# Deep Auto-Set: A Deep Auto-Encoder-Set Network for Activity Recognition Using Wearables

Alireza Abedin Varamin
University of Adelaide
Adelaide, South Australia
alireza.abedinvaramin@adelaide.edu.au

Ehsan Abbasnejad
University of Adelaide
Adelaide, South Australia
ehsan.abbasnejad@adelaide.edu.au

Qinfeng Shi
University of Adelaide
Adelaide, South Australia
javen.shi@adelaide.edu.au

Damith Ranasinghe
University of Adelaide
Adelaide, South Australia
damith.ranasinghe@adelaide.edu.au

Hamid Rezatofighi
University of Adelaide
Adelaide, South Australia
hamid.rezatofighi@adelaide.edu.au

## ABSTRACT

Automatic recognition of human activities from time-series sensory data (referred to as HAR) is a growing area of research in ubiquitous computing. Most recent research in the field adopts supervised deep learning paradigms to automate extraction of intrinsic features from raw signal inputs and addresses HAR as a multi-class classification problem where detecting a single activity class within the duration of a sensory data segment suffices. However, due to the innate diversity of human activities and their corresponding duration, no data segment is guaranteed to contain sensor recordings of a single activity type. In this paper, we express HAR more naturally as a *set prediction problem* where the predictions are *sets* of ongoing activity elements with unfixed and unknown cardinality. For the first time, we address this problem by presenting a novel HAR system that learns to output activity sets using deep neural networks. Moreover, motivated by the limited availability of annotated HAR datasets as well as the unfortunate immaturity of existing unsupervised systems, we complement our supervised set learning scheme with a prior unsupervised feature learning process that adopts convolutional auto-encoders to exploit unlabeled data. The empirical experiments on two widely adopted HAR datasets demonstrate the substantial improvement of our proposed methodology over the baseline models.

## CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous computing**; • **Computing methodologies** → **Supervised learning by classification**; **Unsupervised learning**; **Neural networks**; **Multi-task learning**; **Learning latent representations**;

## KEYWORDS

Activity Recognition, Deep Learning, Time-series Data, Wearable Sensors


*MobiQuitous'18, Nov 2018, New York, USA*

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

**ACM Reference Format:**
Alireza Abedin Varamin, Ehsan Abbasnejad, Qinfeng Shi, Damith Ranasinghe, and Hamid Rezatofighi. 2018. Deep Auto-Set: A Deep Auto-Encoder-Set Network for Activity Recognition Using Wearables. In *Proceedings of 15th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous'18).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.475/123_4

## 1 INTRODUCTION

With the proliferation of low-cost sensing technologies as well as the fast advancements in machine learning techniques, automatic human activity recognition (HAR) using wearable sensors has emerged as a key research area in ubiquitous computing. In this problem, high-level activity information is acquired through analyzing low-level sensor recordings with the goal of providing proactive assistance to users. Having created new possibilities in diverse application domains ranging from health-care monitoring to entertainment industry, HAR has successfully sparked excitement in both academia and industry. Nevertheless, due to the inherent diverse nature of human activities, HAR faces unique methodological challenges such as intra-class variability, inter-class similarity, class imbalance and the Null class problem [4]. Accordingly, it is of great significance to propose systematic approaches towards accurate recognition of activities that triumph over the challenges.

While previous studies have explored both shallow and deep architectures for a diverse range of HAR application scenarios, multi-class classification has been their dominant approach for formulating the problem. As such, sensor time segments obtained from striding a fixed-size sliding window over the sensor data-streams are assigned a single activity class, approximated based on the most [22] or the last [16] observed sample annotations. Such a strategy towards ground-truth approximation is clearly associated with loss of activity information and potentially deludes the supervised training process. This becomes even more problematic since the optimal size for the sliding window is not known a priori [4] and therefore, no segment is guaranteed to contain measurements of a single activity type. Equally important, existing deep HAR systems demand large amounts of annotated training data for enhanced supervised performance. Quiet on the contrary, large-scale annotated HAR datasets are limited and further collection of labeled sensory data is labor intensive, time-consuming and expensive [11]. As opposed to

other domains (*e.g.* image recognition) where human visualization of raw data alleviates the labeling process, manual annotation of sensor signals is a tedious task. Unfortunately, activity recognition systems that leverage the cheaply available unlabeled sensory data are rare in the field which necessitates exploration of effective unsupervised alternatives.

In this paper, we overcome the innate limitations of multi-class formulated HAR by expressing the problem more naturally as a *set prediction problem*. As such, the goal is to predict the *set* of ongoing activity elements (whose cardinality is unknown and unfixed beforehand) within the duration of a time segment. For instance, considering a sensory time segment in which the subject of interest is initially walking but then suddenly stops moving, the system is expected to output the set {walk,stand} to capture the underlying activity transition. Similarly, an output empty set {} intuitively expresses a time segment in which the activities of interest did not occur. Inspired by Rezatofighi *et al.* [20], for the first time we develop an HAR system that performs activity set learning and inference in a principled fashion using deep paradigms. By contrast to conventional multi-label solutions, our principled methodology omits threshold heuristics and instead exploits activity cardinality information to generate outputs. Further motivated by the scarcity of annotated HAR datasets, we complement our supervised training scheme with a prior unsupervised feature learning step that exploits unlabeled time-series data. Through empirical experiments on widely adopted public HAR datasets, we demonstrate the significant improvement of our proposed deep learning based methodology, the *Deep Auto-Set* network (depicted in Fig. 1), over the baseline models. The main contributions of this paper are summarized as follows:

- For the first time, we investigate a novel formulation of HAR where the predictions for sensory time segments are expressed as *activity sets*. Our novel formulation naturally handles sensory segments with varying number of activities and thus, bypasses the conventional ground-truth approximations.
- We present deep Auto-Set: a unified deep learning paradigm that (a) seamlessly functions on raw multi-modal sensory segments, (b) exploits unlabeled data to uncover effective feature representations, and (c) incorporates set objective to learn mappings from input sensory data to target activity sets.
- We demonstrate the effectiveness of our deep Auto-set network through empirical experiments on two HAR representative datasets. We further examine the components of our proposed methodology in isolation, to present insights on their contribution to an enhanced recognition performance.

