Deep Learning Methods for Human Activity Recognition using Wearables

by

Alireza Abedin

Dissertation submitted for the degree of

Doctor of Philosophy

in

School of Computer Science
Faculty of Engineering, Computer & Mathematical Sciences
The University of Adelaide

July 2020

Supervisors:

Associate Professor Damith Chinthana Ranasinghe,
School of Computer Science,
The University of Adelaide

Professor Javen Qinfeng Shi,
School of Computer Science,
The University of Adelaide

Doctor Seyed Hamid Rezatofighi,
School of Computer Science,
The University of Adelaide

Contents

Abstract

Wearable sensors provide an infrastructure-less multi-modal sensing method. Current trends point to a pervasive integration of wearables into our lives with these devices providing the basis for wellness and healthcare applications across rehabilitation, caring for a growing older population, and improving human performance. Fundamental to these applications is our ability to automatically and accurately recognise human activities from often tiny sensors embedded in wearables. In this dissertation, we consider the problem of human activity recognition (HAR) using multi-channel time-series data captured by wearable sensors.

Our collective know-how regarding the solution of HAR problems with wearables has progressed immensely through the use of deep learning paradigms. Nevertheless, this field still faces unique methodological challenges. As such, this dissertation focuses on *developing end-to-end deep learning frameworks to promote HAR application opportunities using wearable sensor technologies and to mitigate specific associated challenges*. In our efforts, the investigated problems cover a diverse range of HAR challenges and spans from fully supervised to unsupervised problem domains.

In order to enhance automatic feature extraction from multi-channel time-series data for HAR, the *problem of learning enriched and highly discriminative activity feature representations with deep neural networks* is considered. Accordingly, novel end-to-end network elements are designed which: (*a*) exploit the latent relationships between multi-channel sensor modalities and specific activities, (*b*) employ effective regularisation through data-agnostic augmentation for multi-modal sensor data streams, and (*c*) incorporate optimization objectives to encourage minimal intra-class representation differences, while maximising inter-class differences to achieve more discriminative features.

In order to promote new opportunities in HAR with emerging battery-less sensing platforms, *the problem of learning from irregularly sampled and temporally sparse readings captured by passive sensing modalities* is considered. For the first time, an efficient set-based deep learning framework is developed to address the problem. This framework is able to learn directly from the generated data, bypassing the need for the conventional interpolation pre-processing stage.

In order to address the multi-class window problem and create potential solutions for the challenging task of concurrent human activity recognition, *the problem of enabling simultaneous prediction of multiple activities for sensory segments* is considered. As such, the flexibility provided by the emerging set learning concepts is further leveraged to introduce a novel formulation of HAR. This formulation treats HAR as a set prediction problem and elegantly caters for segments carrying sensor data from multiple activities. To address this set prediction problem, a unified deep HAR architecture is designed that: (*a*) incorporates a set objective to learn mappings from raw input sensory segments to target activity sets, and (*b*) precedes the supervised learning phase with unsupervised parameter pre-training to exploit unlabelled data for better generalisation performance.

In order to leverage the easily accessible unlabelled activity data-streams to serve downstream classification tasks, *the problem of unsupervised representation learning from multi-channel time-series data* is considered. For the first time, a novel recurrent generative adversarial (GAN) framework is developed that explores the GAN's latent feature space to extract highly discriminating activity features in an unsupervised fashion. The superiority of the learned representations is substantiated by their ability to outperform the de facto unsupervised approaches based on autoencoder frameworks. At the same time, they rival the recognition performance of fully supervised trained models on downstream classification benchmarks.

In recognition of the scarcity of large-scale annotated sensor datasets and the tediousness of collecting additional labelled data in this domain, *the hitherto unexplored problem of end-to-end clustering of human activities from unlabelled wearable data* is considered. To address this problem, a first study is presented for the purpose of developing a stand-alone deep learning paradigm to discover semantically meaningful clusters of human actions. In particular, the paradigm is intended to: (*a*) leverage the inherently sequential nature of sensory data, (*b*) exploit self-supervision from reconstruction and future prediction tasks, and (*c*) incorporate clustering-oriented objectives to promote the formation of highly discriminative activity clusters. The systematic investigations in this study create new opportunities for HAR to learn human activities using unlabelled data that can be conveniently and cheaply collected from wearables.

Statement of Originality

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

The author acknowledges that copyright of published works contained within this dissertation resides with the copyright holder(s) of those works.

I give permission for the digital version of my dissertation to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

	23 / 07 / 2020
Signed	Date

Acknowledgements

This is to express my sincere gratitude towards the people whom without, this dissertation may not have been possible.

First and foremost, I would like to express my appreciation to my Ph.D. supervisors, Associate Professor Damith Ranasinghe, Professor Javen Qinfeng Shi, and Assistant Professor Hamid Rezatofighi for their generous support, tremendous dedication, and endless source of inspiration.

I would like to thank my principle supervisor Dr. Ranasinghe, for motivating me to pursue research in the rewarding area of artificial intelligence (AI) with a focus on solving real-world application problems. I was a foreigner to the realm of AI prior to my candidature. He made me feel welcome, and trusted in my curiosity to explore this truly intriguing domain. Thank you for motivating me to grow into an independent researcher and supporting me all along the way Damith.

I would like to thank Dr. Shi, for getting my hands dirty with support vector machines early days into my candidature, challenging and appreciating my understanding of convex optimisation problems. He truly appreciates the elegance of mathematically supported concepts and encourages adopting this mindset. I am grateful for all your support during my Ph.D. candidature Javen.

I would like to thank Dr. Rezatofighi, for patiently mentoring me through my AI journey in going from strength to strength. He exhibits a strong personality, encourages thinking out of the box and motivates shooting for the skies. Beyond a supervisor, he has been a true friend for me. During challenging times, I have shared my concerns with him and benefited from his advice. I wish you all the best in your new exciting role at Monash University and am confident you will shine as always Hamid.

My appreciation goes to the University of Adelaide for awarding me with a PhD scholarship to pursue higher degree by research. I would like to thank the School of Computer Science, for the gracious assistance in every matter, helping with conference travels and also for the endless supply of milk and coffee beans to fuel my research.

I am grateful for my lovely friends here in Adelaide: Adel, Alireza, Amin, Ghazal, Hirad, Maryam, Mitra, Mohammadreza, Omid, Renato, Sadaf, Sandi, Vahid. You are the family I chose by heart and I cherish our great memories together.

I would like to express my very great appreciation to my inspiring parents, Marzieh and Morteza. I have felt your love, support and presence each and every single day despite the physical distance between us. You are my role models and I feel grateful to have you. I would also like to thank Elham, Elahe, Mohammadreza and Masoud for being the most fun siblings and a big part of my life.

I would like to express my gratitude to my parents-in-law for their love, support, and most importantly for bringing Mahsa into this world. A very special word of thanks goes to Mahsa, my best friend, my soul-mate and the most gorgeous girl in the world. I fell for you ten years ago and you have changed my life ever since. You have been there for me during the most challenging times, comforting me with your cuteness and inspiring me to move forward. I am the luckiest to have you by my side.

Dissertation Conventions

The following conventions have been adopted in this dissertation:

Typesetting

This document was compiled using LaTeX2e. Texstudio 2.12.22 was used as a text editor interfaced to LaTeX2e. Inkscape 0.92.3 was used to produce schematic diagrams and other drawings.

Spelling

Australian English spelling conventions have been used, as defined in the Macquarie English Dictionary—A. Delbridge (Ed.), Macquarie Library, North Ryde, NSW, Australia, 2001.

Referencing

The Harvard reference style is used for referencing and citation in this dissertation.

System of Units

The units comply with the international system of units recommended in an Australian Standard: AS ISO 1000-1998 (Standards Australia Committee ME/71, Quantities, Units and Conversions 1998).

Acronyms

AE	Auto-Encoder
CNN	Convolutional Neural Network
DNN	Deep Neural Network
GAN	Generative Adversarial Network
GRU	Gated Recurrent Unit
HAR	Human Activity Recognition
IMU	Inertial Measurement Unit
KL	Kullback-Leilbler
LSTM	Long Short Term Memory
MAP	Maximum A Posteriori
PCA	Principle Component Analysis
RBM	Restricted Boltzmann Machine
ReLU	Rectified Linear Unit
RFID	Radio Frequency Identification
RNN	Recurrent Neural Network
t-SNE	t-distributed Stochastic Neighbor Embedding
VAE	Variational Auto-Encoder

Publications

Conference Articles

[1] **Abedin, A.**, Motlagh, F., Shi, Q., Rezatofighi, H., Ranasinghe, D.C. 2020. Towards Deep Clustering of Human Activities from Wearables. *Proceedings of the International Symposium on Wearable Computers (ISWC)—CORE Rank: A*.*

[2] **Abedin, A.**, Rezatofighi, H., Shi, Q., Ranasinghe, D.C. 2019. SparseSense: Human activity recognition from highly sparse sensor data-streams using set-based neural networks. *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 5780–5786—CORE Rank: A*, acceptance rate: 17.9%, 850 accepted papers out of 4,752 submissions.

[3] **Abedin, A.**, Abbasnejad, E., Shi, Q., Ranasinghe, D.C, Rezatofighi, H. 2018. Deep Auto-Set: A deep auto-encoder-set network for activity recognition using wearables. *Proceedings of the 15th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous)*, pp. 246–253—CORE Rank: A.

Under-Review Articles

[4] **Abedin, A.**, Ehsanpour, M., Shi, Q., Rezatofighi, H., Ranasinghe, D.C. 2020. Attend and discriminate: Beyond the state-of-the-art for human activity recognition using wearable sensors. *Submitted to the ACM international joint conference on pervasive and ubiquitous computing (UbiComp)—CORE Rank: A*,* (revise and re-submit notification: 08/07/2020).

[5] **Abedin, A.**, Rezatofighi, H., Ranasinghe, D.C. 2020. Guided-GAN: Geometrically-guided adversarial representation learning for activity recognition with wearables. *Submitted to the international conference on Information Processing in Sensor Networks (IPSN)—CORE Rank: A*.*

List of Figures

List of Tables

Chapter 1

Introduction

THIS introductory chapter presents a brief overview of human activity recognition (HAR) in ubiquitous computing, and discusses the scope of the research problems investigated in this dissertation. We specifically clarify the motivations and research objectives in relation to each studied problem, as well as highlighting the contributions of the present research. The chapter is concluded by providing a guide to the structural organisation of the thesis.

1.1 Introduction

Automatic *human activity recognition* (HAR) using wearable sensors has emerged as a key research area in ubiquitous computing Bao and Intille (2004) with thriving development of low-cost sensing technologies as well as the fast advancements in machine learning techniques. In this problem, high-level activity information is acquired by analysing raw low-level sensor data-streams, with the goal of providing proactive yet unobtrusive assistance to users. Having created new possibilities in diverse application domains including health-care Torres-Huitzil and Alvarez-Landero (2015); Subasi et al. (2018); Chesser et al. (2019), smart-homes Zheng, Wang and Black (2008); Wang et al. (2011); Bianchi et al. (2019), manufacturing Günther, Kärcher and Bauernhansl (2019), sports and the entertainment industry Kunze et al. (2006); Ladha et al. (2013); Zhuang and Xue (2019), HAR has successfully sparked excitement in both academia and industry. Fundamental to realising these applications is our ability to automatically and accurately recognise human activities from, often, tiny sensors embedded in wearables. This forms the driving motivation behind the research conducted in this dissertation.

Traditionally, the standard activity recognition pipeline for time-series sensory data involved sliding window segmentation, manual hand-crafted feature design, and subsequent activity classification with classical machine learning algorithms Bulling, Blanke and Schiele (2014); we illustrate the *traditional HAR* work-flow in Fig. 1.1. Studies along these lines have extensively explored hand-crafted features including statistical Bao and Intille (2004); Ravi et al. (2005), basis transform Huynh and Schiele (2005), multi-level Zhang and Sawchuk (2012a), and bio-mechanical Wickramasinghe et al. (2017) features; and have employed shallow classifiers including decision trees

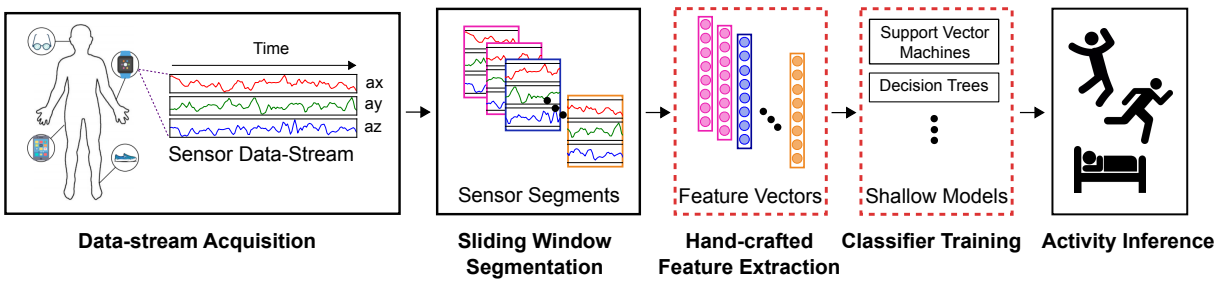


Figure 1.1: Traditional HAR pipeline. We illustrate the traditional human activity recognition pipeline which is characterised by hand-crafted feature engineering and adoption of classical machine learning algorithms in separate stages, highlighted by the red boxes.

Bao and Intille (2004), support vector machines Bulling, Ward and Gellersen (2012), joint boosting Lara et al. (2012) and graphical models Shinmoto Torres et al. (2017a). While this manually tuned procedure has successfully acquired satisfying results for relatively simple recognition tasks, its generalisation performance is limited by heavy reliance on domain expert knowledge to engineer effective features.

Recently, the emerging paradigm of deep learning has demonstrated unparalleled performance in various research areas including computer vision, natural language processing and speech recognition LeCun, Bengio and Hinton (2015). When applied to sensor-based HAR, deep learning allows for automated end-to-end feature extraction, largely alleviating the need for laborious feature engineering procedures; we illustrate the *deep learning HAR* work-flow in Fig. 1.2. Moreover, the adoption of deep neural networks for HAR has successfully created pipelines for end-to-end learning of activity recognition models yielding state-of-the-art performance for diverse applications in ubiquitous computing Ordóñez and Roggen (2016); Hammerla, Halloran and Plötz (2016); Yao et al. (2018); Murahari and Plötz (2018). Consequently, we observe a shift in the research efforts from traditional methods towards deep learning paradigms in addressing complex human activity recognition problems in recent years.

Despite the progress towards deep learning architectures for achieving state-of-the-art performance on HAR problems, this field still faces many unique methodological challenges leaving room for further improvements. Accordingly, it is of great significance to propose systematic approaches towards accurate recognition of activities that triumph over the challenges. In light of the advantages brought about by the introduction of deep neural networks for HAR problems, this dissertation focuses on *developing end-to-end deep learning frameworks to promote HAR opportunities using*

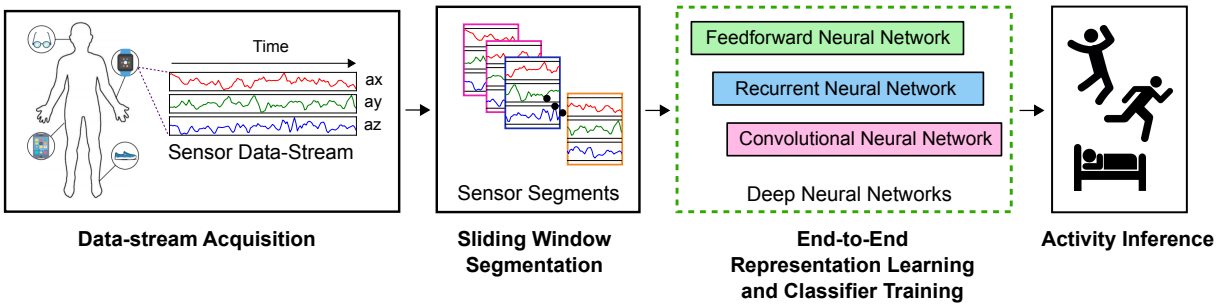


Figure 1.2: Deep learning HAR pipeline. We illustrate the deep learning human activity recognition pipeline which is characterized by end-to-end feature representation learning and simultaneous classifier training, highlighted by the green box.

pervasive sensor technologies and to mitigate specific associated challenges. In what follows, we elaborate on the key challenges explored in this thesis:

- **Learning Highly Discriminative Activity Features.** Wearable sensing devices capture individuals' activity dynamics by continuously recording measurements. They do this using various sensor channels over time, generating *multi-channel time-series* data-streams. In order to cater for the unique characteristics of the generated data, it is essential to design deep learning modules that: (*a*) seamlessly operate on raw time-series data without relying on hand-crafted feature engineering procedures; (*b*) capture the inherent temporal dependencies between samples for sequence modelling; and (*c*) take into account the relationships among the multitude of incorporated sensor channels used for data acquisition. Additionally, human activities are inherently diverse in nature. Consequently, it is challenging to learn feature representations that uniquely represent distinct human actions. Further, *intra-class variability* and *inter-class similarity* pose two fundamental challenges for HAR based on wearables. The former phenomenon refers to the fact that different individuals may execute the same activity differently (*e.g.*, different walking patterns). The latter challenge arises when different classes of activities reflect very similar sensor patterns (*e.g.*, walking upstairs and walking downstairs). In order to accurately classify the actions embedded in the generated sequences, it is of crucial importance to incorporate deep neural architectures and learning strategies that encourage achieving discriminative activity representations.

- **Learning Human Activities from Emerging Passive Sensor Devices.** With the technological advances in sensing platforms, an increasing number of battery-less—so-called, *passive*—wearables are providing the opportunity to collect fine-grained physiological information on human activities. They are able to do this conveniently, at a low cost. In particular, passive sensing modalities Chen et al. (2015); Lemey et al. (2016); Jayatilaka et al. (2019) operating on harvested energy provide maintenance-free, unobtrusive, lightweight and often disposable devices. These characteristics make such devices highly desirable to both older people and healthcare providers Gövercin et al. (2010); Torres et al. (2017). Despite their compelling propositions for sensing applications Philipose et al. (2005), the data-streams collected from these sensors are characterised by

high temporal sparsity. This makes it extremely challenging for conventional deep neural networks to learn from such irregularly sampled sensor streams.

- **Multi-class Window Problem.** Regardless of adopting the traditional or deep learning pipelines, HAR problem formulations often rely on a fixed duration sliding window and predict a single activity class for all samples within the partitioned segment. However, human actions exhibit great diversity in their duration. Consequently, deciding on an optimal size for the sliding window in advance is difficult. As a result, we inevitably observe data segments that contain sensor samples of multiple activity labels; namely, the multi-class window problem Yao et al. (2018). However, the dominant multi-class problem formulation of HAR mistreats such segments by approximating the segment annotations to either the most Yang et al. (2015) or the last Ordóñez and Roggen (2016) observed sample annotations. This strategy towards ground-truth approximation is clearly associated with a loss of activity information and potentially deludes the supervised training process. Moreover, the multi-class formulation of HAR fails to cater for the complex nature of human activities, where actions are not only performed sequentially but are also carried-out simultaneously—the so called concurrent activity recognition problem.

- **Annotated Data Scarcity.** The process of data collection and annotation may be retrospective in the case of vision-based sensing modalities where visual inspections of, for example, video frames provide the basis for ground truth. However, the parallel task with wearables is nearly impossible. Moreover, such methods cannot be easily scaled to gather the large datasets often necessary for deep learning frameworks. In the absence of a reliable visualisation to establish ground truth, acquisition of labelled sensory data is labour-intensive, time-consuming and clearly not scalable to large datasets. This shortcoming poses a significant challenge to the development of deep learning frameworks for HAR problems involving wearable sensors that have predominantly been studied under *supervised learning* regimes. Accordingly, it is crucially important to consider *unsupervised learning* scenarios and explore systematic solutions. This will allow us to benefit from conveniently collectable activity data-streams that lack human data annotations.

1.2 Summary of Original Contributions

This dissertation delivers several original contributions to the field of human activity recognition based on wearable sensors in ubiquitous computing. These contributions focus on developing deep learning paradigms in diverse problem settings, spanning from supervised learning regimes to fully unsupervised training scenarios. The contributions can be summarised as follows:

1. The problem of learning highly discriminative and generalisable activity representations from raw multi-modal data-streams is considered. The study proposes novel deep learning architectural elements to: (*a*) enrich convolutional feature-map representations by exploiting latent correlations between sensor channels; (*b*) incorporate centre-loss to alleviate dealing with intra-class variations of activities; and (*c*) augment multi-channel time-series data with mixup for better generalisation. The contributions from the design concepts are validated through exhaustive quantitative and qualitative experiments, including activity misalignment measures, and ablation studies. This work is currently under-review in the *ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)* under the title "Attend and Discriminate: Beyond the State-of-the-Art for Human Activity Recognition using Wearable Sensors".

2. The problem of activity recognition from temporally sparse data-streams captured by passive wearables is considered. For the first time, the study develops an end-to-end human activity recognition framework to learn directly from temporally sparse data-streams using set-based deep neural networks. Previous studies rely on interpolation pre-processing to synthesise sensory partitions with fixed temporal context. In contrast, the proposed *SparseSense* network seamlessly operates on sparse segments with a potentially varying number of sensor readings and delivers highly accurate predictions despite some missing sensor observations. Extensive experiments on publicly available HAR datasets shows that the proposed novel treatment for sparse data-stream classification results in recognition models that significantly outperform deep learning based HAR models relying on interpolation pre-processing to address sparsity. It also incurs notably lower real-time prediction delays. We believe that this work will provide a new method for understanding human motion data

collected using passive wearables for health-care applications. This work has been published in the *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)* under the title "SparseSense: Human Activity Recognition from Highly Sparse Sensor Data-streams Using Set-based Neural Networks" (Abedin et al., 2019).

3. The inevitable multi-class window problem arising from the dominantly incorporated sliding window segmentation approach is considered. To address this problem, the task of human activity recognition is expressed more naturally as a *set prediction problem*. Within this definition, the predictions are *sets* of ongoing activity elements with unfixed and unknown cardinality that can handle sensor segments with multiple activities. For the first time, the multi-class window problem is addressed by presenting a novel HAR approach that learns to output activity sets using deep neural networks. Moreover, motivated by the limited availability of annotated HAR datasets, the supervised set learning scheme is complemented with a prior unsupervised feature learning process that adopts convolutional auto-encoders to exploit unlabelled data. The empirical experiments on two widely adopted HAR datasets demonstrate the substantial improvement of the proposed methodology over the baseline models. This work has been published in the *Proceedings of the 15th EAI International Conference on Mobile and Ubiquitous Systems (MobiQuitous)* under the title "Deep Auto-Set: A Deep Auto-Encoder-Set Network for Activity Recognition Using Wearables" (Abedin et al., 2018).

4. In order to leverage the easily accessible unlabeled activity data-streams for downstream classification tasks, the problem of learning unsupervised activity representations from multi-channel time-series data is considered. For the first time, the study proposes a bidirectional GAN (BiGAN) framework comprising a recurrent generator, encoder and discriminator that can communicate in a unified architecture to learn unsupervised feature representations. Moreover, a novel strategy is proposed to alleviate the burden on the discriminator to uncover the generator's inverse mapping function by seeking additional feedback from geometric distance penalisation in data and latent manifolds. Interestingly, incorporation of the geometric terms is found to be a vital necessity for successful training of BiGAN in the sequential domain, where (in contrast to the visual domain) extensive training guidelines may be missing. The

unsupervised learned features are evaluated on three downstream sequence classification benchmarks, outperforming existing unsupervised approaches while closely approaching fully supervised performance. This work is currently under-review for the *International Conference in Sensor Networks (IPSN)* under the title "Guided-GAN: Geometrically-Guided Adversarial Representation Learning for Activity Recognition with Wearables".

5. In recognition of the scarcity of large scale annotated datasets, the hitherto unexplored problem of end-to-end clustering of human activities from unlabelled wearable using a deep learning paradigm is considered. To the best of knowledge, this constitutes the first study to investigate and develop a novel deep clustering architecture for HAR problems involving sensor data, with the aim of alleviating the reliance on human data annotations. The systematic experiments demonstrate the effectiveness and generalisability of the proposed approach for clustering human activities across three diverse HAR benchmark datasets. Further, additional insights are shared by: (*a*) examining the unsupervised learned representations from sequential sensor data; and (*b*) an ablation study to validate the network design thinking. We believe this study makes a significant advancement to the learning of human activities from unlabelled data that can be conveniently and cheaply collected from wearables. This work has been accepted to be published in the *Proceedings of the International Symposium on Wearable Computers (ISWC)* under the title "Towards Deep Clustering of Human Activities from Wearables" with an extended version prepared for submission to the *Pattern Recognition Journal* under the title "Deep Sensory Clustering: Unsupervised Learning of Human Activities from Wearables".

1.3 Dissertation Structure

We organise the technical contributions of this dissertation into eight chapters. The organisational structure is outlined in Figure 1.3, and briefly described in the subsequent text.

1. Chapters 1 and 2 provide a brief introduction and background to human activity recognition (HAR) with pervasive sensing technologies; in particular, wearable sensors or so-called wearables.

Introduction	*Chapter 1*	• Introduction to Human Activity Recognition (HAR) with wearables. • Investigated problems, motivations and technical contributions made.
	Chapter 2	• Common sensor modalities. • Popular deep learning architectural components in HAR. • Performance metrics to evaluate activity recognition models.
Supervised HAR	*Chapter 3*	• Investigating new opportunities to improve upon automated feature extraction for achieving highly discriminative activity features. • Development of novel network elements that enhance convolutional feature-maps by capturing the latent relationships among sensor channels, encourage discriminative representations and regularize HAR models.
	Chapter 4	• Investigating the problem of learning activity recognition models from irregular data captured by emerging batteryless sensors. • Development of a set-based HAR system to handle temporally sparse input data-streams.
	Chapter 5	• Investigating the multi-class window problem for sensor segments with data from multiple activities. • Development of a novel set formualtion for HAR to handle sensor segments with multiple activities. • Development of an end-to-end HAR model mapping raw sensory segments to target activity sets. • Development of a convolutional autoencoder to exploit unlabeled data for unsupervised pre-training of network parameters.
Unsupervised HAR	*Chapter 6*	• Investigating the problem of unsupervised activity representation learning from unlabeled activity data-streams. • Rigorous examination of generative adversarial network's latent feature space for unsupervised representation learning in HAR. • Design of recurrent generator, encoder and joint discriminator to cooperate in a unified bidirectional GAN framework.
	Chapter 7	• Investigating the problem of unsupervised activity clustering from wearables. • Exploring the effectiveness of multi-tasks autoencoding objectives for unsupervised initialization of feature space in HAR. • Incorporation of clustering-oriented criteria to learn semantically meaningful groupings of human activities from unlabeled data-streams. • Development of a novel deep clustering architecture for end-to-end clustering of human activities embedded in wearable sensor sequences.
Conclusion	*Chapter 8*	• Summary • Future research directions

Figure 1.3: Thesis structure. We illustrate the organisational structure incorporated in this dissertation.

2. Chapter 3 studies new opportunities to improve upon the automated feature representation learning process for wearable activity data. It addresses key under-explored dimensions with great potential to learn enriched and highly discriminating activity representations. In particular, it discusses systematic strategies to: (*a*) learn to exploit the latent relationships between multi-channel sensor modalities and specific activities, (*b*) leverage the effectiveness of data-agnostic augmentation for multi-modal sensor data-streams to regularise deep HAR models, and (*c*) incorporate a classification loss criterion to encourage minimal intra-class representation differences whilst maximising inter-class differences, to achieve more discriminative features.

3. Chapter 4 investigates the problem of learning activity recognition models from *irregular* and *temporally sparse data* captured by battery-less sensing modalities. Here, the time intervals between sensor readings are irregular, while the number of readings per unit of time are often limited. Bypassing the need for interpolation preprocessing, an efficient set-based deep learning paradigm is proposed to learn directly from sparse data in an end-to-end manner.

4. Chapter 5 investigates a novel formulation of human activity recognition as a *set prediction task*. Such a formulation may serve to overcome the multi-class window problem and the fact that conventional HAR models can only output a single activity label for a given sensor segment. This new formulation allows sensory segments to be associated with a set of activities and, thus, naturally handles segments with multiple activities. In a unified architecture, a deep HAR system is proposed that (*a*) incorporates a set objective to learn mappings from input sensory segments to target activity sets, and (*b*) exploits unlabelled data for unsupervised pre-training of network parameters to achieve better generalisation performance.

5. Chapter 6 examines the problem of *unsupervised activity representation learning* from multi-channel time-series data through generative adversarial networks (GANs). Here, the aim is to leverage cheaply accessible unlabelled data to learn unsupervised feature representations that may serve subsequent downstream sequence classification tasks. To this end, (*a*) a novel bidirectional GAN framework comprised of a recurrent generator, encoder and joint discriminator is designed, and (*b*) a stable training strategy is efficiently implemented by augmenting adversarial feedback with geometric manifold distance guidance.

6. Chapter 7 studies the problem of *unsupervised activity clustering* from unlabeled data-streams captured by wearable sensors. Here, the goal is to uncover semantically meaningful clusters of activity data in an unsupervised manner. For the first time, an end-to-end deep clustering architecture is developed that (*a*) leverages the inherently sequential nature of sensory data, (*b*) exploits self-supervision from reconstruction and future prediction tasks, and (*c*) incorporates a clustering-oriented objective to promote the formation of highly discriminative activity clusters.

7. Chapter 8 summarises the research conducted in this dissertation, and outlines potential future research directions.

Chapter 2

Background

THIS chapter provides a brief background on sensor-based human activity recognition (HAR) with deep learning paradigms. A generic formulation for the problem is presented and the notations incorporated across the chapters are introduced for clarity. Common sensor modalities used for data acquisition are discussed and representative HAR benchmark datasets are reviewed. In addition, this chapter provides an overview of popular deep learning building blocks incorporated in HAR frameworks to learn from raw multi-channel time-series data in an end-to-end manner. Further, the predominantly used evaluation metrics are introduced to ground our work and quantify the experimental results.

2.1 Notations

For notational consistency, we present a generic formal definition of the human activity recognition problems considered in this dissertation. Except where specifically stated, these notations are applicable across chapters. We formally introduce the notations and describe the process whereby the collected raw sensor data-streams are partitioned into smaller chunks to provide the input for end-to-end HAR frameworks. We further define the two broad categories into which the explored HAR problems in this dissertation fall, based on the availability of sensor data annotations—*supervised* and *unsupervised* HAR problems.

Data-streams. We consider developing deep learning-based HAR systems to help us understand a diverse set of k human actions in a predefined activity space $\mathcal{A} = \{a_i\}_{i=1}^{k}$ using multi-modal HAR sensing platforms. Without loss of generality, we assume a hardware-specific sampling rate for the wearable sensors, denoted by f. Such devices continuously record measurements through different sensor channels over time and generate *multi-channel time-series* data. Accordingly, body-worn sensors yield the collected data-stream of raw time-series samples $\mathbf{X}_{\text{stream}}$ and their corresponding

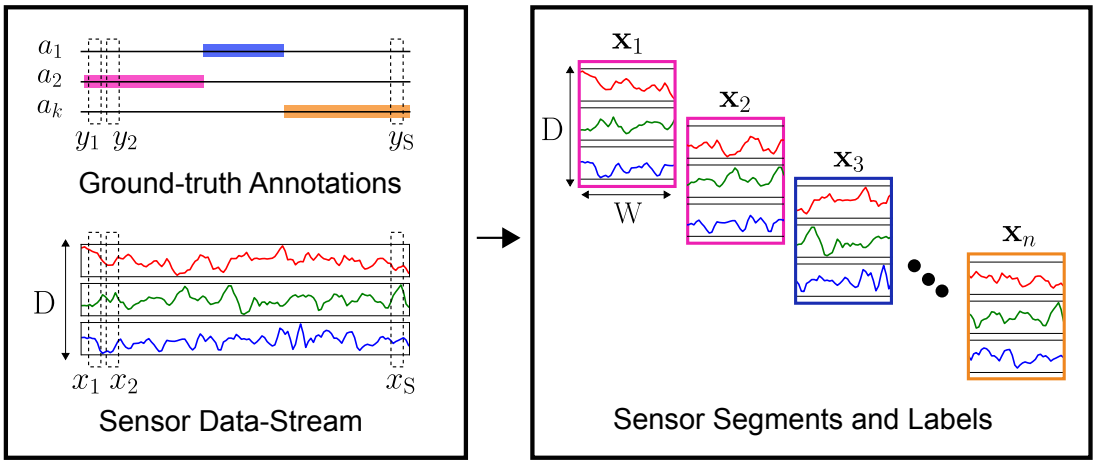


Figure 2.1: Sensor data-stream acquisition and segmentation. (a) Initially, raw multi-channel sensor data-streams—$\mathbf{X}_{\text{stream}}$ and $\mathbf{Y}_{\text{stream}}$—are collected from various sensing modalities over time, and (b) subsequently, a sliding window is adopted to partition the continuous data-stream into a dataset of sensory segments and their corresponding activity labels—$\mathbf{X}_{\text{segment}}$ and $\mathbf{Y}_{\text{stream}}$. Here, we have color-coded the activity labels with colored boxes around the sensor data segments.

activity labels $\mathbf{Y}_{\text{stream}}$

$$\begin{aligned} \mathbf{X}_{\text{stream}} &= (x_1, x_2, \ldots, x_S), \\ \mathbf{Y}_{\text{stream}} &= (y_1, y_2, \ldots, y_S), \end{aligned} \tag{2.1}$$

where $x_t \in \mathbb{R}^D$ denotes the multi-dimensional vector that contains sample measurements over D distinct sensor channels at time step t, $y_t \in \mathcal{A}$ is the corresponding activity annotation, and S denotes the total length of the recorded sequence.

Data Stream Segmentation. We apply the commonly adopted time-series segmentation technique Bulling, Blanke and Schiele (2014) of *a sliding window of fixed temporal duration* δt to partition the acquired sensor stream into a set of n sensor segments $\mathbf{X}_{\text{segment}}$ and corresponding activity labels $\mathbf{Y}_{\text{segment}}$

$$\begin{aligned} \mathbf{X}_{\text{segment}} &= (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n), \\ \mathbf{Y}_{\text{segment}} &= (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n), \end{aligned} \tag{2.2}$$

where $\mathbf{x} \in \mathbb{R}^{D \times W}$ is a slice of captured time-series data with D denoting the number of sensor channels used for data acquisition and $W = f\delta t$ representing the choice for the window duration. In addition, $\mathbf{y}_i$ denotes the last activity annotation observed during the sliding window lifetime for a given data segment $\mathbf{x}_i$. For the purpose of simplicity, we illustrate the process of data-stream acquisition and segmentation in Fig. 2.1 for a triaxial accelerometer (*i.e.*, $D = 3$).

HAR Learning Regimes. Notably, this dissertation explores the development of HAR frameworks under both *supervised* and *unsupervised* problem settings:

- *Supervised Learning Regime.* When operating in the supervised domain, we assume access to a labelled training dataset $\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ comprising pairs of sensor segments and their corresponding activity labels during training. In particular, the HAR problems studied from Chapter 3 to Chapter 5 assume access to annotated sensor data for training HAR models. As such, they fall under the supervised learning category.

- *Unsupervised Learning Regime.* For unsupervised scenarios, we assume access to an unlabelled training dataset $\mathcal{U} = \{\mathbf{x}_i\}_{i=1}^n$ that is composed of sensor data

segments, but which lacks the corresponding annotations during training. In recognising the scarcity of data annotations for HAR, Chapter 6 and Chapter 7 consider HAR problem scenarios which only have access to unlabelled data for the purpose of model development.

2.2 Sensor Modalities

An increasing number of sensing technologies are providing the opportunity to conveniently collect fine-grained physiological information at low-cost. Such information can serve to inform our understanding of human activities. As a consequence of this technological development, diverse sensing modalities for data acquisition are being incorporated into sensor-based human activity recognition. The relevant sensors are mainly classified into three categories: *wearable sensors*, *object sensors*, and *ambient sensors* Chavarriaga et al. (2013). Below, we present a brief description of each category.

- **Wearable sensors.** Body-worn sensors are found in devices such as smart-phones, watches and garments equipped with accelerometers, magnetometers, gyroscopes and barometers. These sensors measure human motion data directly and conveniently, and constitute the most common sensor modalities used for human activity recognition. In particular, *accelerometers* measure acceleration, *gyroscopes* measure angular velocity, and *magnetometers* report changes in the magnetic field. Often, these devices are gathered and used together in an *Inertial Measurement Unit (IMU)*.

- **Object sensors.** These sensors are attached to objects of interest, recording their movement data. Accordingly, human interactions with these objects provide context and localisation information for inferring human activities.

- **Ambient sensors.** These sensors include WiFi, radar, sound, pressure and temperature sensors embedded in smart environments to report state changes in the environment for inferring human actions.

This dissertation primarily conducts experimentation with public HAR datasets of multi-channel time-series data captured by body-worn sensor modalities—*wearable sensors*—due to their ease of deployment and popularity in ubiquitous computing.

The explored HAR datasets are extensively adopted in HAR studies for benchmarking purposes and exhibit great diversity in terms of activities they cover and their application scenarios. To provide an overview, we present a brief description of representative datasets investigated in this dissertation.

Skoda Dataset Stiefmeier et al. (2008). This dataset covers the problem of recognising the activities of assembly-line workers in a manufacturing scenario. For data acquisition, 20 triaxial accelerometers with a sampling rate of 98Hz were worn by a subject on both arms while performing manual quality checks of newly constructed cars. The dataset is annotated with 10 manipulative gestures of interest, including `writing in a notepad`, `checking the steering wheel`, `opening and closing the boot`, `doors`, and `engine bonnet`, as well as a `null class` to identify non-relevant activities. In this dissertation, Chapter 3 studies the corresponding activity recognition task in a supervised learning scenario while Chapter 7 considers the problem in an unsupervised setting.

WISDM Dataset Kwapisz, Weiss and Moore (2011). This dataset contains acceleration measurements from 36 volunteers performing a specific set of activities. The data were collected under controlled, laboratory conditions. The collected dataset delivers high quality data and has frequently been used in HAR studies for benchmarking purposes. The sensing device used for data acquisition is an Android mobile phone with a constant sampling rate of 20 Hz, placed in the subjects' front trouser pocket. The sensor samples carry annotations from `walking`, `jogging`, `climbing up stairs`, `climbing down stairs`, `sitting` and `standing`. This dataset is explored in Chapter 4 to synthesize sparse data-streams as well as in Chapter 5 to detect multiple activities within a segment.

Opportunity Dataset Chavarriaga et al. (2013). This dataset comprises annotated recordings from a wide variety of on-body sensors including IMUs and triaxial accelerometers. For data acquisition, four subjects were instructed to carry out naturalistic kitchen routines. Each sample in the resulting dataset corresponds to 113 real-valued signal measurements recorded at a frequency of 30 Hz. The dataset offers different sets of annotations to address two distinct activity recognition problems: *gesture recognition* and *locomotion recognition*. The former is examined in Chapter 3

with 18 sporadic gestures, and the latter is investigated in Chapter 5 with five modes of locomotion. Notably, both activity tasks require recognition of the `null class`—periods without the activities of interest—challenging the performance of HAR models.

UCI HAR Dataset Anguita et al. (2013). This dataset targets the problem of recognising activities of daily life (ADLs). The data was collected from 30 volunteers wearing a smartphone at waist level in controlled, laboratory conditions while undertaking six physical activities: `laying`, `walking`, `walking upstairs`, `walking downstairs`, `sitting`, and `standing`. During the experiments, the acceleration and angular velocity readings were recorded at a constant rate of 50Hz using the phone's embedded accelerometer and gyroscope, resulting in 9-dimensional measurements. The gathered data has been manually annotated using the video recordings from the experiments. This dataset is adopted to evaluate recognition performance of HAR models for unsupervised representation learning and activity clustering in Chapter 6 and Chapter 7, respectively.

