https://doi.org/10.48550/arXiv.2104.09072,
https://doi.org/10.1109/GCWkshps56602.2022.10008537

Peer reviewed version

Link to published version (if available):
10.48550/arXiv.2104.09072
10.1109/GCWkshps56602.2022.10008537

Link to publication record in Explore Bristol Research
PDF-document

This is the accepted author manuscript (AAM). The final published version (version of record) is available online via IEEE at https://doi.org/10.1109/GCWkshps56602.2022.10008537. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research
General rights
This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/
Self-Supervised WiFi-Based Activity Recognition

Mohammud J. Bocus, Hok-Shing Lau, Ryan McConville, Robert J. Piechocki and Raul Santos-Rodriguez

School of Computer Science, Electrical and Electronic Engineering, and Engineering Maths,
University of Bristol, UK.

{junaid.bocus, sk19152, ryan.mcconville, eerjp, enrsr}@bristol.ac.uk.

Abstract—Traditional approaches to activity recognition involve the use of wearable sensors or cameras in order to recognise human activities. In this work, we extract fine-grained physical layer information from WiFi devices for the purpose of passive activity recognition in indoor environments. While such data is ubiquitous, few approaches are designed to utilise large amounts of unlabelled WiFi data. We propose the use of self-supervised contrastive learning to improve activity recognition performance when using multiple views of the transmitted WiFi signal captured by synchronised receivers deployed in different positions. We conduct extensive Human Activity Recognition (HAR) experiments in two furnished office rooms, whereby six participants of different age groups performed six day-to-day activities. We compare the proposed contrastive learning system with conventional non-contrastive systems and observe significant improvement on the task of WiFi based activity recognition under few-shot learning scenarios. Namely, contrastively pretraining with an AlexNet-based backbone encoder led to a 22% increase in macro $F_1$ score when only 1.29% of labelled training samples are considered in the fine-tuning stage.

I. INTRODUCTION

With the emergence of the Internet of Things (IoT), the development of contextual human sensing applications has become increasingly popular. Research in this area has made significant progress recently. For example, applications such as health monitoring within indoor environments based on Human Activity Recognition (HAR) has become feasible [1]. Several WiFi sensing techniques have the potential for achieving non-restrictive, privacy-friendly indoor activity sensing when compared to current technologies such as cameras or wearable sensors [2]. Among the various WiFi sensing techniques, WiFi Channel State Information (CSI) has gained popularity over the past few years due to its extensive coverage in modern indoor settings, while providing fine-grained physical layer information that is useful for activity recognition [3]. The idea behind radar sensing involves detecting and identifying physical layer changes (e.g., Doppler and phase shifts, multipath propagation, signal attenuation, etc.) in the radar signal.

Recently, research in radio based human sensing has moved towards deep learning approaches which have demonstrated success due their ability to learn in complex environments. One of the most popular deep learning architectures used for radar and WiFi classification is the Convolutional Neural Network (CNN). CNNs have shown much success in radar and WiFi sensing applications, such as activity recognition [4]–[6], localisation [7], and gesture recognition [8]. However, these approaches typically require significant amounts of training data, which can be costly to obtain. Furthermore, generalisation to different environments has proved to be particularly challenging for radar and WiFi classification as the multipath signal propagation is environment dependent [9]. The main contributions of this work are the following:

- We realize a self-supervised contrastive learning network for passive WiFi-based HAR where the user is oblivious to the Radio-Frequency (RF) sensing process (i.e., no body sensors worn). Basically, we leverage multiple views of synchronised WiFi CSI signals and implement a self-supervised HAR model by innovatively combining a contrastive learning framework with multi-view representation.
- Through numerous experiments, we show that the proposed contrastively trained system can increase the activity recognition performance using a suitable network model architecture over a non-contrastive baseline. Basically, we assess the effectiveness of the proposed self-supervised system with limited labelled examples of each activity demonstrating that high performance can be achieved with relatively few labelled data points in a few-shot activity recognition scenario.