## 2 RELATED WORK

The well-established activity recognition pipeline for time-series sensory data involves sliding window segmentation, feature extraction, and activity classification [4]. In this regard, adopting hand-crafted features (e.g. statistical features [19], basis transform features [10] and multi-level features [24]) coupled with employment of shallow classifiers (e.g. support vector machines [5], decision trees [3], joint boosting [14] and multi-layer perceptrons

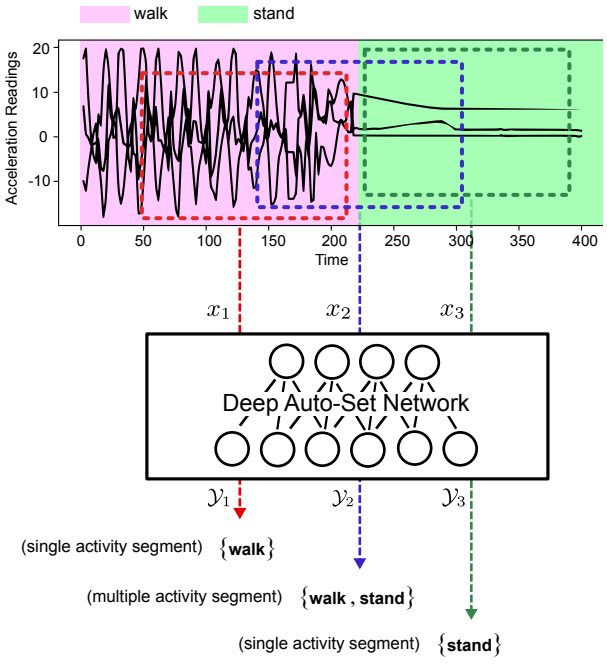


**Figure 1: An illustration of our novel *Deep Auto-Set* network to perform precise activity recognition from time-series data. Our network consumes windowed raw sensory excerpts ($x$), automatically extracts distinctive features and outputs corresponding *sets of activities* ($\mathcal{Y}$) with various cardinalities.**

[18]) has been extensively explored as the traditional approach to HAR [13]. While this manually tuned procedure has successfully acquired satisfying results for relatively simple recognition tasks, its generalization performance is limited by heavy reliance on domain expert knowledge to design distinctive features.

Recently, the emerging paradigm of deep learning has presented unparalleled performance in various research areas including computer vision, natural language processing and speech recognition [15]. When applied to sensor-based HAR, deep learning allows for automated end-to-end feature extraction and thus, largely alleviates the need for laborious feature engineering procedures. Motivated by these, an inertia towards the adoption of deep learning paradigms in HAR has been witnessed [9]. In this regard, convolutional neural networks (CNNs) have appeared as the most popular choice for automatic extraction of effective high-level features. Research in this line includes [22, 23] where raw sensory data were processed by convolutional layers to extract discriminative features. Going beyond CNNs, Hammerla *et al.* [9] conducted extensive experiments to investigate suitability of various deep architectures for HAR using wearables and concluded guidelines for hyper-parameter tuning in different application scenarios. Ordóñez and Roggen [16] developed a recurrent-based neural network (RNN) for wearable sensors and reported state-of-the-art performance on a representative HAR dataset. However, existing proposed supervised solutions are based on a strict assumption that all samples within a sliding