This dissertation further investigates the *Hospital* (Yao et al. (2018); Chapter 3), *Clinical Room* (Torres et al. (2013); Chapter 4), *USC-HAD* (Zhang and Sawchuk (2012b); Chapter 6), and *MHEALTH* (Banos et al. (2014); Chapter 7) datasets for activity recognition in health-care scenarios as well as *PAMAP2* (Reiss and Stricker (2012); Chapter 3) for understanding common activities of daily life.

2.3 Deep Learning Models

Over the past years, human activity recognition has greatly advanced with the introduction of end-to-end deep learning paradigms. Here, we introduce the core deep learning building blocks that have mainly been adopted in the development of HAR frameworks in recent literature. These include *multi-layer perceptrons*, *recurrent neural networks*, *convolutional neural networks*, *autoencoders* and combinations of these components.

Multi-Layer Perceptrons (MLPs). Multi-layer perceptrons adopt fully-connected topologies with computational nodes—namely, neurons—arranged into layers and inter-connected using trainable parameters. Internally, each neuron learns a non-linear

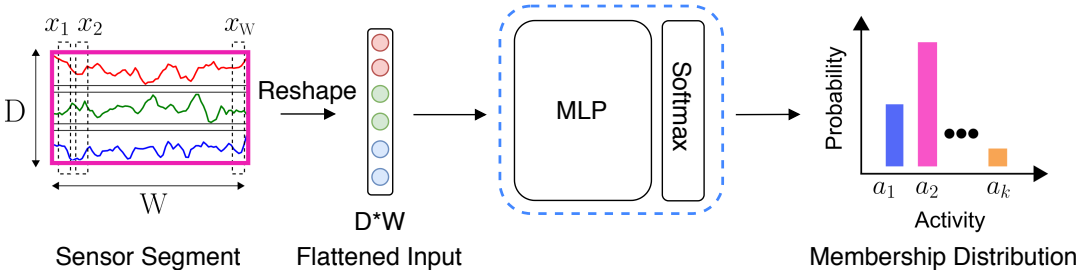


Figure 2.2: Multi-Layer Perceptrons (MLPs). We illustrate an overview of HAR frameworks designed solely using MLPs. Notably, MLPs leverage fully connected structures and require their input to have a flattened representation for processing.

projection of the activations from its preceding layer. Given their dense design structure, MLPs significantly increase the number of network parameters, making them computationally expensive to train. Accordingly, they often serve only as the classification component of HAR frameworks, with the deepest layer generating activity membership distributions. Nevertheless, some early studies have relied on HAR frameworks developed entirely using MLPs Vepakomma et al. (2015); Walse, Dharaskar and Thakare (2016). In Fig. 2.2, we illustrate the work-flow of an HAR framework built upon MLPs to perform classification of human activities from raw time-series segments. Initially, the input segment is reshaped to achieve a flattened vectorised representation. Subsequently, the flattened input is processed through a series of fully connected hidden layers to ultimately generate classification decisions over the activity space.

Recurrent Neural Networks (RNNs) Rumelhart, Hinton and Williams (1986). Recurrent neural networks integrate neurons with recurrent feedback in order to model temporal dependencies in sequential data. In a recursive manner, the output at each time-step is computed as a function of the current input and the hidden state of the network from previous time-steps. In this regard, incorporation of long-short term memory (LSTM) units Hochreiter and Schmidhuber (1997) and gated recurrent units (GRU) Cho et al. (2014) in recurrent neural networks form two extended flavours of RNNs that leverage gating mechanisms to implement memory cells and facilitate the learning of long-range temporal dependencies.

In the context of HAR problems, RNNs and their variants (*i.e.*, LSTMs and GRUs) are directly applicable to raw multi-channel time-series data and allow automatic extraction of the temporal correlations between individual sample measurements

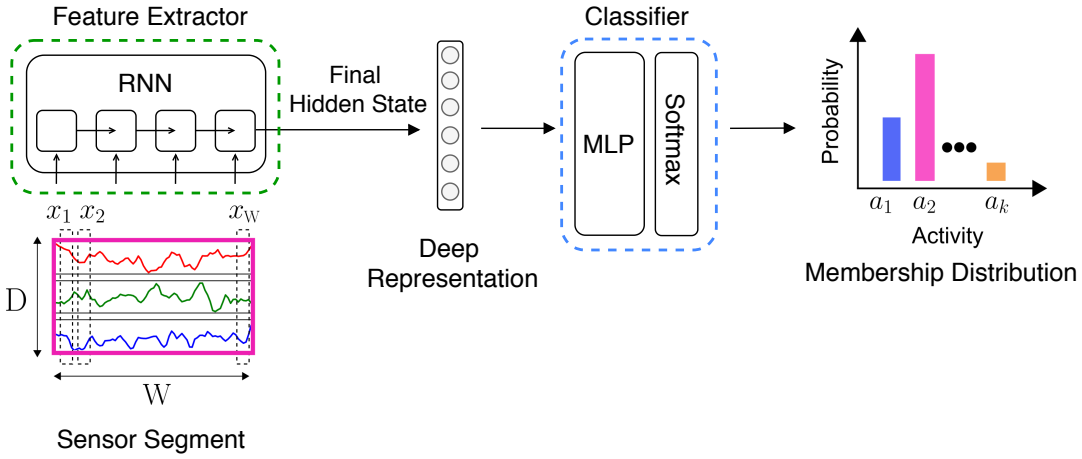


Figure 2.3: Recurrent Neural Networks (RNNs). We illustrate an overview of HAR frameworks incorporating RNNs for temporal modelling of multi-channel time-series data.

at the lowest possible level. We depict the work-flow of an HAR framework incorporating RNNs for feature extraction in Fig. 2.3. Initially, the raw input segment is sequentially processed by the recurrent component for temporal modelling. The final representation achieved after processing the entire input sequence serves as a holistic summary of the input, and is subsequently fed to an MLP to produce activity classification scores.

Convolutional Neural Networks (CNNs) LeCun, Bengio and Hinton (2015). Convolutional neural networks constitute the most popular choice of deep neural networks for the purpose of automating feature extraction from sensor data in human activity recognition studies. In general, CNNs are formed through a stack of convolutional operators with small filters which are capable of automatically capturing salient features at progressively more abstract resolutions. In the context of HAR problems, CNNs predominantly employ 1D convolutional filters. These filters are directly applied along the temporal dimension of sensor channel data to capture the local dependencies in a hierarchical manner. The acquired feature-maps are ultimately unified and mapped into activity class scores using a jointly trained classifier. We illustrate this process in Fig. 2.4.

Hybrid Neural Networks. Based on the successful independent applications of CNNs and RNNs as a way to develop effective HAR models, efforts have been made to combine the CNN-based representation learning approach with an RNN-based

temporal modelling strategy. Within the developed frameworks, the CNN module extracts local features from individual sensor modalities and hierarchically merges them into global features. Subsequently, the RNN module exploits the learned features and extracts temporal relationships at a more abstract representation level.

AutoEncoder Neural Networks. Through stacked hidden layers of encoding-decoding operations, autoencoders provide an effective means to learn unsupervised feature representations from input data. They comprise an encoder neural network and a decoder neural network; the encoder extracts features from unlabelled data (often in a low-dimensional space) and the decoder network attempts to reproduce the original data using the learned features with minimal error. As the unsupervised training process progresses and the corresponding reconstruction error is reduced, the network uncovers better feature representations of the data without relying on data annotations.

In the context of sensor-based activity recognition, autoencoders have been successfully applied to exploit unlabelled activity data for *unsupervised pre-training* and *unsupervised representation learning* tasks. Notably, the encoder and decoder components can incorporate different variations of deep neural networks—*i.e.*, MLPs, CNNs, or RNNs—to uncover representations from sequential sensory data. We present an overview of autoencoder frameworks for unsupervised feature extraction from raw multi-channel time-series data captured using wearables in Fig. 2.4. The reconstruction of unlabelled sensory data by imposing a bottleneck layer in a

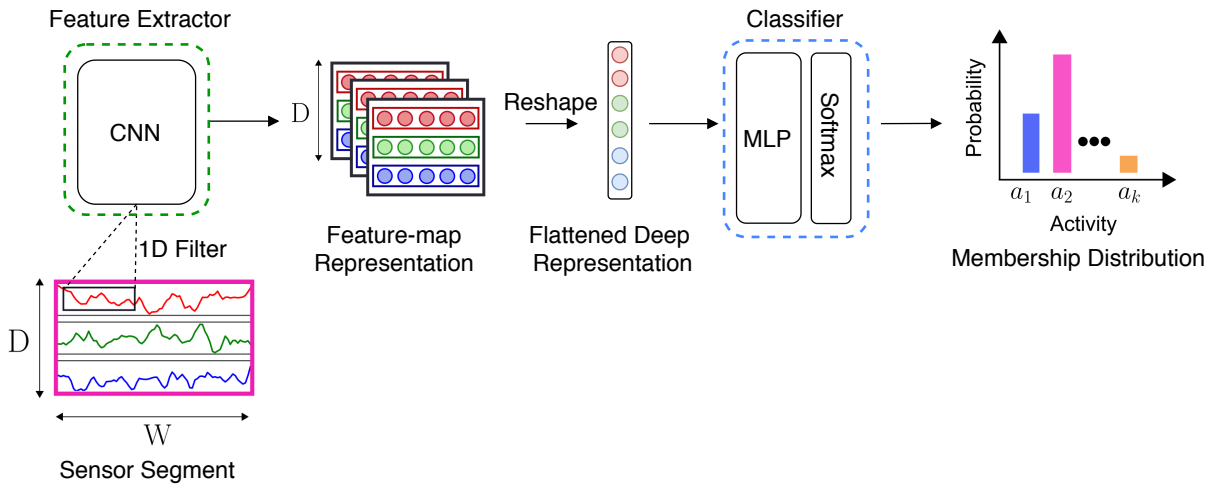


Figure 2.4: Convolutional Neural Networks (CNNs). We illustrate an overview of HAR frameworks integrating CNNs for automatic feature extraction from multi-channel time-series data.

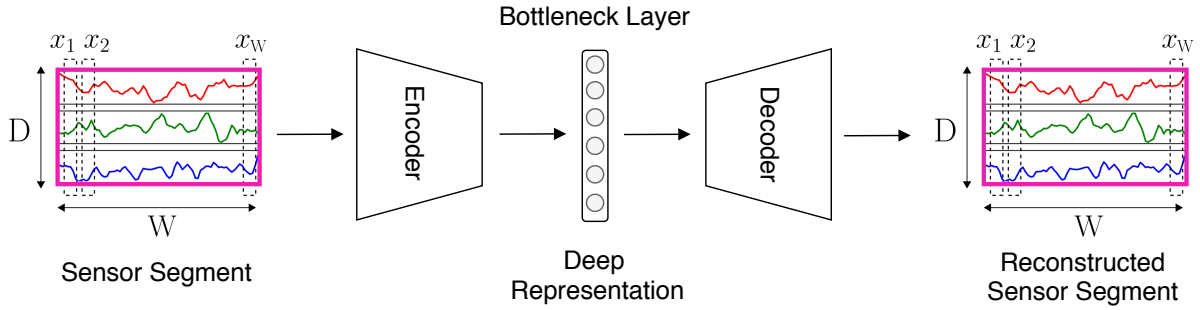


Figure 2.5: AutoEncoder Neural Networks. We illustrate an overview of HAR frameworks adopting autoencoders for unsupervised feature learning from multi-channel time-series data.

low-dimensional space encourages the network to capture only the most salient activity features of the input. These features are critical to the successful reconstruction of the data and, thus, are expected to encode enriched representations.

2.4 Evaluation Metrics

We introduce the evaluation metrics which are most commonly used in the HAR literature to ground studies and quantify experimental results. First, we provide a definition for the primitive terminologies of *True Positive (TP)*, *False Positive (FP)*, *True Negative (TN)*, and *False Negative (FN)*:

- *True Positive (TP).* This term indicates a correct prediction of an activity label that has indeed occurred.
- *False Positive (FP).* This term indicates an incorrect prediction of an activity label that has not occurred.
- *True Negative (TN).* This term indicates a correct rejection of an activity label that has not occurred.
- *False Negative (FN).* This term indicates an incorrect rejection of an activity label that has occurred.

Leveraging these terms, we present formal definitions for *accuracy*, *precision*, *recall* and *F1-score*. These definitions help us to evaluate the performance of the developed HAR frameworks in terms of the relevant activity recognition tasks.

Accuracy (Acc). For a given activity class $a_i \in \mathcal{A}$, we compute the HAR model's accuracy—denoted by Acc_i—to quantify the proportion of correctly classified activity segments (*i.e.*, $\text{TP}_i + \text{TN}_i$) over the total number of activity occurrences in the predictions and ground-truth (*i.e.*, $\text{TP}_i + \text{TN}_i + \text{FP}_i + \text{FN}_i$),

$$\text{Acc}_i = \frac{\text{TP}_i + \text{TN}_i}{\text{TP}_i + \text{TN}_i + \text{FP}_i + \text{FN}_i}. \tag{2.3}$$

Precision. For a given activity class $a_i \in \mathcal{A}$, we compute the HAR model's precision—denoted by Precision_i—to quantify the proportion of correctly predicted activity occurrences (*i.e.*, TP_i) over the total number of activity occurrences in the predictions (*i.e.*, $\text{TP}_i + \text{FP}_i$),

$$\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}. \tag{2.4}$$

Recall. For a given activity class $a_i \in \mathcal{A}$, we compute the HAR model's recall—denoted by Recall_i—to quantify the proportion of correctly predicted activity occurrences (*i.e.*, TP_i) over the total number of label occurrences in the ground-truth (*i.e.*, $\text{TP}_i + \text{FN}_i$),

$$\text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}. \tag{2.5}$$

F-score. For a given activity class $a_i \in \mathcal{A}$, we compute the HAR model's F-score—denoted by F-score_i—by taking into account both precision and recall values and computing their harmonic mean,

$$\text{F-score}_i = 2\left(\frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}\right). \tag{2.6}$$

Notably, the evaluation metrics presented above provide a means to quantify the recognition performance of an HAR framework for a specific activity category. In order to aggregate the recognition performance achieved across all activity

classes, we compute their corresponding *class-average*—denoted by $(.)_m$—and *weighted-average*—denoted by $(.)_w$—to serve as global evaluation metrics

$$\begin{aligned}
\text{Acc}_m &= \frac{1}{k}\sum_{i=1}^{k}\text{Acc}_i & \text{Acc}_w &= \sum_{i=1}^{k} w_i \times \text{Acc}_i \\
\text{Precision}_m &= \frac{1}{k}\sum_{i=1}^{k}\text{Precision}_i & \text{Precision}_w &= \sum_{i=1}^{k} w_i \times \text{Precision}_i \\
\text{Recall}_m &= \frac{1}{k}\sum_{i=1}^{k}\text{Recall}_i & \text{Recall}_w &= \sum_{i=1}^{k} w_i \times \text{Recall}_i \\
\text{F-score}_m &= \frac{1}{k}\sum_{i=1}^{k}\text{F-score}_i & \text{F-score}_w &= \sum_{i=1}^{k} w_i \times \text{F-score}_i
\end{aligned} \tag{2.7}$$

where, w_i is the ratio of sensor segments belonging to the activity class $a_i \in \mathcal{A}$. It is important to note that the *weighted-average* takes into account the frequency of samples belonging to a specific activity class. As such, it is highly affected by the distribution of activity labels. In contrast, the *class-average* weights each activity category equally and reflects the ability of the HAR model to recognise every activity category, regardless of its prevalence in the collected data.

Chapter 3

Supervised Learning of Enriched Activity Feature Representations

THIS chapter considers the problem of automatic feature representation learning using multi-channel time-series data captured by wearable sensors for supervised Human Activity Recognition (HAR). Although our collective know-how to solve HAR problems with wearables has progressed immensely with end-to-end deep learning paradigms, several fundamental opportunities remain overlooked. This chapter rigorously explores these new opportunities to learn enriched and highly discriminating activity representations. This chapter proposes: (*a*) learning to exploit the *latent* relationships between multi-channel sensor modalities and specific activities; (*b*) investigating the effectiveness of *data-agnostic augmentation* for multi-modal sensor data streams to regularise deep HAR models; and (*c*) incorporating a classification loss criterion to encourage minimal intra-class representation differences whilst maximising inter-class differences to achieve more discriminative features. The contributions from the design concepts are validated through extensive experiments, including *activity misalignment* measures, *ablation* studies and insights shared through both quantitative and qualitative studies.

3.1 Motivation and Contribution

Recently, the adoption of deep neural networks for sensor-based human activity recognition (HAR) has created effective pipelines for end-to-end learning of activity recognition models from raw multi-channel time-series data. Despite the progress made towards addressing supervised HAR problems, this chapter particularly discusses opportunities to improve upon the automated feature representation learning process. In our efforts, key under explored dimensions are uncovered with significant potential to effectively enrich activity feature-maps, achieve more discriminative representation, and obtain better generalisation performance in recognition of human activities. In particular:

- HAR data acquisition often involves recording of motion measurements over number of sensors and channels. Therefore, we can expect the capability of different sensor modalities and channels to capture and encode some activities better than others whilst having complex interactions between sensors, channels and activities. Thus, we hypothesise that learning to exploit the relationships between multi-channel sensor modalities and specific activities can contribute to learning enriched activity representations—*this insight remains unstudied*.

- Human actions, for example walking and walking up-stairs, exhibit significant intra-class variability and inter-class similarities. This suggests imposing optimisation objectives for training that not only ensure class separability but also encourage compactness in the established feature space. However, *the commonly adopted cross-entropy loss function does not jointly accommodate both objectives*.

- Due to the laborious process of collecting annotated sequences with wearables, sensor HAR datasets are often small in size. While expanding the training data with virtual samples has proved beneficial in achieving better *generalization* performance for general machine learning problems, exploration of data augmentation for HAR has been largely limited to hand-crafted techniques that alter the data sequences with the assumption of being able to preserve the activity label semantics. However, achieving label-preserving transformations for wearable HAR sensor data is not obvious and intuitively recognizable Um et al. (2017). Thus, the augmented data may not necessarily preserve salient characteristics embedded within the original data, leading to alteration of the activity labels and potentially deluding the supervised training process. For

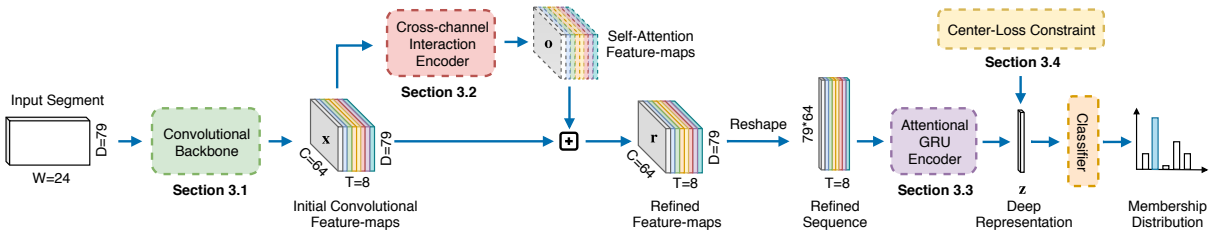


Figure 3.1: Enhanced feature extraction work-flow for wearable activity data. We aggregate the proposed components for achieving highly discriminative and generalisable activity feature representation in a unified HAR framework that seamlessly operates on raw multi-channel time-series data captured by wearables.

instance, in the image domain, a flipped image of a person is still a meaningful illustration of the person concept whilst applying the same method and flipping sensor channels of an inertial sensor leads to a completely different signal.

Motivated by these opportunities, the *key contribution* in this chapter is to propose novel end-to-end trainable components to enhance automatic feature extraction from multi-channel time-series data captured by wearables. We substantiate the effectiveness and generalisability of the proposed elements in a unified HAR framework—illustrated in Fig. 3.1—in achieving more discriminative and generalisable activity feature representations across multiple diverse wearable sensor datasets. The key contributions are summarised below:

1. This chapter *proposes and designs* a *cross-channel interaction encoder* to incorporate a self-attention mechanism to learn to exploit the different capabilities of sensor modalities and latent interactions between multiple sensor channels capturing and encoding activities. The encoder module captures latent correlations between multi-sensor channels to generate self-attention feature maps and enrich the convolutional feature representations (**Section 3.3.2** and **Fig. 3.9**).

2. Temporal attention layers were recently shown in Murahari and Plötz (2018) to improve performance by capturing temporal context in a network constructed using LSTM (long short-term memory) layers capable of learning dependencies in sequences. Therefore, this chapter *designs* an *attentional GRU (gated recurrent unit) encoder* to enhance the sequence of self-attention enriched features by further capturing the relevant temporal context (**Section 3.3.3**). Compared with LSTMs, GRUs are easier to train and leverage fewer parameters Zheng et al. (2018).

3. In recognizing the intra-class variations of HAR activities, this chapter proposes adopting the *center-loss criterion* to encourage minimal intra-class representation differences whilst maximising inter-class differences to achieve more discriminative features and demonstrate the effectiveness of center-loss penalisation for learning highly discriminative activity representations for wearable HAR problems. (**Section 3.3.4**).

4. In recognizing the difficulty of ensuring label-preserving augmentation with hand-crafted approaches in wearable HAR problems, this chapter proposes adopting *mixup* method to take into account both data and label information for multi-modal sensor data augmentation, investigates the effectiveness of the method and demonstrates the seamless integration of mixup for wearable HAR problems (**Section 3.3.5**). Importantly unlike existing augmentation approaches that are dataset dependent and thus require domain expert knowledge for effective adoption, mixup is domain independent and simple to apply; an important consideration for wearable HAR problems based on multiple different sensor modalities and sensor specific semantic and signal characteristics.

5. Under a unified evaluation protocol, the effectiveness and generalisability of the incorporated components is substantiated on diverse HAR datasets (**Section 3.4.4**). Further, the key insights gained from the study in this chapter are shared through extensive quantitative and qualitative results as well as an ablation study to comprehensively demonstrate the contributions made by the architectural elements (**Section 3.4.5**).

3.2 Related Work

3.2.1 Automatic Feature Learning in HAR

Over the past years, the emerging paradigm of deep learning has largely alleviated the need for laborious feature engineering procedures and provided an effective means for automated end-to-end feature extraction from multi-channel time-series data in HAR.

Pioneering studies in the field have explored Restricted Boltzmann Machines (RBMs) for automatic representation learning Plötz, Hammerla and Olivier (2011); Hammerla et al. (2015); Zhang, Wu and Luo (2015); Alsheikh et al. (2016). Recently,

deep architectures based on convolutional neural networks (CNNs) have been predominantly leveraged to automate feature extraction from sensor data streams while mutually enhancing activity classification performance Zeng et al. (2014); Yang et al. (2015); Ronao and Cho (2015); Bhattacharya and Lane (2016). These studies typically employ a cascaded hierarchy of 1D convolution filters along the temporal dimensions to capture salient activity features at progressively more abstract resolutions. The acquired latent features are ultimately unified and mapped into activity class scores using a fully connected network. In particular, Zeng et al. (2014) adopts 1D convolutional filters along the temporal dimension of accelerometer signals to capture local dependencies and scale invariant features. Similarly, Yang et al. (2015) proposes a CNN architecture that employs convolution and pooling layers to capture salient sensor signal patterns at different time scales. In Ronao and Cho (2015), a study is conducted to compare recognition performance of end-to-end trained CNNs against traditional HAR algorithms. In another study, Chen and Xue (2015) investigates the effective kernel width for convolutional operators to automate feature extraction in HAR. Taking a different approach, Yao et al. (2018) develops a fully convolutional HAR architecture that leverages 2D filters to simultaneously detect temporal and spatial feature representations from input sensor data.

Another popular architecture design for HAR adopts deep recurrent neural networks (RNNs) that leverage memory cells to directly model temporal dependencies between subsequent sensor samples. In particular, Inoue, Inoue and Nishida (2018) develops a deep RNN for activity recognition from raw accelerometer data. In Hammerla, Halloran and Plötz (2016), authors investigate forward and bi-directional long short-term memory (LSTM) networks to capture temporal dynamics in both forward and backward directions. In another study, Guan and Plötz (2017) demonstrates how multiple LSTM models can be ensembled to achieve superior recognition performance. In order to reduce the computational complexity, Edel and Köppe (2016) develops a binarised LSTM network for activity recognition on resource-constrained devices.

Combining the representational power of CNNs with RNNs, Ordóñez and Roggen (2016) proposes *DeepConvLSTM* by pairing convolutional and recurrent networks in order to model the temporal correlations at a more abstract representation level. This concept is further extended in Murahari and Plötz (2018), where the recurrent network of DeepConvLSTM is expanded with attention layers to model the relevant temporal context of sensor data. Similarly, Yao et al. (2017) develops *deepsense* by integrating

convolutional layers together with a GRU network. Within the framework, the CNN module extracts local features from individual sensor modalities and hierarchically merges them into global features. Subsequently, the GRU module exploits the learned features and extracts temporal relationships.

3.2.2 Data Augmentation

Data augmentation constitutes an explicit approach to effectively regularise deep neural networks and improve their generalisation performance through artificial expansion of the training dataset Simonyan and Zisserman (2014); He et al. (2016); Zhong et al. (2020). Despite its demonstrated effectiveness for general machine learning problems, considerably limited research efforts in HAR have focused on investigating systematic data augmentation techniques for wearable sensor data. In particular, Um et al. (2017) investigates hand-crafted augmentation approaches including jittering, scaling, cropping, permutation and axis rotations for monitoring of Parkinson's disease using wearable sensors with convolutional neural networks. In Mathur et al. (2018), data augmentation is applied to sensor data in order to specifically counter sampling-jitters resulting from software and hardware heterogeneity in diverse sensing devices. In another study, Faridee et al. (2019) explores a series of sequentially applied transformations—rotation, time-warp, scaling and jittering—in a semi-supervised transfer learning framework for complex human activity recognition.

While the use of data augmentation in these studies consistently demonstrates improved generalisation to unseen data, the incorporated strategies are dataset-dependent and rely on the use of domain expert knowledge for effective and meaningful adoption; *e.g.* it is not straightforward and clear what degree of sensor data scaling is considered reasonable to apply without altering the semantic activity label of the original data. This becomes even more problematic as wearable data are often captured over multitude of sensor channels with diverse magnitudes and innate properties; thus, complicating manual design of label-preserving sensor augmentations. This necessitates investigation of data-agnostic augmentation approaches for multi-channel times-series data in HAR that can be applied to effectively expand the training data captured by diverse sensing modalities without reliance on domain expert knowledge.

Summary. Despite the great progress in the field, we can see that the unique opportunities discussed in Section 3.1 for learning from multi-channel time-series data generated by body-worn sensors remain. Conventionally, the feature-maps generated by convolutional layers are trivially vectorised and fed to fully connected layers or recurrent networks to ultimately produce classification outcomes. However, such manipulation of the convolutional feature-maps fails to explicitly capture and encode the inter-channel interactions that can aid accurate recognition of activities. Moreover, regardless of the architectural designs, cross-entropy loss constitutes the common choice for supervised training of deep HAR models. Yet, this optimisation objective alone does not cater for the need to achieve minimal intra-class compactness of feature representations Wen et al. (2016) necessary to counter the significant intra-class variability of human activities. In addition, while data augmentation has shown great potential for regularising deep neural networks in the computer vision domain, the effectiveness of data-agnostic augmentation for multi-channel time-series data captured by wearables remains under-utilised for HAR.

3.3 Proposed Methodology

The goal is to develop an end-to-end deep HAR model that directly consumes raw sensory data captured by wearables and seamlessly outputs precise activity classification decisions. In our proposed methodology, a network composed of 1D convolutional layers serves as the backbone feature extractor in order to automatically extract an initial feature representation for each sensory segment. Subsequently, a two-staged refinement process is proposed to enrich the initial feature representations prior to classification that allows the model to *i*) effectively uncover and encode the underlying sensor channel interactions at each time-step, and *ii*) learn the relevant temporal context within the sequence of refined representations. Moreover, we encourage intra-class compactness of representations with centre-loss while regularising the network with mixup data augmentation during training. In what follows, the network components illustrated in Fig. 3.1 are elaborated.

3.3.1 1D Convolutional Backbone

Following the sliding window segmentation, the input to the network is a slice of the captured time-series data $\mathbf{x} \in \mathbb{R}^{D \times W}$, where D denotes the number of sensor channels

used for data acquisition and W represents the choice for the window duration. For automatic feature extraction, the input is then processed by a convolutional backbone operating along the temporal dimension. Given the 1D structure of the adopted filters, progressively more abstract temporal representations are learned from nearby samples without fusing features in-between different sensor channels. Ultimately, the backbone yields a feature representation $\bar{\mathbf{x}} \in \mathbb{R}^{\mathrm{C}\times\mathrm{D}\times\mathrm{T}}$, where in each of the C feature maps, the sensor channel dimension D is preserved while the temporal resolution is down-sampled to T. Without loss of generality, in this chapter we employ the convolutional layers of a state-of-the-art HAR model Ordóñez and Roggen (2016) as the backbone feature extractor; the input segment is successively processed by four layers, each utilising 64 one-dimensional filters of size 5 along the temporal axis with ReLU non-linearities.

3.3.2 Cross-Channel Interaction Encoder (CIE)

Accurate realisation of fine-grained human actions using wearables is often associated with utilising multitude of on-body sensing devices that capture activity data across multiple channels. Measurements captured by different sensor channels provide different views of the same undergoing activity and are thus, inherently binded together in an unobservable latent space. Accordingly, we seek to design an end-to-end trainable module that takes as input the initial convolutional feature-maps at each time-step, learns the interactions between any two sensor channels within the feature-maps, and leverages this overlooked source of information to enrich the sensory feature representations for HAR.

Motivated by the emerging successful applications of self-attention Vaswani et al. (2017); Wang et al. (2018b); Zhang et al. (2019) in capturing global dependencies by computing relations at any two positions of the input, here a *Cross-Channel Interaction Encoder (CIE)* is designed that adopts self-attention mechanism to effectively process the initial feature representations and uncover the latent channel interactions. To this end, the normalised correlations are first computed across all pairs of sensor channel features $\bar{\mathbf{x}}_t^d$ and $\bar{\mathbf{x}}_t^{d'}$ using the embedded Gaussian function at each time-step t,

$$\mathbf{a}_t^{d,d'} = \frac{\exp\left(f(\bar{\mathbf{x}}_t^d)^\intercal g(\bar{\mathbf{x}}_t^{d'})\right)}{\sum_{d'=1}^{\mathrm{D}} \exp\left(f(\bar{\mathbf{x}}_t^d)^\intercal g(\bar{\mathbf{x}}_t^{d'})\right)}, \tag{3.1}$$

where $\mathbf{a}_t^{d,d'}$ indicates the attendance of the model to the features of sensor channel d' when refining representations for sensor channel d. Subsequently, the extracted correlations are leveraged in order to compute the response for the d^{th} sensor channel features $\mathbf{x}_t^d \in \mathbb{R}^{\text{C}}$ and generate the corresponding self-attention feature-maps $\mathbf{o}_t^d$ at each time-step

$$\mathbf{o}_t^d = v\Big(\sum_{d'=1}^{\text{D}} \mathbf{a}_t^{d,d'} h(\mathbf{x}_t^{d'})\Big). \tag{3.2}$$

Technically, the self-attention in the CIE module functions as a non-local operation which computes the response for sensor channel d at each time-step by attending to all present sensor channels' representations in the feature-maps at the same time-step. In the above, f, g, h, and v all represent linear embeddings with learnable weight matrices ($\in \mathbb{R}^{\text{C}\times\text{C}}$) that project feature representations into new embedding spaces where computations are carried out. Having obtained the self-attention feature-maps, the initial feature-maps are then added back via a residual link (indicated by $\bigoplus$ in Fig. 3.1) to encode the interactions and generate the refined feature representations $\mathbf{r}_t$,

$$\mathbf{r}_t^d = \mathbf{o}_t^d + \mathbf{x}_t^d. \tag{3.3}$$

With the residual connection in place, the model can flexibly decide to use or discard the correlation information. During training, the HAR model leverages the CIE module to capture the interactions between different sensor channels. The discovered correlations are encoded inside the self-attention weights and leveraged at inference time to help support the model's predictions.

3.3.3 Attentional GRU Encoder (AGE)

As a result of employing the CIE module, the feature-maps generated at each time-step are now contextualised with the underlying cross-channel interactions. As shown in Fig. 3.1, the representations at each time-step are vectorised to obtain a sequence of refined feature vectors $(\mathbf{r}_t \in \mathbb{R}^{\text{CD}})_{t=1}^{\text{T}}$ ready for sequence modelling. Given that not all time-steps equally contribute in recognition of the undergoing activities, it is crucial to learn the relevance of each feature vector in the sequence when representing activity categories. In this regard, applying attention layers to model the relevant temporal context of activities has proved beneficial in recent HAR studies Murahari and Plötz

(2018). Adopting a similar approach, we utilise a 2-layer *attentional GRU Encoder (AGE)* to process the sequence of refined representations and learn soft attention weights for the generated hidden states $(\mathbf{h}_t)_{t=1}^{T}$. In the absence of attention mechanism in the temporal domain, classification decision would only be based on the last hidden state achieved after observing the entire sequence. By contrast, empowering the GRU encoder with attention alleviates the burden on the last hidden state and instead, allows learning a holistic summary $\mathbf{z}$ that takes into account the relative importance of the time-steps

$$\mathbf{z} = \sum_{t} \beta_t \mathbf{h}_t, \tag{3.4}$$

where β_t denotes the computed attention weight for time-step t. Technically, attention values are obtained by first mapping each hidden state into a single score with a linear layer and then normalising these scores across the time-steps with a softmax function.

3.3.4 Centre Loss Augmented Objective

Intra-class variability and inter-class similarity are two fundamental challenges of HAR with wearables. The former phenomena occurs since different individuals may execute the same activity differently while the latter challenge arises when different classes of activities reflect very similar sensor patterns. To counter these challenges, the training objective should encourage the model to learn discriminative activity representations; *i.e.*, representations that exhibit large inter-class differences as well as minimized intra-class variations.

Existing HAR architectures solely rely on the supervision signal provided by the cross-entropy loss during their training phase. While optimising for this criteria directs the training process towards yielding inter-class separable activity features, it does not explicitly encourage learning intra-class compact representations. To boost the discriminative power of the deep activity features within the learned latent space, this chapter proposes to incorporate center-loss Wen et al. (2016) for training the HAR model. The auxiliary supervision signal provided by centre-loss penalises the

distances between activity representations and their corresponding class centres and thus, reduces intra-class feature variations. Formally, centre-loss is defined as

$$\mathcal{L}_c = \frac{1}{2}\sum_{i=1}^{n} \|\mathbf{z}_i - \mathbf{c}_{y_i}\|_2^2, \tag{3.5}$$

where $\mathbf{z}_i \in \mathbb{R}^z$ denotes the deep representation for sensory segment $\mathbf{x}_i$, and $\mathbf{c}_{y_i} \in \mathbb{R}^z$ denotes the y_ith activity class centre computed by averaging the features of the corresponding class. This criteria is enforced on the activity representations obtained from the penultimate layer of the network to effectively pull the deep features towards their class centres.

In each iteration of the training process, the joint supervision of cross-entropy loss together with centre-loss is leveraged to simultaneously update the network parameters and the class centres $\mathbf{c}_y$ in an end-to-end manner. Hence, the aggregated optimisation objective is formulated as

$$\Theta^* = \arg\min_{\Theta} \mathcal{L} + \gamma\mathcal{L}_c, \tag{3.6}$$

where $\mathcal{L}$ represents the cross-entropy loss, γ is the balancing coefficient between the two loss functions, and Θ denotes the collection of all trainable parameters.

3.3.5 Mixup Data Augmentation for HAR

Due to the laborious task of collecting annotated datasets from wearables, current HAR benchmarks are characterised by their limited sizes. Therefore, introducing new modules and increasing the network parameters without employing effective regularisation techniques, makes the model prone to overfitting and endangers its generalisation. In this regard, while extending the training data with augmented samples achieved by *e.g.* slight rotations, scaling, and cropping has consistently led to improved generalisation performance for computer vision applications, these methods are not directly applicable to multi-channel time-series data captured by wearables.

