $|c|$ indicates number of activity classes in the dataset and $\mathrm{prec_c}$ and $\mathrm{recall_c}$ are the precision and recall for class c. We compare the evaluation result with state-of-the-art methods for different supervision types as shown in Tab. I. It shows that our proposed model is not only the best in self-supervised learning methods but also in supervised methods.

The confusion matrix is shown in Fig. 4 and Fig. 5.

We also conduct experiments about the ability of the Transformer to deal with long sequences as shown in Tab. II.

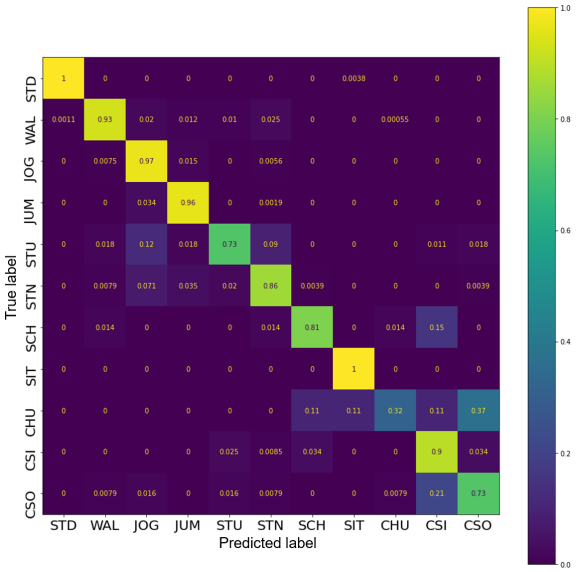

Fig. 4: Confusion Matrix for MobiAct: The class name stands for standing, walking, jogging, jumping, stairs up, stairs down, stand to sit, sitting on a chair, sit to stand, car step-in, and car step-out

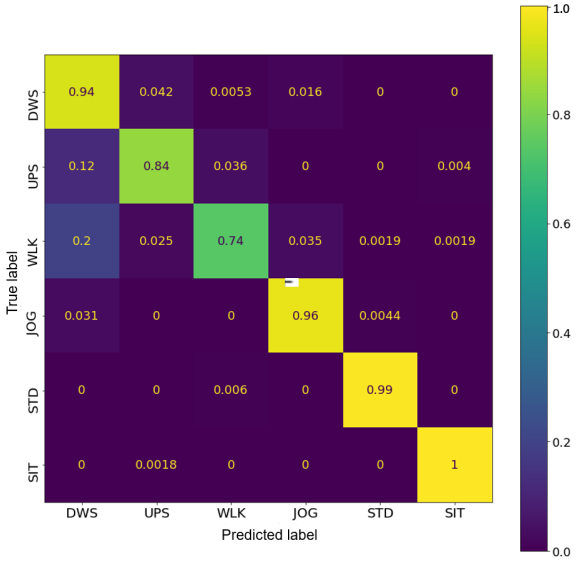

Fig. 5: Confusion Matrix for MotionSense: The class name stands for down stairs, up stairs, walking, jogging, standing, and sitting

C. Results for Temporal Action Localization

Following the workflow of [12], we use structured segment networks to perform the TAL task, while we replace the feature encoder with our model. We apply 0-1 detection that only outputs the interval of an action instance. The evaluation metric is average precision (AP) calculated based on the confusion matrix and Intersection over Union (IoU) shown

TABLE II: Experiment Results for Encoder Architecture: We use different input sequence sizes and encoder structures to evaluate the performance of the Transformer encoder

Encoder	Input Sequence Size	
	200	400
RNN	80.17	81.38
Transformer	81.09	83.84

TABLE III: Experiment Results for TAL: The evaluation metrics is the mean average precision from IoU thresholds from 0.2 to 0.5.