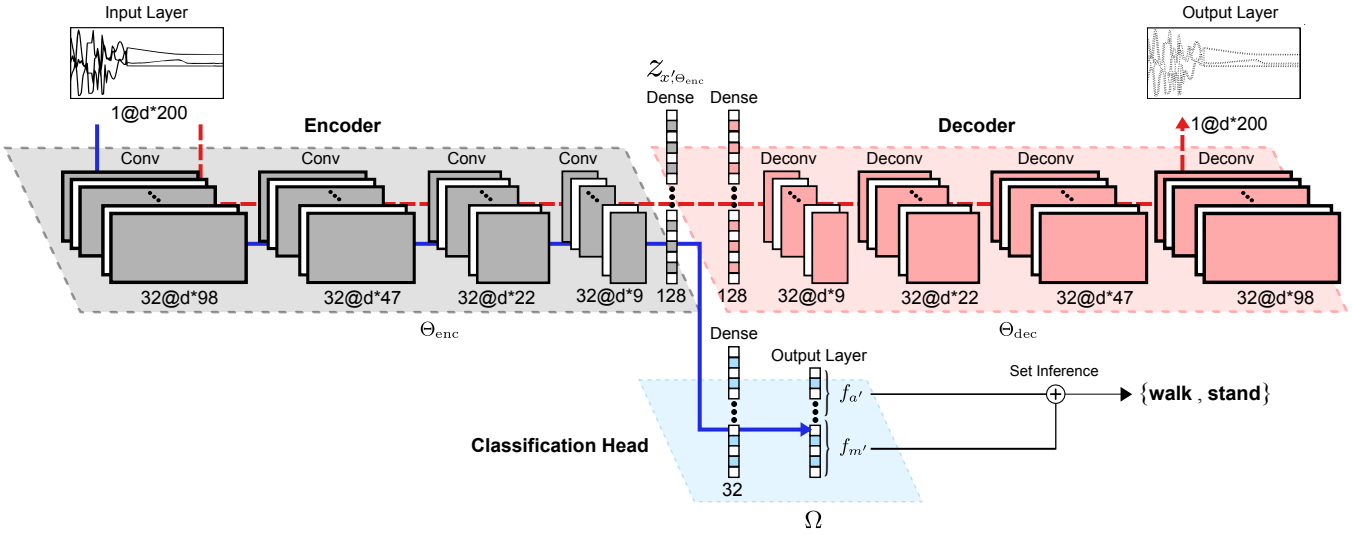


**Figure 2: Unified architecture of our deep *Auto-Set* network. The tags above the feature maps refer to the corresponding layer operations. The numbers before and after "@" respectively correspond to the number of generated feature maps and their dimensions in each layer. In this architecture, all convolution (and deconvolution) layers apply a filter of width 5 (as in [16]) and stride 2 (for down-sampling) along the temporal dimension of the feature maps. For the unsupervised step, starting from the input layer, layer operations on the dashed arrow are consecutively applied on the generated feature maps of previous layers to output the reconstructed segment; these operations correspond to the convolutional auto-encoder network parameterized by $\Theta_{enc}$ and $\Theta_{dec}$. Similarly for the supervised step, operations on the solid arrow correspond to the activity set network parameterized by $\Theta_{enc}$ and $\Omega$. Once the network parameters are optimized, set inference (as described in Section 3.2) is carried out to generate activity set predictions.**

window segment share the same activity annotation. We argue that such assumption is counter-intuitive to the diverse nature of human activities with varying durations and hinders accurate analysis of segments with multiple activities. In this paper, we lift the conventional restriction in HAR problem and present a novel network that naturally allows segmented sensory data be associated with a set of activity elements. Moreover, most existing HAR research solely rely on supervised training for feature extraction. In the absence of sufficiently large annotated datasets this leads to poor generalization performances. Taking into account the apparent scarcity of annotated HAR datasets, we exploit unlabeled time-series data to learn useful feature representations by adopting convolutional auto-encoders. In this regard, the most relevant study to ours is [2] where layer-wise pre-training of fully connected deep belief networks is adopted and the recognition problem is limited to preprocessed spectrograms of acceleration measurements. By contrast, our proposed unsupervised methodology substitutes the layer-by-layer pre-training with an end-to-end optimization of reconstruction objective and is also seamlessly applied on raw multi-modal sensor data.

## 3 DEEP AUTO-SET FOR HUMAN ACTIVITY RECOGNITION

Here we elaborate on our novel methodology towards addressing HAR as a set prediction problem, which we refer to as the deep *Auto-Set*. The working flow of our proposed solution involves an unsupervised feature learning step (described in Section 3.1) that exploits cheaply accessible unlabeled sensor measurements followed by a supervised fine-tuning step (detailed in Section 3.2) that leverages valuable label information to extract more discriminative features while simultaneously training the network to generate activity sets for the given sensory data. Noting that our methodology is not confined to a specific network architecture, we carry out both supervised and unsupervised tasks by adopting a CNN architecture employed in [16] as the core of our network and apply modifications to suit our problem settings; this architecture comprises four convolutional layers followed by two dense layers that apply rectified linear units (ReLUs) for non-linear transformation as well as a softmax logistic regression output layer to yield the classification outcome. Specifically for the unsupervised feature learning step, we construct a symmetric convolutional auto-encoder by arranging a chain of deconvolutional operations in the decoder network symmetric to the convolutional layers in the encoder network. This choice is grounded over the success of auto-encoders in improving generalization performance through unsupervised feature learning [7]. In addition, for the supervised activity set learning step, the encoder network is augmented with a multi-label classification head and the output layer is adjusted to suit the set formulation. The overall architecture of our deep *Auto-Set* network is illustrated in Fig. 2. In the proposed architecture, all convolution (and deconvolution) operations are applied along the temporal dimension of the feature maps, automatically uncovering temporal signal patterns within the timespan of the filters.