Accordingly, we explore the effectiveness of a recently proposed data-agnostic augmentation strategy, namely *mixup* Zhang et al. (2018a), for time-series data in order to regularise the deep HAR model. This approach has demonstrated its potential in significantly improving the generalisation of deep neural networks by encouraging simple linear behaviour in-between training data. In addition, unlike

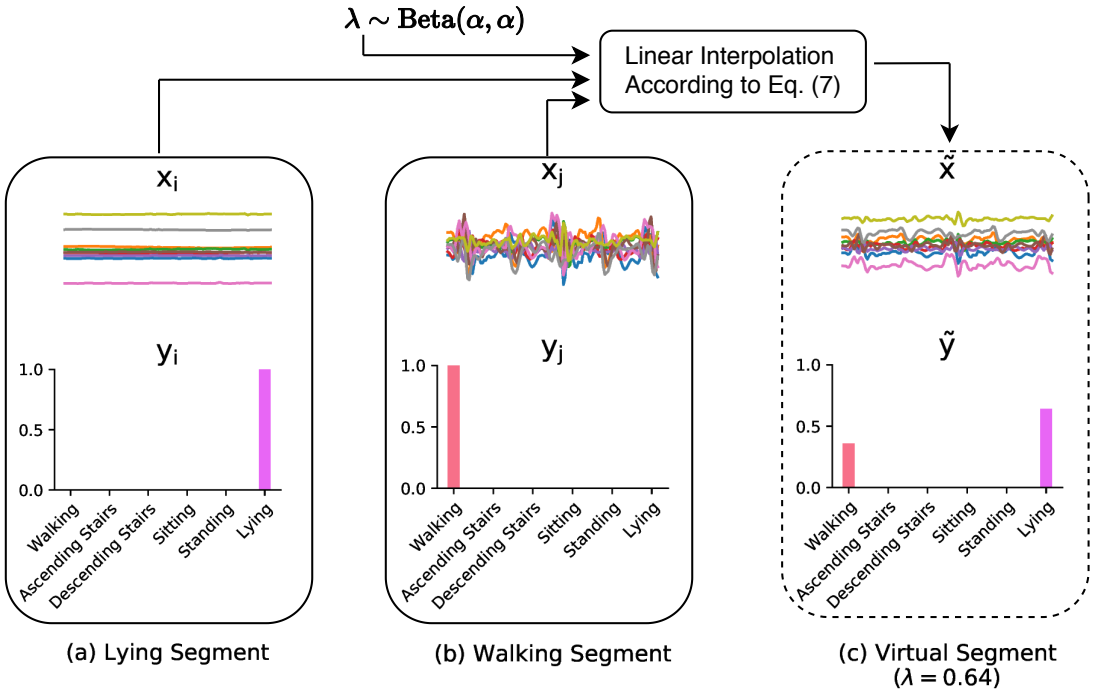


Figure 3.2: Mixup multi-channel time-series data augmentation. We leverage mixup data augmentation technique to generate virtual sequences during training. We interpolate in-between samples. Here, we visualise (a) a sequence of sensor data from the training split corresponding to the `lying` activity and its one-hot encoded label representation, (b) a training sensor data segment corresponding to the `walking` activity, and (c) a *virtual* or generated sequence and its target label according to Eq. 3.7 with a drawn λ value of 0.64 (sampled from a Beta distribution). The visualised data corresponds to a subset of sensor channels in the PAMAP2 dataset Reiss and Stricker (2012).

existing augmentation approaches that are dataset dependent and thus require domain expert knowledge for effective adoption, mixup strategy is domain independent and simple to apply. In essence, mixup yields augmented virtual example $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ through linear interpolation of training example pairs $(\mathbf{x}_i, \mathbf{y}_i)$ and $(\mathbf{x}_j, \mathbf{y}_j)$,

$$\begin{aligned} \tilde{\mathbf{x}} &= \lambda \mathbf{x}_i + (1-\lambda)\mathbf{x}_j \\ \tilde{\mathbf{y}} &= \lambda \mathbf{y}_i + (1-\lambda)\mathbf{y}_j, \end{aligned} \tag{3.7}$$

where λ sampled from a Beta(α, α) distribution is the mixing-ratio and α is the mixup hyper-parameter controlling the strength of the interpolation. Notably, mixup augmentation allows efficient generation of virtual examples on-the-fly by randomly picking pairs from the same minibatch in each iteration. In this work, we adopt mixup strategy to augment the time-series segments in each mini-batch and train the model end-to-end with the generated samples. We visually explain the augmentation process

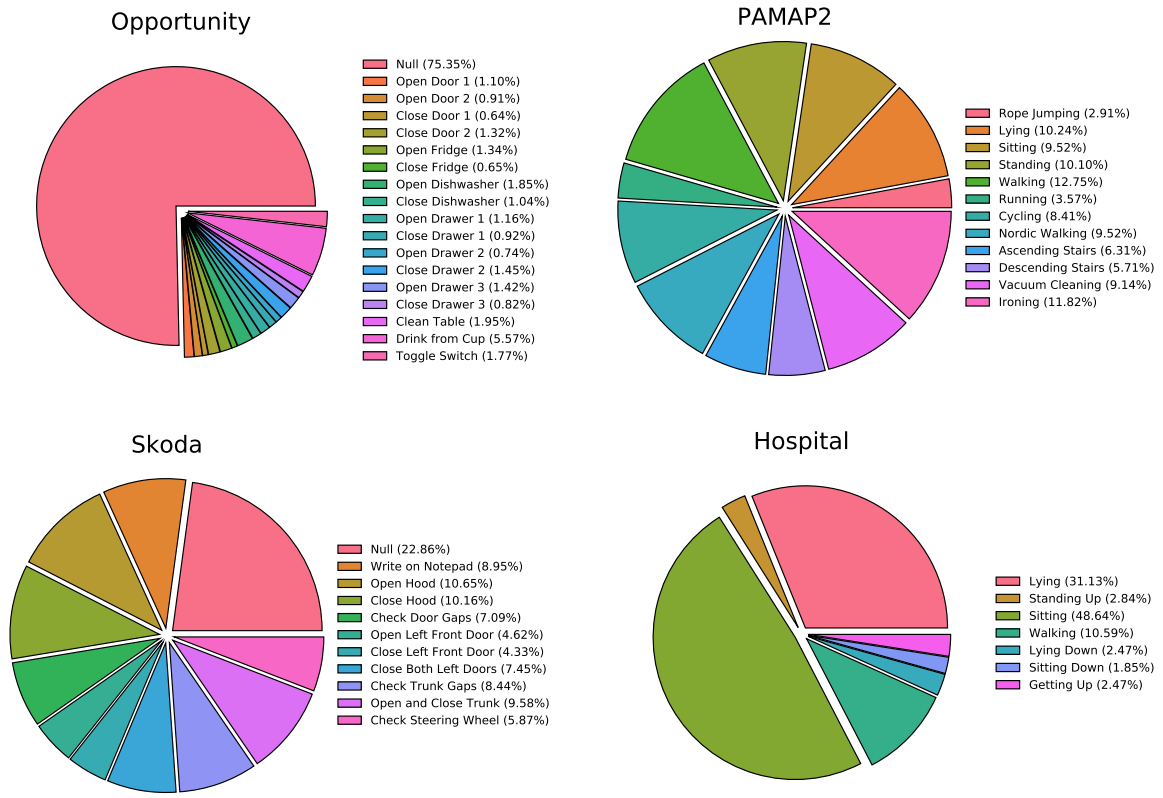


Figure 3.3: Activity distributions in benchmark HAR datasets. We illustrate the activity categories covered and their corresponding distributions within each dataset.

with an example in Fig. 3.2, where a pair of randomly drawn training data sequences are linearly interpolated to yield a novel virtual sequence.

3.4 Experiments and Results

3.4.1 Datasets

To validate the effectiveness of the incorporated network elements and provide empirical evidence of their generalisability, four HAR benchmarks are employed that exhibit great diversity in terms of the sensing modalities used and the activities to be recognised. A brief description of the datasets is provided in what follows.

Opportunity Dataset Chavarriaga et al. (2013). This dataset is captured by multiple body-worn sensors. Four participants wearing the sensors were instructed to carry out naturalistic kitchen routines. The data is recorded at a frequency of 30 Hz and

is annotated with 17 sporadic gestures as well as a `Null` class. Following Hammerla, Halloran and Plötz (2016), the 79 sensor channels not indicating packet-loss are used. For evaluation, we use runs 4 and 5 from subjects 2 and 3 as the holdout test-set, run 2 from participant 1 as the validation-set, and the remaining data as the training-set.

PAMAP2 Dataset Reiss and Stricker (2012). This dataset is aimed at recognition 12 diverse activities of daily life. Data was recorded over 52 channels with annotations covering prolonged household and sportive actions. Replicating Hammerla, Halloran and Plötz (2016), we use runs 1 and 2 from subject 6 as the holdout test-set, runs 1 and 2 from subject 5 as the validation-set, and the remaining data for training.

Skoda Dataset Stiefmeier et al. (2008). The dataset covers the problem of recognising 10 manipulating gestures of assembly-line workers in a manufacturing scenario. Following Guan and Plötz (2017), we use the data recorded over 60 sensor channels collected from the right arm, utilise the first 80% of each class for the training-set, the following 10% for validation and the remainder as the test-set.

Hospital Dataset Yao et al. (2018). This dataset is collected from 12 hospitalised older patients wearing an inertial sensor over their garment while performing 7 categories of activities. All the data is recorded at 10 Hz. Following Yao et al. (2018), data from the first eight subjects are used for training, the following three for testing, and the remaining for the validation set.

3.4.2 Experimental Setup

To ensure a fair comparison, we directly adopt the evaluation protocol and metrics used in the recent literature Hammerla, Halloran and Plötz (2016); Guan and Plötz (2017); Murahari and Plötz (2018); Haresamudram, Anderson and Plötz (2019). Where possible, sensor data are down-sampled to 33 Hz to achieve a consistent temporal resolution with the Opportunity dataset. Each sensor channel is normalised to zero mean and unit variance using the training data statistics. The training data is partitioned into segments using a sliding window of 24 samples (*i.e.*, W=24) with 50% overlap between adjacent windows. For a realistic setup, sample-wise evaluation is adopted to compare the performance on the test-set; thus, a prediction is made for

Table 3.1: Hyper-parameters. Summary of hyper-parameter values selected per dataset. All other hyper-parameters are kept constant across all datasets.

Hyper-parameter	Opportunity	PAMAP2	Skoda	Hospital
Dropout ratio p_{feat}	0.5	0.9	0.5	0.5
Dropout ratio p_{cls}	0.5	0.5	0.0	0.5
Weighting coefficient γ	3×10^{-4}	3×10^{-3}	3×10^{-1}	3×10^{-1}

every sample of the test sequence as opposed to every segment. Given the imbalanced class distributions in the datasets (see Figure 3.3), as in Guan and Plötz (2017); Hammerla, Halloran and Plötz (2016); Murahari and Plötz (2018), the class-average F-score

$$\text{F-score}_m = \frac{1}{k} \sum_{i=1}^{k} \text{F-score}_i \tag{3.8}$$

is used as the evaluation metric to reflect the ability of the HAR model to recognise every activity category regardless of its prevalence in the collected data. Here, k denotes the number of activity classes and F-score_i is the harmonic mean of precision and recall terms computed for activity class a_i according to Eq. 2.6.

3.4.3 Implementation Details

The experiments are implemented using Pytorch Paszke et al. (2017). The entire network is trained end-to-end for 300 epochs by back-propagating the gradients of the loss function based on mini-batches of size 256 and in accordance with the Adam Kingma and Ba (2015) update rule. The learning rate is set to 10^{-3} and decayed every 10 epochs by a factor of 0.9. For *mixup* augmentation, we fix α=0.8. All these hyper-parameters are kept constant across all datasets. For each dataset, we choose a dropout probability $p \in \{0, 0.25, 0.5, 0.75, 0.9\}$ for the refined feature-maps (p_{feat}) and the feature vectors fed to the classifier (p_{cls}), and select the centre-loss weighting coefficient $\gamma \in 3 \times \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, as summarised in Table 3.1.

3.4.4 Results

Classification Measure. We compare the effectiveness of our learned activity feature representations against state-of-the-art HAR models on four standard benchmarks

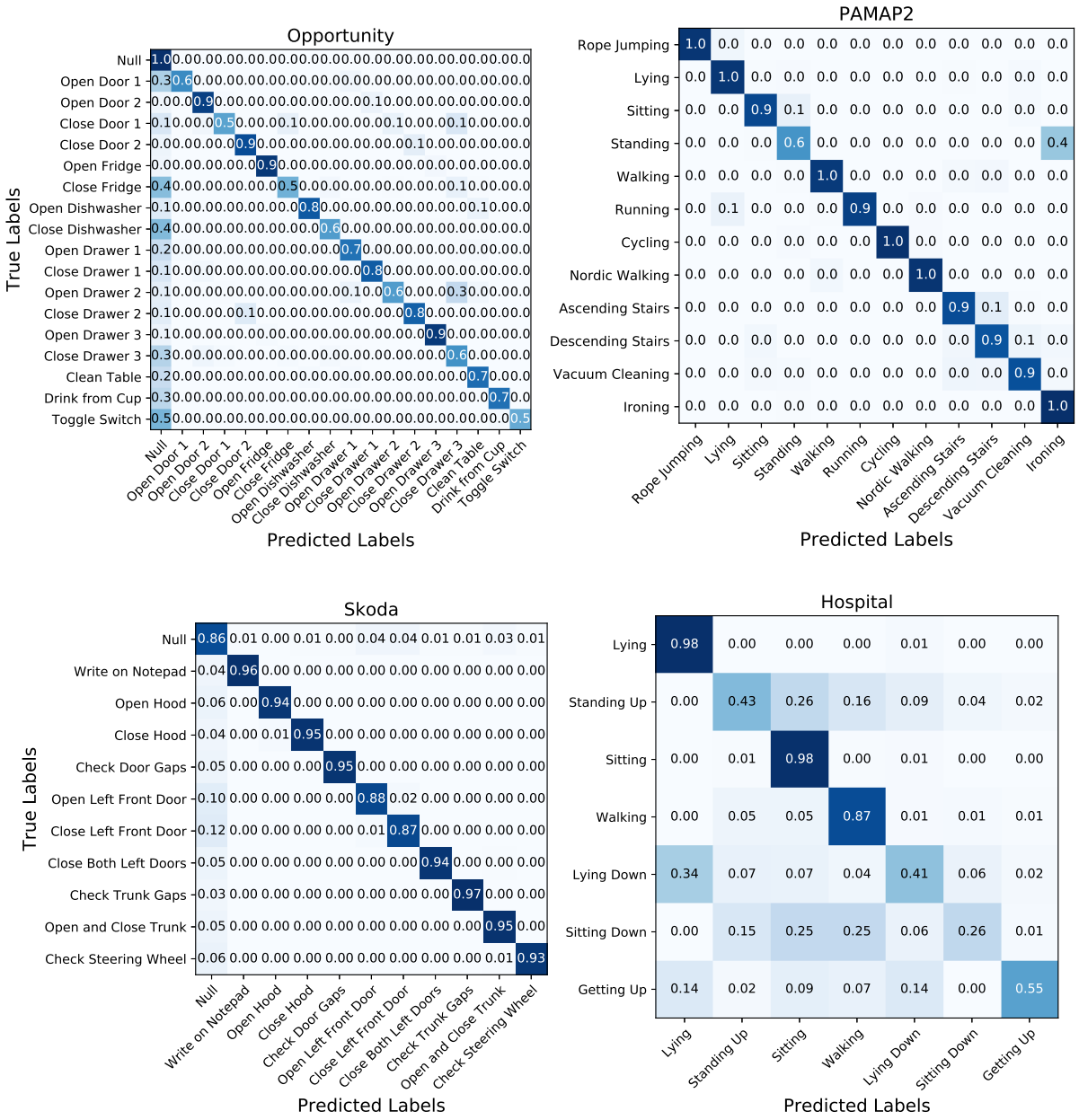


Figure 3.4: Class-specific recognition performance. We illustrate the confusion matrices highlighting the class-specific recognition performance for the testing splits of Opportunity, PAMAP2, Skoda, and Hospital HAR datasets. The vertical axis represents the ground-truth activity categories and the horizontal axis denotes the predicted activities.

Table 3.2: Hold-out evaluation. We present a comparison of sample-wise activity recognition performance based on class-average F-scores on the holdout test sequences. The baseline results are quoted from Guan and Plötz (2017); Murahari and Plötz (2018), except for (*) where the publications' code is used.

HAR Model	Opportunity	PAMAP2	Skoda	Hospital*
LSTM Learner Baseline Guan and Plötz (2017)	65.9	75.6	90.4	62.7
DeepConvLSTM Ordóñez and Roggen (2016)	67.2	74.8	91.2	62.8
b-LSTM-S Hammerla, Halloran and Plötz (2016)	68.4	83.8	92.1	63.6
Dense Labelling Yao et al. (2018)*	62.4	85.4	91.6	62.9
Att. Model Murahari and Plötz (2018)	70.7	87.5	91.3	64.1
Ours	**74.6**	**90.8**	**92.8**	**66.6**
(Improvement over Runner-up)	**(5.52%)**	**(3.77%)**	**(0.76%)**	**(3.9%)**

in Table 3.2. As elucidated in Section 3.4.2, every baseline generates sample-wise predictions on the entire holdout test sequence and the performance is judged based on the acquired class-average F-score (F-score$_m$). The baseline results are directly quoted from Guan and Plötz (2017); Murahari and Plötz (2018), except where indicated by (*), where the published code is used with the standard evaluation protocol.

In Table 3.2, we can see that the introduced network elements consistently yield significant recognition improvements over the state-of-the-art models. Interestingly, we observe the highest performance gain of 5.52% on the Opportunity dataset characterised by *i*) the largest number of incorporated sensor channels; *ii*) the greatest diversity of the actions to recognise; and *iii*) the highest ratio of class imbalance. The experimental results highlight the significant contribution made by the integrated components in dealing with challenging activity recognition tasks. Notably, the network still achieves a moderate performance improvement on the *performance saturated* Murahari and Plötz (2018) Skoda dataset.

For further insights, the class-specific recognition results from the proposed model are summarised by presenting confusion matrices for the four recognition tasks in Fig. 3.4. We can see that for Opportunity and Skoda datasets with the inclusion of a `Null` class in the annotations, most of the confusions occur in distinguishing between the ambiguous `Null` class and the activities of interest. This can be understood since the `Null` class represents an infinite number of irrelevant activity data for the HAR problem in hand; thus, explicitly modelling this unknown space is a difficult problem.

Table 3.3: Cross-fold evaluation. We present a comparison of segment-wise activity recognition performance based on class-averaged f1-scores with cross-fold evaluation.

HAR Model	Opportunity	Opportunity (*w/o* Null)	PAMAP2	Skoda	Skoda (*w/o* Null)	Hospital
LSTM Learner Baseline	75.6 ± 0.7	70.2 ± 0.7	97.8 ± 0.1	90.9 ± 0.6	82.6 ± 0.6	71.5 ± 2.1
DeepConvLSTM	73.0 ± 0.8	67.7 ± 0.8	97.9 ± 0.1	90.8 ± 0.2	83.2 ± 0.2	72.1 ± 2.4
b-LSTM-S	77.2 ± 1.1	71.8 ± 1.1	97.9 ± 0.1	90.9 ± 0.2	83.5 ± 0.2	72.4 ± 1.4
Dense Labeling	78.5 ± 0.4	73.1 ± 0.4	98.4 ± 0.1	92.1 ± 0.3	84.1 ± 0.3	70.3 ± 0.7
Att. Model	78.1 ± 0.2	72.3 ± 0.6	98.4 ± 0.1	90.4 ± 0.5	82.8 ± 0.4	72.5 ± 1.7
Ours	$\mathbf{81.1 \pm 0.2}$	$\mathbf{75.7 \pm 0.1}$	$\mathbf{98.7 \pm 0.1}$	$\mathbf{93.2 \pm 0.4}$	$\mathbf{85.3 \pm 0.3}$	$\mathbf{73.1 \pm 1.9}$

For completeness, additional extensive cross-fold evaluations are performed across all benchmark datasets in Table 3.3 to complement the hold-out evaluations presented in Table 3.2. Following Jordao et al. (2018), fully non-overlapping windows are employed to generate sensor segments with no temporal overlaps. This is to guarantee that the segment contents do not simultaneously appear both in training and testing splits and prevent data leakage from the training set to the test sets. Subsequently, 3-fold stratified cross-validation is adopted on the datasets to produce the training and testing splits while preserving activity class distributions across all folds. Each constructed fold is in turn utilized once for testing while the remaining folds constitute the training data. The resulting class-average F-score is reported in Table 3.3 for the benchmark datasets and the corresponding HAR models. In the case of the Opportunity and Skoda datasets, the recognition performance is reported both including and ignoring the `Null` class (*w/o Null* columns in Table 3.3) during inference. Inclusion of the `Null` class may result in an overestimation of the recognition performance due to its large prevalence and thus, providing both results gives better insights into the nature of the errors made by the models Ordóñez and Roggen (2016).

Consistent with the observations of hold-out evaluations, the proposed framework in this chapter presents superior performance in identifying human activity classes from raw sensor data across all benchmarks as compared with the baseline HAR models in Table 3.3. Interestingly, a comparison of results between the two evaluation methods—*i.e.,* hold-out evaluation in Table 3.2 and cross-fold evaluation in Table 3.3—*indicates significantly higher recognition performance for the latter*. This is mainly due to the fact that with cross-fold evaluation, the activity data captured by a subject may appear both in the training and testing folds (despite not having any temporal overlap

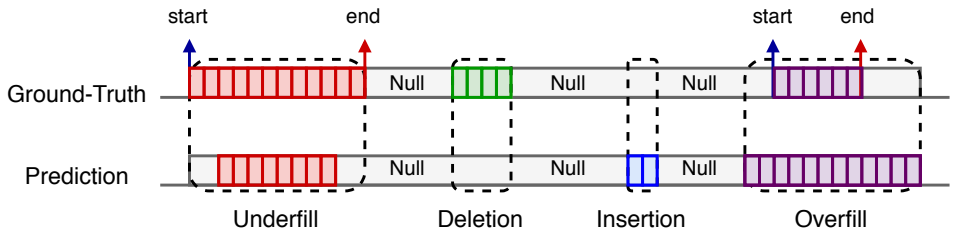


Figure 3.5: Missalignment measures. We illustrate different categories of misalignment measures investigated in this paper. Here, presence of different activity classes is represented with distinct colors and the Null class is denoted with gray. The sequence of continuous sample predictions is compared against the ground-truth sample labels to compute the underfill, overfill, insertion and deletion misalignment measures.

of the sensor data), thus leading to better generalisation performance of the trained models on the testing sets.

Misalignment Measure. In addition to the reported classification metrics, we further report on the explicitly designed misalignment measures of *overfill/underfill*, *insertion*, and *deletion* Ward, Lukowicz and Gellersen (2011) and provide comparisons with the state-of-the-art HAR model Murahari and Plötz (2018) in Table 3.4. These metrics characterise *continuous* activity recognition performance and provide finer details on temporal prediction misalignment with respect to ground truth as illustrated in Fig. 3.5. Specifically:

- *Overfill* and *Underfill* indicate errors when the predicted start or end time of an activity are earlier or later than the ground-truth timings.
- *Insertion* errors refer to incorrectly predicting an activity when there is `Null` activity.
- *Deletion* represents wrongly predicting `Null` class when an activity exists.

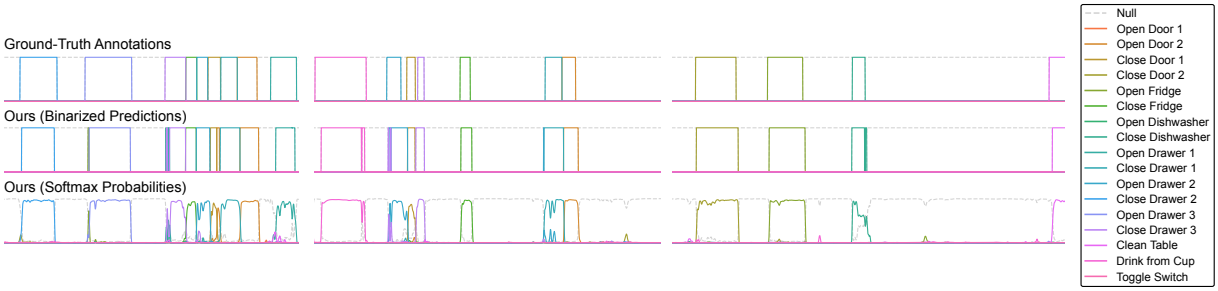
Since some measures require the existence of `Null` class by definition, results on Opportunity and Skoda datasets are reported. The quantitative results in Table 3.4 indicate the improved capability of the proposed model to predict a continuous sequence of activity labels that more accurately aligns with ground-truth timings and better recognises existence or absence of activities of interest.

Table 3.4: Misalignment measures comparison. We report explicitly designed metrics to measure misalignments for continuous activity recognition analysis. (*) denotes the best performing state-of-the-art recognition model Murahari and Plötz (2018) according to Table 3.2.

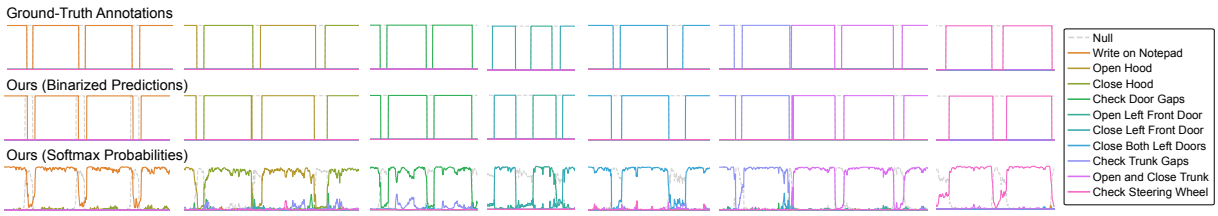
	Opportunity		Skoda	
Alignment Measures	Ours	SoA*	Ours	SoA*
Deletion (↓)	**0.62**	0.69	**0.04**	**0.04**
Insertion (↓)	**2.87**	3.34	**2.01**	3.34
Underfill/Overfill (↓)	**3.71**	4.15	5.33	**5.17**
True Positives (↑)	**92.8**	91.82	**92.62**	91.45

Further, fragments of sensor recordings from these datasets are visualised in Fig. 3.6 for qualitative assessment. The Skoda dataset includes repetitive execution of quality check gestures while the Opportunity dataset is characterised by short duration and sporadic activities. We present the ground-truth annotations (top rows), the proposed model's softmax output probabilities (last rows) and the binarised sequence of predictions (middle rows) obtained after applying argmax operation on the soft scores for each time-step. At every time-step, we colour-code and plot the output class probabilities for each activity category, where we observe a strong correspondence between the ground-truth annotations, activity duration and the predicted activity scores.

Efficiency Analysis In Fig. 3.7, we present a computational complexity comparison among activity recognition models explored in this study. In particular, we illustrate the number of trainable network parameters associated with the HAR baselines for each activity recognition benchmark dataset in Fig. 3.7-a. Clearly, FCN Yao et al. (2018) demonstrates significantly lower number of parameters as compared with the other models due to its fully convolutional structure and abandoning fully connected layers entirely. This is beneficial for realizing activity recognition on edge devices where storage constraints may be a concern. As for the baseline LSTM learner Guan and Plötz (2017), we observe a similar number of parameters across all benchmark datasets. This is due to the fact that the number of parameters within the LSTM networks are heavily influenced by the number of adopted hidden units. This also holds for the b-LSTM-S Hammerla, Halloran and Plötz (2016) architecture, however employing roughly twice the number of learnable parameters due to the bi-directional connections. The



(a) Opportunity Dataset



(b) Skoda Dataset

Figure 3.6: Visualisation of network predictions on holdout test fragments. The proposed HAR model accurately localises and classifies short duration gestures embodied in sequences of sensor signals captured by wearables. We visualise the model's predictions against the ground-truth annotations for sequence fragments of Opportunity and Skoda datasets which include a Null class label representing activities of non-interest.

remaining HAR frameworks—DeepConvLSTM Ordóñez and Roggen (2016), Att. Model Murahari and Plötz (2018), and *Ours*—employ identical backbone feature extractors and mainly differ in terms of the recurrent networks and the attentional components. *Notably, by replacing the LSTM recurrent network with the GRU modules, our proposed HAR framework reduces the number of trainable parameters within the architecture.*

Additionally, in Fig. 3.7-b, we analyze the inference time required by each HAR model to process a single sensory window of 24 samples for all benchmark datasets. To simulate a real-time deployment scenario, the holdout test sets of the benchmark datasets are segmented and sequentially processed—*i.e.*, with a batch size of one—by the HAR baselines and the corresponding total processing time is divided by the total of number of processed segments to generate the results in Fig. 3.7-b. While all frameworks are suitable for real-time predictions—*i.e.*, consuming a processing time of approximately 0.8-1.9 milliseconds—the b-LSTM-S network demonstrates the highest inference time.

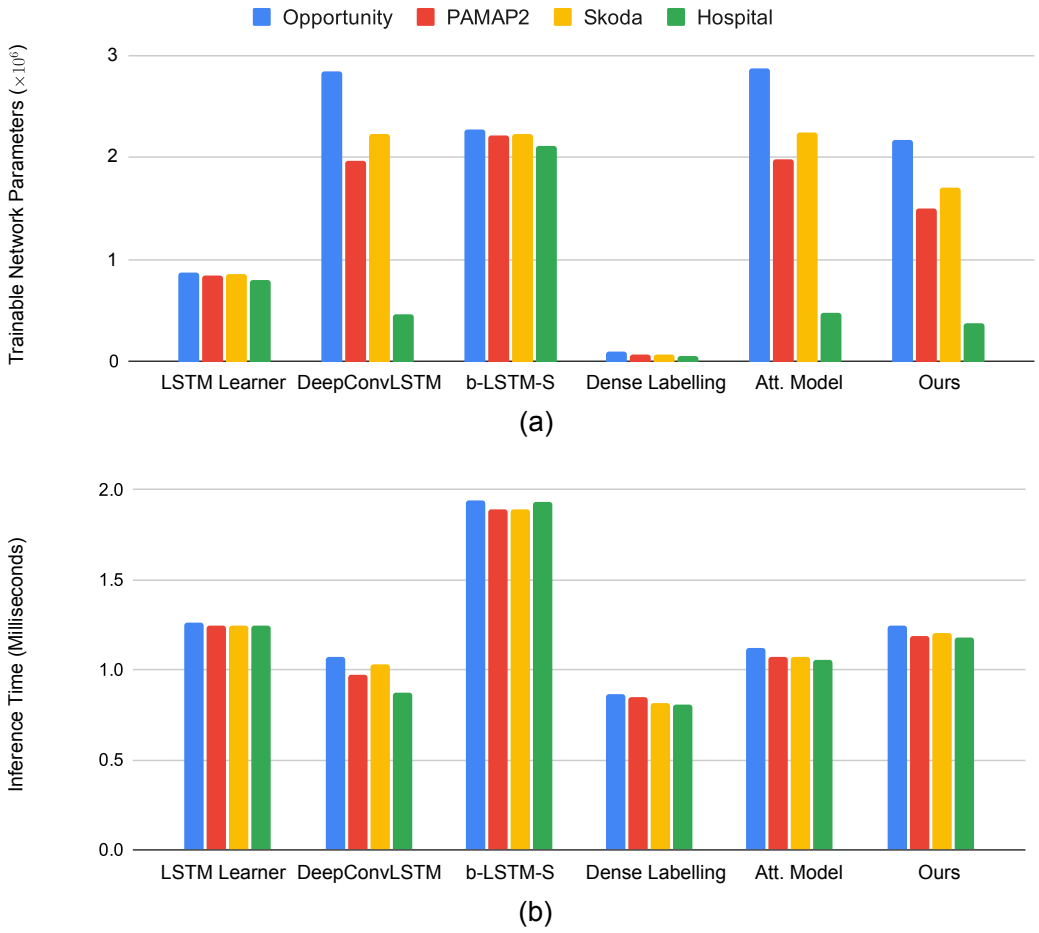


Figure 3.7: Efficiency analysis. A computational complexity comparison among activity recognition models explored in this study: (a) the number of trainable network parameters, and (b) the inference time required for processing a single sensory segment associated with the HAR baselines for each activity recognition benchmark dataset.

3.4.5 Ablation Studies and Insights

Given that the proposed HAR model integrates several key ideas into a single framework, we conduct an ablation study on the Opportunity dataset to understand the contribution made by the various components for the human activity recognition task in Table 3.5. For each ablated experiment, we remove specific modules of our framework and as a reference we include DeepConvLSTM—the backbone of our network as illustrated in Fig. 3.1. Unsurprisingly, removing any component handicaps the HAR model and reduces performance (to 67.2%—see DeepConvLSTM baseline performance) while incorporating all components together yields the highest performing HAR model (74.6%—see mixup+Centre-loss+CIE+AGE).

Table 3.5: Ablation study. We investigating the contribution of integrated modules through an ablation study conducted on the Opportunity dataset.

HAR Model	F_m
DeepConvLSTM Baseline	67.2
Ours (mixup)	70.7
Ours (mixup + Centre-loss)	72.2
Ours (mixup + AGE)	71.7
Ours (mixup + CIE)	73.0
Ours (mixup + Centre-loss + AGE)	72.3
Ours (mixup + Centre-loss + CIE)	73.2
Ours (mixup + CIE + AGE)	74.0
Ours (mixup + Centre-loss + CIE + AGE)	**74.6**

Table 3.6: Data augmentation analysis. A comparison of activity recognition performance on Opportunity dataset based on the class-averaged f1-scores achieved while employing different data augmentation strategies.

	No Augmentation	Jittering	Scaling	Magnitude Warping	Mixup
F_m	70.2	69.4	70.4	70.3	74.6

Notably, the effectiveness of mixup augmentation in regularizing models learnt from time-series wearable HAR data is demonstrated by the significant relative improvement of 5.2% over the DeepConvLSTM Baseline compared to employing mixup alone (an improvement from 67.2% to 70.7%). The virtual multi-channel time-series data attained through in-between sample linear interpolations expand the training data and effectively improve the generalisation of learned activity features to unseen test sequences.

We also investigated conventional hand-crafted augmentation strategies adopted for wearable HAR. We adopt the recent methods studied in Um et al. (2017) including jittering, scaling and magnitude warping using the officially provided implementations in Table 3.6. Here, jittering simulates additive sensor noise, scaling changes the magnitude of segment data by multiplying by a random scalar, and magnitude warping convolves the segment with a random sinusoidal curve using arbitrary amplitude, frequency and phase.

According to the results, mixup clearly outperforms existing augmentation methods by a large margin. Moreover, in line with the observations made in Um et al. (2017), we see that data augmentation techniques may adversely affect recognition performance if not carefully tuned and applied, as is the case here for the jittering approach. We argue that depending on the target task—*i.e.,* the activities to be recognized, sensor channel characteristics, intra-class variations and inter-class similarities—hand-crafted augmentation methods demand domain expert knowledge for effective adoption. In particular, for the Opportunity dataset with 79 sensor channels and 18 fine-grained activity classes, it is not trivial and straight-forward to design channel specific augmentations.

Most importantly, as opposed to the conventional hand-crafted augmentation strategies, mixup augmentation takes into account both *data* and *label* information (see Eq. 3.7) when generating novel samples. This approach largely alleviates the concerns regarding the label-preservation of transformations, and allows simple adoption for diverse activity recognition problem scenarios. This is substantiated in Table 3.6 comparisons of recognition performance on the Opportunity dataset, where mixup provides improved results over no augmentation whilst hand-crafted augmentation strategies such as jittering negatively impacts performance and other methods provides approximately similar results to those achieved with no augmentation.

As hypothesised, encouraging minimal intra-class variability of representations with *centre-loss* consistently improves the recognition performance for activities (mixup+Centre-loss). In addition, while both *CIE* and *AGE* modules allow learning better representations of activities reflected by the enhanced metrics for mixup+CIE+AGE compared to mixup (4.7% relative improvement), we observe a larger performance gain when incorporating *CIE* module as compared with *AGE*; the former encodes the cross-channel sensor interactions with self-attention while the latter learns the relevance of time-steps with temporal attention. Presumably this is due to the fact that within the Opportunity challenge setup, the sequence of representations fed to the GRU is quite short in length (*i.e.*, T=8) and therefore, the last hidden state alone captures most of the information relevant to the activity. In order to verify this, we extract the learned attention scores β_t corresponding to the hidden states $(\mathbf{h}_t)_{t=1}^{\mathrm{T}=8}$ of the GRU encoder and present an illustration for every activity category of Opportunity dataset in Fig. 3.8. In line with the observations made in Murahari and Plötz (2018), the recurrent neural network progressively becomes more informed about the activity

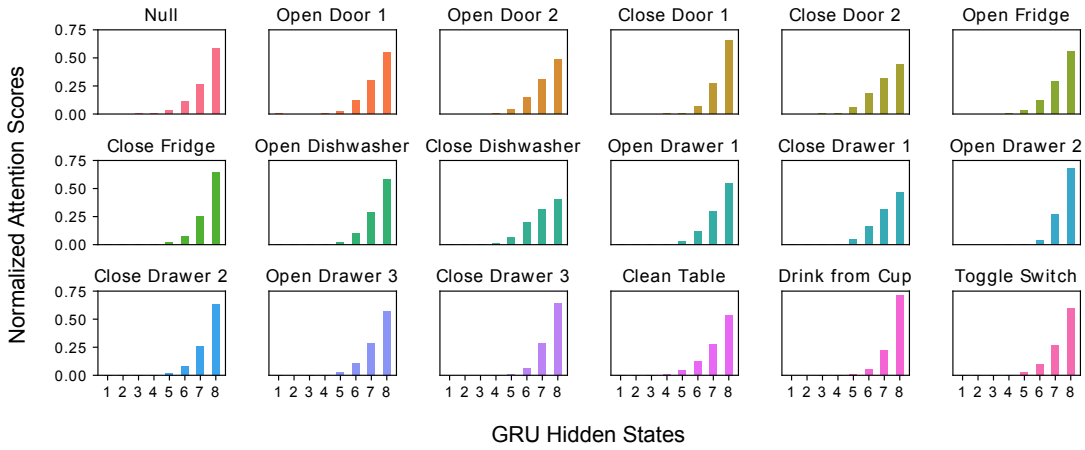


Figure 3.8: Learned temporal attention scores. A visualisation of discovered temporal attentions by the proposed AGE module. The vertical axis represents the normalised attention scores and the horizontal axis denotes hidden states $(\mathbf{h}_t)_{t=1}^{\mathrm{T}=8}$ of the GRU encoder. The GRU becomes progressively more informed about the activity and thus, places higher attention on the few last hidden states with the last state dominating the attention weights.

and thus, proportionally places higher attention on the few last hidden states with the last state dominating the attendance.

On the other hand, we observed exploiting latent channel interactions to significantly improve activity representations as highlighted in ablation results in Table 3.5. To visually explain the learned self-attention correlations from the proposed cross-channel encoder, we graph two segments associated with activity classes of `drinking from cup` and `cleaning table`. The CIE module consumes an input sequence and generates a normalised score matrix of size D×D, corresponding to the attention between each pair of D=79 channels. In Fig. 3.9, we present the normalised self-attention scores, $\mathbf{a} \in \mathbb{R}^{79\times79}$ (attained from softmax operation) in Eq. 3.1, where each column in the attention matrix indicates the extent that a particular sensor channel attends to available sensor channels.

We observe a clear and meaningful focus on a subset of channels vital to the recognition of activities indicated by dark rows in the matrices. For example, we notice high attendance: (*a*) to the inertial measurement units (IMUs) on the right arm when *right hand is being used for* `drinking from cup`; and (*b*) to the IMUs placed on the back and left-upper arm when *upper-body is bent during* `cleaning table`. Thus, the explicit modelling of sensor channel interactions not only leads to improved recognition performance as substantiated by the ablation study in Table 3.5, but also facilitates visual explanation through interpretable scores.

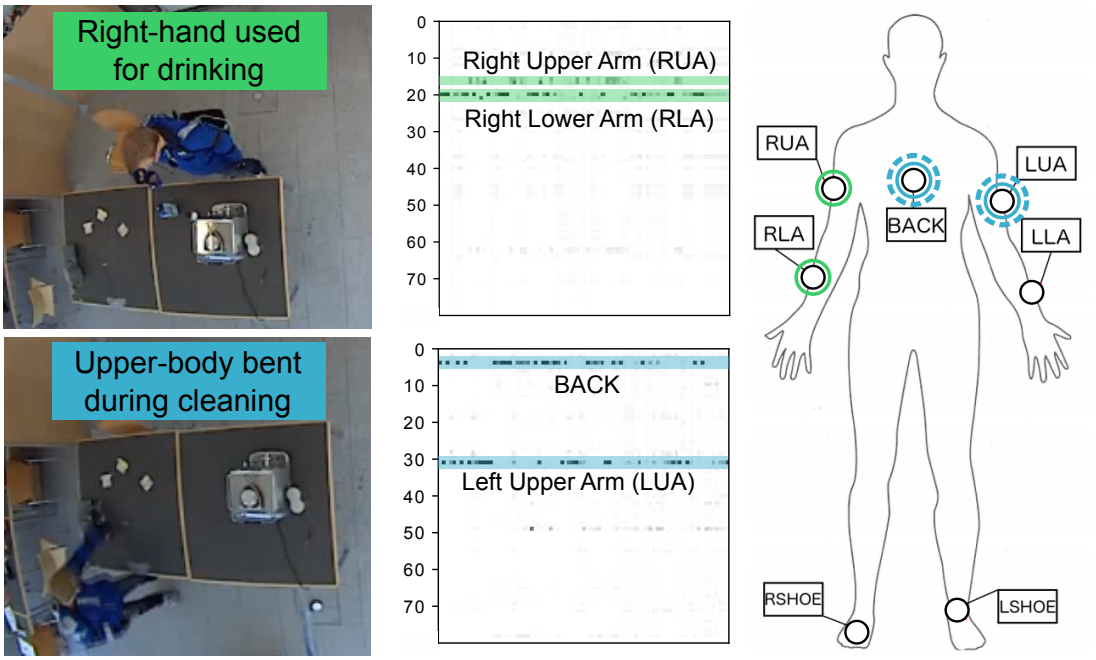


Figure 3.9: Learned self-attention correlations. A visualization of learned self-attention correlations by our CIE module. (a) Subject engaged in two activities; (b) discovered cross-channel correlations by our model selected for *Right-hand used for* `drinking from cup` (dark shaded marks along the rows highlighted in green) and *Upper-body bent during* `cleaning table` (dark shaded marks along the rows highlighted in blue) as shown in video snapshots recorded during the data collection process and shown in (a); and (c) highly attended sensor locations for each activity—color-coded to match green and blue highligts in (b)—in the Opportunity dataset.