	IoU Thresholds				
Model	0.2	0.3	0.4	0.5	Average
RNN with CPC	56.04	49.76	42.86	36.68	46.34
Proposed	**66.22**	**58.66**	**46.98**	**39.10**	**52.74**

as follows:

$$\mathbf{AP} = \sum_{i=0}^{n-1} (R_i - R_{i+1}) P_i, \tag{6}$$

$$R_n = 0, P_n = 1, \tag{7}$$

where n is the numbers of pre-set IoU thresholds, R_i and P_i is the recall and precision at threshold i.

We perform the TAL task on EVARS dataset [1]. The result is shown in Tab. III. We can see that the learned representation from sensor data has the ability to distinguish actions from inactive background data. And compared with the current state-of-the-art model [3] in human activity recognition, our proposed model still performs better, which demonstrates the generality and effectiveness of the representations learned by our framework.

D. Discussion

Fig. 4 and Fig. 5 show that the same problem exists on both datasets that similar activities are hard to be distinguished. For example, activities like 'sit to stand' and 'car step-out', are so close that 'car step-out' can be considered as one kind of 'sit to stand', and the only difference is the subtle movement in the horizontal level, which is difficult to be detected by an IMU sensor in the pocket. If we can add one sensor in the wrist, such kind of movements can be captured. Other than these cases, the model can effectively recognize these daily activities with very limited supervised training.

Although temporal action localization is a relatively difficult task, we obtained quite good results without action classification. This suggests that our proposed framework is generalizable and can be used in a variety of different scenarios.

V. Conclusions

In this paper, We proposed a Transformer-based self-supervised learning framework for sensor data representation learning. It firstly applies contrastive learning to the Transformer to extract informative representation from unlabeled data. The framework design focuses on the inherent characteristic of sensor data, which can be concluded as consistency and temporal contextual correlation.

We evaluate our representation learning framework on human activity recognition and temporal action localization. With our experiment results, we demonstrate the effectiveness and generality of our method. The pre-trained model can capture the latent information in temporal IMU sensor data and has the ability to be applied to the different downstream tasks.

However, sensor data has limitations from the physical level. Compared with other information-rich data like video, the knowledge contained in wearable sensor data is restricted to the body parts where the sensor is located. To compensate for this gap, we can use several IMU sensors located in different body parts to obtain a wealth of information. This brings the relationship of multiple sensors into consideration, which we believe will also profoundly reflect the characteristics of human activity.

References

[1] Yuzhe Hao et al. "EvIs-Kitchen: Egocentric Human Activities Recognition with Video and Inertial Sensor data". In: *International conference on multimedia modeling* (2023).

[2] Harish Haresamudram, David V Anderson, and Thomas Plötz. "On the role of features in human activity recognition". In: *Proceedings of the 23rd International symposium on wearable computers*. 2019, pp. 78–88.

[3] Harish Haresamudram, Irfan Essa, and Thomas Plötz. "Contrastive predictive coding for human activity recognition". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5.2 (2021), pp. 1–26.

[4] Harish Haresamudram et al. "Masked reconstruction based self-supervision for human activity recognition". In: *Proceedings of the 2020 international symposium on wearable computers*. 2020, pp. 45–49.

[5] Mohammad Malekzadeh et al. "Mobile sensor data anonymization". In: *Proceedings of the international conference on internet of things design and implementation*. 2019, pp. 49–58.

[6] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748* (2018).

[7] Francisco Javier Ordóñez and Daniel Roggen. "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition". In: *Sensors* 16.1 (2016), p. 115.

[8] Attila Reiss and Didier Stricker. "Introducing a new benchmarked dataset for activity monitoring". In: *2012 16th international symposium on wearable computers*. IEEE. 2012, pp. 108–109.

[9] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. "Multi-task self-supervised learning for human activity detection". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3.2 (2019), pp. 1–30.

[10] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[11] George Vavoulas et al. "The mobiact dataset: Recognition of activities of daily living using smartphones". In: *International Conference on Information and Communication Technologies for Ageing Well and e-Health*. Vol. 2. SciTePress. 2016, pp. 143–151.

[12] Yue Zhao et al. "Temporal action detection with structured segment networks". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2914–2923.