In order to provide a clear formulation of the problem, here we introduce the notations used throughout this paper. In this paper, we use $\mathcal{Y}$ for a set with unknown cardinality and $\mathcal{Y}^m$ for a set with known cardinality $m$. Let us define the set of $M$ supported activity elements by $\mathcal{A} = \{a_i\}_{i=1}^{M}$. Consider a collected data stream which contains raw time-series recordings from $d$ distinct sensor channels. We assume that for a subset of the recordings, sample annotation is not provided. Accordingly, adopting time-series segmentation with a sliding window size of $w$ on the data stream results in 1) a labeled training dataset $\mathcal{S} = \{(\boldsymbol{x}_i, \mathcal{Y}_i^{m_i})\}_{i=1}^{N_1}$ of size $N_1$, where each training instance is a pair consisting of a sensory segment $\boldsymbol{x}_i \in \mathbb{R}^{d\times w}$ with a fixed 2D representation and a target activity set $\mathcal{Y}_i^{m_i} = \{a_1, \ldots, a_{m_i}\} \subseteq \mathcal{A}, |\mathcal{Y}_i| = m_i, m_i \in \mathbb{Z}^+$, as well as 2) an unlabeled dataset $\mathcal{V} = \{\bar{\boldsymbol{x}}_i\}_{i=1}^{N_2}$ of size $N_2$, where each instance is an unlabeled sensory segment $\bar{\boldsymbol{x}}_i \in \mathbb{R}^{d\times w}$. In order to leverage a larger number of segments for the unsupervised feature learning task, we define the unlabeled training dataset $\mathcal{U} = \{\boldsymbol{x}_i'\}_{i=1}^{N_1+N_2} = \mathcal{V} \cup \{\boldsymbol{x}_i\}_{i=1}^{N_1}$ where each training instance $\boldsymbol{x}_i' \in \mathbb{R}^{d\times w}$ is either a segment whose target activity set was not provided in the first place or a segment whose target set was intentionally discarded to augment the unlabeled dataset.

## 3.1 Unsupervised Feature Learning

Through stacked hidden layers of encoding-decoding operations, auto-encoder learns latent representations of the sensory data in an unsupervised fashion. The reconstruction of unlabeled segments captures the process in which the sensor signals are generated and allows for the correlation between various sensor channels be captured. Thus, the latent representations learned by the auto-encoder serve as efficient features that are highly effective in discriminating activity patterns. Formally, the input to the convolutional auto-encoder network is an unlabeled sensory time segment $\boldsymbol{x}' \in \mathcal{U}$ on which the encoder network $f_{\text{enc}} : \mathbb{R}^{d\times w} \rightarrow \mathbb{R}^p$ (parameterized by $\Theta_{\text{enc}}$) is firstly applied to obtain the latent representation $\boldsymbol{z}_{\boldsymbol{x}',\Theta_{\text{enc}}}$, *i.e.*

$$\boldsymbol{z}_{\boldsymbol{x}',\Theta_{\text{enc}}} = f_{\text{enc}}(\boldsymbol{x}'; \Theta_{\text{enc}}). \tag{1}$$

The resulting latent representation $\boldsymbol{z}_{\boldsymbol{x}',\Theta_{\text{enc}}} \in \mathbb{R}^p$ is then utilized by the decoder network $f_{\text{dec}} : \mathbb{R}^p \rightarrow \mathbb{R}^{d\times w}$ (parameterized by $\Theta_{\text{dec}}$) to reconstruct the input. Noting that the generated reconstruction is directly influenced by the values of $\Theta_{\text{enc}}$ and $\Theta_{\text{dec}}$, we define the loss incurred by the output of auto-encoder network (illustrated by the dashed path in Fig. 2) given the unlabeled segment $\boldsymbol{x}'$ as

$$\mathcal{L}_{\text{auto}}(\boldsymbol{x}'; \Theta_{\text{enc}}, \Theta_{\text{dec}}) = \|\boldsymbol{x}' - f_{\text{dec}}(\boldsymbol{z}_{\boldsymbol{x}',\Theta_{\text{enc}}}; \Theta_{\text{dec}})\|^2. \tag{2}$$

We adopt an end-to-end approach towards training the convolutional auto-encoder parameters by minimizing the reconstruction objective on the unlabeled dataset $\mathcal{U}$

$$(\Theta_{\text{enc}}^*, \Theta_{\text{dec}}^*) = \arg\min_{\Theta_{\text{enc}},\Theta_{\text{dec}}} \sum_{i=1}^{N_1+N_2} \mathcal{L}_{\text{auto}}(\boldsymbol{x}_i'; \Theta_{\text{enc}}, \Theta_{\text{dec}}). \tag{3}$$

In this architecture, the encoder network extracts features from unlabeled data and the decoder network uses the learned features to reconstruct the input. As the unsupervised training process progresses and the corresponding reconstruction loss is reduced, the network uncovers better feature representations of the sensory data. As a result, the acquired encoder network weights ($\Theta_{\text{enc}}^*$) can later be adopted in favor of a better guided supervised training.

## 3.2 Supervised Activity Set Learning and Inference

Using the labeled training dataset $\mathcal{S} = \{(\boldsymbol{x}_i, \mathcal{Y}_i^{m_i})\}_{i=1}^{N_1}$, the goal here is to train an activity set network that predicts a set of activity elements $\mathcal{Y}^+ = \{a_1, \ldots, a_m\}$ with unknown and unfixed cardinality $m$ for a given test sensor segment $\boldsymbol{x}^+$. In our architecture, this is carried out by optimizing a *set objective* through tuning the activity set network parameters which include weights corresponding to the encoder layers ($\Theta_{\text{enc}}$) as well as the extra dense layers ($\Omega$) in the classification head. Similar to [20], in this paper we adopt joint learning and inference to learn and predict activity sets for HAR which we describe in what follows.