3.5 Conclusions

The emergence of deep learning paradigms has facilitated development of end-to-end human activity recognition frameworks and has created a growing number of possibilities for HAR applications. Despite the great progress in this emerging field, this chapter explored solutions to unique and fundamental challenges that benefit from further investigations. We presented network architectural elements to enrich activity feature representations and demonstrated its generalisability by evaluations across four diverse benchmarks. In particular, systematic solutions were discussed to: (*a*) enrich activity representations by exploiting latent correlations between sensor channels; (*b*) incorporate centre-loss to alleviate dealing with intra-class variations of activities; and (*c*) augment multi-channel time-series data with mixup for better generalisation beyond training data. We hope to see the incorporation

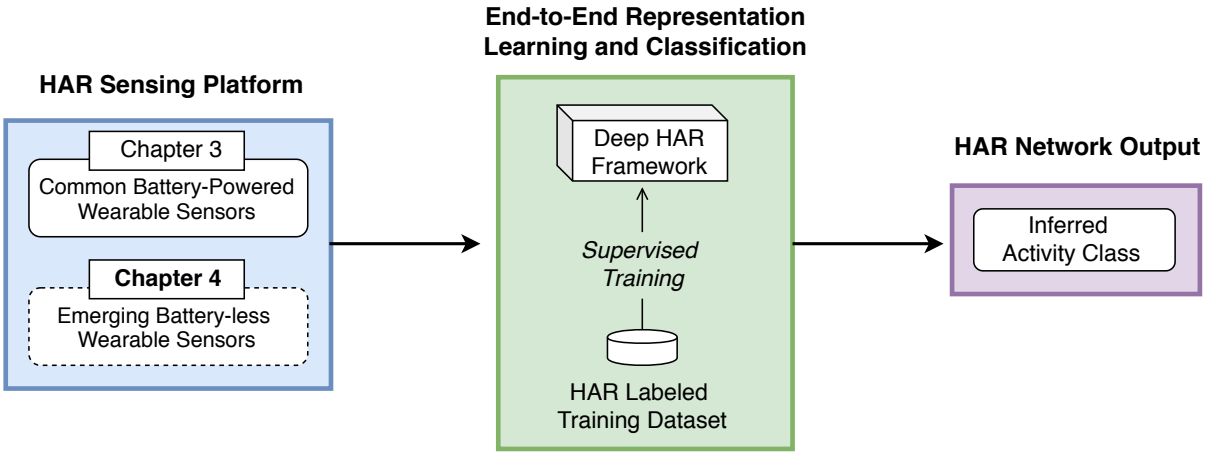


Figure 3.10: Upcoming chapter sneak peek.

of the introduced components in follow-up studies for effective training of activity recognition systems.

The next chapter focuses on an emerging generation of battery-less wearable computing platforms to enable low-cost, maintenance-free and unobtrusive monitoring of human activities. Investigating this problem is of importance given its great application opportunities in the health-care sector and in particular for older people Jayatilaka et al. (2019). Hence, we will discuss the unique challenges associated with adoption of passive sensors, review existing technological solutions and finally present a novel framework to tackle the problem for the first time in an end-to-end fashion. For clarification, we contrast the presented HAR problem in this chapter against the corresponding problem scope explored in Chapter 4 in Fig. 3.10.

Chapter 4

Learning from Sparse Passive Sensor Data-streams

THIS chapter considers the problem of developing end-to-end human activity recognition systems using emerging battery-less—passive—body worn sensor technologies. These *passive* wearables are providing new and innovative methods for human activity recognition (HAR), especially in healthcare applications for older people. Passive sensors are low cost, lightweight, unobtrusive and desirably disposable; attractive attributes for healthcare applications in hospitals and nursing homes. Despite the compelling propositions for sensing applications, the data streams from these sensors are characterised by *high sparsity*—the time intervals between sensor readings are irregular while the number of readings per unit time are often limited. This chapter rigorously explores the problem of learning activity recognition models from temporally sparse data. We describe how to learn directly from sparse data using a deep learning paradigm in an end-to-end manner, and demonstrate significant classification performance improvements on real-world passive sensor datasets from older people over the state-of-the-art deep learning human activity recognition models. Further, insights into the model's behaviour are provided through complementary experiments on a benchmark dataset and visualisation of the learned activity feature spaces.

4.1 Motivation and Contribution

Increasing plethora of wearables are providing the opportunity to conveniently and at low-cost collect *fine-grained* physiological information to understand human activities. While wearables predominantly employ battery powered devices, new opportunities for human activity recognition applications, especially in healthcare, are being created by battery-less or *passive* wearables operating on harvested energy Ranasinghe et al. (2014); Chen et al. (2015); Lemey et al. (2016); Shinmoto Torres et al. (2017b). In particular, older people have expressed a preference for unobtrusive and wearable sensing modalities Gövercin et al. (2010); Torres et al. (2017). Accordingly, this chapter rigorously investigates the problem of human activity recognition in the context of battery-less sensing modalities, discussing their unique challenges and ultimately, presenting an end-to-end solution to learn about human activities through their lens.

In contrast to using often bulky and obtrusive battery powered wearables, passive sensing modalities provide maintenance-free, often disposable, unobtrusive and lightweight devices highly desirable to both older people and healthcare providers. However, the very nature of these sensors leads to new challenges; *i.e.*, the process of operating a battery-less sensor and transmitting the data captured is reliant on harvested power. Due to variable times to harvest adequate energy to operate sensors, the data-streams generated are highly sparse with variable inter-sample times. We illustrate the problem in Fig. 4.1 for a data stream captured by a body-worn passive

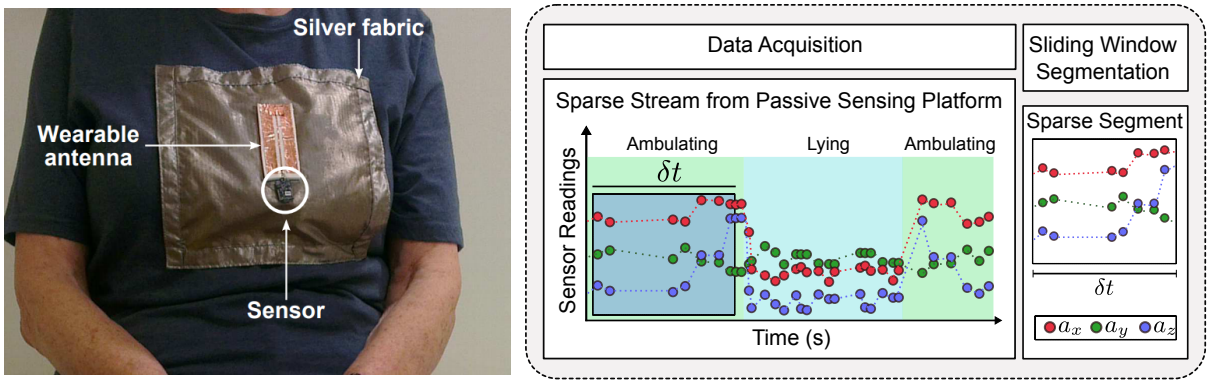


Figure 4.1: Data acquisition from passive wearable sensors. Left: Older volunteer wearing a passive sensor over their clothing in the *clinical rooms* public datasets used in this work (datasets and figure from Torres et al. (2013)). Right: Due to variable times to harvest adequate energy to operate sensors, the data-streams generated are highly sparse with variable time intervals between sensor data reporting times. Thus, adopting the sliding window segmentation results in sensor partitions with variable number of motion measurements.

sensor. We can see two key artefacts: *i*) the variable time intervals between sensor data reporting times; and *ii*) the relatively low average sampling rate. Accordingly, this chapter considers the problem of deriving human activity recognition (HAR) models from *sparse data-streams* using a deep learning paradigm in an *end-to-end* manner.

The dominant human activity recognition pipeline uses fixed duration sliding window to partition wearable time-series data-streams and feed neural networks during both training and inference stages Wang et al. (2019); Guan and Plötz (2017); Ordóñez and Roggen (2016); Hammerla, Halloran and Plötz (2016); Yang et al. (2015); Zeng et al. (2014). When dealing with sparse data partitions, a common remedy is to rely on interpolation techniques as a pre-processing step to synthesise sensor observations to obtain a fixed size representation from time-series partitions as illustrated in Fig. 4.2 Wickramasinghe and Ranasinghe (2015); Gu et al. (2018). However, we recognise two key issues with an interpolated sparse data-stream:

- Interpolating between sensor readings that are temporarily distant can potentially lead to poor approximations of missing measurements and contextual activity information. Accordingly, adoption of convolutional filters or recurrent layers to extract temporal patterns from the poorly approximated measurements may potentially propagate the estimation errors to the activity recognition model—we substantiate this through extensive experiments in Section 4.3.3.
- Interpolation is as an intermediate processing step that prevents end-to-end learning of activity recognition models directly from raw data and introduces real-time inference delays in time critical applications—we demonstrate the time overheads imposed in Section 4.3.3.

In this study, instead of relying on the naturally poor temporal correlations between consecutively received samples in sparse data-streams, we consider incentivising the activity recognition model to uncover discriminative representations from the input sensory data partitions of various sizes to distinguish different activity categories. Our intuition is that a few information bearing sensor samples, although not temporally consistent, can capture adequate amount of information. Therefore, this chapter proposes learning HAR models directly from sparse data-streams. An illustrative summary of the proposed methodology for sparse data-stream classification in comparison with the conventional treatment is presented in Fig. 4.2. Given that we no longer rely on often poor temporal information, we represent sparse data stream

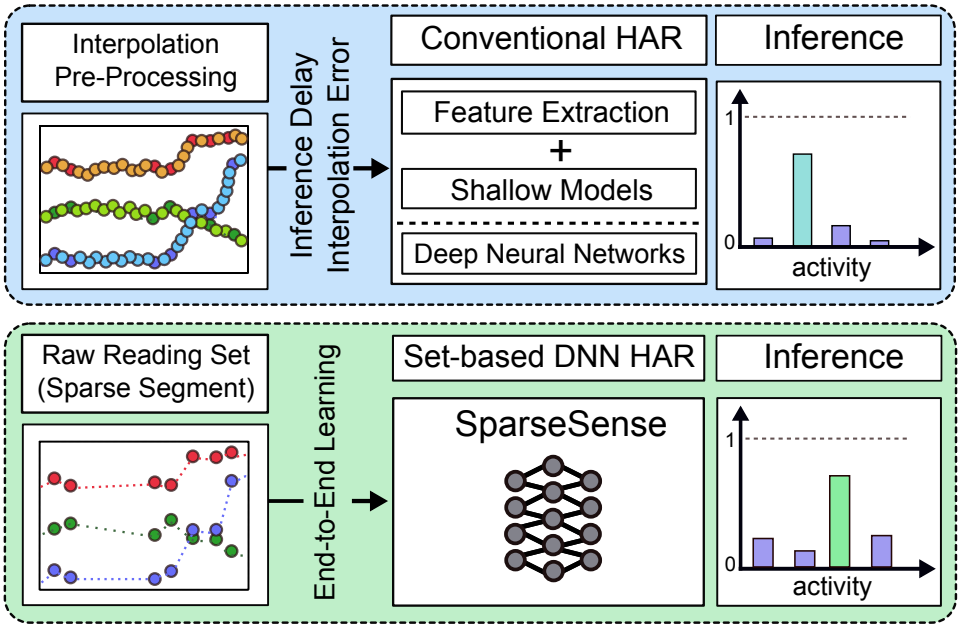


Figure 4.2: Sparse data-stream classification pipelines. We present an overview of the conventional sparse data-stream classification pipeline (blue plane) versus our novel set-based deep learning pipeline (green plane). Top: The conventional pipeline applies interpolation pre-processing on the sparse segments to synthesise fixed temporal context for model training and inference. bottom: In contrast, our proposed approach elegantly allows end-to-end learning of activity recognition models directly from sparse segments to deliver highly accurate classification decisions.

partitions as unordered sets with various cardinalities from which embeddings capable of discriminating activities can be learned. This approach is inspired by recent research efforts to investigate set-based deep learning paradigms to address a new family of problems where inputs Qi et al. (2017); Zaheer et al. (2017) of the task are naturally expressed as sets with unknown and unfixed cardinalities. Therefore, the strategy here is to develop activity recognition models that can learn and predict from incomplete sets of sensor observations, without requiring any extra interpolation efforts.

The main contributions of this chapter are summarised in what follows. In particular:

1. A new problem is formulated and addressed with a deep neural network formulation—learning from sparse sensor data-streams in an end-to-end manner.

2. We show that set learning can tolerate missing information which otherwise would not be possible with conventional DNN.

3. It is substantiated that the proposed novel treatment of the problem yields significantly outperforming recognition models with lower inference delays

compared with the state-of-the-art on naturally sparse public datasets—over 4% improvement in the best case. In addition, further comparisons with a benchmark HAR dataset are presented to provide deeper insights into the performance improvements obtained from the proposed approach.

4.2 Proposed Methodology

We first present a formal description of the human activity recognition problem with sparse data-streams and introduce the notations used throughout this chapter before elaborating on the proposed activity recognition framework to learn directly from sparse data-streams in an end-to-end manner.

4.2.1 Problem Formulation

Consider a collected data-stream of raw time-series samples from body-worn sensors of the form $\mathbf{X}_{stream} = (x_1, x_2, ..., x_S)$, where $x_t \in \mathbb{R}^D$ is a multi-dimensional vector that contains sample measurements over D distinct sensor channels at time step t and S is the total length of the sequence. Without loss of generality, we assume a hardware-specific sampling rate for the wearable sensors, denoted by f.

HAR with Uniform Time-series Data. In an ideally controlled laboratory setup, sensor samples are constantly taken at regular intervals of $\frac{1}{f}$ seconds. In such case, applying the commonly adopted time-series segmentation technique with a sliding window of fixed temporal context δt yields the labelled dataset

$$\mathcal{D}_{\text{uniform}} = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), ..., (\mathbf{x}_n, \mathbf{y}_n)\}, \tag{4.1}$$

where $\mathbf{x}_i = [x_i, ..., x_{i+W-1}] \in \mathbb{R}^{D \times W}$ is a *fixed size* segment of captured sensor readings, $W = f\delta t$ is the constant number of received samples, and $\mathbf{y}_i$ denotes the corresponding one-hot encoded ground-truth from the pre-defined activity space $\mathcal{A} = \{a_1, ..., a_k\}$. The acquired dataset can then be utilised to train activity recognition models using out-of-the-box machine learning techniques.

HAR with Sparse Time-series Data. Unfortunately, sparse time-series data often found in real-world deployment settings, especially with passive sensors have variable inter-sensor observation intervals.

In this case, utilising a fixed time sliding window approach to segment the sparse data-stream results in the labelled dataset:

$$\mathcal{D}_{\text{sparse}} = \{(\mathcal{X}_1^{m_1}, \mathbf{y}_1), (\mathcal{X}_2^{m_2}, \mathbf{y}_2), ..., (\mathcal{X}_n^{m_n}, \mathbf{y}_n)\}, \tag{4.2}$$

where $\mathcal{X}_i^{m_i} = \{x_i, ..., x_{i+m_i-1}\} \in \overbrace{\mathbb{R}^D \times ... \times \mathbb{R}^D}^{m_i}$ is a set of sparse sensor observations during a timed window, $m_i \in \mathbb{N}$ is the cardinality of the obtained observation set, and $\mathbf{y}_i$ denotes the corresponding activity class. We emphasise that the number of received sensor readings in the time interval δt is *unfixed* for different sensory segments and upper bounded by the sensor sampling rate; *i.e.*, for any given sensory segment $\mathcal{X}_i^{m_i}$, we have $m_i \leq \mathrm{f}\delta t$.

Having acquired the training dataset of sparse sensory segments $\mathcal{D}_{\text{sparse}} = \{(\mathcal{X}_i^{m_i}, \mathbf{y}_i)\}_{i=1}^n$, the goal here is to directly learn a mapping function $\mathcal{F}_{\Theta^*} : 2^{\mathbb{R}^D} \rightarrow \mathcal{A}$, that operates on input sensory sets with unfixed cardinalities and accurately predicts the underlying activity classes,

$$\mathbf{y}_i = \mathcal{F}_{\Theta^*}(\mathcal{X}_i^{m_i}) = \mathcal{F}_{\Theta^*}(\{x_i, ..., x_{i+m_i-1}\}), \ \ \forall i \in \{1, ..., n\}.$$

4.2.2 SparseSense Framework

Our work is built upon the insight that incorporating interpolation techniques to recover the missing measurements across large temporal gaps between received sensor observations in sparse data-streams leads to poor estimations and therefore, significant interpolation errors. As we demonstrate in Section 4.3.3, the adoption of convolutional filters or recurrent layers to extract temporal patterns from the poorly approximated measurements can potentially propagate the estimation errors to the activity recognition model.

Instead of forcing the network to exploit the potentially weak temporal correlations in sparse data-streams, this chapter proposes learning global embeddings from sets that encode aggregated information related to an activity. Therefore, we propose

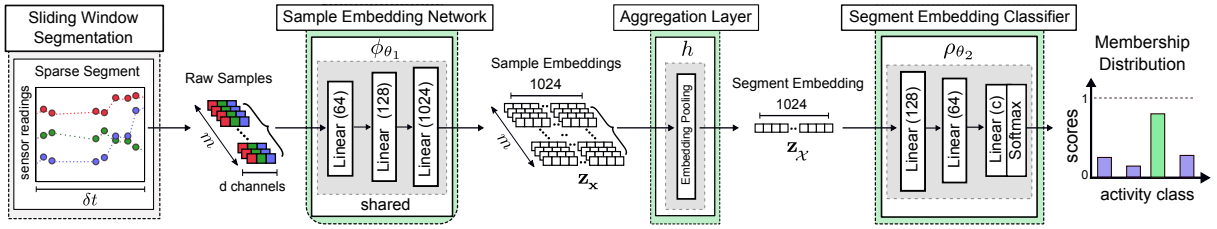


Figure 4.3: SparseSense network architecture. The proposed network consumes sets of raw sensor observations with potentially varying cardinalities, uncovers latent projections for individual samples, aggregates sample embeddings into a global segment embedding, and maps the acquired segment embedding to its corresponding activity category. The number of neurons constituting the linear layers are outlined in parenthesis. All layers utilise ReLUs for non-linear transformation of activations except for the last layer which leverages a softmax activation function.

formulating sparse segments as unordered sets with unfixed and unknown number of sensor readings. Hence, we design *SparseSense* as a set-based activity recognition framework for the HAR task that directly manipulates sets of received sensor readings with irregular inter-sample observation intervals and outputs the corresponding activity membership distributions. The proposed approach provides a complete end-to-end learning method that incentivises the activity recognition model to uncover globally discriminative representations for the input sparse segments with variable number of samples, and distinguish different activity categories accordingly.

Network Architecture

The overall architecture of the proposed *SparseSense* network is illustrated in Fig. 4.3. Essentially, we approximate the optimal mapping function $\mathcal{F}_{\Theta^*}$ through training of a deep neural network parameterised by Θ. The primary task for integrating set learning into deep neural networks is employing a *shared network* to map each set element independently into a higher dimensional embedding space (to facilitate class separability) and adopting a symmetric operation across the element embeddings to generate a global representation for the entire set that does not rely on the set element orderings. This pipeline is incorporated into the building blocks of our network as elucidated in what follows:

Input. Adopting sliding window segmentation over the sparse data-stream yields sets of sparsely received sensor observations $\mathcal{X}$ in the pre-defined temporal window δt, with potentially varying cardinalities.

Shared Sample Embedding Network. The embedding network $\phi_{\theta_1} : \mathbb{R}^D \rightarrow \mathbb{R}^z$ parameterised by θ_1, operates identically and independently on each sample measurement x within the received observation set $\mathcal{X}$ and learns a corresponding higher dimensional projection $\mathbf{z}_x \in \mathbb{R}^z$ to alleviate separability of activity features in the new embedding space; *i.e.*, $\mathbf{z}_x = \phi_{\theta_1}(x), \forall x \in \mathcal{X}$. Technically, ϕ_{θ_1} is a standard multi-layer perceptron (MLP) whose parameters are *shared* between the sensor sample readings; *i.e.*, all samples undergo the same layer operations and are therefore processed identically through a copy of the MLP.

Aggregation Layer. Described by $h : \mathbb{R}^z \times \ldots \times \mathbb{R}^z \rightarrow \mathbb{R}^z$, the aggregation layer applies a symmetric operation across the latent representations of individual sensor samples and extracts a fixed size global embedding $\mathbf{z}_{\mathcal{X}} \in \mathbb{R}^z$ to represent the sensory segment as a whole. Thus, for a given sensory segment $\mathcal{X}_i$, we have

$$\mathbf{z}_{\mathcal{X}_i} = h(\{\mathbf{z}_{x_i}, \ldots, \mathbf{z}_{x_{i+m_i-1}}\}). \tag{4.3}$$

Notably, the shared sample embedding network coupled with the symmetric aggregation layer allow summarizing sparse segments with effective high-dimensional projections that *i*) do not rely on the weak temporal ordering of the sparse samples, and, *ii*) ensure fixed size tensor representations independent of the number of received readings. Inspired by Qi et al. (2017), in this chapter, we set h to incorporate a feature-wise maximum pooling across sample embeddings which promises robustness against set element perturbations.

Segment Embedding Classifier. Described by $\rho_{\theta_2} : \mathbb{R}^z \rightarrow \mathcal{A}$ parameterised by θ_2 is trained to exploit the segment embeddings $\mathbf{z}_{\mathcal{X}}$ through multiple layers of non-linearity and predict the corresponding activity class probability distributions $\hat{\mathbf{y}}$; *i.e.*, $\hat{\mathbf{y}} = \rho_{\theta_2}(\mathbf{z}_{\mathcal{X}})$.

Here, a softmax activation function governs the output of our network to yield posterior probability distributions over the activity space $\mathcal{A}$.

Summary. Now, we can express the mathematical operations constituting the forward pass of the proposed activity recognition model for a given sparse sensory segment $\mathcal{X}_i$ as:

$$\mathcal{F}_\Theta(\mathcal{X}_i^{m_i}) = \rho_{\theta_2}\left(h(\{\phi_{\theta_1}(x_i), \ldots, \phi_{\theta_1}(x_{i+m_i-1})\})\right), \tag{4.4}$$

where Θ denotes the collection of all network parameters; *i.e.*, $\Theta = (\theta_1, \theta_2)$.

Network Training and Activity Inference

During the training process, the goal is to learn the network parameters Θ such that the disagreement between the network outputs and the corresponding ground-truth activities is minimised for the training dataset. We can precisely express this discrepancy minimisation by adopting an end-to-end optimisation of the negative log-likelihood loss function $\mathcal{L}_{\text{NLL}}$ on the training dataset $\mathcal{D}_{\text{sparse}}$; *i.e.*,

$$\Theta^* = \arg\min_{\Theta} \sum_{i=1}^{n} \mathcal{L}_{\text{NLL}}\left(\mathcal{F}_{\Theta}(\mathcal{X}_i^{m_i}), \mathbf{y}_i\right). \tag{4.5}$$

As the training process progresses and the corresponding objective function is minimised, the SparseSense network uncovers highly discriminative embeddings for sparse segments that allow effective separation of classes in the activity space.

Once the training procedure converges and the optimal network parameters Θ^* are learned from the training dataset, we adopt a maximum a posteriori (MAP) inference to promote the most probable activity category for any given set of sparse sensor readings; *i.e.*, the highest scoring class in the softmax output of the network is chosen to be the final prediction.

4.3 Experiments and Results

4.3.1 Datasets

To ground our study, we evaluate the proposed framework on two naturally sparse public datasets collected in clinical rooms with older people using a body-worn battery-less sensor intended for ambulatory monitoring in hospital settings. For further insights, extensive empirical analysis of the proposed approach are presented on a HAR benchmark dataset with synthesised sparsification and comparisons are provided against the state-of-the-art deep learning based HAR models.

Clinical Room Datasets Torres et al. (2013). The dataset is collected from fourteen older volunteers, with a mean age of 78 years, performing a set of broadly scripted

activities while wearing a W^{2}ISP over their attire at the sternum level (see Fig. 4.1). The W^{2}ISP is a passive sensor-enabled RFID (Radio Frequency Identification) device that operates on harvested electromagnetic energy emitted from nearby RFID antennas to send data with an upper-bound sampling rate of 40 Hz. Data collection was carried out in two clinical rooms with two different antenna deployment configurations to power the sensor and capture data; resulting in *Roomset1* and *Roomset2* datasets. Each sensor observation in the obtained datasets records triaxial acceleration measurements as well as contextual information from the RFID platform indicating the antenna identifier and the strength of the received signal from the sensor. These recordings were manually annotated with `lying on bed`, `sitting on bed`, `ambulating` and `sitting on chair` to closely simulate hospitalised patients' actions.

WISDM Dataset Kwapisz, Weiss and Moore (2011). This dataset contains acceleration measurements collected through controlled, laboratory conditions and covers the activities of `walking`, `jogging`, `climbing up stairs`, `climbing down stairs`, `sitting` and `standing`. The collected dataset delivers high quality data and has frequently been used in HAR studies for benchmarking purposes. Accordingly, we find this dataset a suitable choice for thorough investigation of our SparseSense network under different levels of synthesised data sparsification.

4.3.2 Experimental Setup

In our investigations, we initially perform per-feature normalisation to scale real-valued observation attributes to the $[0, 1]$ interval. We consider a fixed temporal context δt and obtain sensory partitions by sliding a window over the recorded data-streams. The acquired segments are assumed to reflect adequate information related to a wearer's current activity and are thus, assigned a categorical activity label based on the most observed sample annotation in the time-span of the sliding window.

The experiments are implemented in Pytorch Paszke et al. (2017) deep learning framework on a machine with a NVIDIA GeForce GTX 1060 GPU. The SparseSense deep human activity recognition model is trained in a fully-supervised fashion by back-propagating the gradients of the loss function in mini-batches of size 128; *i.e.*, the network parameters are iteratively adjusted according to the RMSProp Tieleman and Hinton (2012) update rule in order to minimize the negative log-likelihood loss

using mini-batch gradient descent. The optimiser learning rate is initialised with 10^{-4}, reduced by a factor of 0.1 after 100 epochs, and the optimisation is ceased after 150 epochs. Further, a weight decay of 10^{-4} is imposed as L_2 penalty for regularisation. Following previous studies, we employ 7-fold stratified cross-validation on the datasets and preserve activity class distributions across all folds. Each constructed fold is in turn utilised once for validation while the remaining six folds constitute the training data.

4.3.3 Baselines and Results

Clinical Room Experiments

In Table 4.1, we report the class-average F-score (F-score$_m$) as the widely adopted evaluation metric and compare SparseSense with activity recognition models previously studied for the naturally sparse clinical room datasets as well as solid deep learning based HAR models. Wickramasinghe and Ranasinghe (2015) has explored shallow models including support vector machines (SVM^{lin} and SVM^{rbf}) and conditional random fields (*CRF*) trained using hand-crafted features extracted from either raw or interpolated sparse segments. In addition, we investigate the effectiveness of *Bi-LSTM* Hammerla, Halloran and Plötz (2016), *DeepCNN* and *DeepConvLSTM* Ordóñez and Roggen (2016) as solid deep learning baselines representing popular state-of-the-art frameworks for HAR applications.

In particular, Bi-LSTM leverages bidirectional LSTM recurrent layers to directly learn the temporal dependencies of samples within the sensory segments. Both DeepCNN and DeepConvLSTM adopt four layers of 1D convolutional filters along the temporal dimension of the fixed size segmented data to automatically extract feature representations. However, DeepCNN is then followed by two fully connected layers to aggregate the feature representations while DeepConvLSTM utilises a two layered LSTM to model the temporal dynamics of feature activations prior to the final softmax layer. We refer interested readers to the original papers introducing the HAR models for further details and network specifications.

Following Wickramasinghe and Ranasinghe (2015), for each baseline we explore progressively increasing window duration, *i.e.* $\delta t \in \{2, 4, 8, 16\}$, adopt per-channel interpolation schemes (*linear*, *cubic*, *quadratic* and *previous*) to compensate for the missing acceleration data and report the highest achieving configurations in Table 4.1

for all competing approaches. In this regard, *cubic* and *quadratic* interpolation schemes respectively refer to a spline interpolation of second and third order, and the *previous* scheme fills missed values with the previously received sensor readings.

From the outlined results, we observe that the SparseSense network outperforms all the baseline models with a large margin in the task of sparse data-stream classification. Notably, the baselines are: *i*) well-engineered shallow models that require a large pool of domain expert hand-crafted features; and *ii*) state-of-the-art deep learning HAR models that demand interpolation techniques to synthesise regular sensor sampling rates. In contrast, SparseSense seamlessly operates on sparse sets of sensory observations without requiring any extra interpolation efforts or manually designed features, and automatically extracts highly discriminative embeddings for the classification task in an end-to-end framework.

Table 4.1: Performance comparison for the naturally sparse clinical room datasets. Reported results and design choices for the baselines with asterisks are quoted directly from Wickramasinghe and Ranasinghe (2015). The remaining baselines are replicated following their original paper descriptions. To ensure a fair comparison, the reported results are the highest achieving combination of window sizes (explored in the range of $\{2, 4, 8, 16\}$ seconds) and interpolants (*linear*, *cubic*, *quadratic* and *previous*) selected for all the competing approaches.

Dataset (clinical room)	**HAR Model**	**Interpolant** (acceleration)	**Input**	**Window Size** (δt)	**Precision**$_m$ (mean±std)	**Recall**$_m$ (mean±std)	**F-score**$_m$ (mean±std)
Roomset1	SVM^{lin*}	Cubic	Hand-crafted features	4 seconds	87.87±2.55	83.44±1.72	84.96±1.23
	SVM^{rbf*}	None	Hand-crafted features	8 seconds	90.39±2.70	87.42±1.42	88.45±1.68
	CRF*	Linear	Hand-crafted features	2 seconds	85.97±2.43	82.35±3.08	83.73±2.40
	Bi-LSTM	Linear	Raw sensor readings	2 seconds	89.97±0.78	85.11±0.99	86.96±1.06
	DeepCNN	Quadratic	Raw sensor readings	4 seconds	92.43±1.21	87.93±1.74	89.73±1.55
	DeepConvLSTM	Linear	Raw sensor readings	4 seconds	91.87±1.43	88.88±1.79	90.42±1.54
	(Ours) SparseSense	None	Raw sensor readings	2 seconds	**95.0±0.75**	**94.08±0.78**	**94.51±0.62**
Roomset2	SVM^{lin*}	Cubic	Hand-crafted features	2 seconds	87.06±4.10	84.00±2.90	84.97±3.74
	SVM^{rbf*}	None	Hand-crafted features	8 seconds	90.97±4.11	83.88±2.04	85.53±2.86
	CRF*	None	Hand-crafted features	16 seconds	83.68±6.50	78.29±3.58	79.99±4.76
	Bi-LSTM	Previous	Raw sensor readings	2 seconds	92.38±0.91	91.4±0.62	91.78±0.58
	DeepCNN	Linear	Raw sensor readings	4 seconds	93.11±0.94	91.7±1.18	92.36±0.99
	DeepConvLSTM	Previous	Raw sensor readings	4 seconds	94.16±0.52	93.05±0.78	93.77±0.63
	(Ours) SparseSense	None	Raw sensor readings	2 seconds	**97.07±0.52**	**96.88±0.34**	**96.97±0.37**

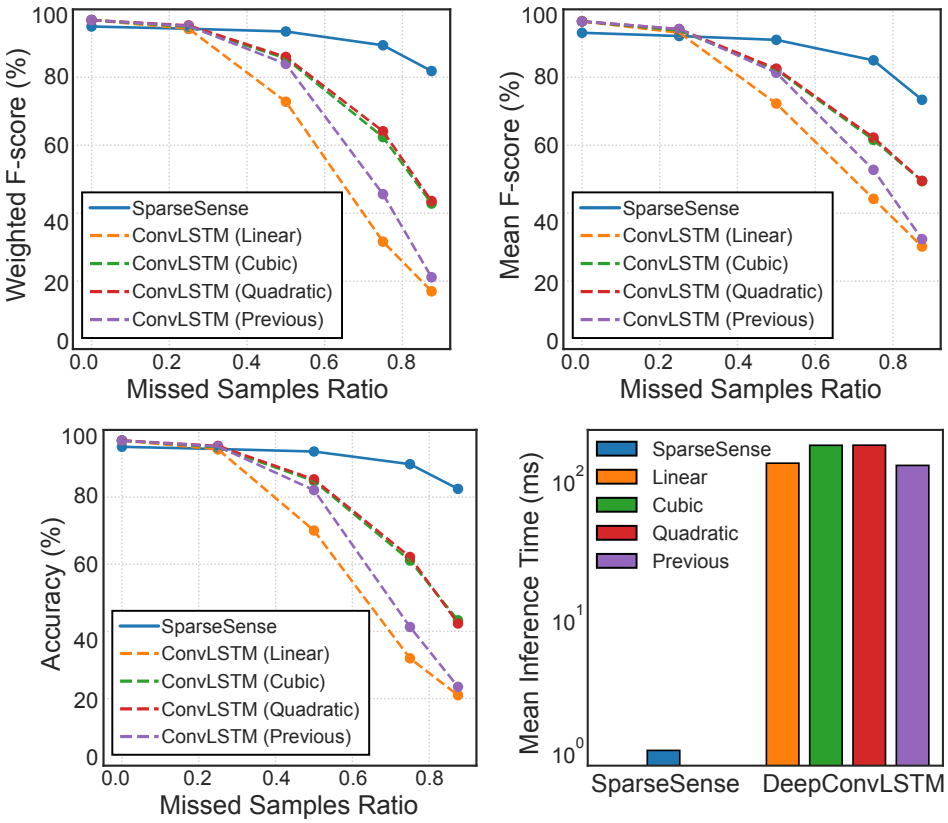
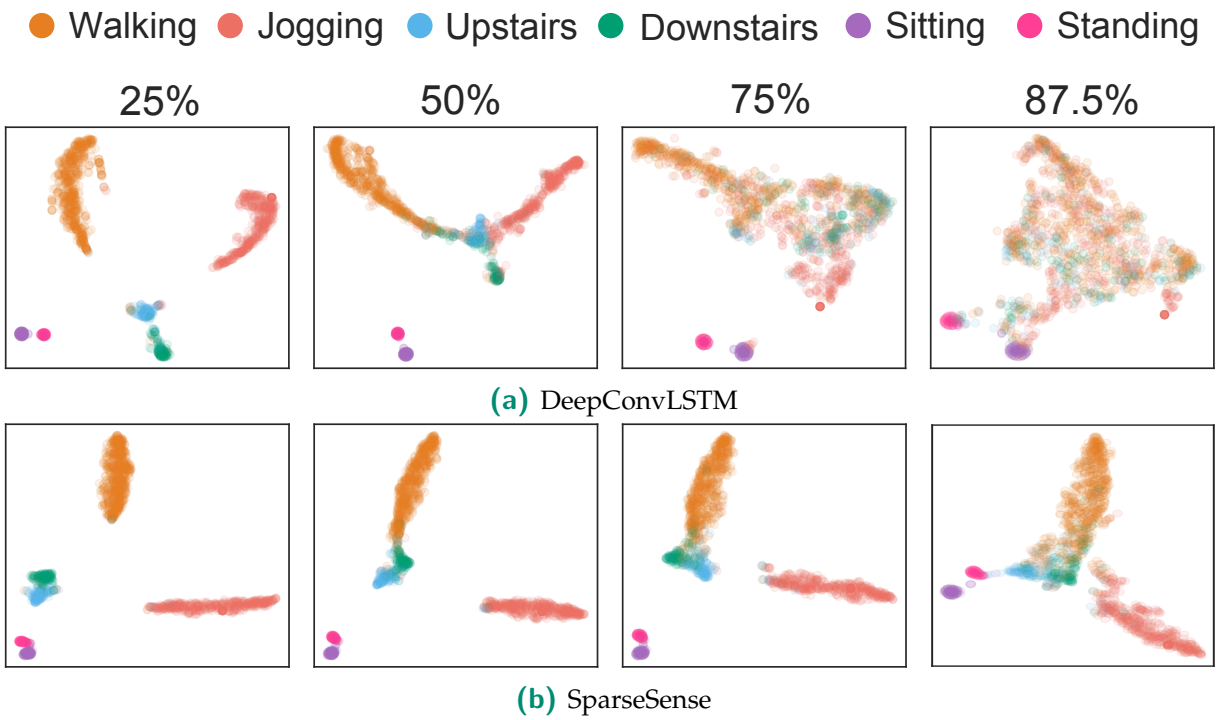


Figure 4.4: Performance comparison analysis. We present a comparison on activity recognition performance and computational complexity of the proposed SparseSense framework for sparse data-stream classification against the state-of-the-art DeepConvLSTM HAR model. DeepConvLSTM is tailored for sensory segments of fixed size temporal context and thus, requires sparse segments be re-sampled through adoption of interpolation methodologies prior to inputting them.

WISDM Benchmark Experiments

To provide additional insights onto the model's behaviour, we conduct experiments on WISDM benchmark dataset and analyse the network's classification performance under different levels of synthesised data sparsification. Taking into account the superior performance of DeepConvLSTM among the baselines in Table 4.1, here we only present comparisons with this model. Following Kwapisz, Weiss and Moore (2011); Alsheikh et al. (2016), we partition the data-streams into fixed size sensory segments using a sliding window of 10 seconds duration (corresponding to 200 sensor readings) and train the HAR models on the acquired segmented data. Subsequently at test time, we drop sensor readings at random time-steps in order to generate synthetic sparse segments.



(a) DeepConvLSTM

(b) SparseSense

Figure 4.5: 2D visualization of the learned feature spaces. We present t-SNE visualisations in 2D for (a) DeepConvLSTM and (b) SparseSense under different data sparsity levels indicated by the percentage of artificially imposed missed readings. SparseSense learns robust embeddings that maintain cluster separation even under significant missing sample settings.

Tolerance to Data Sparsity and Delays

In Fig. 4.4, the obtained evaluation measures are plotted for both HAR models under different sparsification settings. When data segments are received in full, DeepConvLSTM performs better than SparseSense due to its ability in capturing temporal dependencies between consecutive sensor readings. However, as the data sparsity increases and the temporal correlation weakens, we observe a significant drop in classification performance of DeepConvLSTM. Notably, with large temporal gaps between sensor observations, interpolation techniques cannot produce good estimations of the missing samples and fail to recover the original acceleration measurements which in turn impacts the classification decisions of DeepConvLSTM. In contrast, not only does SparseSense achieve comparable classification results for completely received sensory segments, but it also displays great robustness to data sparsity by making accurate decisions for incomplete segments of sensor data. In addition, we show in the bar plot the mean processing time required by the HAR models to make predictions on a mini-batch of 128 segments. Clearly, our framework

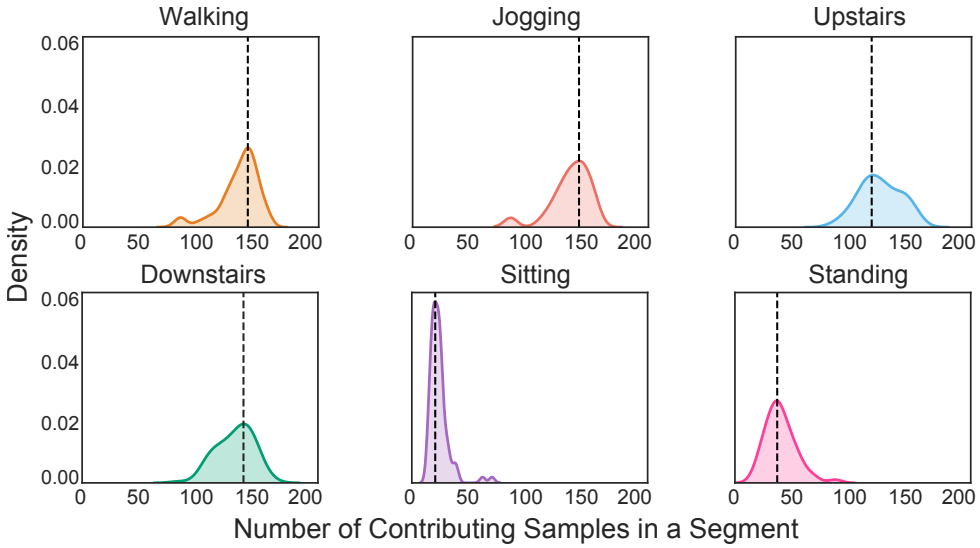


Figure 4.6: Contributing samples analysis. We illustrate the density plots for the number of contributing samples that constitute the aggregated segment embeddings for each activity category of WISDM dataset.

demonstrates a significant advantage over other HAR models for real-time activity recognition using sparse data-streams by removing the need for prior interpolation pre-processing.