The remainder of this paper is organised as follows: An overview of related work is provided in Section II. Section III introduces the concept of representation learning with contrastive losses. Section IV provides a description of the contrastive learning framework which is used as a pretraining process for WiFi-based HAR. The experimental setup and system parameters are described in Section V. In Section VI, we perform experiments to evaluate the effect of the pretraining process on the activity recognition performance ($F_1$ score) under different conditions, such as using different encoder networks and sampling in the fine-tuning stage. Finally conclusions are drawn at the end of this paper.

II. RELATED WORK

Recently, the simple framework for contrastive learning, coined SimCLR [10], has been investigated for HAR using wearable sensor data [11] where the authors use several combinations of signal transformations for augmenting the time-series data (acquired from accelerometer and gyroscope
sensors on a mobile device [12]) instead of image augmentation operators. The representations learned by the contrastive learning framework led to better performance compared to fully-supervised training when the models were fine-tuned with labels. However, additional investigation is needed to explore under which conditions or signal transformations for data augmentation, SimCLR is beneficial for HAR systems based on wearable sensor data. In the same context, the authors of [13] investigated the effectiveness of Contrastive Predictive Coding (CPC) framework for HAR when applied to various wearable sensor datasets. CPC, which is used as a self-supervised pretraining step (pretext task), is intended to capture the long-term temporal structure of sensor data streams. After the pretraining stage, fine-tuning is performed with small amounts of labelled training data, which involves performing activity recognition on the learned representations using a classifier. Compared to other state-of-the-art supervised and unsupervised approaches, the results with CPC pretraining showed improved activity recognition performance (best mean $F_1$ score of 89.05% on MotionSense dataset [12]). The authors of [14] use a combination of a 1D CNN with a transformer encoder as a backbone model as part of the SimCLR framework to learn feature representations from unlabelled wearable sensor data. They propose to use random combinations of simple time-series augmentations (to enhance the quality of learnt embeddings) within a random augmentation module to obtain multiple views originating from the initial time-series instances during the pretraining stage. Their contrastive self-supervised learning approach achieves the best mean $F_1$ score of 91.14% on the UCI-HAR dataset [15] when compared to other self-supervised methods. This score is also higher than the one achieved by DeepConvLSTM [16] which is trained in a supervised manner on the same dataset (mean $F_1$ score of 82.83%).

III. REPRESENTATION LEARNING WITH CONTRASTIVE LOSSES

The goal of contrastive learning is to build a better data representation via judicious design of auxiliary tasks (here contrastive losses). Crucially, the entire pipeline is completely automated (‘no labels’). In essence, we are attempting to build the function $h: X \rightarrow Y$, which maps the input data $X$ to some latent compact representation space $Y$. The mapping should maximise the mutual information $I(X; Y)$, while attempting to minimise the size of $|Y|$. As a result, $Y$ is made to capture the salient information about the data, while removing all spurious redundancy, so that subsequent classification tasks can be achieved relatively easily with just a handful of labelled examples (this is referred to as the fine-tuning stage). However, direct maximisation of mutual information is a computationally intractable problem, and instead [17] shows that lower bound on mutual information can be maximised by minimizing the contrastive loss i.e. $-L_{\text{cont}} < I(X; Y)$.

The goal of contrastive loss $L_{\text{cont}}$ is to minimise the difference between the learned representations of positive pairs of data and maximise the difference between the negative pairs [10]. Generally, the positive pair consists of two samples of the same data point, which differ in some way, while the negative pair consists of two samples belonging to a different data point. For example, SimCLR [10] is a popular contrastive learning approach, originally proposed in computer vision under a self-supervised setting and has been shown to be very successful. Under the paradigm, a batch of images undergoes augmentation, such as rotation, crop and colouration. During the training, the two augmented samples originating from the same image are described as the positive pair, while the two augmented samples originating from different images are identified as the negative pair. Contrastive learning has been used to learn crossmodal representations of audio-visual information [18] and to learn spatio-temporal features of the scenes [19].

IV. CONTRASTIVE WiFi-BASED HAR SYSTEM

Related works on contrastive learning when applied to wearable sensor data have considered various transformation functions for data augmentation. The latter is required in the pretraining stage and the type of data augmentation chosen ultimately affects the performance of the network. In contrast, we do not perform any kind of data augmentation as we leverage synchronised WiFi data that is collected simultaneously from different receivers (views), as the pairs. The training objective is to encourage the representations of the synchronised data, the different views, to be closer in the embedded space. These views represent the identical semantics of the environment and the activity at a given time-point. We propose that by training the model contrastively in this way, the model learns the inherent features that are invariant to the noise involved in such systems. Furthermore, we propose that this form of self-supervision will better utilise the data collected and improve activity recognition performance using WiFi-CSI.