*3.2.1 Set Learning.* In order to develop an accurate HAR system that meets the application demands, the network is required to correctly predict both the set cardinality (number of ongoing activities) as well as the set elements (activity types) given a sensory segment. Formally, given an input segment $\boldsymbol{x}$, the output of our activity set network comprises a 1) *set cardinality* term $f_{m'}(\boldsymbol{x})$ with log softmax activation which produces cardinality scores, as well as a 2) *set element* term $f_{a'}(\boldsymbol{x})$ with sigmoid activation which produces scores for the set elements (activity types). In order to compute the loss incurred by the output of the activity set network (shown by the solid path in Fig. 2) given a labeled segment $\boldsymbol{x}$ with the target set $\mathcal{Y}^m$, we define our set objective as

$$\begin{aligned}\mathcal{L}_{\text{set}}(\boldsymbol{x}, \mathcal{Y}^m; \Theta_{\text{enc}}, \Omega) = &\sum_{a\in\mathcal{Y}} \ell_{\text{bce}}(a, f_{a'}(\boldsymbol{x}; \Theta_{\text{enc}}, \Omega)) \\ &+ \ell_{\text{nll}}\left(m, f_{m'}(\boldsymbol{x}; \Theta_{\text{enc}}, \Omega)\right),\end{aligned} \tag{4}$$

where $\ell_{\text{nll}}$ and $\ell_{\text{bce}}$ denote the negative log likelihood loss and the binary cross entropy loss, respectively. We consider the same *i.i.d* assumption adopted in [20] for the set elements and perform MAP estimate to train the network parameters by minimizing the set objective on the labeled dataset $\mathcal{S}$, *i.e.*

$$(\Theta_{\text{enc}}^*, \Omega^*) = \arg\min_{\Theta_{\text{enc}},\Omega} \sum_{i=1}^{N_1} \mathcal{L}_{\text{set}}(\boldsymbol{x}_i, \mathcal{Y}_i^{m_i}; \Theta_{\text{enc}}, \Omega). \tag{5}$$

As such, $\Theta_{\text{enc}}^*$ and $\Omega^*$ are estimated by computing the partial derivatives of the objective function in Eq. (4) and employing standard backpropagation in order to learn the network parameters.

*3.2.2 Set Inference.* During the prediction phase for a given time segment $\boldsymbol{x}^+$, the goal is to predict the most likely set of activity elements $\mathcal{Y}^* = \{a_1, \ldots, a_m\}$. Using the optimal parameters $(\Theta_{\text{enc}}^*, \Omega^*)$ learned from the training dataset $\mathcal{S}$, a MAP inference is adopted to output the most likely activity set as

$$\begin{aligned}\mathcal{Y}^* = \arg\max_{m', \mathcal{Y}^{m'}} \; &f_{m'}\left(\boldsymbol{x}^+; \Theta_{\text{enc}}^*, \Omega^*\right) + m' \log U \\ &+ \sum_{a'\in\mathcal{Y}^{m'}} \log f_{a'}(\boldsymbol{x}^+; \Theta_{\text{enc}}^*, \Omega^*),\end{aligned} \tag{6}$$

where $U$, estimated from the validation set of the data, is a normalization constant that allows comparison between sets with different

cardinalities. We derive the optimal solution for the above problem by solving a simple linear program as suggested in [20].

# 4 EXPERIMENTS

## 4.1 Datasets

For the evaluation of our approach, we adopt two widely used public HAR datasets that present both periodic and static activities. These benchmarks are elaborated as follows:

- **WISDM Actitracker dataset** [12]: This dataset contains 1,098,207 triaxial accelerometer readings gathered from 36 users which reflect activity patterns of *walking*, *jogging*, *sitting*, *standing*, and *climbing stairs*. The acceleration measurements are collected with Android mobile phones at a constant sampling rate of 20 Hz. We randomly select recordings from 8 users as the testing set and use the remaining data as our training and validation sets.
- **Opportunity dataset** [6]: This dataset comprises annotated recordings from a wide variety of on-body sensors configured on four subjects while carrying out morning activities. The annotations include several modes of locomotion along with a *Null* activity (referring to non-relevant activities) which makes the recognition problem much more challenging. For data collection, subjects were instructed to perform five Activities of Daily Living (ADL) runs as well as a drill session with 20 repetitions of a predefined sequence of activities. Each sample in the resulting dataset corresponds to 113 real valued signal measurements recorded with a sampling rate of 30 Hz. We employ the same subset of data as in the Opportunity challenge [6] for training and testing purposes: ADL runs 4 and 5 collected from subjects 2 and 3 compose our testing set, and the remainder of the recordings from subjects 1,2 and 3 form our training and validation sets.

## 4.2 Data Preparation

The preparation process involves performing per channel normalization to scale real valued attributes to [0,1] interval as well as segmentation and ground-truth generation, as described in what follows:

**Time-series Segmentation:** Following the experiments in [2, 12], we fix the sliding window size $w$ to incorporate 200 samples for both datasets (i.e, segments of 10 and 6.67 seconds duration for Actitracker and Opportunity dataset, respectively). However, since using non-overlapping sliding windows hinders real-time recognition of human activities, we set the sliding window stride to 20 samples. Such a deployment setting leads to generating predictions every second for the Actitracker dataset and every 0.67 seconds for the Opportunity dataset, holding the potential for novel online applications in HAR.

**Set Ground-Truth Preparation:** Considering the sample annotations of a windowed sensory excerpt, the goal is to prepare the corresponding target set of activity elements as the training data. To this end, we consider a minimum *expected recognition length* denoted by $r$, based on which we include activities in the target set. As such, if a minimum of $r$ sample annotations from a specific activity are observed in a time segment, the activity label appears in the target set. If no activity persists for the duration of $r$, the target activity set is considered as an empty set {}, representing the Null activity segment. In our experiments, we set $r$ to half the sensor sampling rates; *i.e.*, 10 and 15 for Actitracker and Opportunity datasets, respectively.