SparseSense Model Behaviour

We visualise the learned feature spaces for both models in 2D space using t-distributed stochastic neighbor embedding (t-SNE) Maaten and Hinton (2008) in Fig. 4.5. In the absence of significant data sparsity, the segment embeddings belonging to each activity category are clustered together while different activities are separated in the feature space. However, while SparseSense is able to maintain this cluster separation for severely missed sample ratios and incomplete observation sets, DeepConvLSTM clearly struggles to discriminate between the interpolated segments. Technically, the symmetric max pooling operation in the aggregation layer of SparseSense incentivises our HAR model to summarise sensory segments using only the most informative readings present in the segment. We refer to these information bearing samples as the *contributing samples*.

In Fig. 4.6, we provide density plots for the number of sensor readings that ultimately contribute to the aggregated segment embeddings for each activity category of the WISDM dataset. We observe that SparseSense intelligently summarises the segments through discarding potentially redundant information in the neighbouring samples

when windows of fully received samples ($m = 200$) are presented to the network—see the density plots where the tails towards 200 contributing samples have a probability of zero. More interestingly, the network displays a clear distinction in its behaviour towards learning embeddings for static activities (*i.e.*, sitting and standing) as opposed to dynamic activities (*i.e.*, walking, jogging and climbing stairs) by exploiting far fewer number of sensor observations out of the $m = 200$ received samples in the window. This can be intuitively understood as static activities reflect signal patterns with small changes in sensor measurements of a timed window as compared with dynamic activities and thus, can be summarised with smaller number of observations.

4.4 Conclusion

A large body of the literature in ubiquitous computing explores human activity recognition (HAR) in the context of battery-powered sensor platforms. In contrast, this chapter examined alternative technological solutions that enable integration of emerging passive wearable devices for unobtrusive activity monitoring, in particular for healthcare applications. Accordingly, an end-to-end human activity recognition framework was presented to learn directly from temporally sparse data-streams using set-based deep neural networks. In contrast to previous studies that rely on interpolation pre-processing to synthesise sensory partitions with fixed temporal context, the proposed *SparseSense* network seamlessly operates on sparse segments with potentially varying number of sensor readings and delivers highly accurate predictions in the presence of missing sensor observations. Through extensive experiments on publicly available HAR datasets, it was substantiated how our novel treatment for sparse data-stream classification problem results in activity recognition models that significantly outperform solid deep learning methods that rely on interpolation pre-processing, while incurring notably lower real-time prediction delays. The method developed herein provides insights into an effective approach for understanding human motion data using passive wearables, particularly for health-care applications.

Interestingly, the next chapter discusses how the flexibility provided by the powerful set learning frameworks can further be leveraged to address a natural problem arising from the conventional multi-class classification formulation of human activity recognition problems. In a nutshell, adopting the sliding window segmentation

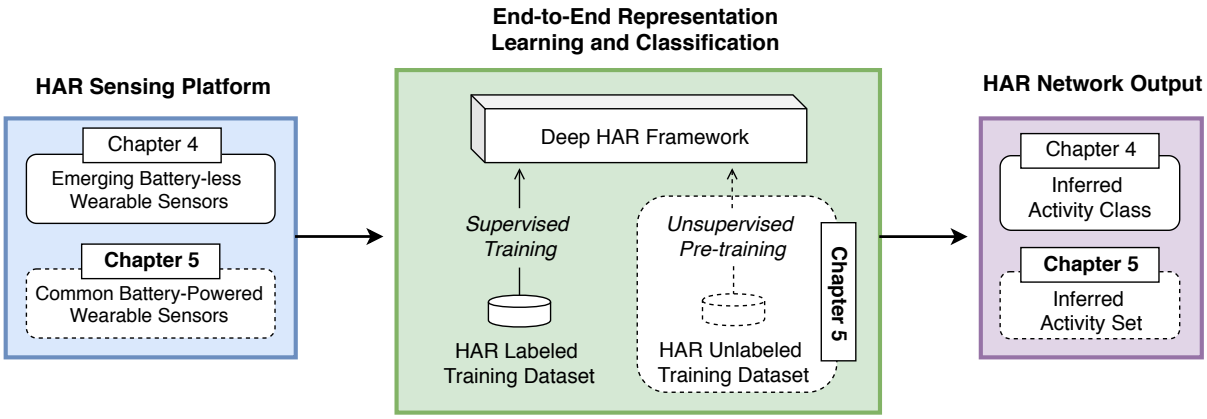


Figure 4.7: Upcoming chapter sneak peek.

technique inevitably results in sensor partitions that carry motion data beyond a single activity category. However, the conventional HAR formulation casts activity recognition as a multi-class classification problem where each sensor segment is assumed to be associated with a single activity category. This assumption is relaxed in the next chapter and a novel formulation of HAR is presented to elegantly handle simultaneous prediction of multiple activities. Moreover, through preliminary exploitation of unlabelled data, next chapter serves as the transition point from fully supervised training regimes discussed so far to unsupervised development of HAR frameworks in this dissertation. We illustrate the explored HAR problem in this chapter in comparison against the investigated problem in Chapter 5 in Fig. 4.7.

Chapter 5

Learning to Predict Activity Sets from Wearable Sensor Data-streams

MOST recent research in the field of human activity recognition (HAR) adopts supervised deep learning paradigms to automate extraction of intrinsic features from raw signal inputs and addresses HAR as a multi-class classification problem, as in Chapter 3 and Chapter 4; here, 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. This chapter expresses 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, this problem is addressed by presenting a novel HAR approach that learns to output activity sets using deep neural networks. Moreover, motivated by the limited availability of annotated HAR datasets as well as the current immaturity of existing unsupervised deep learning methods, the supervised set learning scheme is preceded with a prior unsupervised feature learning process that adopts convolutional auto-encoders to exploit unlabelled data. The empirical experiments on two widely adopted HAR datasets demonstrate the substantial improvement of the proposed methodology over comparable methods.

5.1 Motivation and Contribution

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 Yang et al. (2015) or the last Ordóñez and Roggen (2016) 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 Bulling, Blanke and Schiele (2014) and therefore, no segment is guaranteed to contain measurements of a single activity type Yao et al. (2018)—the so called *multi-class window problem*. We illustrate the problem in Fig. 5.1, where the acquired sliding window segment carries sensor samples from both walking and standing activities. However, the conventional HAR formulation approximates the ground-truth for this segment with either *walking* or *standing* activity, respectively corresponding to the most and the last observed sample annotations.

Equally important, existing deep HAR systems demand large amounts of annotated training data for enhanced supervised performance. However, large-scale annotated HAR datasets are limited. Further, collection of labelled sensory data is labour intensive, time-consuming and expensive Kim, Helal and Cook (2010). As opposed to other domains (*e.g.* image recognition) where human visualisation of raw data alleviates the labelling process, manual annotation of sensor signals is a tedious task. Unfortunately, activity recognition systems that leverage the cheaply available

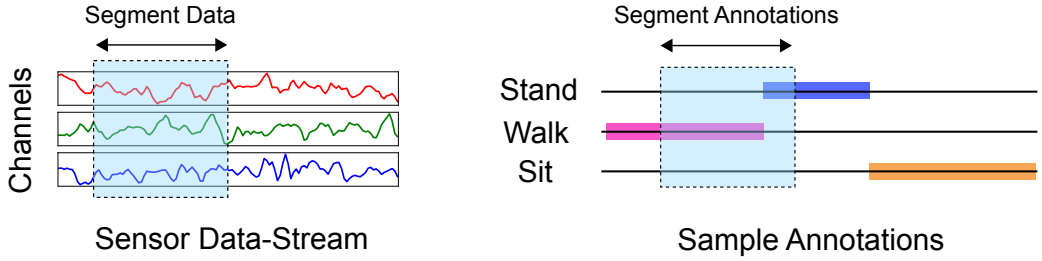


Figure 5.1: Multi-class window problem. The multi-class formulation of HAR approximates the ground-truth for each sensor segment with a single activity label based on either the most or the last observed sample annotations; here, this respectively corresponds to *walking* and *standing* activities.

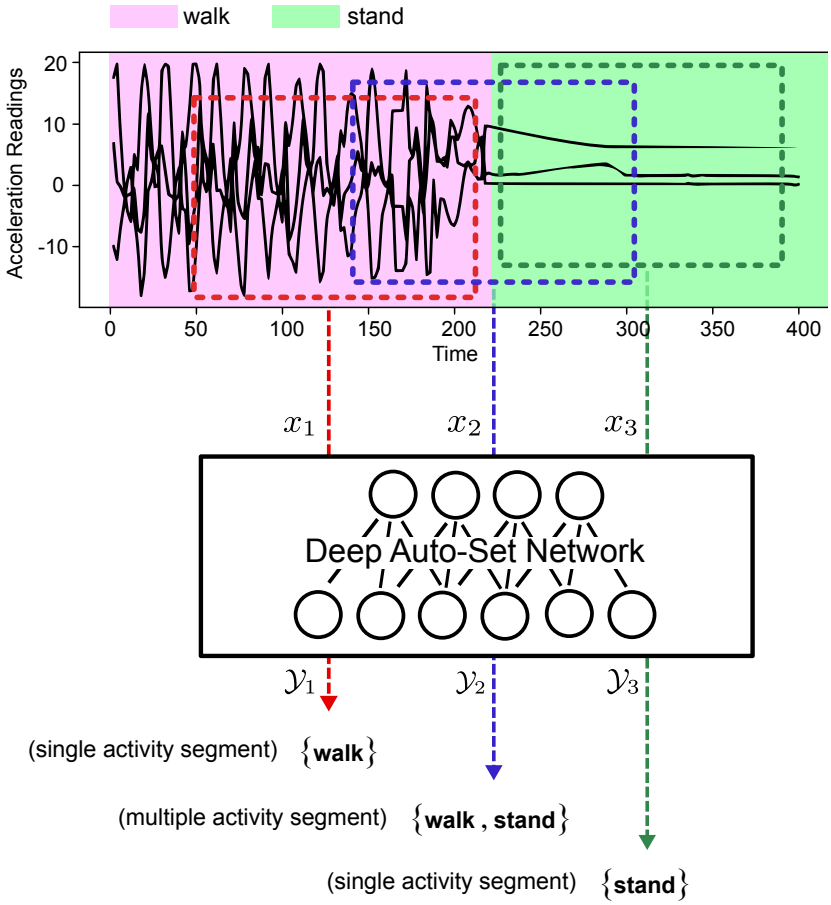


Figure 5.2: High-level overview of Deep Auto-Set. We illustrate the novel *Deep Auto-Set* network to perform precise activity recognition from time-series data. The network consumes windowed raw sensory excerpts ($\mathbf{x}$), automatically extracts distinctive features and outputs corresponding *sets of activities* ($\mathcal{Y}$) with various cardinalities.

unlabelled sensory data are rare in the field and, therefore, necessitates the exploration of effective unsupervised alternatives.

This chapter overcomes 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 the study in Rezatofighi et al. (2018), for the first time we develop a HAR system that performs

activity set learning and inference in a systematic fashion using deep paradigms. In contrast to conventional multi-label approaches, our methodology omits heuristic thresholding methods for selecting activity labels and instead learns to predict cardinality in addition to the activity labels. Further, motivated by the scarcity of annotated HAR datasets, we complement our supervised training scheme with a prior unsupervised feature learning step that exploits unlabelled time-series data. Through experiments on widely adopted public HAR datasets, we demonstrate the significant improvement achieved from the proposed deep learning based methodology, the *Deep Auto-Set* network (depicted in Fig. 5.2), over the baseline models.

The main contributions of this chapter are summarised as follows:

1. For the first time, a novel formulation of a human activity recognition problem from body worn sensor data streams is investigated where the predictions for sensor segments are expressed as *activity sets*. The proposed formulation naturally handles sensory segments with varying number of activities and thus, avoids the potential loss of information from conventional ground-truth approximations necessary during model training.

2. We present Deep Auto-Set: a unified deep learning paradigm that (a) seamlessly functions on raw multi-modal sensory segments, (b) exploits unlabelled data to uncover effective feature representations, and (c) incorporates set objective to learn mappings from input sensory data to target activity sets.

3. The effectiveness of the proposed Deep Auto-Set network is demonstrated through empirical experiments on two HAR representative datasets. In addition, the components of the proposed methodology are examined in isolation, to present insights on their contribution to an enhanced recognition performance.

5.2 Related Work

Motivated by the unparalleled performance of end-to-end learning in diverse application domains, we are seeing an increasing adoption of deep learning paradigms in HAR Zeng et al. (2014); Yang et al. (2015); Hammerla, Halloran and Plötz (2016); Ordóñez and Roggen (2016); Yao et al. (2018). 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 Zeng et al. (2014); Yang et al. (2015) where raw sensory data were processed by convolutional layers to extract discriminative features. Going beyond CNNs, Hammerla *et al.* Hammerla, Halloran and Plötz (2016) 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 Ordóñez and Roggen (2016) developed a recurrent-based neural network (RNN) for wearable sensors and reported state-of-the-art performance on a representative HAR dataset. Except for the dense labelling and prediction approach in Yao et al. (2018), existing supervised solutions are based on the assumption that all samples within a sliding window segment share the same activity annotation. We argue that such an assumption is counter-intuitive to the diverse nature of human activities with varying duration and hinders accurate analysis of segments with multiple activities. In this chapter, we present a novel network that naturally allows segmented sensory data to 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 generalisation performance. Taking into account the scarcity of annotated HAR datasets and the difficulty of doing so, we exploit unlabelled time-series data to learn useful feature representations by adopting convolutional auto-encoders. In this regard, the most relevant study to ours is Alsheikh et al. (2016) where layer-wise pre-training of fully connected deep belief networks is adopted and the recognition problem is limited to pre-processed spectrograms of acceleration measurements. In contrast, our proposed unsupervised methodology substitutes the layer-by-layer pre-training with an end-to-end optimisation of the reconstruction objective and is also seamlessly applied on raw multi-modal sensor data.

5.3 Proposed Methodology

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 the proposed solution involves an unsupervised feature learning step (described in Section 5.3.1) that exploits cheaply accessible unlabelled sensor measurements followed by a supervised fine-tuning step (detailed in Section 5.3.2) 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 Ordóñez and Roggen (2016) as the core of our network and apply modifications to suit our problem settings; this architecture comprises of 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 generalisation performance through unsupervised feature learning Erhan et al. (2010).

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. 5.3. 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 time span of the filters.

In order to provide a clear formulation of the problem, here we introduce the notations used throughout this chapter. In this chapter, we use $\mathcal{Y}$ for a set with unknown cardinality and $\mathcal{Y}^m$ for a set with known cardinality m. We define the set of k supported activity elements by $\mathcal{A} = \{a_i\}_{i=1}^{k}$. 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:

- A labelled training dataset $\mathcal{S} = \{(\mathbf{x}_i, \mathcal{Y}_i^{m_i})\}_{i=1}^{n_1}$ of size n_1, where each training instance is a pair consisting of a sensory segment $\mathbf{x}_i \in \mathbb{R}^{\mathrm{D}\times\mathrm{W}}$ with a fixed 2D representation and a target activity set $\mathcal{Y}_i^{m_i} = \{a_1, \ldots, a_{m_i}\} \subseteq \mathcal{A}, |\mathcal{Y}_i| = m_i, m_i \in \mathbb{Z}^+$.

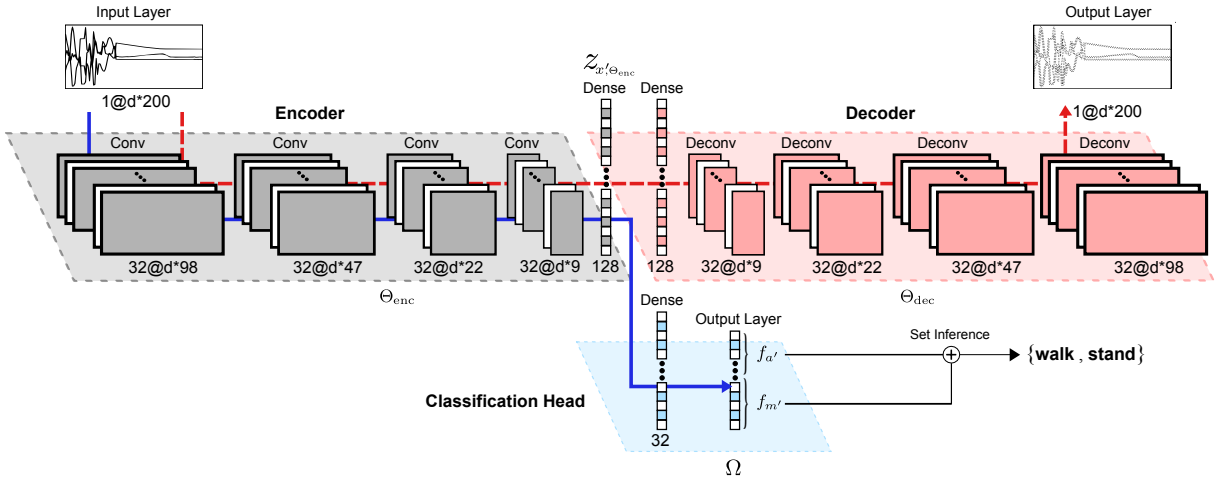


Figure 5.3: Unified architecture of Deep Auto-Set network. The tags above the feature maps refer to the corresponding layer operations. The numbers before and after "@" respectively correspond to the number of generated feature-maps and their dimensions in each layer. In this architecture, all convolution (and deconvolution) layers apply a filter of width 5 (as in Ordóñez and Roggen (2016)) 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 parameterised by Θ_{enc} and Θ_{dec}. Similarly for the supervised step, operations on the solid arrow correspond to the activity set network parameterised by Θ_{enc} and Ω. Once the network parameters are optimised, set inference (as described in Section 5.3.2) is carried out to generate activity set predictions.

- An unlabelled dataset $\mathcal{V} = \{\bar{\mathbf{x}}_i\}_{i=1}^{n_2}$ of size n_2, where each instance is an unlabelled 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 unlabelled training dataset $\mathcal{U} = \{\mathbf{x'}_i\}_{i=1}^{n_1+n_2} = \mathcal{V} \cup \{\mathbf{x}_i\}_{i=1}^{n_1}$ where each training instance $\mathbf{x'}_i \in \mathbb{R}^{D \times W}$ 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 unlabelled dataset.

5.3.1 Unsupervised Feature Learning

Through stacked hidden layers of encoding and decoding operations, auto-encoder learns latent representations of the sensory data in an unsupervised fashion. The reconstruction of unlabelled segments captures the process in which the sensor signals

are generated and allows for the correlations between various sensor channels to 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 unlabelled sensory time segment $\mathbf{x}' \in \mathcal{U}$ on which the encoder network $f_{\text{enc}} : \mathbb{R}^{\text{D}\times\text{W}} \rightarrow \mathbb{R}^p$ (parameterised by Θ_{enc}) is firstly applied to obtain the latent representation $z_{\mathbf{x}',\Theta_{\text{enc}}}$, *i.e.*

$$\mathbf{z}_{\mathbf{x}',\Theta_{\text{enc}}} = f_{\text{enc}}(\mathbf{x}';\Theta_{\text{enc}}). \tag{5.1}$$

The resulting latent representation $\mathbf{z}_{\mathbf{x}',\Theta_{\text{enc}}} \in \mathbb{R}^p$ is then utilised by the decoder network $f_{\text{dec}} : \mathbb{R}^p \rightarrow \mathbb{R}^{\text{D}\times\text{W}}$ (parameterised by Θ_{dec}) to reconstruct the input. Noting that the generated reconstruction is directly influenced by the values of Θ_{enc} and Θ_{dec}, we define the loss incurred by the output of auto-encoder network (illustrated by the dashed path in Fig. 5.3) given the unlabelled segment $\mathbf{x}'$ as

$$\mathcal{L}_{\text{auto}}(\mathbf{x}';\Theta_{\text{enc}},\Theta_{\text{dec}}) = \|\mathbf{x}' - f_{\text{dec}}(\mathbf{z}_{\mathbf{x}',\Theta_{\text{enc}}};\Theta_{\text{dec}})\|^2. \tag{5.2}$$

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

$$(\Theta^*_{\text{enc}},\Theta^*_{\text{dec}}) = \arg\min_{\Theta_{\text{enc}},\Theta_{\text{dec}}} \sum_{i=1}^{n_1+n_2} \mathcal{L}_{\text{auto}}(\mathbf{x}'_i;\Theta_{\text{enc}},\Theta_{\text{dec}}). \tag{5.3}$$

In this architecture, the encoder network extracts features from unlabelled 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 (Θ^*_{enc}) can later be adopted in favour of a better guided supervised training.

5.3.2 Supervised Activity Set Learning and Inference

Using the labelled training dataset $\mathcal{S} = \{(\mathbf{x}_i, \mathcal{Y}_i^{m_i})\}_{i=1}^{n_1}$, 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 $\mathbf{x}^+$. In our

architecture, this is carried out by optimising a *set objective* through tuning the activity set network parameters which include weights corresponding to the encoder layers (Θ_{enc}) as well as the extra dense layers (Ω) in the classification head. Similar to Rezatofighi et al. (2018), in this chapter we adopt joint learning and inference to learn and predict activity sets for HAR which we describe in what follows.

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 $\mathbf{x}$, the output of our activity set network comprises of: i) a *set cardinality* term $f_{m'}(\mathbf{x})$ with log softmax activation which produces cardinality scores; as well as ii) a *set element* term $f_{a'}(\mathbf{x})$ 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. 5.3) given a labelled segment $\mathbf{x}$ with the target set $\mathcal{Y}^m$, we define our set objective as

$$\begin{aligned}\mathcal{L}_{\text{set}}(\mathbf{x}, \mathcal{Y}^m; \Theta_{\text{enc}}, \Omega) = \sum_{a \in \mathcal{Y}} \ell_{\text{bce}}(a, f_{a'}(\mathbf{x}; \Theta_{\text{enc}}, \Omega)) \\ + \ell_{\text{nll}}\left(m, f_{m'}(\mathbf{x}; \Theta_{\text{enc}}, \Omega)\right),\end{aligned} \tag{5.4}$$

where ℓ_{nll} and ℓ_{bce} denote the negative log likelihood loss and the binary cross entropy loss, respectively. We consider the same *i.i.d* assumption adopted in Rezatofighi et al. (2018) for the set elements and perform MAP estimate to train the network parameters by minimising the set objective on the labelled dataset $\mathcal{S}$, *i.e.*

$$(\Theta^*_{\text{enc}}, \Omega^*) = \arg\min_{\Theta_{\text{enc}}, \Omega} \sum_{i=1}^{n_1} \mathcal{L}_{\text{set}}(\mathbf{x}_i, \mathcal{Y}_i^{m_i}; \Theta_{\text{enc}}, \Omega). \tag{5.5}$$

As such, Θ^*_{enc} and Ω^* are estimated by computing the partial derivatives of the objective function in Eq. (5.4) and employing standard back-propagation in order to learn the network parameters.

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

$$\begin{aligned}\mathcal{Y}^* = \arg\max_{m', \mathcal{Y}^{m'}} \; & f_{m'}\left(\mathbf{x}^+; \Theta^*_{\text{enc}}, \Omega^*\right) + m' \log U \\ & + \sum_{a' \in \mathcal{Y}^{m'}} \log f_{a'}(\mathbf{x}^+; \Theta^*_{\text{enc}}, \Omega^*),\end{aligned} \tag{5.6}$$

where U, estimated from the validation set of the data, is a normalisation 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 Rezatofighi et al. (2018).

5.4 Experiments and Results

5.4.1 Datasets

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

WISDM dataset Kwapisz, Weiss and Moore (2011). 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. The recordings from 8 subjects are used as the holdout testing set and the remaining data constitute the training and validation sets.

Opportunity dataset Chavarriaga et al. (2013). 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 Chavarriaga et al. (2013) 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.

5.4.2 Data Preparation

The preparation process involves performing per channel normalisation to scale real valued attributes to [0,1] interval as well as segmentation and ground-truth generation, as described below.

Time-series Segmentation. Following the experiments in Kwapisz, Weiss and Moore (2011); Alsheikh et al. (2016), 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 WISDM 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 WISDM dataset and every 0.67 seconds for the Opportunity dataset.

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 class` activity segment. In our experiments, we set r to half the sensor sampling rates; *i.e.*, 10 and 15 for WISDM and Opportunity datasets, respectively.

5.4.3 Evaluation Metrics

We employ the widely used HAR evaluation measures to report the performance of the baselines and our *Deep Auto-Set* network. We compute per-class *precision*, *recall* and

F-score according to Eq. 2.4, Eq. 2.5, and Eq. 2.6, respectively. For a specific activity label, precision is defined as the ratio of the correctly predicted label occurrences over the total number of label occurrences in the predictions. Similarly, 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-class F-score corresponds to the harmonic mean of precision and recall. Accordingly, we aggregate the per-class measures by reporting the corresponding class-average values of Precision_m, Recall_m and F-score_m.

We also use the overall *exact match ratio* (MR), as adopted in Guo and Gu (2011); Alessandro et al. (2013), to report a harsh evaluation of performance. This metric requires the predicted activity set to exactly match the corresponding target set (both in terms of the set cardinality and the set elements) and therefore, does not tolerate 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 MR_c; *i.e*, for instance 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.

5.4.4 Implementation Details

The experiments are implemented using PyTorch Paszke et al. (2017) 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 optimiser with weight decay and initial learning rate respectively set to $5 \cdot 10^{-5}$ and 10^{-4}, on mini-batches of size 64 by back-propagating the gradients of corresponding loss functions. For the supervised training step, the optimiser 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 WISDM and Opportunity datasets. We refer interested readers to Hammerla, Halloran and Plötz (2016) for excellent guidelines on setting architecture and optimiser hyper-parameters.

Table 5.1: Evaluation of multi-class formulated baseline. Performance evaluation of the baseline CNN architecture Ordóñez and Roggen (2016) trained with multi-class formulated objective against both the approximated ground truth (equivalent to the last observed sample annotation) as well as the actual ground-truth for Opportunity dataset.

Model	Network Prediction	Evaluation Ground Truth	F-score$_m$	MR
CNN Ordóñez and Roggen (2016)	Single activity label	Last sample's label	0.890	87.4%
		Actual labels	0.793	54.7%

5.4.5 Results

A key motivation for our work is the activity information loss that is incurred by conventional ground truth approximations in multi-class problem formulations. In order to verify this, we conform to the conventional multi-class formulation of HAR and train the CNN in Ordóñez and Roggen (2016) by minimising the multi-class classification objective. In Table 5.1, we report performance of the resulting HAR system by comparing the generated predictions against both the *approximate* ground truth (obtained from the last observed sample annotation) as well as the *actual* multi-label ground truth for Opportunity dataset. To clarify, consider the scenario where a sensory segment of interest initiates with measurements of walking and terminates with standing. Thus, the approximate ground truth would be `standing` but the actual ground truth labels are the set {`walking`,`standing`}. Assuming that the network solving the multi-class formulated problem predicts the underlying activity to be `standing`, in our evaluation against the actual ground truth represented by the set of labels {`walking`, `standing`}, the predicted class `standing` is treated as a true positive whereas the missing class `walking` is considered as a false negative.

In Table 5.1, the lower performance measures obtained from the evaluation against the actual ground truth labels as compared with the approximated ground truth suggest that there are sensory segments in the HAR dataset that convey measurements of multiple activities in the time span of the sliding window—see the result for MR in Table 5.1. For these segments, approximating the ground-truth can lead to missed activity information for a multi-class formulation of HAR, especially in the presence of short duration activities such as activity transitions Yao et al. (2018). In contrast, a set-based formulation allows capturing the presence of multiple activity labels in the ground truth. Although we have shown a comparison for a multi-class problem formulation commonly employed for HAR, we can see that it is not possible to make a

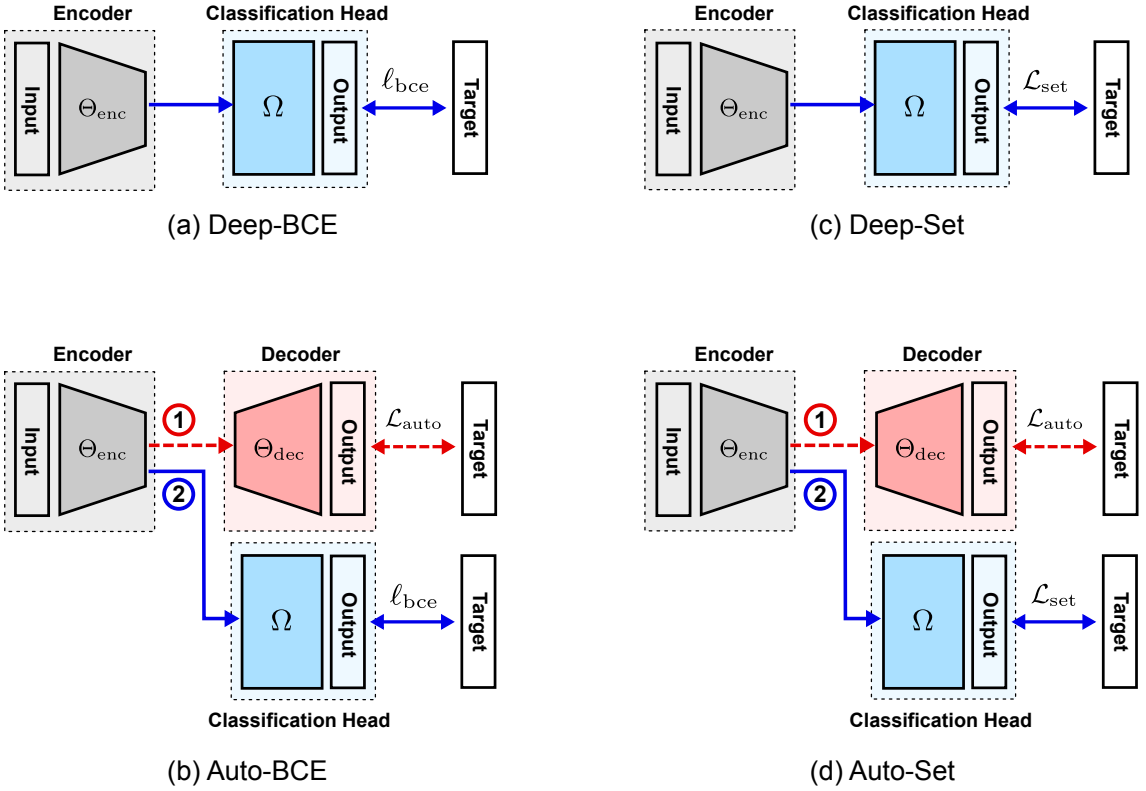


Figure 5.4: Investigated Frameworks. We present an overview of different activity recognition models explored in this chapter.

fair comparison with our set-based formulation beyond what we have observed here. Therefore, 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.

Activity Recognition Models. In Fig. 5.4, we illustrate the schematic architectures for investigated frameworks in this chapter and provide a brief description in what follows:

- *Deep-BCE*: A conventional multi-label model that follows a purely supervised minimisation of binary cross entropy loss (ℓ_{bce}) for training and heuristic thresholding of activity scores for inference.

- *Auto-BCE*: A conventional multi-label model that leverages a prior unsupervised feature learning step via minimisation of reconstruction objective ($\mathcal{L}_{auto}$) as well as a supervised optimisation of binary cross entropy loss.

Table 5.2: Exact match ratio evaluation of multi-label formulated frameworks. We present a comparison of the proposed *Deep Auto-Set* network against the baselines according to the obtained exact match ratio for each dataset. The best results are highlighted with boldface. Note that for the WISDM 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
WISDM	(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%**

- *Deep-Set*: A set-based model that follows a purely supervised optimisation of the set objective ($\mathcal{L}_{\text{set}}$) proposed in Eq. (5.4) for training and the MAP inference introduced in Eq. (5.6) for set inference.

- *Auto-Set*: The proposed Deep Auto-Set model elaborated in Section 5.3.

Notably, 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 within a segment, all recognition models adopted in this chapter are capable of predicting multiple activities for a given sensory segment. We adopt the same layer operations presented in Fig. 5.3 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 5.2 and Table 5.3 for different evaluation metrics. From the reported results, we can see that our novel *Deep Auto-Set* network consistently outperforms the baselines on WISDM and Opportunity datasets in terms of both F-score and exact match ratio performance metrics. Moreover, the match ratios in Table 5.2 suggest that *Deep 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

Table 5.3: F-score, precision and recall evaluation of multi-label formulated frameworks. We provide a comparison of the proposed *Deep Auto-Set* network against the baselines according to the obtained class-average F-score (F-score_m), precision (Precision_m) and recall (Recall_m) for each dataset. The best results are highlighted with boldface.

Dataset	Model	F-score_m	Precision_m	Recall_m
WISDM	(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**

values); as well as *iii*) recognising short duration human activities (implied from MR_3 values). 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 set loss function for the underlying problem.

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

- Activity recognition systems that leverage unlabelled data present better performance over their solely supervised variants; *e.g.*, note the improved performance of *Auto-BCE* over *Deep-BCE*.
- Compared with a conventional multi-label formulation: i) incorporating set loss into the training process can allow the network to learn multiple activities represented in the ground truth data of a given segment more accurately; and ii) the set inference procedure can jointly exploit cardinality and set element scores to generate predictions instead of empirically determined thresholds; *e.g.*, note the performance improvement of *Deep-Set* over *Deep-BCE*.
- While each component of the 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.

5.5 Conclusion

This chapter contrasted the de facto HAR problem formulation with a novel set prediction formulation. As opposed to the conventional multi-class treatment of HAR problems, the intuitive formulation developed herein allows sensory segments to be associated with a set of activities and thus, naturally handles segments with multiple activities. In a unified architecture, the corresponding activity recognition problem was addressed by developing a deep HAR system that: (*a*) exploits unlabelled data to uncover effective feature representations; and (*b*) incorporates a set objective to learn mappings from input sensory segments to target activity sets. To provide insights on how each component of the proposed methodology contributes to enhance recognition performance in isolation, three different multi-label activity recognition models were explored as the baselines. Finally, through empirical experiments on HAR representative datasets, the effectiveness of the proposed *Deep Auto-Set* network for human activity recognition was demonstrated.

While not explored in this chapter, the proposed set-based methodology naturally provides an elegant solution for the challenging problem of *concurrent human activity recognition*; here, the task aims to recognise not only the sequentially occurring actions but also the co-occurring activities. Hence, we leave it for future work to explore the effectiveness of the proposed set-based framework to tackle complex concurrent activity recognition scenarios.

Up until this chapter, we investigated *supervised training regimes* to learn activity representations from raw multi-channel time-series data; *i.e.*, in Chapter 3 and Chapter 4 the entire network parameters were trained with full supervision from labelled HAR datasets, and in this chapter (*a*) the network weights were pre-trained through unlabelled data and subsequently (*b*) all parameters (including those corresponding to the feature extractor) were again finetuned with full supervision with annotated data. As emphasised, the process of collecting annotated sensory datasets is tedious, time-consuming and not scalable to large volumes of data. On the contrary, unlabelled data acquisition is cheap and feasible at large-scale; *e.g.*, physical activities of athletes, factory workers, or hospitalised patients can continuously be recorded through low-cost and unobtrusive embedded sensors throughout the day without demanding a human workforce to provide online or post-hoc annotations. Accordingly, it is crucially important to comprehend the potentials of *unsupervised learning* alternatives in ubiquitous computing and explore

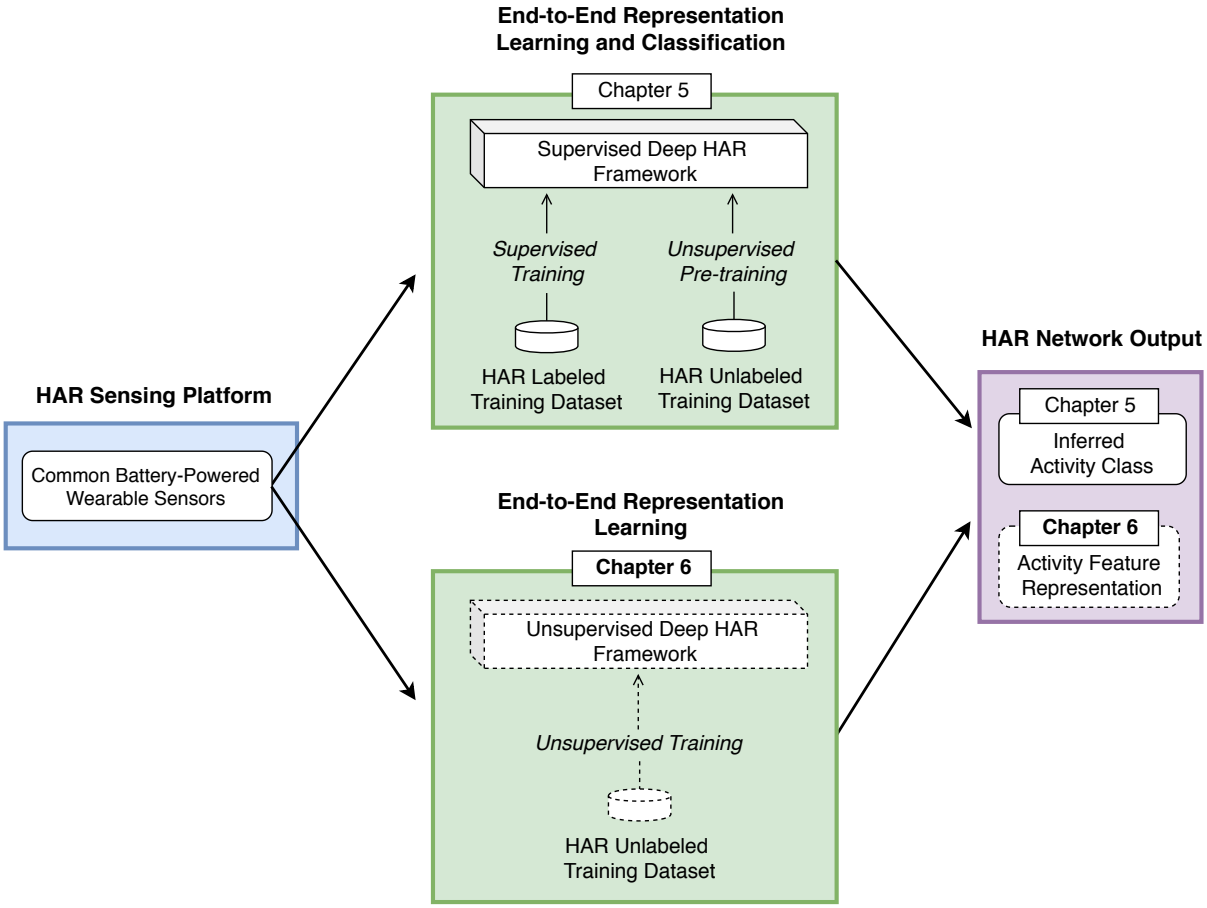


Figure 5.5: Upcoming chapter sneak peek.

systematic approaches to exploit vast amounts of easily accessible unlabelled data to promote new HAR application opportunities. This forms the core motivation for the problems investigated in Chapter 6 and Chapter 7, where we respectively study *unsupervised representation learning* and *clustering* of human actions from wearable sensor data using deep learning paradigms. In Fig. 5.5, we illustrate the examined HAR problem in this chapter in comparison against the explored unsupervised problem in Chapter 6.

Chapter 6

Unsupervised Representation Learning with Generative Adversarial Networks

GENERATIVE adversarial networks (GANs) are emerging as state-of-the-art for diverse visual benchmarks from synthetic data generation to unsupervised representation learning for natural images. This chapter considers the problem of building a generative adversarial network architecture for multi-modal sequential data capable of communicating in a bidirectional GAN framework to learn unsupervised feature representations from easily accessible unlabelled activity data. Addressing this problem results in acquiring enriched feature representations that can effectively serve subsequent downstream classification tasks. The proposed network formulation, *Guided-GAN*, alleviates the burden on the discriminator in achieving inverting generators and encoders by seeking to augment feedback from geometric distance penalisation in data and latent manifolds. Interestingly, we discover that the proposed formulation is vital for successfully training a bidirectional GAN framework in the sequential domain. In addition, the quality of features learned in the unsupervised setting are evaluated on three downstream classification benchmarks, outperforming existing unsupervised approaches whilst closely approaching the performance of fully supervised learned representations.