For each received packet, the raw physical layer CSI measurement is obtained from the Intel 5300 Network Interface Card (NIC) as a 3D matrix of complex values, $n_t \times n_r \times N_{sc}=3 \times 3 \times 30$, where $n_t$ is the number of transmit antennas, $n_r$ is the number of receive antennas and $N_{sc}$ is the number of Orthogonal Frequency Division Multiplexing (OFDM) subcarriers [20]. The CSI is the channel estimate which is used during the equalisation stage in the WiFi receiver to reverse the adverse effects of the channel on the transmitted signal. We use a similar signal processing pipeline as in our previous works [6], [21], [22] to convert the raw CSI data into an image-like format known as spectrogram (time-frequency representation of human activities). Typically, the main signal processing steps consist of (i) denoising the raw WiFi CSI amplitude data using Discrete Wavelet Transform (DWT), (ii) segmenting the meaningful CSI variations caused by a human activity into fixed duration windows using ground truth activity labels and moving variance, (iii) dimensionality reduction using Principal Component Analysis (PCA) and (iv) conversion to spectrograms using Short-Time Fourier Transform (STFT).
The generated spectrograms serve as input to our contrastive learning system. We realize our self-supervised contrastive learning HAR network from multiple synchronised views of WiFi signals by innovatively introducing and combining SimCLR [10] and the Contrastive Multiview Coding [23] in a CSI perspective. Fig. 1 shows the overview of the pretraining contrastive learning stage. With the multi-view spectrograms as input, we use the Normalised Temperature Cross-entropy (NT-Xent) proposed by [10] to calculate the contrastive loss, which works as follows. As we train with mini-batches, we obtain 2N projections that result from applying a projection network to our embeddings of each pair, from each mini-batch with N samples with two views. We then form positive pairs if the two projections originate from the same time-point but different views, and negative pairs otherwise. Under the mini-batch, each projection forms exactly one positive pair and 2N − 1 negative pairs. To calculate the loss, we first compute the pairwise similarity $s$ for every available pair of projections $z_i$ and $z_j$ as follows:

$$s_{i,j} = \frac{x_i^T \cdot z_j}{\|x_i\| \|z_j\|}.$$  \hfill (1)

We apply an exponential function on each pair, and divide the positive pair by the sum of the negative pairs. Then we take the mean value of the losses to backpropagate through the networks. After pretraining, we discard the projection heads, freeze the encoders, add a classification network on top of the encoders and fine-tune them with labelled activity samples.

V. Experiments

A. Experimental setup

During the data collection in each experiment room, we used two WiFi CSI receivers, primary and secondary, as depicted by the green boxes CSI-RX1 and CSI-RX2 (with corresponding $x$-$y$ coordinates) in Fig. 2, respectively (the complete dataset details are available in [24]). The WiFi CSI transmitter is depicted by the red box in each room. To record the CSI, the injector mode was used whereby a Next Unit of Computing (NUC) device was configured as the transmitter (injector) while two other NUCs were configured as receivers and they were monitoring the channel into which packets were injected. This method requires that both the transmitter and receiver be equipped with the Intel 5300 NIC. The parameters of the CSI system are summarized in Table I. The working principles of the CSI system are thoroughly described in [22].

As shown in Fig. 2, the monitoring areas are of dimensions 4.46 m $\times$ 4.06 m and 4.06 m $\times$ 4.53 m for Rooms 1 and 2, respectively. Six participants performed six different activities within these enclosed spaces, namely, laying down (“laydown”), sitting on a chair (“sit”), standing from chair (“stand”), standing from the floor (“standff”), walking (“walk”) and body rotating (“bodyrotate”). The subjects were free to choose where they want to perform the activities within the monitoring spaces. Furthermore, we captured the participants’ motion and natural behaviour as would be the case in a real-world situation.