## 4.3 Evaluation Metrics

We employ the widely used HAR evaluation measures to report the performance of the baselines and our deep *Auto-Set* network: mean per-label *precision* ($P_{\text{mean}}$), *recall* ($R_{\text{mean}}$) and *f1-score* ($F_{\text{mean}}$). For a specific activity label, label-based precision is defined as the ratio of the correctly predicted label occurrences over the total number of label occurrences in the predictions. Similarly, per-label recall is defined as the ratio of the correctly predicted label occurrences over the total number of label occurrences in the ground-truth. In this regard, per-label f1-score corresponds to the harmonic mean of precision and recall. Accordingly, $P_{\text{mean}}$, $R_{\text{mean}}$ and $F_{\text{mean}}$ are calculated by averaging their corresponding per-label values. We also use the overall *exact match ratio* (MR), as adopted in [1, 8], to report a harsh evaluation of models' performance. This metric requires the predicted activity set exactly match the corresponding target set (both in terms of the set cardinality and the set elements) and therefore, does not appreciate partially correct predictions. For instance, no credit is considered for a predicted set of {walk} when the target set is {walk,stand}. We further decompose this measure over different activity set cardinalities $c$ and additionally report $\text{MR}_c$; *i.e*, for instance $\text{MR}_2$ corresponds to the number of correctly predicted activity sets with cardinality of 2 over the total number of target sets with this cardinality.

## 4.4 Implementation Details

In this paper, the experiments are implemented using Pytorch [17] as the deep learning framework and are run on a machine with a single GPU (NVIDIA GeForce GTX 1060). The network parameters are learned using ADAM optimizer with weight decay and initial learning rate respectively set to $5 \cdot 10^{-5}$ and $10^{-4}$, on mini-batches of size 64 by backpropagating the gradients of corresponding loss functions. For the supervised training step, the optimizer learning rate is scheduled to gradually decrease after each epoch. Moreover, training is stopped if validation objective does not decrease for 5 subsequent epochs. Accordingly, the corresponding weights for the epoch with the best validation performance are applied to report performance on the testing sets. The hyper-parameter $U$ is set to be 2.5 and 3.4, respectively adjusted on the validation sets of Actitracker and Opportunity datasets. We refer interested readers to [9] for excellent guidelines on setting architecture and optimizer hyper-parameters.

## 4.5 Baselines and Results

A big motive for our work is the activity information loss that is incurred by conventional multi-class ground truth approximations. In order to verify this, we conform to the conventional multi-class formulation of HAR and train the CNN adopted in [16] by minimizing the multi-class classification objective. In Table 1, we report performance of the resulting HAR system by comparing the generated predictions against both the *approximate* multi-class ground truth (as reported in [16]) as well as the *exact* multi-label ground

**Table 1: Apparent performance degradation of the baseline CNN architecture [16] trained with multi-class objective when tested against multi-label (exact) ground-truth for Opportunity dataset.**

| Model | Training Ground Truth | Inference Ground Truth | $F_{\text{mean}}$ | MR |
|---|---|---|---|---|
| CNN [16] | Multi-class | Multi-class | 0.890 | 87.4% |
| | | Multi-label | 0.793 | 54.7% |

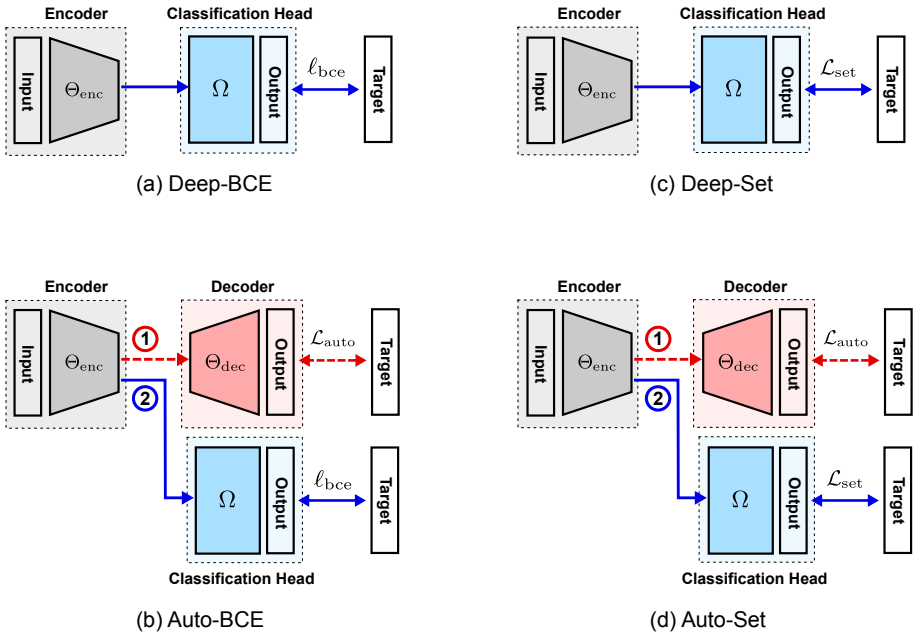


**Figure 3: An overview of different activity recognition models explored in this paper. Note that all models adopt the same network architecture to generate classification outputs and thus, share the same number of parameters. Therefore, the enhanced recognition performance is a product of effective unsupervised feature learning as well as incorporating novel loss functions for the underlying problem.**

truth for Opportunity dataset. The severe degradation of evaluation measures when tested against the exact ground truth confirms that the multi-class formulation of HAR reflects a poor representation of the ongoing activities in reality. As hypothesized, where multiple activities are present in a window, approximating the sensory segment's ground-truth introduces detrimental effects on the recognition system's performance. Grounded over these observations, we omit empirical comparisons with existing multi-class based solutions and instead present evaluation against multi-label based activity recognition systems that can handle segments with multiple activities in what follows.