6.1 Motivation and Contribution

Generative adversarial networks (GANs) Goodfellow et al. (2014) built upon deep convolutional neural networks have emerged as the state-of-the-art in diverse natural image generation benchmarks ranging from image-to-image translation Liu and Tuzel (2016); Zhu et al. (2017); Taigman, Polyak and Wolf (2017); Park et al. (2019) to super-resolution Sønderby et al. (2017); Ledig et al. (2017). Despite their demonstrated empirical strength in the visual domain to capture semantic variations of data distributions, the adoption of GANs for the challenging task of unsupervised representation learning for sequential multi-modal data remains. In our efforts to bridge the gap between unsupervised representation learning and GANs for multi-modal sequential data, this chapter considers unsupervised investigation of GAN's latent feature space to uncover discriminative sequence representations.

Sequential data are uniquely characterized by the inherent sample dependencies across time and demand architecture designs beyond convolutional operators for temporal modelling. In addition, GANs are notorious for their unstable training process and sensitivity to hyper-parameter selections. Although, the immense community effort in the visual domain has resulted in well-established guidelines for architectural designs—weight initialisations and hyper-parameter settings for stable training of convolutional GANs Radford, Metz and Chintala (2015); Salimans et al. (2016); Miyato et al. (2018a); Karras et al. (2018)—the same exploration is lacking in the sequential domain. Moreover, examinations of GANs for temporal data are predominantly confined to synthesising artificial sequences that resemble the original data Mogren (2016); Esteban, Hyland and Rätsch (2017); Alzantot, Chakraborty and Srivastava (2017); Moshiri et al. (2020); Wang et al. (2018a); Yoon, Jarrett and van der Schaar (2019), while investigation of GAN's latent feature space for unsupervised learning remains.

Grounded on the immense success of generative models in the visual domain, this chapter explores the GAN's latent feature space to offer an appealing alternative to the de facto autoencoder-based frameworks Freitag et al. (2018); Bai et al. (2019); Haresamudram, Anderson and Plötz (2019) for sequence representation learning; this chapter proposes a GAN formulation in our efforts to bridge the gap between unsupervised representation learning and GANs for multi-modal sequential data. Our network architecture draws inspiration from BiGANs Donahue, Krähenbühl and Darrell (2017); most significantly we present a critical new formulation to enable its *robust* application for sequential data. Firstly, we design a sequential generator,

encoder and joint discriminator architectures armed with recurrent neural networks that can cooperate in a unified framework. Second, an intuitive extension for effective training of our recurrent framework is proposed. Interestingly, the generator and encoder in BiGAN do not directly communicate; instead, a discriminator receiving pairs of data and latents conducts the discrimination in the joint space and is responsible for encouraging the encoder to uncover the generator's inverse mapping. Consequently, BiGAN relies solely on the discriminator's guidance to match the joint data-latent distributions and achieve inverting generator and encoder components at the adversarial game's theoretical solution. *Unfortunately, convergence to the optimal theoretical solution is difficult to meet in practice and thus, encoder does not necessarily uncover the inverse mapping* Dumoulin et al. (2017); Zhang et al. (2018b). To address these difficulties, *we alleviate the burden on the discriminator by additionally aligning the data and latent manifolds through geometric distance penalisation*. Our intuitive approach measures and attempts to minimise the errors associated with reconstructing data and latent samples and is efficiently implemented through re-using existing components and *weight sharing* as illustrated in Figure 6.3. The key contributions made in this chapter are summarised as follows:

1. A first rigorous study of generative adversarial frameworks for unsupervised representation learning from sequential multi-modal data is presented.

2. A *novel* unsupervised representation learning framework with symmetrically orchestrated recurrent generator and encoder components is developed. The proposed framework augments adversarial feedback with geometric distance guidance to encourage the encoder to invert the generator mappings. Exploiting the symmetrical architecture, the Guided-GAN is efficiently implemented with weight sharing and re-using the generator and the encoder for the reconstruction tasks.

3. A series of systematic experiments are conducted to demonstrate the effectiveness and generalisability of the proposed approach.

6.2 Related Work

Recently, generative adversarial networks (GANs) Goodfellow et al. (2014) built upon deep convolutional neural networks have demonstrated great success in

approximating arbitrary complex data distributions and thus, have emerged as the state-of-the-art for realistic data generation on variety of benchmarks Denton et al. (2015); Radford, Metz and Chintala (2015); Brock, Donahue and Simonyan (2019). The incorporated adversarial training pits a generative network against a discriminative model in a minimax game. As opposed to its traditional counterparts (*e.g.* variational autoencoders Kingma and Welling (2014)), GAN's optimisation objective directly aims for plausible data generation instead of enforcing element-wise reconstruction; casting it as an elegant framework within the research community for capturing high-level semantics.

Beyond data generation, recent research efforts highlight investigations into the latent representations learnt using this powerful framework for unsupervised learning of enriched feature representations. However, the standard GAN framework lacks an inference network; *i.e.* while the generator learns mappings from the latent space to the data space, directly inverting this process is not accounted for. Accordingly, the visual perception arena has witnessed recent attempts Chen et al. (2016); Zhang et al. (2018b); Donahue, Krähenbühl and Darrell (2017); Dumoulin et al. (2017); Perarnau et al. (2016); Donahue and Simonyan (2019) to extend the GAN framework with an encoder network responsible for projecting natural images back into their corresponding latent feature representations. In Chen et al. (2016), mutual information maximisation is adopted to infer and gain control over a subset of latent features. In order to achieve full inference on the latents in Perarnau et al. (2016); Zhang et al. (2018b), the generator output is directly fed to an encoder network that is trained to reconstruct the latents entirely. In particular, Donahue, Krähenbühl and Darrell (2017); Dumoulin et al. (2017) take an interesting approach for learning the generator's inverse mapping and propose the bidirectional GAN (BiGAN) as an effective means for visual representation learning on `ImageNet` Deng et al. (2009); within the framework, the encoder and the generator do not directly communicate. However, the discriminator receives pairs of data and latents to conduct the discrimination task in the joint space. The resulting adversarially learned representations demonstrate state-of-the-art performance when transferred for auxiliary supervised discrimination tasks on natural images.

In contrast, interest in the adoption of GANs for sequential multi-modal data has predominantly focused on realistic sequence generation. In a preliminary work, Mogren (2016) proposes a generative adversarial model that operates on continuous sequential data and applies the framework on a collection of classical music. The

resulting GAN adopts LSTM networks for generator and discriminator to generate polyphonic music. Applying architectural modifications, the authors in Esteban, Hyland and Rätsch (2017) develop a recurrent GAN to produce synthetic medical time-series data. The approach is further extended to exploit data annotations for conditional generation in order to substitute sensitive patient records. In Alzantot, Chakraborty and Srivastava (2017), authors generate synthetic sensor data preserving statistics of smartphone accelerometer sensor traces. Similarly in Wang et al. (2018a), authors attempt to synthesise sensory data captured by wearable sensors for human activity recognition. In order to cover the multi-modal distribution of human actions, independent activity-specific GANs are developed. We observe that deep generative models have demonstrated significant potential in unsupervised learning of enriched features for natural images in the visual domain, however, investigations into the sequential domain remains.

6.3 Background and Methodology

Multi-modal sensing platforms continuously record measurements through different sensor channels over time and generate sequential multi-modal data. The acquired stream is then partitioned into segments $\mathbf{x} \in \mathbb{R}^{D \times W}$ using a sliding window, where D denotes the number of sensing modalities used for data acquisition and W represents the choice for the window duration. Here, the goal is to learn unsupervised representations enriched with distinctive features that can subsequently benefit classification of generated sequences. In what follows, we first discuss how existing GAN frameworks can be adopted for unsupervised representation learning of such sequences (Section 6.3.1). Highlighting the existing challenges, we then introduce our novel framework to uncover unsupervised representations with higher correspondence to class semantics (Section 6.3.2).

6.3.1 Unsupervised Representation Learning with GAN frameworks

Recurrent Generative Adversarial Networks

The standard GAN Goodfellow et al. (2014) comprises of two parameterised feed-forward neural networks—a generator $\mathcal{G}_\phi$ and a discriminator $\mathcal{D}_\omega$—competing against one another in a minimax game. Ultimately, the goal is for the generator to

capture the underlying data distribution p_x. To this end, the generator exploits a simple prior distribution p_z to produce realistic samples that trick the discriminator. On the contrary, the discriminator is trained to distinguish between the real and the generated samples. The resulting adversarial game optimises

$$\min_{\mathcal{G}_\phi} \max_{\mathcal{D}_\omega} \mathbb{E}_{\mathbf{x}\sim p_x}[\log \mathcal{D}_\omega(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_z}[\log(1 - \mathcal{D}_\omega(\mathcal{G}_\phi(\mathbf{z})))], \tag{6.1}$$

where, $\mathbf{x} \sim p_x$ represent the mini-batch training samples and $\mathbf{z} \sim p_z$ denote the drawn latents.

Extending the vanilla GAN to generate sequences of real-valued data, Esteban, Hyland and Rätsch (2017) substitutes both the generator and discriminator with recurrent neural networks and develops the Recurrent GAN (RGAN) for medical time-series generation. Within the resulting framework depicted in Fig. 6.1-a, the generator $\mathcal{G}_\phi$ takes a latent sample $\mathbf{z}$ and sequentially generates multi-channel data for each time-step. Similarly, the discriminator $\mathcal{D}_\omega$ consumes an input sequence and delivers per time-step classification decisions. We visualise the internal structure of these components in Fig. 6.2.

While the focus in Esteban, Hyland and Rätsch (2017) is solely on sequence generation, in our experimental study, re-investigate the framework for the purpose of unsupervised feature learning for sequences; the intermediate representations from the trained discriminator of a GAN are found to capture useful feature representations for related supervised tasks Radford, Metz and Chintala (2015). Intuitively, these set of features are attained free of cost and encoded in the discriminator weights when distinguishing real sequences from generated sequences during training. Notably,

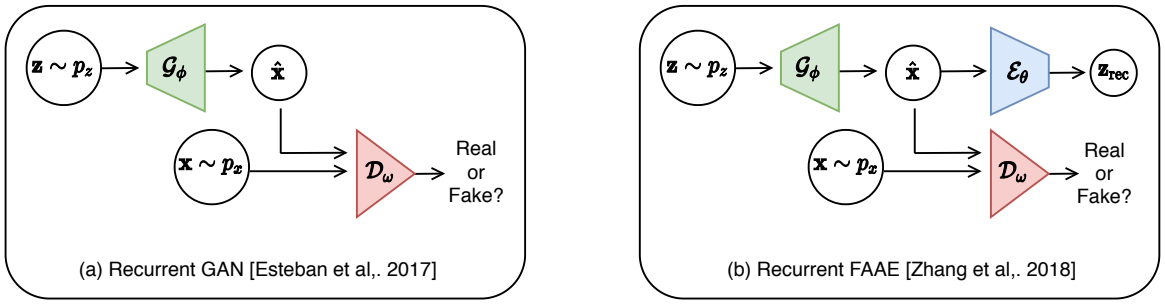


Figure 6.1: Baseline generative framework pipelines. We visualise the workflow for (a) RGAN Esteban, Hyland and Rätsch (2017), and (b) Recurrent adaptation of flipped adversarial autoencoder proposed in Zhang et al. (2018b).

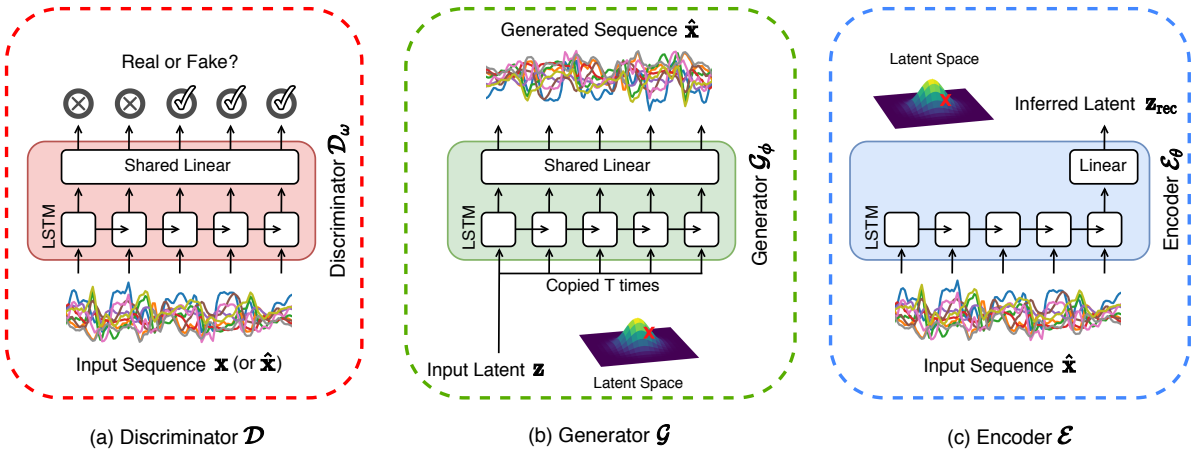


Figure 6.2: Baseline generative framework building blocks. We illustrate the baselines' recurrent building blocks: (a) Discriminator functions in the data space to produce real vs. fake classification scores at each time-step. (b) Generator consumes a latent input repeated to the sequence length and produces a synthetic sample in the sequential data space. (c) Encoder serves as the inference machine and projects an input sequence into its corresponding latent representation. Notably, we depict a slightly modified version of Esteban, Hyland and Rätsch (2017); *i.e.*: i) a shared linear layer is added on top of the recurrent networks; and ii) instead of sampling independent latents, a single latent is sampled and replicated to the sequence length.

RGAN provides the arguably most straightforward extension of a regular GAN for the sequential domain and thus, we base our own investigations by building upon this framework.

Recurrent Flipped Adversarial AutoEncoder

Despite their empirical strength to model arbitrary data distributions, the vanilla GAN and in turn the RGAN, naturally lack an inference mechanism to directly infer the latent representation $\mathbf{z}$ for a given data sample $\mathbf{x}$. Accordingly, Zhang et al. (2018b) proposes a natural extension of GANs to jointly train an encoder network that embodies the inverse mapping and coins the name Flipped Adversarial AutoEncoder (FAAE). Visualised in Fig. 6.1-b, the resulting framework exploits the adversarial guidance of a discriminator in the data space exactly identical to a regular GAN. In order to train the encoder, it additionally minimises the reconstruction error associated with reproducing the latent representations. The training objective thus translates to:

$$\min_{\mathcal{G}_\phi, \mathcal{E}_\theta} \max_{\mathcal{D}_\omega} \mathbb{E}_{\mathbf{x}\sim p_x}[\log \mathcal{D}_\omega(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_z}[\log(1 - \mathcal{D}_\omega(\mathcal{G}_\phi(\mathbf{z})))] + \|\mathbf{z} - \underbrace{\mathcal{E}_\theta(\mathcal{G}_\phi(\mathbf{z}))}_{\mathbf{z}_{\text{rec}}}\|_2^2, \quad (6.2)$$

where, $\mathcal{E}_\theta$ denotes the parameterised encoder network and $\mathbf{z}_{\text{rec}} = \mathcal{E}_\theta(\mathcal{G}_\phi(\mathbf{z}))$ is the reconstructed latent representation.

Since there exists no previous exploration of FAAE framework for the sequential domain, we design and augment RGAN with a recurrent encoder $\mathcal{E}_\theta$ depicted in Fig. 6.2. In particular, the encoder reads through the generated sequences from the generator and updates its internal hidden state according to the received measurements at each time step. Ultimately, the final hidden state after processing the entire sequence is exploited to regress the latent representations. In addition, the encoder and generator receive adversarial feedback from the discriminator for parameter updates during training.

Bi-directional Generative Adversarial Networks

While the FAAE framework paves the way for learning the inverse mapping function, the encoder $\mathcal{E}_\theta$ performance is heavily reliant on the quality and diversity of generator's produced samples. Essentially, the encoder is never exposed to the original data from the training set and thus, its learned feature representations are handicapped by the generator's performance. Accordingly, Donahue, Krähenbühl and Darrell (2017); Dumoulin et al. (2017) propose the BiGAN with a novel approach to integrate efficient inference. We illustrate BiGAN while contrasting it with the proposed formulation, Guided-GAN, in the gray shaded area in Fig. 6.3; *i.e.*, the discriminator is modified to discriminate not only in the data space, but rather in the joint data-latent space between $(\mathbf{x}, \mathcal{E}_\theta(\mathbf{x}))$ and $(\mathcal{G}_\phi(\mathbf{z}), \mathbf{z})$ pairs. Hence, the corresponding minimax objective is defined as

$$\min_{\mathcal{G}_\phi, \mathcal{E}_\theta} \max_{\mathcal{D}_\omega} \mathbb{E}_{\mathbf{x}\sim p_x}[\log \mathcal{D}_\omega(\mathbf{x}, \mathcal{E}_\theta(\mathbf{x}))] + \mathbb{E}_{\mathbf{z}\sim p_z}[\log(1 - \mathcal{D}_\omega(\mathcal{G}_\phi(\mathbf{z}), \mathbf{z}))]. \tag{6.3}$$

To satisfy the objective, the generator is motivated to produce samples resembling the real data distribution and the encoder is incentivised to output latent representations matching with the prior latent distribution. It is shown in Donahue, Krähenbühl and Darrell (2017) that the theoretical optimal solution to this adversarial game leads to the encoder and generator inverting one another while the joint distributions are aligned. Importantly, the encoder in BiGAN has the luxury of directly learning from real samples $\mathbf{x} \sim p_x$.

6.3.2 The Proposed Framework

We formulate a novel framework to uncover unsupervised representations with higher correspondence to class semantics by drawing inspiration from the BiGAN as visualised in Fig. 6.3. Our work addresses the problem of convergence and enables the application of bi-directional GANs for sequential multi-modal data. We observe, that the generator and encoder within the BiGAN framework do not directly communicate. Consequently, the discriminator alone bears the burden of matching the joint data-latent distributions and guiding the encoder and generator components towards inverting one another at the optimal solution. Unfortunately, converging to the optimal theoretical solution is difficult to achieve in practice; thus, the encoder and the generator do not necessarily invert one another Dumoulin et al. (2017); Zhang et al. (2018b). We observed this behaviour to be even more problematic in the sequential data domain. In response, a new intuitive extension is proposed to alleviate the training convergence observed in BiGANs and an efficient implementation architecture is designed as detailed in what follows.

Geometrically-Guided Adversarial Feedback

In addition to the adversarial feedback provided by the discriminator to match the joint data-latent distribution, we optimise geometric distance functions to match the marginal manifolds independently; *i.e.*, we receive gradients from aligning: **i)** the original data manifold with generator's induced output manifold; and **ii)** the prior latent manifold with the encoder's output manifold. In particular for early stages of training, geometric distance optimisation usually provides much easier training gradients Che et al. (2017). We discovered this to be a vital necessity for successful training of a BiGAN in the sequential domain where GAN heuristics may be missing[1]. To this end, we measure and minimise the reconstruction errors associated with reproducing both the data and the latent representations. Hence, the proposed

[1]Our attempts to train recurrent BiGANs without the proposed manifold distance minimisation terms were unsuccessful. Specifically, the encoder did not learn useful representations (resulting in extremely low downstream classification performance), and failed to uncover the generator's inverse mapping function.

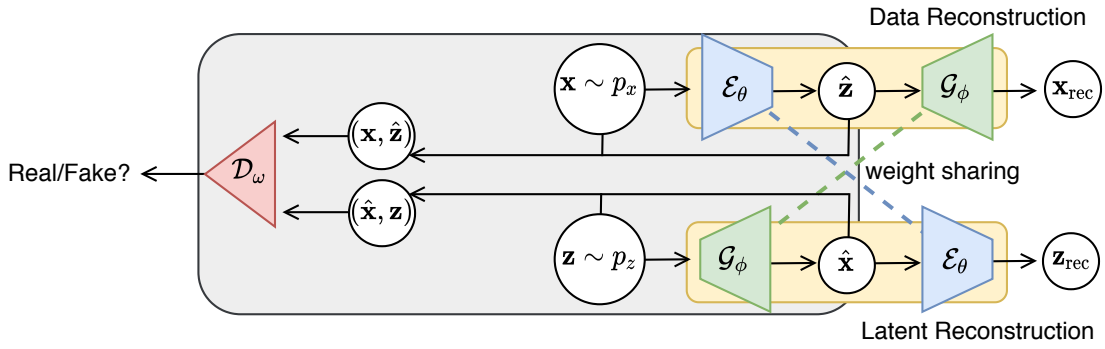


Figure 6.3: Guided-GAN framework pipeline. We present an overview of the proposed adversarial game for unsupervised representation learning from sequential multi-modal data. Compared to Donahue, Krähenbühl and Darrell (2017) (highlighted by the gray box), Guided-GAN proposes: (*a*) incorporating additional gradient feedback from geometric distance minimisation in both data and latent manifolds, (*b*) the efficient architecture implementation through parameter sharing (indicated by dashed lines), and (*c*) integration of our designed recurrent building blocks (depicted in Fig. 6.4) for temporal modelling.

minimax adversarial game is formulated as:

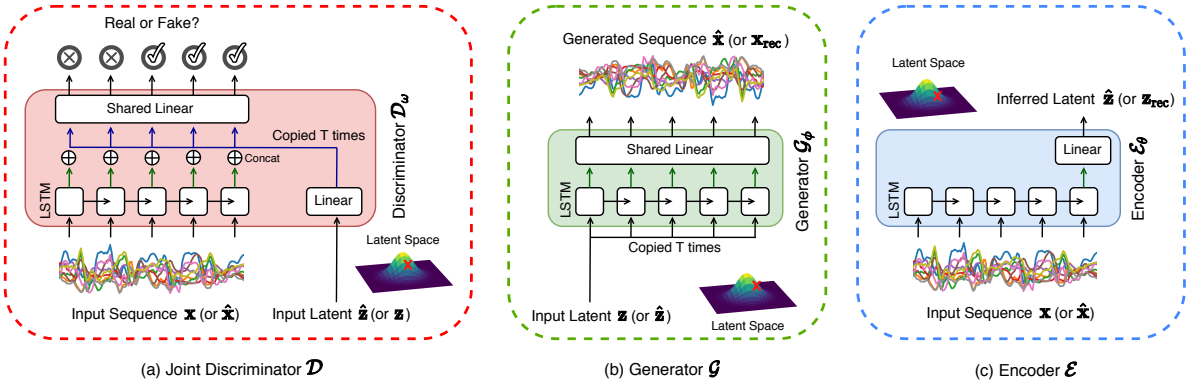
$$\begin{aligned}\min_{\mathcal{G}_\phi, \mathcal{E}_\theta} \max_{\mathcal{D}_\omega} \mathbb{E}_{\mathbf{x}\sim p_x}[\log \mathcal{D}_\omega(\mathbf{x}, \mathcal{E}_\theta(\mathbf{x}))] + \mathbb{E}_{\mathbf{z}\sim p_z}[\log(1 - \mathcal{D}_\omega(\mathcal{G}_\phi(\mathbf{z}), \mathbf{z}))] \\ + \lambda_x \|\mathbf{x} - \underbrace{\mathcal{G}_\phi(\mathcal{E}_\theta(\mathbf{x}))}_{\mathbf{x}_{\text{rec}}}\|_2^2 + \lambda_z \|\mathbf{z} - \underbrace{\mathcal{E}_\theta(\mathcal{G}_\phi(\mathbf{z}))}_{\mathbf{z}_{\text{rec}}}\|_2^2,\end{aligned} \tag{6.4}$$

where $\|.\|_2^2$ imposes the ℓ_2 reconstruction retrieval in the data and latent feature spaces. Notably, the reconstruction errors are efficiently computed at no extra model complexity cost through weight sharing as depicted in Fig. 6.3. Here, λ_x and λ_z denote the loss balancing coefficients.

Recurrent Symmetrical Adversarial Framework

In order to exploit the sequential nature of multi-modal data streams, the core building blocks of our framework are empowered by recurrent units with memory cells, as depicted in Fig. 6.4. Moreover, the generator and encoder communicate in a *symmetrical orchestration* to measure the manifold distances in data and latent spaces.

Recurrent Joint Discriminator. The discriminator $\mathcal{D}_\omega$ serves in the joint data-latent space attempting to differentiate joint input samples of $(\mathbf{x}, \mathcal{E}_\theta(\mathbf{x}))$ against $(\mathcal{G}_\phi(\mathbf{z}), \mathbf{z})$



(a) Joint Discriminator $\mathcal{D}$ (b) Generator $\mathcal{G}$ (c) Encoder $\mathcal{E}$

Figure 6.4: Guided-GAN framework building blocks. We present a visualisation of our developed recurrent building components (a) Discriminator receives pairs of data-latents, learns a sequence of aggregated feature representations and delivers per time-step classification scores, (b) Generator and (c) Encoder leverage the same architecture design as recurrent GAN and FAAE. However, they are additionally trained to minimise the associated reconstruction errors in data and latent spaces, respectively. For discriminator, $\bigoplus$ represents the concatenation operation of projected data and latent features.

pairs. Internally, the multi-modal stream is initially processed by an LSTM network yielding a sequence of hidden state representations. Simultaneously, the latent input is linearly projected and concatenated to the LSTM hidden states at each time-step. The resulting sequence aggregates the learned features from both data and latent spaces, and is used to construct per time-step classification decisions.

Symmetrical Generator and Encoder. The generator $\mathcal{G}_\phi$ and encoder $\mathcal{E}_\theta$ of our framework share the same internal structure with the recurrent GAN and FAAE. However, they are symmetrically connected and serve augmented responsibilities: ***i)*** the generator is additionally exposed to encoded latents from the encoder's posterior distribution ($\hat{\mathbf{z}} = \mathcal{E}_\theta(\mathbf{x}); \mathbf{x} \sim p_x$) and is trained to reconstruct the input sequence; and ***ii)*** the encoder observes the generator's outputs ($\hat{\mathbf{x}} = \mathcal{G}_\phi(\mathbf{z}); \mathbf{z} \sim p_z$) and learns to regress the corresponding latent representation. *Notably, given the symmetrical architecture design, our encoder now has access to both the original data samples (in contrast to FAAE) as well as the novel generated data samples (as opposed to BiGAN) to uncover generalisable feature representations and serve as a fully fledged feature extractor.*

6.4 Experiments and Results

6.4.1 Datasets

To validate our framework and provide empirical evidence towards its generalisability, we investigate the effectiveness of our unsupervised representations on two downstream classification tasks: *a)* sequential digit classification, and *b)* human activity recognition (HAR) from body worn multi-modal sensors. Notably, enabling unsupervised HAR solutions creates new opportunities, especially in the health domain, where collection of annotated data is tedious, time-consuming and not scalable to large volumes of data.

Sequential MNIST Dataset LeCun et al. (1998). For visual interpretation, we re-purpose the popular MNIST hand-written digits dataset as sequential multi-channel dataset; *i.e.*, each 28×28 gray-scale digit is treated as a sequence of 28 time-steps, with each time-step carrying values over 28 channels. This dataset contains 10 distinct digit categories and offers standard train and test splits with respectively 60000 and 10000 samples.

UCI HAR Dataset Anguita et al. (2013). This dataset is collected with 30 volunteers wearing a smartphone at the waist whilst undertaking six physical activities. The embedded accelerometer and gyroscope resulted in 9-dimensional readings recorded at a constant rate of 50 Hz. This dataset provides standard train/test splits with 70% of the volunteers used for training and the remaining for the test split.

USC-HAD Dataset Zhang and Sawchuk (2012b). The dataset is aimed at recognition of fine-grained daily activities in health-care scenarios and consists of 12 activities collected from 14 subjects over 6 sensor channels. Replicating the protocol from Haresamudram, Anderson and Plötz (2019), data from the first ten participants constitute the training split, the following two participants form the validation set, and the remaining is used as the test split.

Following Haresamudram, Anderson and Plötz (2019), sensory data are down-sampled to 33 Hz, and per-channel normalisation is adopted using the training data statistics to scale the values into the range $[-1, 1]$. Subsequently, the data-streams are partitioned into segments using a sliding window of 30 samples (*i.e.*, W=30) with 50% overlap between adjacent windows.

6.4.2 Unsupervised Activity Representation Learning Baselines

We briefly introduce the alternative approaches that serve as concrete baselines for the task of unsupervised activity representation learning:

Recurrent Autoencoder (RAE) Haresamudram, Anderson and Plötz (2019). The framework comprises of a deterministic encoder $\mathcal{E}$ and a decoder $\mathcal{G}$ trained directly to minimise ℓ_1 or ℓ_2 element-wise reconstruction error in the *data space*.

Motion2Vector (M2V) Bai et al. (2019). This framework includes a decoder $\mathcal{G}$, however with a stochastic encoder $\mathcal{E}$ parameterising an Isotropic Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The framework is trained with reconstruction error in the *data space* as well as KL-divergence additionally optimised for the *latent space*; KL-divergence is incorporate to match the encoder output distribution with the standard Gaussian prior.

Recurrent Generative Adversarial Network (RGAN) Esteban, Hyland and Rätsch (2017). The framework trains a generator $\mathcal{G}$ and a discriminator $\mathcal{D}$ by optimising the standard adversarial loss in the *data space* according to Eq. (6.1).

Recurrent Flipped Adversarial Autoencoder (RFAAE). We adapt the approach proposed in Zhang et al. (2018b) to a recurrent framework. In addition to the generator $\mathcal{G}$ and discriminator $\mathcal{D}$ of a standard GAN, this baseline jointly trains an additional encoder $\mathcal{E}$ to regress the latents; according to Eq. (6.2), the adversarial loss is optimised in the *data space*, and the ℓ_2 reconstruction error is minimised in the *latent space* between the encoder outputs and the sampled priors.

6.4.3 Experimental Setup

We follow the standard evaluation protocol in Zhang, Isola and Efros (2016) to assess the quality of unsupervised learned sequence representations of different baselines. First, we train all network modules using the unlabelled train-split sequences. Subsequently, we freeze the *feature extractor* parameters and leverage the training labels to train a single linear classifier on the learned representations. Except for the RGAN baseline Esteban, Hyland and Rätsch (2017), the *encoder* network $\mathcal{E}$ within the frameworks serves as the *feature extractor*. For the RGAN baseline lacking an encoder module, the penultimate representations from the discriminator network $\mathcal{D}$ are used as the unsupervised features. The trained classifier is then evaluated on the held-out test split sequences and we report the achieved classification performance.

Table 6.1: Unsupervised representation learning comparison. We present a comparison of unsupervised learned representations when transferred for use on downstream classification tasks. The classification performance is judged based on the attained accuracy and class-averaged F-scores (values in parenthesis) on the holdout test sequences. All of the methods employ recurrent neural networks, except for Zhang et al. (2018b), where we employ a recurrent adaptation of the proposed method.

Recurrent Representation Learning Model	Sequential MNIST	UCI HAR	USC-HAD
RAND	51.4 (48.8)	51.6 (44.3)	34.3 (21.2)
M2V Bai et al. (2019)	82.8 (82.6)	70.9 (69.8)	54.1 (44.0)
RGAN Esteban, Hyland and Rätsch (2017)	80.2 (79.8)	76.3 (75.6)	50.7 (42.8)
RFAAE (our recurrent adaption of Zhang et al. (2018b))*	94.0 (93.9)	88.5 (88.4)	64.4 (57.2)
RAE-ℓ_1 Haresamudram, Anderson and Plötz (2019)	95.5 (95.5)	87.1 (87.1)	63.8 (54.8)
RAE-ℓ_2 Haresamudram, Anderson and Plötz (2019)	95.7 (95.7)	87.2 (87.2)	65.2 (56.0)
(Ours) Guided-GAN	**97.3 (97.3)**	**89.0 (88.9)**	**67.2 (59.9)**
SUP	99.1 (99.1)	91.1 (91.0)	68.6 (62.6)

6.4.4 Implementation Details

We implement our experiments using PyTorch Paszke et al. (2017) deep learning framework and *employ the following parameters described for all the methods*. All network parameters are trained end-to-end for 500 epochs by back-propagating the gradients of the corresponding loss functions averaged over mini-batches of size 64 and in accordance with the Adam Kingma and Ba (2015) update rule. The learning rate for Adam is set to 10^{-3} and the beta values $\beta_1 = 0.5$, $\beta_2 = 0.999$ are used. The fixed prior distribution p_z for deep generative models is set to be a 100-dimensional isotropic Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$. To ensure a fair comparison, the encoder $\mathcal{E}$, generator (interchangeably decoder in autoencoder frameworks) $\mathcal{G}$, and discriminator $\mathcal{D}$ constitute a single-layer uni-directional LSTM with 100 hidden neurons to process the input sequences. For our Guided-GAN, the loss weighting coefficients $\lambda_z = 1$ and $\lambda_x = 0.01$ are kept constant across the sensory datasets.

6.4.5 Results

Downstream Sequence Classification. As elaborated in Section 6.4.3, we transfer the unsupervised learned representations and investigate their effectiveness for downstream supervised classification tasks. Notably, the feature extractor parameters are frozen and only the parameters corresponding to a single classifier are trained using supervision from labelled data.

In Table 6.1, we summarise the downstream sequence classification performance by reporting the classification accuracy together with the class-averaged F-score (F_m) for each baseline. Given the potential imbalanced class distributions in the datasets, the latter metric reflects the ability of the HAR model to recognise every activity category regardless of its prevalence in the collected data. *In addition to the discussed unsupervised baselines, we further present results from a fully supervised trained feature extractor (**SUP**) as well as a frozen randomly initialised feature extractor (**RAND**) on each dataset for the readers' reference.*

Across the three datasets, the large performance gap between RGAN (80.2%, 76.3% and 50.7% respectively) against RFAAE (94%, 88.5% and 64.4%) highlights the significance of incorporating an inference network for effective representation learning in generative adversarial frameworks. While the discriminator's delegated task of distinguishing between real and generated sequences benefits the penultimate representations (see the superiority of RGAN over RAND baseline), learning an inverse mapping to the latent feature space through encoder results in significantly more effective features. However, the encoder in RFAAE is only trained on synthetic sequences and never encounters real data samples. Accordingly, its performance as a feature extractor is heavily reliant on the quality and diversity of the generator's sequences. In contrast, the encoder in our framework is exposed to both real data sequences as well as generated ones, which evidently offers feature representations of higher quality, achieving 3.51%, 0.56% and 4.35% relative improvements respectively on Sequential `MNIST`, `UCI HAR` and `USC-HAD` datasets.

Comparing the lower performance levels of Motion2Vector (M2V) against its non-variational counterparts (RAE-ℓ_1 and RAE-ℓ_2), we observe that its ability to sample new sequences comes at the cost of harming the feature representation qualities. However, the proposed Guided-GAN framework bridges this shortcoming by allowing the generation of realistic synthetic data while simultaneously achieving higher quality representations. In fact, the proposed approach not only outperforms

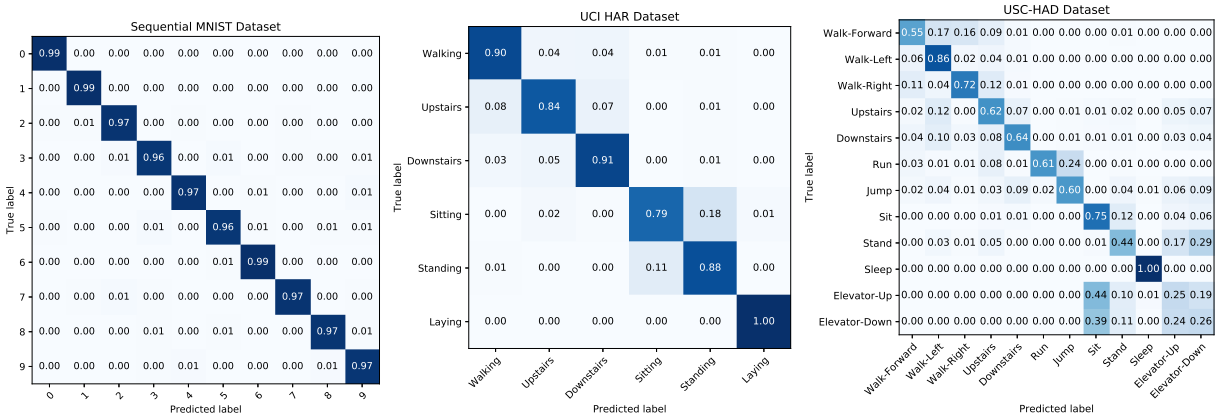


Figure 6.5: Class-specific recognition performance. We present the confusion matrices highlighting the class-specific recognition performance for the testing splits of `Sequential MNIST`, `UCI HAR`, and `USC-HAD` benchmarks. The vertical axis represents the ground-truth labels and the horizontal axis denotes the predicted labels.

Table 6.2: Classification network parameter specifications. We present a detailed description of trainable and frozen parameters for downstream classification evaluation on the three benchmarks.

Parameter Count	Sequential MNIST	UCI HAR	USC-HAD
Total	63110	55106	54512
Frozen	62100	54500	53300
Trainable	1010	606	1212
(Trainable Ratio)	(**1.6%**)	(**1.1%**)	(**2.2%**)

existing unsupervised baselines with a large margin but it also closely approaches the fully supervised baseline–*SUP*–performance.

Notably, we summarise the number of parameters corresponding to the feature extractor (*frozen*), classifier (*trainable*) and the ratio of trainable parameters to the total number of network parameters (*trainable ratio*) in Table 6.2. *We can observe the proposed Guided-GAN to achieve comparable classification performance to that of the fully supervised baseline (SUP) whilst having access to only* ***1.6%****,* ***1.1%****, and* ***2.2%*** *of the parameters for training on the three datasets, respectively.*

For reference, we summarise the class-specific recognition results from the Guided-GAN's unsupervised features by presenting confusion matrices for the downstream classification tasks in Fig. 6.5.

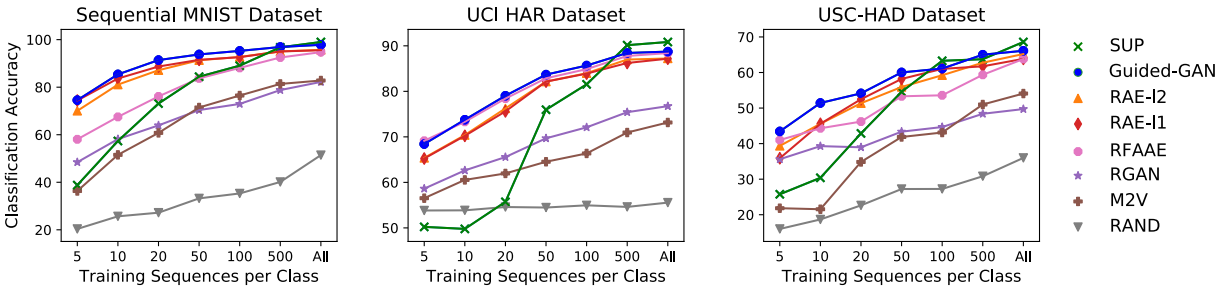
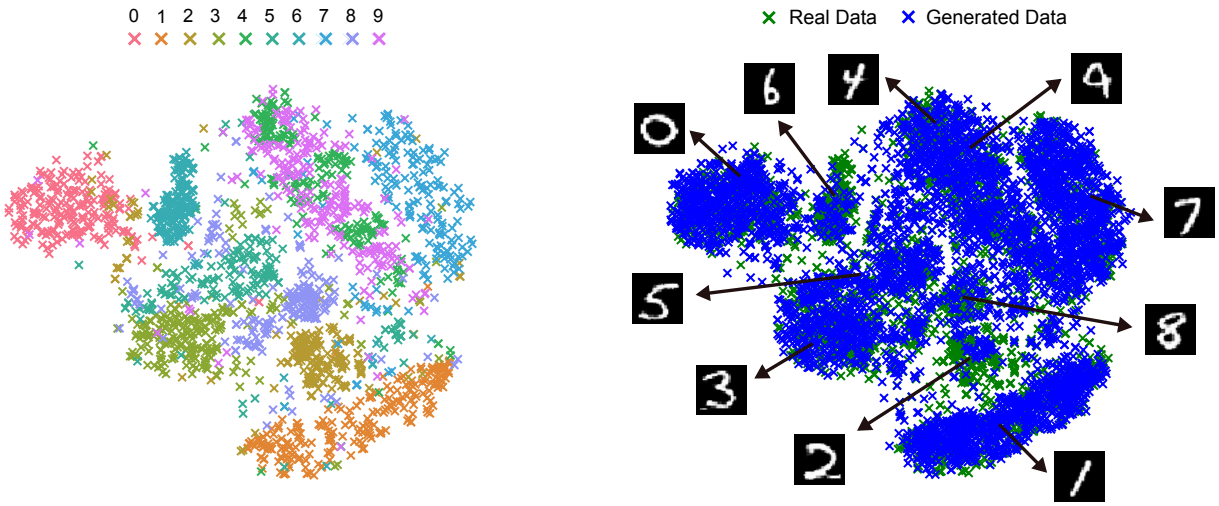


Figure 6.6: Effect of labelled training data size. We present a comparison of classification performance for learned sequence representations with varying sizes of annotated training data. Results are averaged over five different sets of training data and reported over the entire held-out test splits.