B. Contrastive Model Details

For the experimental evaluations, we test three popular CNN architectures as the encoder (backbone model) in the

![Fig. 1. An overview of the WiFi based contrastive pretraining. In a feed-forward pass with a mini-batch of N spectrogram samples, each sample $x_i$, $i \in \{1, 2, \ldots, N\}$ contains two views $m \in \{1, 2\}$. In the figure, the green colours refer to a positive pair of samples, i.e., the signal at two different synchronised receivers, while the red refers to negative samples belonging to a different time-point. Each view of the sample $x_i^m$ is input into the corresponding encoder network $f_{\theta_i^m}$ to obtain the embedding $h_i^m$, which then further passes through the projection network $g_{\theta_i^m}$ to obtain the projection $z_i^m$. After pretraining, the projection networks are discarded and the weights of the encoder networks frozen. For activity recognition, a classification network is added on top of the encoders and fine-tuned with examples of labelled activities.](image1)

![Fig. 2. Experiment room layouts.](image2)

### Table I WiFi CSI System Parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>WiFi Band</td>
<td>5 GHz (channel 149)</td>
</tr>
<tr>
<td>NIC</td>
<td>Intel 5300</td>
</tr>
<tr>
<td>Subcarriers, $N_{sc}$</td>
<td>30</td>
</tr>
<tr>
<td>Antenna</td>
<td>omni-directional (6 dBi)</td>
</tr>
<tr>
<td>Packet injection rate</td>
<td>1600 Hz</td>
</tr>
<tr>
<td>No. of transmit antennas, $n_t$</td>
<td>3</td>
</tr>
<tr>
<td>No. of receive antennas, $n_r$</td>
<td>3</td>
</tr>
</tbody>
</table>
contrastive learning framework: AlexNet, VGG16, ResNet18 and one shallow network to analyse the effects of model architecture on the contrastive learning performance. The shallow network consists of three convolutional layers with 32, 64 and 96 filters, with kernel sizes of 5, 4 and 3, respectively. Each convolutional layer is followed by a batch normalisation layer and a maximum pooling layer. ReLU is used as the activation function for the first two layers while Tanh is used as the activation function for the third convolutional layer. A global 2D maximum pooling layer is added at the end. Note that the spectrograms are resized to dimensions 224×224 since this is a typical image size that is commonly used in the above-mentioned CNN architectures.

For the pretraining phase, the projection head is a linear layer with a size of 128 units. By default, the following parameters are used during pretraining: A Stochastic Gradient Descent (SGD) optimizer with a learning rate 0.0005, pretraining batch size of 64 and 750 epochs. The temperature parameter, \( \tau \), is set at 0.1. As for the fine-tuning phase, the classification head is a multi-layer perceptron with two hidden layers consisting of 128 and 6 neurons, respectively. Leaky ReLU activation function is applied to the first layer with a dropout rate of 0.1. For the fine-tuning evaluation, a batch size of 64 and Adam optimizer with a learning rate of 0.0005 are used for 200 epochs. In the fine-tuning stage, the projection head is discarded, the layers of the backbone encoder are frozen, and the classification head is attached to the encoder which is fine-tuned with labelled activity samples.

C. Training and Evaluation

We randomly select 80% of the samples in the dataset as the training set and the remaining 20% is used to evaluate the model. All of the training set was used in the pretraining phase for contrastive learning. In the fine-tuning phase, we randomly select one, five and ten labelled examples per class to evaluate the few-shot learning capability of the model. These correspond to 0.26%, 1.29% and 2.58% of labelled training samples, respectively, that are used to train the supervised and self-supervised models in the fine-tuning phase.

VI. Results

We first present the results in Fig. 3 which compares the macro averaged \( F_1 \) score over 200 epochs between the non-contrastively-trained system and contrastively-trained systems.