**Activity Recognition Models:** Summarized in Fig. 3 are the schematic architectures for (*a*) *Deep-BCE*: a conventional multi-label model that follows a purely supervised minimization of binary cross entropy loss ($\ell_{\text{bce}}$) for training and heuristic thresholding of activity scores for inference, (*b*) *Auto-BCE*: a conventional multi-label model that leverages a prior unsupervised feature learning step via minimization of reconstruction objective ($\mathcal{L}_{\text{auto}}$) as well as a supervised optimization of binary cross entropy loss, and (*c*) *Deep-Set*: a set-based model that follows a purely supervised optimization of the set objective ($\mathcal{L}_{\text{set}}$) proposed in Eq. (4) for training and the MAP inference introduced in Eq. (6) for set inference. It should be noted that, as opposed to existing multi-class based HAR systems which are restricted to predict a single activity class even when an activity transition takes place, all recognition models adopted in this paper are capable of predicting multiple activities for a given sensory segment. We adopt the same layer operations presented in Fig. 2 for supervised and unsupervised training steps of the baseline models.

The performance results of our deep *Auto-Set* network and the baseline models on the two HAR representative datasets are shown in Table 2 and Table 3 for different evaluation metrics. From the reported results, it can clearly be seen that our novel deep *Auto-Set* network consistently outperforms the baselines on Actitracker and Opportunity datasets in terms of both F1-score and exact match ratio performance metrics, respectively obtaining a significant improvement of 3.9% and 2.3% over Deep-BCE model on average.

**Table 2: Comparison of our proposed deep *Auto-Set* network against the baselines according to the obtained exact match ratio for each dataset. The superior results are highlighted with boldface. Note that for the Actitracker dataset, sensor segments with cardinality of 0 (corresponding to Null segments) and 3 do not exist.**

| Dataset | Model | MR | $MR_0$ | $MR_1$ | $MR_2$ | $MR_3$ |
|---|---|---|---|---|---|---|
| Actitracker | (Baseline) Deep-BCE | 90.1% | - | 91.1% | 60.2% | - |
| | (Ours) Auto-BCE | 92.9% | - | 93.9% | 62.7% | - |
| | (Ours) Deep-Set | 93.2% | - | 93.9% | 71.5% | - |
| | **(Ours) Auto-Set** | **94.9%** | - | **95.5%** | **75.1%** | - |
| Opportunity (locomotions) | (Baseline) Deep-BCE | 82.0% | 70.7% | 85.0% | 84.9% | 68.3% |
| | (Ours) Auto-BCE | 83.1% | 73.7% | 85.1% | 85.3% | 69.9% |
| | (Ours) Deep-Set | 83.9% | 78.2% | 86.8% | 84.9% | 68.7% |
| | **(Ours) Auto-Set** | **84.9%** | **80.2%** | **87.1%** | **85.6%** | **75.6%** |

**Table 3: Comparison of our proposed deep *Auto-Set* network against the baselines according to the obtained mean f1-score ($F_{\text{mean}}$), precision ($P_{\text{mean}}$) and recall ($R_{\text{mean}}$) for each dataset. The superior results are highlighted with boldface.**

| Dataset | Model | $F_{\text{mean}}$ | $P_{\text{mean}}$ | $R_{\text{mean}}$ |
|---|---|---|---|---|
| Actitracker | (Baseline) Deep-BCE | 0.943 | 0.908 | 0.980 |
| | (Ours) Auto-BCE | 0.966 | 0.949 | 0.983 |
| | (Ours) Deep-Set | 0.961 | 0.943 | 0.980 |
| | **(Ours) Auto-Set** | **0.973** | **0.957** | **0.989** |
| Opportunity (locomotions) | (Baseline) Deep-BCE | 0.927 | 0.901 | 0.954 |
| | (Ours) Auto-BCE | 0.936 | 0.918 | 0.955 |
| | (Ours) Deep-Set | 0.934 | 0.915 | 0.955 |
| | **(Ours) Auto-Set** | **0.943** | **0.927** | **0.960** |

Moreover, the match ratios in Table 2 suggest that *Auto-Set* is a robust activity recognition system capable of *i*) distinguishing different activity classes accurately (implied from $MR_0$ and $MR_1$ values), *ii*) identifying activity transition segments (implied from $MR_2$ values) as well as *iii*) realizing very short appearances of human activities (implied from $MR_3$ values).

We summarize the experimental results on both datasets by concluding that:

- Activity recognition systems that leverage unlabeled data present superior performance over their solely supervised variants; *e.g.*, note the superiority of *Auto-BCE* over *Deep-BCE*.
- Incorporating set loss into the training process yields more accurate activity models as compared with the conventional multi-label loss. This is further complemented by the set inference procedure which omits trivial thresholding and instead jointly exploits cardinality and set element scores to generate predictions; *e.g.*, note the superiority of *Deep-Set* over *Deep-BCE*.
- While each component of our proposed methodology (unsupervised feature learning and supervised set learning) individually introduces performance boost in recognition of human activities, when coupled together in a unified framework, the resulting HAR system proves to be the most effective.