Varying Labelled Dataset Sizes. To gain further insights into the generalisation capability of feature representations learned by the investigated unsupervised approaches, we analyse the classification performance on the entire testing splits for the three datasets while varying the amount of available labelled data for supervised classifier training in Fig. 6.6. The reported results are averaged over five different sets of training data.

We observe the unsupervised baselines to provide an effective means to learn useful feature representations by exploiting unlabelled data in the absence of large amounts of annotated training data, resulting in substantial performance gains over the *RAND* and *SUP* baselines; supervised classifier training on top of a randomly initialised feature extractor (*RAND* baseline) fails to learn clear classification boundaries to discriminate different classes regardless of the amount of available labelled data and the fully supervised baseline struggles to generalise to the unseen test sequences when trained on low volumes of annotated samples. Moreover, our approach consistently offers better generalisation to unseen data compared with existing unsupervised remedies in the presence of extremely limited labelled training data. Further, Guided-GAN achieves higher classification performance when leveraging all labelled training data that it is comparable with the supervised baselines trained end-to-end with *full* supervision.

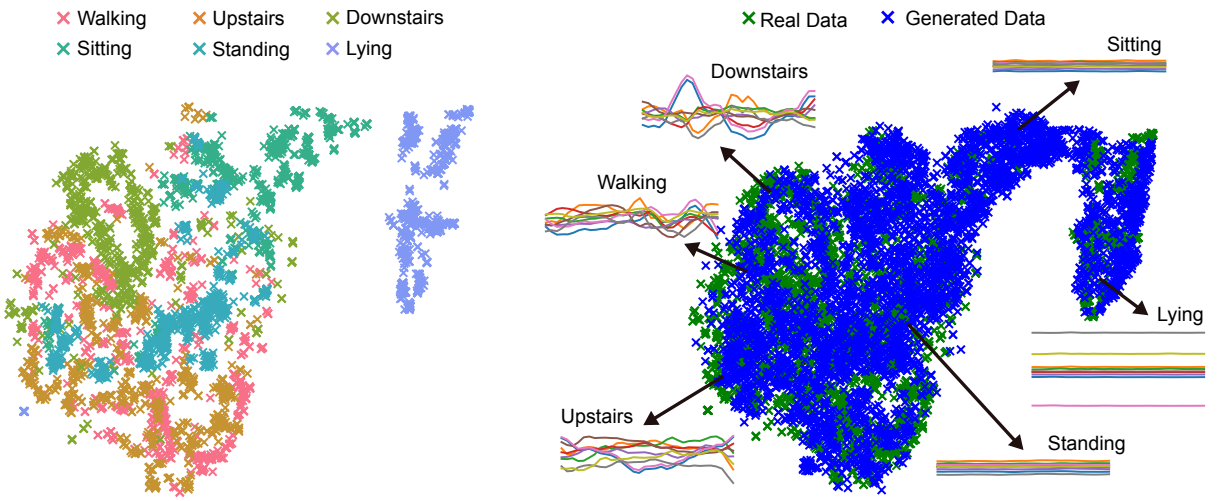
Assessing Multi-modal Sequence Generation. While *unsupervised feature learning* constitutes the main focus of our study, we qualitatively demonstrate the ability of the unconditional generator trained through Guided-GAN is able to generate diverse and realistic sequences. To this end, we visualise the data spaces for both the *real* and



(a) Real Data

(b) Generated Data overlaid on Real Data

Figure 6.7: Sequential MNIST data generation. For `Sequential MNIST` dataset, we present t-SNE visualisation of: (a) real data sequences colour-coded with semantic labels, and (b) generated sequences overlaid on real data samples. Evidently, our Guided-GAN's generator successfully captures semantic variations in the data and aligns with the real data distribution.



(a) Real Data

(b) Generated Data overlaid on Real Data

Figure 6.8: UCI HAR data generation. For `UCI HAR` benchmark, we present t-SNE visualisation of: (a) real data sequences colour-coded with semantic activity labels, and (b) generated sequences overlaid on real data samples. Our Guided-GAN's generator is observed to successfully capture semantic variations embedded in the multi-modal motion sequences.

generated datasets in 2D using t-SNE Maaten and Hinton (2008) in Fig. 6.7 and Fig. 6.8 for `Sequential MNIST` and `UCI HAR` respectively.

We can observe that the *generated* data distribution closely follows the *real* data distribution, as indicated by the dense overlap between their corresponding sequence samples. In addition, we observe a smooth interpolation in the space between different categories for the generated sequences; *e.g.*, the generated samples interconnecting the `Lying` and `Sitting` activity categories in Fig. 6.8-b. Further, we depict a set of generated sequences where visual inspection of generated data for `Sequential MNIST` clearly demonstrates a conformance to the class label semantics.

Investigating the Faithfulness of Reconstructions. It has been reported in Dumoulin et al. (2017); Zhang et al. (2018b) that the reconstructions of data with Bidirectional GAN (BiGAN) Donahue, Krähenbühl and Darrell (2017) and Adversarially Learned Inference (ALI) Dumoulin et al. (2017) are not always faithful reproductions of the inputs; *in extreme cases deviating entirely from the semantic labels.*

We conduct a set of experiments to quantitatively measure the veracity of the sequential multi-modal data reconstructions by our Guided-GAN. To this end, we rigorously explore the downstream classification tasks on `Sequential MNIST` and `UCI HAR` while considering different *development datasets* (denoting the dataset used to train the supervised classifier) and *evaluation datasets* (denoting the dataset used for evaluation):

- **Train**—The standard training split data and labels.
- **Test**—The standard testing split data and labels.
- **Reconstructed Train**—The reconstructed training data attained by applying $\mathbf{x}_{\text{rec}} = \mathcal{G}_\phi(\mathcal{E}_\theta(\mathbf{x}))$ for every sequence in the original **Train** split whilst retaining the original labels.
- **Reconstructed Test**—The reconstructed test data attained by applying $\mathbf{x}_{\text{rec}} = \mathcal{G}_\phi(\mathcal{E}_\theta(\mathbf{x}))$ for every sequence in the original **Test** split whilst maintaining the original labels.

We summarise the corresponding classification performances in Table 6.3. From the results, across both datasets, we observe that Guided-GAN demonstrates

Table 6.3: Quantitative assessment of reconstruction faithfulness. We quantitatively assess the faithfulness of data reconstructions through rigorous evaluation on downstream classification tasks. We report classification accuracy together with class-averaged F-score (value in parenthesis) on the holdout evaluation datasets.

Development Dataset	Evaluation Dataset	Sequential MNIST	UCI HAR
Train	Test	97.3 (97.3)	89.0 (88.9)
Reconstructed Train	Test	95.1 (95.0)	84.2 (84.0)
Train	Reconstructed Test	93.9 (93.8)	83.4 (83.1)
Reconstructed Train	Reconstructed Test	94.6 (94.6)	84.9 (84.7)

reconstructions of reasonable faithfulness to their original semantic categories; *i.e.*, substituting the original data splits—e.g. **Train**—with their corresponding reconstructions—**Reconstructed Train**—still allows learning a classifier with comparable performance to the original data splits. For reference, we further include qualitative samples of data reconstructions in Fig. 6.9.

Training Comparison between Guided-GAN and Recurrent BiGAN. As emphasized, our attempts to train recurrent BiGANs without the proposed manifold distance minimization terms were unsuccessful. For reference, we present empirical

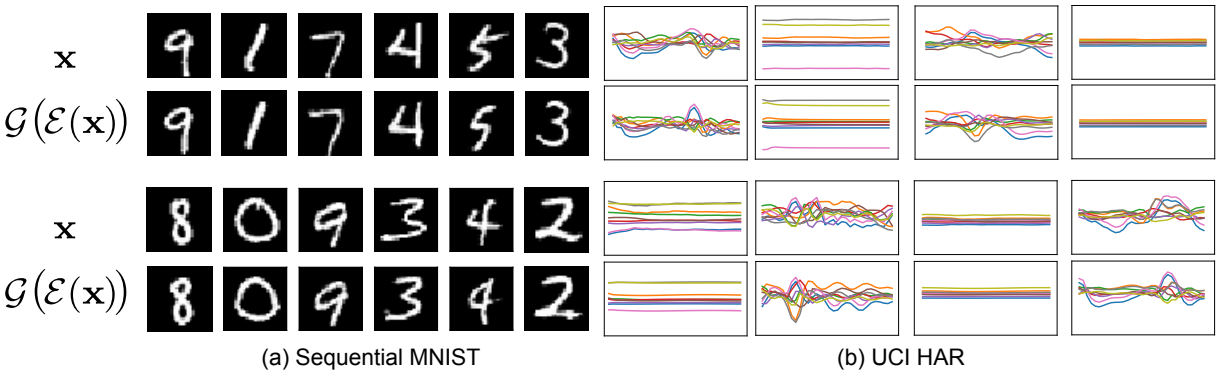


Figure 6.9: Qualitative assessment of reconstruction faithfulness. We present qualitative results for data reconstructions with Guided-GAN for (a) `Sequential MNIST` and (b) `UCI HAR` datasets, where the first odd rows represents the original data $\mathbf{x}$ and the even rows are the corresponding reconstructions $\mathbf{x}_{\text{rec}} = \mathcal{G}_\phi(\mathcal{E}_\theta(\mathbf{x}))$. Interestingly, we observe reconstructions for UCI HAR sequences reflecting different phases and variations of activity sequences whilst preserving the semantics of the reconstructed sequences.

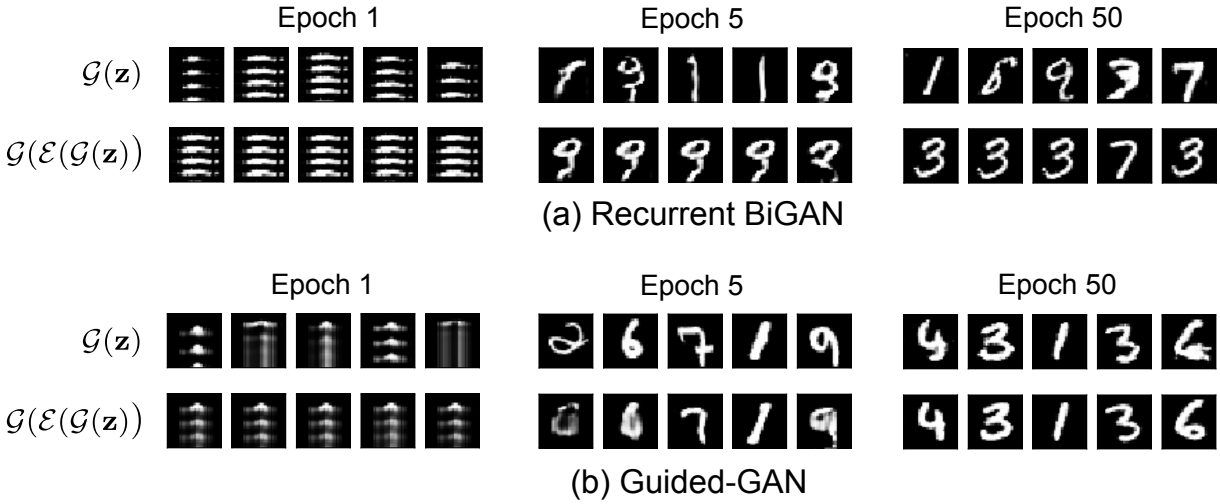


Figure 6.10: Convergence comparison. We present generated digit sequences together with their corresponding reconstructions at different stages of training for (*a*) *recurrent* BiGAN, and (*b*) the proposed Guided-GAN. As illustrated, we observed through multiple experiments, the inability of the recurrent BiGAN to uncover the generator's inverse mapping.

results obtained from training recurrent BiGAN as well as our Guided-GAN on `Sequential MNIST` for ease of visual inspections.

To observe the behaviour of generators and encoders, we visualise randomly generated samples $\hat{\mathbf{x}} = \mathcal{G}_\phi(\mathbf{z})$ at different stages of the training process together with their corresponding reconstructions $\mathbf{x}_{\text{rec}} = \mathcal{G}_\phi(\mathcal{E}_\theta(\hat{\mathbf{x}}))$ for both *recurrent* BiGAN and the proposed Guided-GAN in Fig. 6.10. In particular, we observed that the sole discriminator in BiGAN was not able to guide the recurrent encoder towards uncovering the generator's inverse mapping function. Thus, no useful representations were obtained and accordingly, extremely low downstream classification performance was achieved. In contrast, the proposed Guided-GAN successfully inverts the generator and encoder at the very early stages of training process.

6.5 Conclusion

This chapter examined the problem of extracting unsupervised feature representations from unlabelled activity data using the generative adversarial network's latent feature space. To this end, *Guided-GAN* was proposed in our efforts to bridge the gap between unsupervised representation learning and generative adversarial networks for sequential multi-modal data. In recognition of the inherent temporal dependencies within the sequential data, a recurrent generator, encoder and discriminator

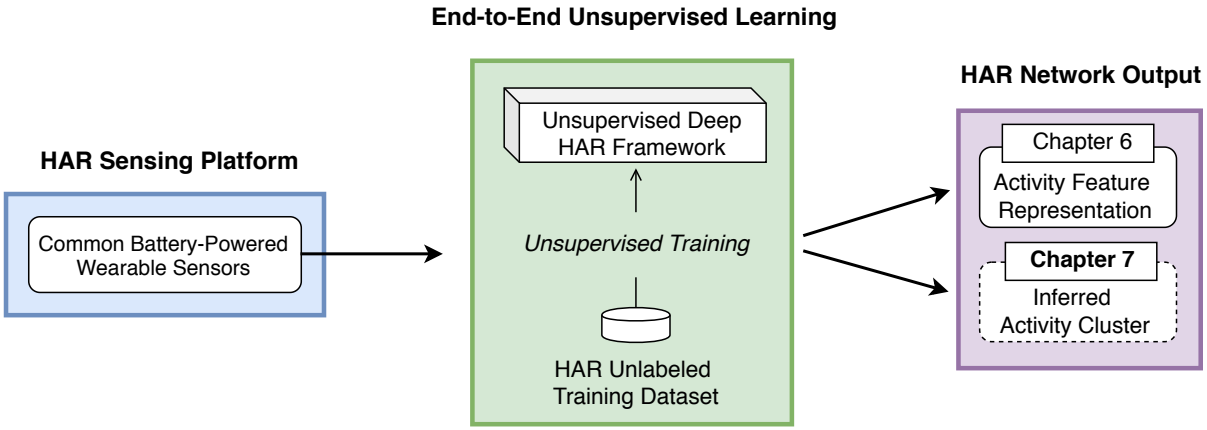


Figure 6.11: Upcoming chapter sneak peek.

architectures were designed cooperating in a bidirectional GAN framework. Further, the key insight in training BiGAN frameworks for sequential data was shared that the discriminator adversarial feedback alone may be insufficient to uncover the generator's inverse mapping. Hence, Guided-GAN leveraged an intuitive formulation to alleviate the burden on the discriminator in achieving inverting generators and encoders by seeking additional training guidance from geometric distance penalisation in data and latent manifolds. *When evaluated on three downstream sequence classification benchmarks, our learned sequence representations outperformed existing unsupervised approaches whilst closely approaching the performance of fully supervised learned features*. Therefore, it was substantiated that the proposed Guided-GAN can effectively leverage unlabelled data to extract discriminative features in an unsupervised fashion.

In recognition of the importance of unsupervised learning regimes in the absence of large-scaled annotated HAR datasets, next chapter studies the challenging problem of unsupervised human activity clustering using end-to-end deep learning paradigms. For the first time, we will develop a stand-alone unsupervised HAR framework that consumes raw multi-channel time-series data and generates cluster assignments with high correspondence to human activity semantics. We depict the examined unsupervised HAR problem in this chapter in comparison against the investigated clustering problem in Chapter 7 in Fig. 6.11.

Chapter 7

Deep Clustering of Human Activity Data-streams from Wearables

THE costly work of gathering and annotating sensory activity datasets is tedious, labour intensive, time consuming and not scalable to large volumes of data. In our efforts to develop unsupervised strategies to exploit unlabelled wearable data, Chapter 6 investigated the problem of unsupervised representation learning in order to learn discriminative activity features from unlabelled data that benefit subsequent downstream classification tasks. In this chapter, we focus on the fundamental problem of human activity clustering from wearables and develop end-to-end strategies using deep learning paradigms. Through exhaustive experiments, the effectiveness of the proposed approach is demonstrated to jointly learn *unsupervised representations* for sensory data and generate *cluster assignments* with strong semantic correspondence to distinct human activities.

7.1 Motivation and Contribution

As discussed in Chapter 3 through Chapter 5, human activity recognition problems have predominantly relied on supervised learning regimes where deep learning paradigms are extremely successful in learning activity representations from annotated data. While the process of collection and annotation may be retrospective with vision based sensing modalities where visual inspections of, for example, video frames provides the basis for ground truth, the parallel task with wearables is nearly impossible. Moreover, such methods cannot be easily scaled to gather large datasets often necessary for deep neural networks (DNNs). In comparison to other domains, generating labelled data to benefit from supervised learning methods to build HAR applications in the absence of a reliable visualisation to establish ground truth is a unique HAR problem with wearable sensors.

Although unsupervised methods provide avenues for learning from unlabelled data, investigations of unsupervised learning from multi-sensor time-series datastreams from wearables remains remains limited to *pre-training* Alsheikh et al. (2016); Abedin et al. (2018) as in Chapter 5 or *unsupervised representation learning* Haresamudram, Anderson and Plötz (2019); Bai et al. (2019) as in Chapter 6. Unsupervised alternatives such as deep clustering exist for classification of image data without requiring any labels, however, these frameworks are tailored for static images and lack the *inherent* capability to learn representations and cluster sequential time-series data recorded by wearables. Therefore, our motivation for this chapter is to investigate and develop a deep clustering architecture for unsupervised HAR for multi-sensor time series data sequences from wearables. In particular, clustering methods operate in the absence of supervisory feedback from annotated data. Therefore, data representation quality has a significant impact on clustering performance or the ability to learn semantically meaningful groupings of input sequences. Our goal in this chapter is to develop a deep clustering architecture that:

- Leverages the inherently sequential nature of sensory data.
- Learns *clustering friendly* representations of activity features in the multi-sensor and multi-channel input signals that offer separation of activity classes in the feature space.
- Promotes the formation of highly discriminative clusters with high semantic correspondence to human activities.

This chapter proposes *Deep Sensory Clustering*—a deep clustering architecture that learns highly discriminative representations using self-supervision with reconstruction and future prediction tasks informed by feedback from a clustering objective to guide the network towards clustering-friendly representations. The self-supervised tasks intend to incentivise the network to learn salient activity features that offer semantic separation in the feature space while simultaneously reducing the risk of collapsing clusters. Further, the optimisation objective is augmented with a clustering-oriented criterion to further refine the feature representations and gradually promote clustering-friendliness in the feature space. The framework design concepts are validated through extensive experiments; The key contributions are summarised below:

1. This chapter develops an *unsupervised* deep learning network architecture for clustering human activities from sequences of wearable sensor data streams. The proposed approach, to the best of knowledge, provides the *first* standalone, end-to-end, deep clustering method for wearables.

2. The proposed novel architecture is built upon insights gained from exploring the effectiveness of multi-task autoencoding objectives augmented with a clustering-oriented criteria to learn semantically meaningful representations from sensory data in an unsupervised fashion. We demonstrate that the *augmentation* of the clustering-oriented criteria—*previously unexplored*—significantly improves the clustering quality as well as learning representations.

3. The effectiveness and generalisability of the proposed approach is demonstrated through a systematic experimental regime conducted on three popular and diverse HAR benchmark datasets (`UCI HAR`, `Skoda`, `MHEALTH`).

4. Insights gained from the study in this chapter are shared through both quantitative and qualitative results from: *i)* an investigation into the effectiveness of the learned representations from the sensor data; and *ii)* an ablation study.

7.2 Related Works

Here, we provide: *i*) a brief overview of the literature on human activity recognition using wearable sensors; and *ii*) review existing clustering approaches that leverage

the high representational power of deep neural networks for the unsupervised task of clustering. In the sequel, we describe the motivation of our work by highlighting the limitations of existing studies.

7.2.1 Human Activity Recognition with Wearable Sensors

Recently, the superior performance of supervised deep neural networks in classification tasks has led to a shift towards the adoption of deep learning paradigms for recognising human activities with wearables Wang et al. (2019); Nguyen et al. (2017); Khan, Roy and Misra (2018); Hossain, Al Haiz Khan and Roy (2018); Abedin et al. (2019). In this line, researchers have explored CNNs Zeng et al. (2014); Yang et al. (2015); Ronao and Cho (2015); Jiang and Yin (2015), RNNs Hammerla, Halloran and Plötz (2016); Guan and Plötz (2017), and a combination of convolutional and recurrent layers Ordóñez and Roggen (2016); Murahari and Plötz (2018) to effectively model the temporal dependencies inherent in sequences captured with sensors whilst mutually enhancing activity classification performance.

Despite the tremendous progress achieved in this emerging area of research, existing deep HAR systems owe their success to large amounts of annotated training data and are only applicable to the supervised learning regime. However, acquisition of labelled sensory data is labour intensive, time-consuming and clearly not scalable to large datasets. On the other hand, unsupervised learning with deep neural networks for HAR has merely been investigated as a means for weight initialisation or unsupervised feature learning prior to supervised fine tuning with labels Alsheikh et al. (2016); Abedin et al. (2018); Haresamudram, Anderson and Plötz (2019); Bai et al. (2019), rather than a standalone end-to-end approach for exploiting cheaply accessible unlabelled data. In particular, Alsheikh et al. (2016) adopts layer-wise pre-training of a fully connected deep belief network using pre-processed spectrograms and Abedin et al. (2018) pre-trains a deep convolutional autoencoder on raw multi-modal sensor data. Both approaches then rely on subsequent optimisation of network weights using labelled data. In contrast to these studies, the focus of this chapter is on developing *end-to-end clustering* solutions that can take as input *raw sensory data* captured by wearables and directly output activity categorisation without relying on any supervision from annotated data; a yet unexplored objective in HAR with wearable sensors.

7.2.2 Clustering with Deep Neural Networks

Clustering is a central concept in unsupervised machine learning and serves a wide range of data-driven application domains. The main objective of clustering is to categorise data into groups of similar samples. Previously, data transformation methods Wold, Esbensen and Geladi (1987); Hofmann, Schölkopf and Smola (2008) and traditional clustering algorithms Arthur and Vassilvitskii (2007); Reynolds (2015) were applied sequentially to project raw data into a feature space where separation would be easier. However, the clustering performance of these traditional approaches were challenged by the complex structure of high-dimensional data.

Recently, the high representational power of deep neural networks has been leveraged in order to achieve clustering-friendly representations and cluster assignments simultaneously in an end-to-end manner; hence the rise of *Deep Clustering* paradigms. In this regard, Xie, Girshick and Farhadi (2016); Guo et al. (2017a,b); Li, Qiao and Zhang (2018) adopt reconstruction task to initialise a feature space for representing images using deep autoencoders. Subsequently, a novel cluster assignment hardening loss is incorporated to iteratively refine the assignment of the representations to clusters. Similar ideas have been investigated in Chen, Lv and Zhang (2017); Ghasedi Dizaji et al. (2017), where reconstruction and clustering-oriented objectives are jointly incorporated to learn cluster-specific representations with deep neural networks. The former defines a locality preserving criteria to learn structure preserving image representations and further penalises embedded features based on their proximity to the cluster centres. The latter minimises a relative cross-entropy loss coupled with a regularisation term to encourage balanced cluster assignments. In Yang et al. (2017), k-means loss and reconstruction objective are jointly incorporated using an alternative optimisation algorithm. Taking a different approach, Hu et al. (2017) proposes an unsupervised discrete representation learning algorithm for deep clustering using a fully connected architecture. This method maximises the mutual information between the input images and output cluster assignments. Avoiding discrete configurations of optimisation objective, Shah and Koltun (2018) formulates clustering as a global continuous objective. Authors in Yang, Parikh and Batra (2016) leverage a convolutional neural network for image feature representation and uses agglomerative clustering objective in a recurrent process. For a comprehensive overview of deep clustering studies, we refer interested readers to Min et al. (2018); Aljalbout et al. (2018).

Summary. Though these methods achieve excellent performances for computer vision applications, existing deep clustering frameworks are tailored for static image datasets and are not directly applicable to the unlabelled sequences captured by wearables; the adopted architectures merely incorporate fully connected layers and convolutional operations which as substantiated by our experiments in Section 7.4.3, inherently suffer from the inability to learn representations for time-series data recorded by wearables and generate clusters of activities. By contrast, this chapter rigorously seeks to develop a recurrent architecture design as well as unsupervised training objectives that can elegantly handle the sequential nature of wearable sensor data, effectively categorising them into semantically meaningful clusters of human actions.

7.3 Proposed Methodology

Consider the problem of clustering a set of n unlabelled segments of sensory readings $\{\mathbf{x}_i\}_{i=1}^{n}$ captured from wearable sensors into k clusters, each representing a semantic human activity category. These sensory partitions are obtained by applying a sliding window of fixed temporal duration δt over D sensor channels of recorded datastreams. Taking into account the absence of supervisory signals from annotated data and the crucial impact of data representation quality on the clustering performance, our goal is to develop an unsupervised framework that *i*) accounts for the inherent sequential nature of sensory data to learn intrinsic representations; and *ii*) induces clustering-friendly spaces with high semantic correspondence to human activities.

To satisfy these requirements, this chapter proposes an unsupervised two-staged *Deep Sensory Clustering* framework, illustrated in Figure. 7.1. In the first stage, we pre-train a multi-task autoencoder with a recurrent structure to jointly perform reconstruction and future anticipation for the input sensory measurements. These self-supervised tasks are intended to incentivise the network to learn salient activity features that offer semantic separation in the feature space while simultaneously reducing the risk of collapsing clusters. Once the feature space is initialised, in the second stage we augment the optimisation objective with a clustering-oriented criteria to further refine the feature representations and gradually promote clustering-friendliness in the space. As substantiated by our experiments, this step significantly improves the achieved clustering performance as well as the quality of the learned feature representations. We elaborate on the workflow of our framework in what follows.

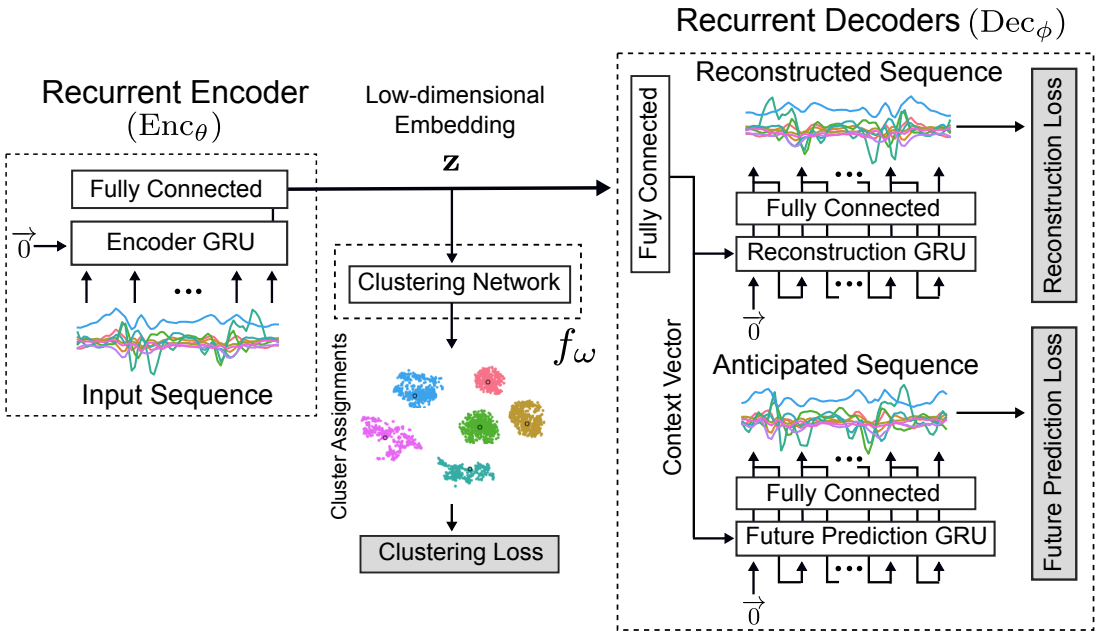


Figure 7.1: Deep Sensory Clustering framework. We illustrate an overview of the proposed pipeline for Deep Sensory Clustering.

7.3.1 Stage (I): Pre-training with Multi-Task Autoencoder

In order to facilitate learning clustering-friendly representations from multi-channel sensory data, we initialise the feature space by pre-training a recurrent autoencoder to accomplish auxiliary tasks in an unsupervised fashion. In order to accomplish the delegated tasks, the network is forced to extract enriched representations from the sensory sequences. As a result, the network learns an initial non-linear mapping from the data space into a feature space with semantically more relevant representations where clustering criterion can be imposed and optimised for additional improvements.

Recurrent Encoder (Enc_θ)

The encoder component of our network consumes a windowed excerpt of raw multi-channel sensory sequence and learns a compact fixed length representation as a holistic summary of the input. In particular, we adopt a bi-directional GRU that reads through the partitioned sensory sequence $\mathbf{x}$ in both forward and backward directions and updates its internal hidden state in each time step according to the received input. The final hidden state obtained after scanning the entire input sequence is reduced in dimensionality through a fully connected layer. The resulting low-dimensional embedded feature, denoted by $\mathbf{z} \in \mathbb{R}^z$, encodes contextual activity information

by modelling the temporal dependencies present in the input sequence of sensory measurements $\mathbf{x}$. Hence, we summarise the parameterised operations associated with encoding the input sequence $\mathbf{x}_i$ as

$$\mathbf{z}_i = \text{Enc}_\theta(\mathbf{x}_i). \tag{7.1}$$

Notably, we impose an under-complete representation learning constraint by restricting the embedding feature dimension to be smaller than that of the input space. This motivates the network to automatically extract the most salient activity features from raw sensory data-streams.

Conditional Recurrent Decoders (Dec_ϕ)

The decoder modules of the framework are structured symmetrical to the encoder component. Firstly, a context vector is achieved by back projecting the embedded representation from the encoder into a higher-dimensional space such that it can be used to initialise the hidden states for the decoders. Two recurrent decoders then jointly exploit the generated context vector to accomplish different self-supervised tasks without requiring any manual supervision. Inspired by Srivastava, Mansimov and Salakhudinov (2015), in this chapter we share the encoder network between decoders with two different expertise; one decoder is specialised to reconstruct the temporally inversed input sequence, while the other one learns to anticipate the future sensory measurements that should follow after, conditioned on the encoded input representation. Hence, not only the network has to learn a representation enriched with sufficient information to reproduce the input sequence, but also encode features that allow extrapolating future measurements.

Initially, a zero input vector is fed to the decoders as a signal for commencing the decoding process. For subsequent time steps, the linearly projected output from the previous time step is fed to the decoders as input. In order to control the information flow to the decoders, we avoid adopting the teacher forcing practice; *i.e.*, we chose not to provide the expected ground-truth from previous time steps as the input to the decoders and instead use the actual decoder outputs. While this strategy substantially adds difficulty to the decoding tasks, it helps uncover stronger representations by

encouraging the model to look deep into the encoder for the required information. We summarise the parameterised decoding process as

$$(\hat{\mathbf{y}}_i^{\text{rec}}, \hat{\mathbf{y}}_i^{\text{fut}}) = \text{Dec}_{\phi}(\mathbf{z}_i), \tag{7.2}$$

where $\hat{\mathbf{y}}_i^{\text{rec}}$ and $\hat{\mathbf{y}}_i^{\text{fut}}$ respectively denote the reconstructed and the anticipated sequences generated from the input $\mathbf{x}_i$ to satisfy the tasks.

Pre-training Optimisation

We pre-train the entire recurrent autoencoder with a joint objective function,

$$\mathcal{L}_{\text{AE}}^{(i)} = \mathcal{L}_{\text{rec}}^{(i)} + \mathcal{L}_{\text{fut}}^{(i)} = \underbrace{\|\mathbf{y}_i^{\text{rec}} - \hat{\mathbf{y}}_i^{\text{rec}}\|^2}_{\text{reconstruction loss}} + \underbrace{\|\mathbf{y}_i^{\text{fut}} - \hat{\mathbf{y}}_i^{\text{fut}}\|^2}_{\text{future prediction loss}}, \tag{7.3}$$

where $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{fut}}$ denote the mean square error between each decoder's generated output sequence (*i.e.*, $\hat{\mathbf{y}}^{\text{rec}}$ and $\hat{\mathbf{y}}^{\text{fut}}$) and the expected ground-truth target sequences (*i.e.*, $\mathbf{y}^{\text{rec}}$ and $\mathbf{y}^{\text{fut}}$). Once the training is complete and the discrepancy between the generated outputs and their corresponding target sequences is minimised, the optimal network parameters, *i.e.*, $(\theta^*, \phi^*) = \min_{\theta,\phi} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_{\text{AE}}^{(i)}$, serve as an initialisation point for the second stage. We empirically show that jointly optimising for these auxiliary tasks results in a low-dimensional and semantically meaningful feature space where we can enforce and optimise clustering objectives to gain additional performance gains.

7.3.2 Stage (II): Representation Refinement with Clustering Criteria

Once the autoencoder becomes proficient in accomplishing the auxiliary tasks and hence the feature space finds a semantic orientation, we extend our framework with a parameterised clustering network $f_{\omega}(.)$ capable of estimating cluster assignment distributions and iteratively optimise a clustering objective $\mathcal{L}_{\text{C}}$ to refine the feature space and guide the network towards yielding clustering-friendly representations. In order to verify the insensitivity of the proposed approach to a specific clustering criterion, this chapter incorporates two representative and diverse state-of-the-art clustering-oriented objectives, namely *Cluster Assignment Hardening* Xie, Girshick and Farhadi (2016) and *Information Maximising Self-Augmented Training* Hu et al. (2017); the former is a centroid-based approach while the latter takes an information-theoretic

stance for clustering. Validated by our experiments, both criteria consistently result in significantly improved representations. Note that our primary goal is to present a proof-of-concept for a previously unexplored objective rather than an exhaustive search over all possible combinations of clustering objectives.

During the refinement stage, both the clustering loss $\mathcal{L}_{\mathrm{C}}$ and the autoencoding objectives $\mathcal{L}_{\mathrm{AE}}$ are jointly incorporated to be optimised. Hence, the aggregated optimisation criterion for instance i is formulated as

$$\mathcal{L}^{(i)} = \gamma \mathcal{L}_{\mathrm{C}}^{(i)} + \mathcal{L}_{\mathrm{AE}}^{(i)}, \tag{7.4}$$

where $\gamma \in [0, 1]$ is the coefficient that controls the balance between the two objectives. Note that we chose not to discard the decoding tasks during the refinement step in order to preserve the local data structure and allow a smoother manipulation of the feature space without distorting the previously established one. Once the network parameters are optimised with respect to the global criterion, *i.e.*, $(\theta^*, \phi^*, \omega^*) = \min_{\theta,\phi,\omega} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}^{(i)}$, the clustering network of our framework directly delivers cluster assignments without requiring a separate clustering algorithm to be run on the representations in a decoupled process. In what follows, we describe the clustering criteria utilised in this work.

Cluster Assignment Hardening (CAH)

This clustering objective leverages the similarities between the data representations and the cluster centroids as a kernel to compute soft cluster assignments. Putting emphasis on the high confidence assignments, it then purifies the clusters and forces the assignments to have stricter probabilities.

To incorporate this method, our clustering network f_ω comprises a single layer which maintains the cluster centroids $(\omega_j \in \mathbb{R}^{\mathrm{z}})_{j=1}^{k}$ as tunable network parameters and generates assignment distributions $Q_i = f_\omega(\mathbf{z}_i)$ for each instance i. This layer follows the Student's t-distribution to measure the similarity of embedded sequence representation $\mathbf{z}_i \in \mathbb{R}^{\mathrm{z}}$ to the k cluster centroids and therefore, obtains the normalised similarities $Q_i = (q_{ij})_{j=1}^{k}$,

$$q_{ij} = \frac{(1 + \|\mathbf{z}_i - \omega_j\|^2)^{-1}}{\sum_{j'=1}^{k} (1 + \|\mathbf{z}_i - \omega_j'\|^2)^{-1}}. \tag{7.5}$$

Through squaring this distribution and then normalising it, an auxiliary target distribution $P_i = (p_{ij})_{j=1}^{k}$ is then defined that leverages high confidence assignments in order to point the learning process towards stricter cluster assignments,

$$p_{ij} = \frac{q_{ij}^2 / \sum_{i=1}^{n} q_{ij}}{\sum_{j'=1}^{k} (q_{ij'}^2 / \sum_{i=1}^{n} q_{ij'})}. \tag{7.6}$$

Subsequently, the soft assignment distribution Q_i is iteratively purified through minimizing the Kullback-Leilbler (KL) divergence between the soft labels and the auxiliary target distribution via training the network parameters,

$$\mathcal{L}_{\mathrm{C}}^{(i)} = \mathrm{KL}(P_i || Q_i) = \sum_{j=1}^{k} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{7.7}$$

This centroid-based approach requires the cluster centres to be initialised *once* at the beginning of the refinement stage. The initial centres are obtained by applying classical clustering algorithms on the acquired representations from the optimal pre-trained parameters; *i.e.*, $\{\mathbf{z}_i = \mathrm{Enc}_{\theta^*}(\mathbf{x}_i)\}_{i=1}^{n}$.

Information Maximising Self-Augmented Training (IMSAT)

As an alternative, we explore incorporating IMSAT for further refinement of the established feature space. This method learns a probabilistic classifier to maximise the mutual information between the inputs and their corresponding cluster assignments in order to achieve statistical dependencies. In addition, the unsupervised training process is regularised via a self-augmentation loss that imposes local invariance on the learned data representations.

To integrate this approach, our parameterised clustering network f_ω constitutes a multi-layer perceptron with a final softmax activation that generates the discrete output distribution $Q_i = f_\omega(\mathbf{z}_i)$ for each instance i. Using these normalised cluster assignment probabilities $Q_i = (q_{ij})_{j=1}^{k}$, IMSAT minimises the following objective,

$$\mathcal{L}_{\mathrm{C}}^{(i)} = \lambda \mathcal{L}_{\mathrm{IM}}^{(i)} + \mathcal{L}_{\mathrm{SAT}}^{(i)}, \tag{7.8}$$

where $\mathcal{L}_{\mathrm{IM}}$ and $\mathcal{L}_{\mathrm{SAT}}$ respectively denote the information maximisation and the self-augmentation loss functions, and λ is the weighting constant for the objectives.