The baseline (non-contrastive) model basically consists of training the backbone encoder and the classifier at the same time (i.e., regular supervised training with cross entropy loss). Both approaches are based on an AlexNet architecture, and consider as input the WiFi CSI spectrograms derived from two different receivers. In this evaluation, we consider 100% of the labelled training samples in the fine-tuning stage. Furthermore, we also include the results with the Supervised Contrastive Learning (SCL) framework, coined SupCon, proposed in [25]. SCL extends the self-supervised SimCLR method to a fully-supervised setting, allowing the effective use of label information [25]. SCL brings together clusters of data points pertaining to the same class in the embedding space while at the same time pushing apart clusters from different classes [25]. The superior performance of SupCon is reflected in Fig. 3 where it achieves a macro \( F_1 \) score of 90.6% at the 200th epoch, compared to 85.1% and 87.1% for SimCLR and the supervised non-contrastive model (with cross entropy loss function), respectively. Referring to Fig. 4(b), we can also observe that in the embedding space, SupCon results in more structured clusters than SimCLR (which presents relatively vague clusters in Fig. 4(a)), as well as the supervised non-contrastive model (Fig. 4(c)). Although SupCon shows a clear benefit over SimCLR, this approach is strictly supervised, which limits its applicability to large amount of labelled data only.
Furthermore, we compare our contrastive learning system with a non-contrastively trained version under different training regimes, as illustrated in Fig. 5. An AlexNet-based backbone encoder is used in this evaluation and 100% of the labelled training samples are considered in the fine-tuning stage. In this experiment, “Non-Contrastive (first)” refers to CSI data from the primary receiver (CSI-RX1) only, while “Non-Contrastive (joint)” refers to CSI data from both receivers (CSI-RX1 and CSI-RX2) but is still not contrastively trained. It can be observed from this figure that the inclusion of the data from the second CSI receiver slightly improves the performance under a normal supervised training regime (87.1% for “first” and 90.9% for “joint”). The contrastive learning system achieves a maximum macro $F_1$ score of 85.1%. We shall see in further evaluations that the self-supervised contrastive learning system is most beneficial under few-shot learning scenarios, that is, when the amount of available labelled training data is extremely limited.

A. Sample Efficiency

Next, we evaluate the performance of the system in one-shot and few-shot activity recognition scenarios using the AlexNet backbone encoder model. The results of this can be seen in Fig. 6, where the number of labelled samples in the classification stage was reduced to one, five and ten examples of each activity, corresponding to 0.26%, 1.29% and 2.58% of labelled training samples, respectively. The results demonstrate that the use of contrastive pretraining on synchronised WiFi CSI from multiple receivers can significantly improve activity recognition performance in one- and few-shot learning scenarios. For instance, the percentage increases in macro $F_1$ score with the AlexNet-based contrastive system are around 24%, 22% and 7% for one, five and ten examples per class, respectively.

Figures 7 and 8 show the confusion matrices of the non-contrastively and contrastively pretrained systems (AlexNet backbone), respectively, when only 1.29% (five examples per class) of labelled training samples are used in the fine-tuning phase. A noticeable increase in accuracy is observed across all six activities, with the activity “sit” achieving the highest improvement of 36%, followed by “walk” (24% increase), “bodyrotate” (19% increase), “stand from floor (standff)” (18% increase), “laydown” (5% increase) and finally “stand” (2% increase). Overall, the introduction of contrastive learning leads to a performance increase in the accuracy of approximately 24% and macro $F_1$ score of 22%.

B. Encoder Architectures

Fig. 9 compares the performance of the non-contrastive and contrastively pretrained models with different encoder architectures. Only 1.29% (five examples per class) of labelled training samples are considered in the fine-tuning phase. From this, we can observe that the difference in macro $F_1$ score between non-pretrained and pretrained models is largest with an AlexNet-based CNN encoder. Furthermore, the same encoder also achieves the best overall activity recognition performance (macro $F_1$ score of 67.3%).
In this work, we propose a system which uses contrastive learning on WiFi data to improve activity recognition performance. Specifically, the system utilises multiple views from synchronised receivers for contrastive learning. We evaluate the performance of our proposed model using experimental data consisting of six activities recorded from six human participants. Through experimental evaluations we show significant improvement in the activity recognition performance using self-supervised contrastive learning when compared to conventional supervised models under few-shot learning scenarios. Specifically, contrastively pretraining with an AlexNet-based encoder lead to a 22% increase in macro $F_1$ score when only 1.29% of labelled training samples are considered in the fine-tuning phase.

Acknowledgements

This work was performed as a part of the OPERA Project, funded by the UK Engineering and Physical Sciences Research Council (EPSRC), Grant EP/R018677/1.

References