## 5 CONCLUSIONS

In this paper, we defined human activity recognition as a set prediction problem in a principled manner. By contrast to the conventional multi-class treatment, our intuitive formulation allows sensory segments be associated with a set of activities and thus, naturally handles segments with multiple activities. In a unified architecture, we addressed the problem by developing a deep HAR system that 1) exploits unlabeled data to uncover effective feature representations and 2) incorporates set objective to learn mappings from input sensory segments to target activity sets. To provide insights on how each component of our proposed methodology contributes to an enhanced recognition performance in isolation, we explored three different multi-label activity recognition models as our baselines. Finally, through empirical experiments on HAR representative datasets, we demonstrated the effectiveness of our proposed deep *Auto-Set* network for human activity recognition.

## REFERENCES


[1] Antonucci Alessandro, Giorgio Corani, Denis Mauá, and Sandra Gabaglio. 2013. An ensemble of Bayesian networks for multilabel classification. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*. 1220–1225.

[2] Mohammad Abu Alsheikh, Ahmed Selim, Dusit Niyato, Linda Doyle, Shaowei Lin, and Hwee-Pink Tan. 2016. Deep Activity Recognition Models with Triaxial Accelerometers. In *Artificial Intelligence Applied to Assistive Technologies and Smart Environments*.

[3] Ling Bao and Stephen S. Intille. 2004. Activity Recognition from User-Annotated Acceleration Data. In *Proceedings of the 2nd International Conference on Pervasive Computing*, Alois Ferscha and Friedemann Mattern (Eds.). 1–17.

[4] Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A Tutorial on Human Activity Recognition Using Body-worn Inertial Sensors. *Comput. Surveys* 46, 3 (2014), 33.

[5] Andreas Bulling, Jamie A. Ward, and Hans Gellersen. 2012. Multimodal Recognition of Reading Activity in Transit Using Body-worn Sensors. *ACM Transactions on Applied Perception* 9, 1 (2012), 2.

[6] Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Tröster, José del R. Millán, and Daniel Roggen. 2013. The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters* 34, 15 (2013), 2033 – 2042.

[7] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. 2010. Why Does Unsupervised Pre-training Help Deep Learning? *Journal of Machine Learning Research* 11 (2010), 625–660.

[8] Yuhong Guo and Suicheng Gu. 2011. Multi-label classification using conditional dependency networks. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*. 1300.

[9] Nils Y. Hammerla, Shane Halloran, and Thomas Plötz. 2016. Deep, Convolutional, and Recurrent Models for Human Activity Recognition Using Wearables. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*. 1533–1540.

[10] Tâm Huynh and Bernt Schiele. 2005. Analyzing Features for Activity Recognition. In *Proceedings Conference on Smart Objects and Ambient Intelligence: Innovative*

*Context-aware Services: Usages and Technologies*. 159–163.
[11] Eunju Kim, Sumi Helal, and Diane Cook. 2010. Human Activity Recognition and Pattern Discovery. *IEEE Pervasive Computing* 9, 1 (2010), 48–53.
[12] Jennifer R. Kwapisz, Gary M. Weiss, and Samuel A. Moore. 2011. Activity Recognition Using Cell Phone Accelerometers. *ACM SigKDD Explorations Newsletter* 12, 2 (2011), 74–82.
[13] O. D. Lara and M. A. Labrador. 2013. A Survey on Human Activity Recognition using Wearable Sensors. *IEEE Communications Surveys Tutorials* 15, 3 (2013), 1192–1209.
[14] Oscar D Lara, Alfredo J Pérez, Miguel A Labrador, and José D Posada. 2012. Centinela: A human activity recognition system based on acceleration and vital sign data. *Pervasive and mobile computing* 8, 5 (2012), 717–729.
[15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.
[16] Francisco Javier Ordóñez and Daniel Roggen. 2016. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors* 16, 1 (2016), 115.
[17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).
[18] C. Randell and H. Muller. 2000. Context awareness by analysing accelerometer data. In *Proceedings of the 4th International Symposium on Wearable Computers*. 175–176.
[19] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L. Littman. 2005. Activity Recognition from Accelerometer Data. In *Proceedings of the 17th Conference on Innovative Applications of Artificial Intelligence*. 1541–1546.
[20] Seyed Hamid Rezatofighi, Anton Milan, Qinfeng Shi, Anthony R. Dick, and Ian D. Reid. 2018. Joint Learning of Set Cardinality and State Distribution. In *AAAI*. (to appear).
[21] A. Wickramasinghe, D. C. Ranasinghe, C. Fumeaux, K. D. Hill, and R. Visvanathan. 2017. Sequence Learning with Passive RFID Sensors for Real-Time Bed-Egress Recognition in Older People. *IEEE Journal of Biomedical and Health Informatics* 21, 4 (2017), 917–929.
[22] Jian Bo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy. 2015. Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition. In *Proceedings of the 24th International Conference on Artificial Intelligence*. 3995–4001.
[23] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu, and J. Zhang. 2014. Convolutional Neural Networks for human activity recognition using mobile sensors. In *Proceedings of the 6th International Conference on Mobile Computing, Applications and Services*. 197–205.
[24] Mi Zhang and Alexander A. Sawchuk. 2012. Motion Primitive-based Human Activity Recognition Using a Bag-of-features Approach. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. 631–640.