The mutual information loss is represented by the difference between the conditional entropy and the marginal entropy,

$$\mathcal{L}_{\text{IM}}^{(i)} = \underbrace{\sum_{j=1}^{k} q_{ij} \log q_{ij}}_{\text{conditional entropy}} - \underbrace{\sum_{j=1}^{k} \sum_{i'=1}^{n} q_{i'j} \log \left(\sum_{i''=1}^{n} q_{i''j} \right)}_{\text{marginal entropy}} \tag{7.9}$$

where minimising the conditional entropy promotes unambiguous cluster assignments while increasing the marginal entropy prevents large clusters to distort the feature space. As for the self-augmented loss, we adopt virtual adversarial training Miyato et al. (2018b) in order to penalise for dissimilarities between the discrete representation of the original data Q_i, and its adversarial perturbed version $P_i(r) = f_\omega\left(\text{Enc}_\theta(\mathbf{x}_i + r)\right)$,

$$\mathcal{L}_{\text{SAT}}^{(i)} = \text{KL}(Q_i \| P_i(r)) \tag{7.10}$$

where perturbation r is chosen to be an adversarial direction for input $\mathbf{x}_i$ and is efficiently approximated by a pair of forward and backward passes,

$$r = \underset{\|r'\|^2 \leq \epsilon}{\arg\max}\, \text{KL}(Q_i \| P_i(r')). \tag{7.11}$$

In Eq. 7.11, ϵ is a hyper-parameter controlling the range of the local perturbation.

7.4 Experiments and Results

We conduct extensive experiments: *i*) to demonstrate the effectiveness of the proposed *Deep Sensory Clustering* approach as an end-to-end deep clustering architecture; and *ii*) validate our network design thinking by presenting an ablation study.

7.4.1 Datasets

To ground our work, we evaluate the proposed framework on three sensor-based HAR benchmarks. The selected datasets exhibit great diversity in terms of the sensing modalities used and the activities to be recognised. We summarise the investigated datasets in Table 7.1 and provide a brief description in what follows.

Table 7.1: HAR datasets specifications. We summarise the datasets explored in this chapter.

Dataset	UCI HAR	Skoda	MHEALTH
Sensor Sampling Rate	50Hz	33Hz	50Hz
Sliding Window Duration (δt)	2.56s	1s	2.56s
Number of Sensor Channels (D)	9	60	23
Number of Activity Categories (k)	6	10	12
Number of Training Segments	7352	5448	4088
Number of Testing Segments	2947	718	1022

UCI HAR Dataset Anguita et al. (2013). This dataset targets the problem of recognising six activities of daily life using a smartphone worn at the waist level. The sensor measurements are collected at a frequency of 50Hz from the phone's embedded accelerometer and gyroscope, ultimately providing 9-dimensional recordings. This dataset provides randomly partitioned train and test splits, where 70% of the volunteers were used for generating the training split and the remaining users for the test split. In addition, a sliding window of 2.56 seconds with 50% overlap between adjacent segments is applied to partition the recorded data-streams.

Skoda Dataset Stiefmeier et al. (2008). This dataset is concerned with recognition of 10 manipulative gestures in a manufacturing scenario. Sensor data-streams are collected from assembly-line workers with body-worn triaxial accelerometers while performing manual quality checks of newly constructed cars. Following Guan and Plötz (2017), data-streams of 60 sensor channels collected from the subject's right arm are downsampled to 33Hz and a fixed-duration sliding window of 1 second is used to obtain sensory segments with 50% overlap between adjacent windows. We follow the same protocol as Anguita et al. (2013) to generate the training and testing data.

MHEALTH Dataset Banos et al. (2014). The dataset is aimed for recognition of 12 physical activities with diverse action intensities and execution speeds collected from 10 subjects. The dataset comprises body motion and vital signs recorded at a sampling rate of 50 Hz over 23 sensor channels. Since, the sampling rate in Anguita et al. (2013) is identical to the MHEALTH dataset, we follow the same protocol to segment the data and generate the train and test splits.

7.4.2 Implementation Details

Data Prepration. Data-streams are initially re-scaled using per-channel normalisation. After adopting the sliding window segmentation technique to partition the continuous data-streams, we consider the first 50% of sensory measurements in each segment to constitute the input sequences to our framework. Accordingly, the temporally inversed version of the input is used as the target sequence for the reconstruction task while the remaining sensory measurements are considered as the target sequence for the future prediction task.

Network Architecture. For the encoder, we leverage a two-layered bi-directional GRU with 256 hidden units. The decoders have an identical structure but utilise uni-directional connections. Considering the lower input dimension for UCI HAR as compared with skoda and MHEALTH datasets, we impose a bottleneck embedding dimension of 64 for the former and 256 for the latter ones in our autoencoder network. The clustering network $f_\omega(.)$ for adopting CAH has a single layer that generates soft cluster assignments according to Eq. 7.5. In addition, $f_\omega(.)$ for incorporating IMSAT uses two stacked fully connected layers with 1200 neurons, followed by a final layer with softmax activation whose dimension is set to be the number of activity categories.

Optimisation Settings. We implement our experiments in PyTorch Paszke et al. (2017) and optimise the network parameters according to ADAM Kingma and Ba (2015) update rule in mini-batches of 256 using an initial learning rate of 10^{-3}, decayed by a factor of 10 after 70 epochs. The network is pre-trained for 100 epochs, and fine tuned with a clustering objective until the cluster assignment changes between two consecutive epochs is less than 0.1%. The weighting coefficients γ and λ are set to be 0.1, and perturbation range ϵ is 0.5. All above parameters are set to achieve a reasonably descent autoencoding loss ($\mathcal{L}_{\text{AE}}$) and are held constant across all datasets to refrain unrealistic parameter tuning.

7.4.3 Clustering

This section focuses on evaluating our deep clustering network architecture for the task of end-to-end *activity clustering* using multi-channel sensory data. The proposed approach is compared against popular centroid-based *k*-means Arthur and Vassilvitskii (2007) clustering as well as representative hierarchical clustering algorithms including agglomerative clustering with average linkage (AC-Average)

Jain, Murty and Flynn (1999), agglomerative clustering with complete linkage (AC-Complete) and Ward agglomerative clustering (AC-Ward). In addition, we present comparisons against existing end-to-end deep clustering methods proposed in Xie, Girshick and Farhadi (2016); Guo et al. (2017a); Hu et al. (2017) and show how they fail to cater for the sequential nature of time-series data. We base our evaluations for clustering on the two widely adopted metrics of *unsupervised clustering accuracy* (*ACC*) and Normalised Mutual Information (*NMI*).

Clustering Performance. In Table 7.2, we evaluate the clustering performance of the baselines on both the: *i*) *data space* using raw input representations: and *ii*) *autoencoding space* using the embedded features $\{\mathbf{z}_i = \text{Enc}_{\theta^*}(\mathbf{x}_i)\}_{i=1}^{n}$ attained by optimising $\mathcal{L}_{\text{AE}}$ in the pre-training stage, and *iii*) compare with the *end-to-end cluster assignments* generated by deep clustering baselines and the proposed *Deep Sensory Clustering*, optimised for CAH or IMSAT criteria. As required by the CAH objective, we report results over two different strategies to initialise the cluster centres *only once* before commencing the refinement stage: *i*) we run *k*-means clustering on the embedded features to obtain *k* centroids; and *ii*) we perform Ward clustering and use the mean representation of the obtained clusters as the initial centres. We also assess performance levels on the test splits in order to evaluate the generalisation of the clustering algorithms beyond the data seen during the training stage.

As indicated by the results, not only our end-to-end approach outperforms traditional clustering algorithms applied on both input data and auto-encoding spaces, but also offers a large performance margin over representative deep clustering baselines originally proposed for image datasets. Without any manual supervision, the proposed unsupervised approach directly delivers cluster assignments that highly correspond to the activities of interest in the explored datasets, indicated by the mean accuracies of 75.33%, 50.97% and 55.13%, respectively obtained on UCI HAR, Skoda and MHEALTH datasets. In addition, the *consistent improvement* of unsupervised metrics across all three HAR datasets using the proposed framework, demonstrates: *i*) the insensitivity to the choice of our optimised clustering loss $\mathcal{L}_{\text{C}}$, CAH or IMSAT; and *ii*) the generalisability to different HAR problems. In the remaining sections, we investigate the clustering-oriented feature space achieved from the CAH variant of our framework with Ward initialisation and report analysis with UCI HAR dataset given that: *i*) the performance gains from the proposed approach is agnostic the clustering objective; and *ii*) the proposed approach generalises well across all three datasets.

Visualisation. In Figure 7.2, we demonstrate the evolution of the feature space towards the ultimate clustering-oriented embedding space achieved with our framework by visualising the data representation for the sequences in UCI HAR using t-SNE Maaten and Hinton (2008). Here, we show the original dataset, the dataset embedded by the encoder after the pre-training stage (autoencoding space) and the final representations after optimising for the aggregated objective function $\mathcal{L}$ in Eq. 7.4 (clustering-oriented space). *Without manual supervision, the framework discovers well-defined and clearly separated clusters of activity segments with strong correspondence to the ground-truth labels*. As a reference, we also present the clustering spaces achieved by the baseline deep clustering methods in Figure 7.3, where the feature spaces achieved fail in correctly discovering activity clusters; *e.g.* static activities of `sitting` and `standing` are recognised as a single cluster, and different `walking` variations are completely mingled. These visualisations highlight the necessity of leveraging recurrent structures within the network architectures and incorporation of effective self-supervised tasks when dealing with time-series data generated by wearables.

Table 7.2: Clustering performance comparison. We present a quantitative comparison of clustering performance on three HAR benchmark datasets based on acquired accuracy (ACC) and NMI.

	UCI HAR Dataset				Skoda Dataset				MHEALTH Dataset			
	Train Split		Test Split		Train Split		Test Split		Train Split		Test Split	
	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC
Traditional Clustering on Input Data Space												
k-means	44.28%	48.25%	42.28%	42.14%	43.41%	41.01%	46.01%	40.67%	49.71%	39.55%	48.37%	42.37%
AC-Average	1.38%	19.16%	1.93%	18.29%	4.61%	14.34%	30.98%	26.04%	4.44%	9.21%	7.55%	9.31%
AC-Complete	3.97%	19.56%	20.04%	31.69%	30.85%	27.48%	39.01%	37.47%	16.42%	11.82%	17.84%	11.15%
AC-Ward	41.07%	42.26%	48.21%	43.26%	46.55%	44.68%	46.92%	41.78%	54.06%	45.16%	56.99%	45.99%
Traditional Clustering on Autoencoding Space												
k-means	51.93%	60.19%	45.49%	55.62%	53.75%	47.56%	50.64%	42.62%	54.86%	43.96%	55.75%	48.24%
AC-Average	45.18%	37.57%	46.41%	34.61%	18.88%	16.96%	38.59%	30.22%	34.54%	20.47%	47.01%	29.26%
AC-Complete	40.66%	40.03%	40.81%	43.67%	32.55%	32.47%	41.57%	35.93%	42.23%	35.05%	44.42%	36.51%
AC-Ward	75.27%	74.78%	52.83%	60.33%	55.81%	51.51%	54.41%	45.96%	61.07%	48.91%	57.04%	46.28%
End-to-End Deep Clustering												
DEC Xie, Girshick and Farhadi (2016)	52.85%	50.45%	53.00%	49.85%	45.32%	40.46%	47.06%	40.25%	51.86%	43.64%	52.38%	44.91%
IDEC Guo et al. (2017a)	54.86%	51.14%	50.47%	50.15%	49.54%	47.41%	47.47%	45.96%	50.89%	42.49%	53.44%	44.72%
IMSAT Hu et al. (2017)	62.92%	67.98%	57.81%	62.78%	45.71%	44.14%	47.66%	42.48%	54.77%	49.44%	55.17%	49.32%
(Ours) Deep Sensory Clustering w/ CAH (*k*-means Init.)	64.75%	64.54%	61.58%	61.28%	56.91%	50.97%	57.01%	50.28%	62.65%	**57.19%**	63.06%	**56.85%**
(Ours) Deep Sensory Clustering w/ CAH (Ward Init.)	76.43%	78.79%	71.25%	75.41%	56.97%	**52.9%**	**59.06%**	**53.48%**	59.42%	51.57%	60.91%	53.33%
(Ours) Deep Sensory Clustering w/ IMSAT	**77.84%**	**82.66%**	**73.65%**	**78.22%**	**57.86%**	49.01%	58.85%	49.03%	**65.67%**	56.61%	**66.72%**	55.95%

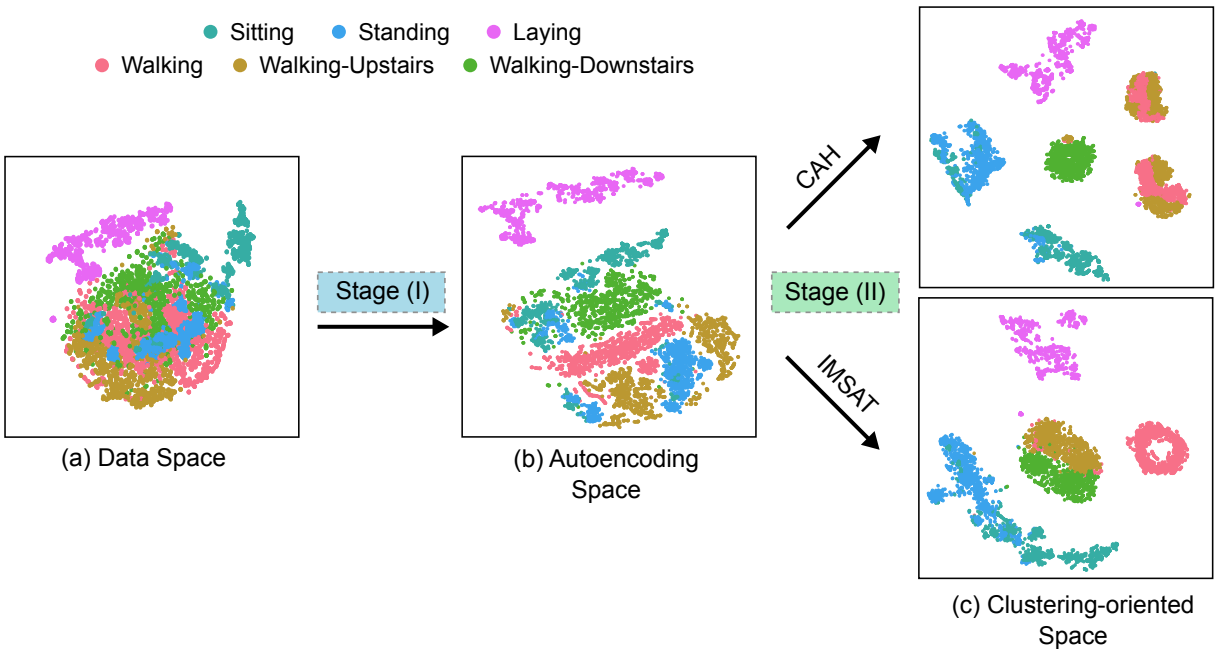
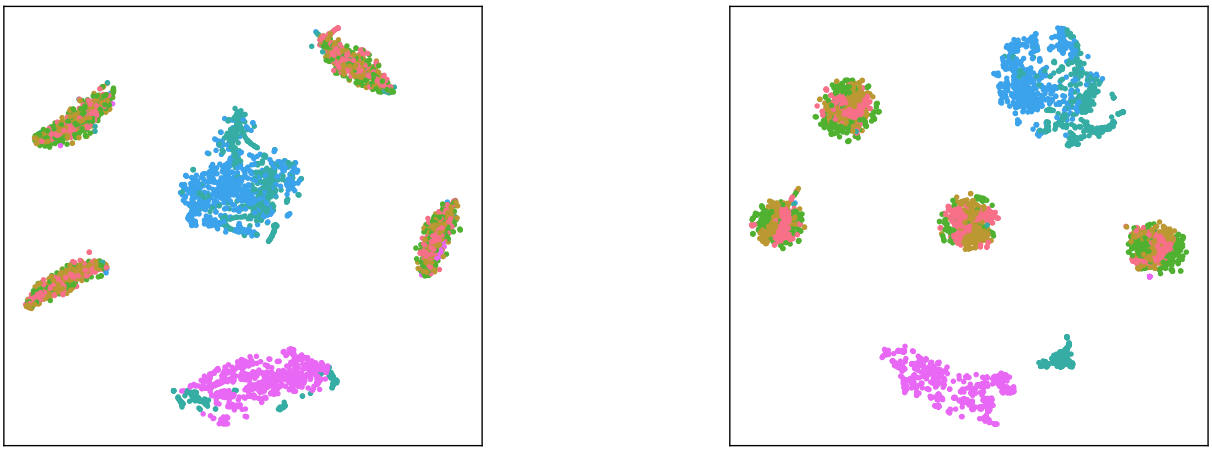


Figure 7.2: Deep Sensory Clustering space visualisations. We illustrate t-SNE visualisations of data representations for UCI HAR dataset in (a) Input data space, (b) Autoencoding space, and (c) Clustering-oriented space achieved with the proposed architecture using CAH and IMSAT criteria. Sequence representations are colour-coded with their corresponding ground-truth activity labels.

Ablation Study. In order to provide insights on the importance of each optimisation criterion adopted in our framework, we perform an ablation study on UCI HAR to assess the relevance of the embedded features for the clustering task after training with regards to each objective. To this end, we pass the dataset through the optimised encoder module and perform classic clustering algorithms on the embedded features. The obtained accuracies in each embedding space are summarised in Figure. 7.4. Consistent with our previous observations, the performance improvements gained in the reduced space produced by the encoder, substantiates the validity of the self-supervised tasks for unsupervised extraction of discriminative features. Moreover, jointly optimising for sequence reconstruction and future prediction tasks provides a more reliable embedding space compared with considering each task individually, as indicated by the higher mean clustering performance. Most importantly, we observe that introducing the clustering criterion results in a clustering-friendly feature space where all algorithms demonstrate a uniform and significantly improved clustering performance. We observed the same trend for the other two HAR datasets.



(a) DEC Xie, Girshick and Farhadi (2016) **(b)** IDEC Guo et al. (2017a)

Figure 7.3: Baseline clustering space visualisations. We visualise the clustering spaces achieved by the baseline deep clustering methods for UCI HAR dataset using t-SNE. Evidently, colour-coded representations from different activities are not effectively separated in the feature space.

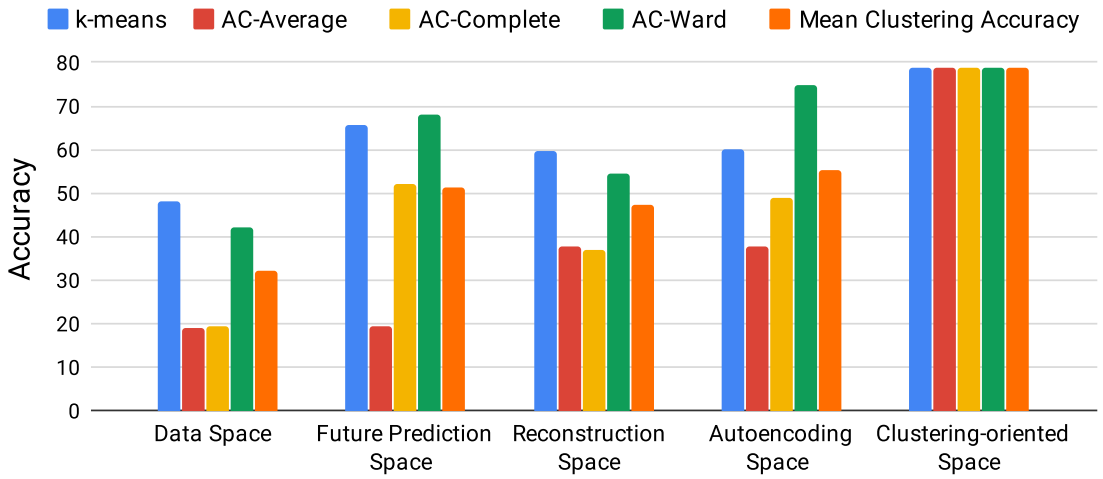


Figure 7.4: Ablation study. We present an ablation study on the validity of optimisation criteria adopted in our architecture for the task of clustering. The future prediction, reconstruction, autoencoding and clustering-oriented spaces respectively denote the embedding spaces achieved after optimising for $\mathcal{L}_{\text{fut}}$, $\mathcal{L}_{\text{rec}}$, $\mathcal{L}_{\text{AE}}$, and the proposed aggregated $\mathcal{L}$ (Eq. 7.4) using the CAH objective.

7.5 Conclusions

In our efforts to study unsupervised learning possibilities in HAR, this chapter examined the hitherto unexplored problem of *end-to-end clustering* of human activities from unlabelled multi-channel time-series data captured by wearables using a deep learning paradigm. For the first time, a novel deep clustering architecture was developed for HAR problems with wearable sensor data that (*a*) leverages the inherently sequential nature of sensory data, (*b*) exploits self-supervision from reconstruction and future prediction tasks, and (*c*) incorporates clustering-oriented objectives to promote the formation of highly discriminative activity clusters. The systematic experimental regime demonstrated the effectiveness and generalisability of the proposed approach for clustering human activities across three diverse HAR benchmark datasets. Further, insights into the proposed approach were shared by examining the unsupervised learned representations from sequential sensor data, and conducting an ablation study to validate the network design thinking.

The next chapter will briefly review the HAR problems explored in this dissertation and share the conclusions made upon the investigations. Moreover, the next chapter will outline potential areas worthy of future research investigations emanating from this dissertation.

Chapter 8

Conclusion

THIS chapter concludes the dissertation and suggests directions for future work.

8.1 Summary

This dissertation focused on emerging research problems concerned with recognition of human activities using often tiny wearable devices and end-to-end deep learning paradigms. The investigated problems covered diverse HAR problem settings ranging from fully supervised to unsupervised problem domains.

We investigated HAR problems under supervised learning settings where data annotations were assumed to be available during training. These investigations are presented in three chapters:

- In Chapter 3, we describe novel deep learning elements for supervised feature extraction. These elements have been developed to learn highly discriminative and generalisable activity feature representations from multi-channel time-series data captured by wearable sensors. The study is motivated by investigating key under-explored dimensions of HAR to improve accurate end-to-end recognition of human activities in diverse ubiquitous application scenarios. In particular, the developed components: (*a*) introduce a new HAR module based on self-attention with the aim of exploiting the latent relationships between multi-channel sensor modalities and human activities, (*b*) demonstrate the effectiveness of mixup data augmentation for sequential data to regularise deep HAR models, and (*c*) jointly incorporates cross-entropy and centre-loss training objectives to elegantly handle the inevitable challenges of intra-class variations and inter-class similarities in human activities. The extensive quantitative experimental results substantiated the effectiveness and generalisability of the introduced network elements on diverse activity recognition problem benchmarks. Accordingly, we hope to see the incorporation of the introduced components in future applications, with the goal of effectively training activity recognition systems.

- In Chapter 4, a first end-to-end deep learning framework—namely, *SparseSense*—is proposed for temporally sparse data. This framework promotes integration of the emerging battery-less wearable devices for unobtrusive activity monitoring, in particular for healthcare applications in hospitals and nursing homes. In order to handle the temporally sparse nature of the acquired data-streams, the framework incorporates set-based neural networks to learn robust activity representations with a high tolerance for missing sensor observations. Notably, the developed method seamlessly operates

on sparse segments with potentially varying numbers of sensor readings and bypasses the need for interpolation pre-processing. Through empirical evaluations, it is demonstrated that our novel treatment for sparse data-stream classification results in activity recognition models that significantly outperform solid HAR baselines while incurring notably lower real-time prediction delays. Consequently, this study provides a deep learning method for the construction of HAR models that enable low-cost, maintenance-free and unobtrusive solutions to understand human motion data captured by battery-less wearable computing platforms.

- In Chapter 5, the flexibility provided by set learning frameworks has been further leveraged to present a novel formulation of HAR. This formulation elegantly handles simultaneous prediction of multiple activities for a given sensor segment. The proposed framework—namely *Deep Auto-Set*—demonstrates an end-to-end strategy for the direct learning of the cardinalities of ongoing activities together with their confidence scores. At inference time, the generated cardinality and confidence scores are jointly taken into account to predict a set of human actions corresponding to the raw input sensor readings. This strategy lifts the limitations of conventional HAR models, which are limited to predicting a single activity label even for multi-class windows. Moreover, in order to facilitate better generalisation performance on unseen test data, the supervised set learning scheme is preceded by an unsupervised pre-training stage that exploits unlabelled data to initialise network parameters. This preliminary exploitation of unlabelled data for parameter pre-training of HAR frameworks sets the scene for our investigations in Chapter 6-7 under fully unsupervised HAR problem settings.

Moreover, under unsupervised learning conditions where manual supervision from data annotations are absent during training, the exploration of HAR problems in this dissertation was organised into two chapters:

- In Chapter 6, for the first time, a recurrent bidirectional GAN—namely, *Guided-GAN*—has been designed to learn unsupervised activity representations from the GAN's latent feature space. The developed framework comprises a recurrent generator, encoder and joint discriminator specifically designed for sequential data. Governed by an adversarial game, these components

communicate in a unified architecture to learn unsupervised feature representations. Notably, the proposed network formulation alleviates the burden on the discriminator in achieving inverting generator and encoders by seeking additional feedback from geometric distance penalisation in data and latent manifolds, efficiently implemented through weight sharing. This strategy is rooted in the key insight that the discriminator's adversarial feedback alone may be insufficient to uncover the generator's inverse mapping. When evaluated on diverse downstream classification benchmarks, it was demonstrated that the obtained unsupervised feature representations outperformed existing unsupervised approaches while closely approaching the performance of *fully* supervised learned representations. In addition, further quantitative and qualitative evaluations validated (*a*) the generator's ability to produce diverse and realistic sequences, and (*b*) the veracity of sequence reconstructions through the learned generator and encoders.

- In Chapter 7, a first stand-alone end-to-end deep learning framework—namely, *Deep Sensory Clustering*—is proposed. The framework is designed to learn to discover semantically meaningful clusters of human actions embedded in unlabelled sensor data-streams. In a novel unified architecture design, the proposed solution (*a*) leverages the inherently sequential nature of sensory data, (*b*) exploits self-supervision from reconstruction and future prediction tasks, and (*c*) incorporates a clustering-oriented objective to promote the formation of highly discriminative activity clusters. Our systematic experimental regime demonstrated the effectiveness and generalisability of the proposed approach for clustering human activities across three diverse HAR benchmark datasets. Further, we shared insights into the proposed approach by examining the unsupervised learned representations from sequential sensor data, and conducting an ablation study to validate the network design thinking. This study creates new opportunities to learn human activities from unlabelled data that can be conveniently and cheaply collected from wearables.

8.2 Future Research Opportunities

The following problems and challenges highlight a number of potential opportunities worthy of future explorations. They have been derived from our analysis of HAR

problems with deep learning paradigms, as well as from all the investigations conducted in this dissertation.

- **Concurrent Human Activity Recognition.** In Chapter 5, a novel formulation of HAR based on set learning is presented. Its purpose is to elegantly handle simultaneous prediction of multiple activities for a given sensor segment. While the evaluations were conducted on popular HAR datasets with subjects performing activities one after another *sequentially*, the proposed set-based methodology also offers an elegant solution for the challenging problem of concurrent human activity recognition. Within this problem, the goal is to recognise not only the sequential activities, but also to recognise the co-occurring activities from raw sensory time-series data. For example, for a sensor segment which has recorded the activities of an individual who is drinking coffee while simultaneously walking on the street, the HAR system is expected to generate the corresponding output set of {walk,drink} to precisely capture the underlying concurrent activities. Notably, as opposed to existing multi-class based HAR systems (which are limited to the prediction a single activity category for a given sensor segment), all recognition models explored in Chapter 5 are capable of predicting multiple activities simultaneously. Thus, as a future direction, the proposed systems can be evaluated for recognition of concurrent human activities. In particular, very limited studies have focused on exploring deep learning solutions for concurrent activity recognition Zhang et al. (2017); Li et al. (2017); Okita and Inoue (2017), leaving further opportunities to tackle this challenging problem.

- **Clustering of Human Activity Data within GAN's Latent Space.** In Chapter 6, we study the GAN's latent space for unsupervised feature learning and derive *Guided-GAN* to extract enriched activity representations from wearable data. Through evaluation of downstream classification tasks, we observe that the GAN's latent space provides a powerful alternative over the de facto autoencoder-based frameworks for extracting unsupervised activity feature representations. Moreover, the promising results from a recent study by Mukherjee et al. (2019) demonstrate that the GAN's latent space can be further adapted for the challenging task of clustering. That is, by wisely replacing the popular standard Gaussian prior with a mixture of discrete and continuous latent variables, we can impose a non-smooth geometry in the GAN's latent space in

order to elegantly represent the distinct cluster categories. A logical follow-up step to our investigations in Chapter 6 is to extend the Guided-GAN framework for the task of clustering. This can be pursued by modifying: (*a*) the recurrent generator so that it can consume a prior that comprises standard Gaussian vectors (to represent continuous random variables) cascaded with one-hot encoded vectors (to represent discrete variables); and (*b*) the recurrent encoder so that it can predict both the discrete and continuous portions of the latent representations. Notably, the virtue of achieving a clustering solution within the GAN's latent space is that the generator can then be used for conditional synthetic sequence generation. That is, by fixing the discrete portion of the latent code and varying the continuous component, we can sample synthetic sequences which correspond to different modes of activity data distribution, in an unsupervised manner. We leave it to future studies to investigate the GAN's latent space for the clustering of human activities with deep generative models.

- **Deep Clustering with Unknown Number of Activities.** In Chapter 7, the problem of activity clustering from unlabelled sensor data-streams captured by wearables is investigated. Briefly, the problem is concerned with categorising a set of n unlabelled segments of sensory readings $\{\mathbf{x}_i\}_{i=1}^{n}$ into k clusters, each representing a semantic human activity category. We present the *Deep Sensory Clustering* framework to address the problem by training a multi-task recurrent autoencoder jointly optimised with feedback from a clustering criterion. The proposed solution, inline with the majority of studies in the area of deep clustering Xie, Girshick and Farhadi (2016); Guo et al. (2017a,b); Li, Qiao and Zhang (2018); Chen, Lv and Zhang (2017); Ghasedi Dizaji et al. (2017); Yang et al. (2017); Hu et al. (2017), relies on prior knowledge of the number of ground-truth clusters. That is, the formulation requires the number of activities k to be set a priori. However, for real-world practical HAR applications, the number of executed human activities during unsupervised data acquisition is often not known in advance. Accordingly, it is of high interest to tackle the problem of deep clustering for human activities without leveraging any prior knowledge on the true number of clusters. Instead k should be inferred as a *random latent variable* in future studies.

Biography

Alireza Abedin received his B.Sc. in Computer Engineering from Khajeh Nasir Toosi University of Technology, Tehran, Iran in 2014, and M.Sc. in Information Technology from Amirkabir University of Technology, Tehran, Iran in 2016, where he respectively ranked *2nd* and *1st* among all program graduates.

In 2017, he was awarded the University of Adelaide International Scholarship to pursue his doctoral degree at the School of Computer Science, The University of Adelaide under the supervision of Associate Professor Damith Chinthana Ranasinghe, Professor Javen Qinfeng Shi, and Doctor Seyed Hamid Rezatofighi.

During his candidature, he secured the third place award at UbiComp 2019 in the *emteq Activity Recognition Challenge* as a member of the team *DeepSense*. In 2019, he was a visiting researcher in Lab-STICC, IMT Atlantique, France, for collection of multi-modal human activity recognition dataset from the Experiement'Haal Living Lab. He was the recipient of the *International Joint Conferences on Artificial Intelligence (IJCAI)* travel grant in 2019, and the *Faculty of Engineering, Computer and Mathematical Sciences Higher Degree by Research (ECMS HDR)* traveling scholarship in 2020. His main research interests lie primarily in the area of deep learning, machine learning and artificial intelligence with applications for Human Activity Recognition (HAR).

Alireza Abedin
alireza.abedinvaramin@adelaide.edu.au

Bibliography

Abedin, A., Abbasnejad, E., Shi, Q., Ranasinghe, D.C. and Rezatofighi, H., 2018. Deep auto-set: A deep auto-encoder-set network for activity recognition using wearables. *Proceedings of the 15th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. (Cited on pages 7, 112, and 114.)

Abedin, A., Rezatofighi, S.H., Shi, Q. and Ranasinghe, D.C., 2019. SparseSense: Human activity recognition from highly sparse sensor data-streams using set-based neural networks. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. (Cited on pages 7 and 114.)

Alessandro, A., Corani, G., Mauá, D. and Gabaglio, S., 2013. An ensemble of bayesian networks for multilabel classification. *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*. (Cited on page 82.)

Aljalbout, E., Golkov, V., Siddiqui, Y., Strobel, M. and Cremers, D., 2018. Clustering with deep learning: Taxonomy and new methods. *arxiv preprint arxiv:1801.07648*. (Cited on page 115.)

Alsheikh, M.A., Selim, A., Niyato, D., Doyle, L., Lin, S. and Tan, H.P., 2016. Deep activity recognition models with triaxial accelerometers. *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*. (Cited on pages 28, 66, 75, 81, 112, and 114.)

Alzantot, M., Chakraborty, S. and Srivastava, M.B., 2017. Sensegen: A deep learning architecture for synthetic sensor data generation. *The IEEE International Conference on Pervasive Computing and Communications Workshops*. (Cited on pages 90 and 93.)

Anguita, D., Ghio, A., Oneto, L., Parra, X. and Reyes-Ortiz, J.L., 2013. A public domain dataset for human activity recognition using smartphones. *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. (Cited on pages 18, 100, and 123.)

Arthur, D. and Vassilvitskii, S., 2007. K-means++: The advantages of careful seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics. (Cited on pages 115 and 124.)

Bai, L., Yeung, C., Efstratiou, C. and Chikomo, M., 2019. Motion2vector: Unsupervised learning in human activity recognition using wrist-sensing data. *Proceedings of the ACM International Symposium on Wearable Computers*. (Cited on pages 90, 101, 102, 112, and 114.)

Banos, O., Garcia, R., Holgado-Terriza, J.A., Damas, M., Pomares, H., Rojas, I., Saez, A. and Villalonga, C., 2014. mhealthdroid: A novel framework for agile development of mobile health applications. *Ambient Assisted Living and Daily Activities*. Springer International Publishing. (Cited on pages 18 and 123.)

Bao, L. and Intille, S.S., 2004. Activity recognition from user-annotated acceleration data. *Proceedings of the 2nd International Conference on Pervasive Computing*. (Cited on pages 2 and 3.)

Bhattacharya, S. and Lane, N.D., 2016. Sparsification and separation of deep learning layers for constrained resource inference on wearables. *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*. (Cited on page 29.)

Bianchi, V., Bassoli, M., Lombardo, G., Fornacciari, P., Mordonini, M. and De Munari, I., 2019. Iot wearable sensor and deep learning: An integrated approach for personalized human activity recognition in a smart home environment. *IEEE Internet of Things Journal*. IEEE. (Cited on page 2.)

Brock, A., Donahue, J. and Simonyan, K., 2019. Large scale gan training for high fidelity natural image synthesis. *The International Conference on Learning Representations*. (Cited on page 92.)

Bulling, A., Blanke, U. and Schiele, B., 2014. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys*. (Cited on pages 2, 15, and 72.)

Bulling, A., Ward, J.A. and Gellersen, H., 2012. Multimodal recognition of reading activity in transit using body-worn sensors. *ACM Transactions on Applied Perception*. (Cited on page 3.)

Chavarriaga, R., Sagha, H., Calatroni, A., Digumarti, S.T., Tröster, G., R. Millán, J. del and Roggen, D., 2013. The opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters*. (Cited on pages 16, 17, 37, 80, and 81.)

Che, T., Li, Y., Jacob, A.P., Bengio, Y. and Li, W., 2017. Mode regularized generative adversarial networks. *The International Conference on Learning Representations*. (Cited on page 97.)

Chen, D., Lv, J. and Zhang, Y., 2017. Unsupervised multi-manifold clustering by learning deep representation. *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*. (Cited on pages 115 and 136.)

Chen, S.J., Fumeaux, C., Ranasinghe, D.C. and Kaufmann, T., 2015. Paired snap-on buttons connections for balanced antennas in wearable systems. *IEEE Antennas and Wireless Propagation Letters*. (Cited on pages 4 and 54.)

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I. and Abbeel, P., 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in Neural Information Processing Systems*. (Cited on page 92.)

Chen, Y. and Xue, Y., 2015. A deep learning approach to human activity recognition based on single accelerometer. *The IEEE International Conference on Systems, Man, and Cybernetics*. (Cited on page 29.)

Chesser, M., Jayatilaka, A., Visvanathan, R., Fumeaux, C., Sample, A. and Ranasinghe, D.C., 2019. Super low resolution rf powered accelerometers for alerting on hospitalized patient bed exits. *IEEE International Conference on Pervasive Computing and Communications*. (Cited on page 2.)

Cho, K., Van Merriënboer, B., Bahdanau, D. and Bengio, Y., 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arxiv preprint arxiv:1409.1259*. (Cited on page 19.)

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009. ImageNet: A Large-Scale Hierarchical Image Database. *Conference on Computer Vision and Pattern Recognition*. (Cited on page 92.)

Denton, E., Chintala, S., Szlam, A. and Fergus, R., 2015. Deep generative image models using a laplacian pyramid of adversarial networks. *Proceedings of the 28th International Conference on Neural Information Processing Systems*. (Cited on page 92.)

Donahue, J., Krähenbühl, P. and Darrell, T., 2017. Adversarial feature learning. *The International Conference on Learning Representations*. (Cited on pages 90, 92, 96, 98, and 107.)

Donahue, J. and Simonyan, K., 2019. Large scale adversarial representation learning. *Advances in Neural Information Processing Systems*. (Cited on page 92.)

Dumoulin, V., Belghazi, I., Poole, B., Mastropietro, O., Lamb, A., Arjovsky, M. and Courville, A., 2017. Adversarially learned inference. *The International Conference on Learning Representations*. (Cited on pages 91, 92, 96, 97, and 107.)

Edel, M. and Köppe, E., 2016. Binarized-blstm-rnn based human activity recognition. *The International Conference on Indoor Positioning and Indoor Navigation*. (Cited on page 29.)

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.A., Vincent, P. and Bengio, S., 2010. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*. (Cited on page 76.)

Esteban, C., Hyland, S.L. and Rätsch, G., 2017. Real-valued (medical) time series generation with recurrent conditional gans. *arxiv preprint arxiv:1706.02633*. (Cited on pages 90, 93, 94, 95, 101, and 102.)

Faridee, A.Z.M., Khan, M.A.A.H., Pathak, N. and Roy, N., 2019. Augtoact: Scaling complex human activity recognition with few labels. *Proceedings of the 16th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. (Cited on page 30.)

Freitag, M., Amiriparian, S., Pugachevskiy, S., Cummins, N. and Schuller, B., 2018. audeep: Unsupervised learning of representations from audio with deep recurrent neural networks. *Journal of Machine Learning Research*. (Cited on page 90.)

Ghasedi Dizaji, K., Herandi, A., Deng, C., Cai, W. and Huang, H., 2017. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. *The IEEE International Conference on Computer Vision*. (Cited on pages 115 and 136.)

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems*. (Cited on pages 90, 91, and 93.)

Gövercin, M., Költzsch, Y., Meis, M., Wegel, S., Gietzelt, M., Spehr, J., Winkelbach, S., Marschollek, M. and Steinhagen-Thiessen, E., 2010. Defining the user requirements for wearable and optical fall prediction and fall detection devices for home use. *Informatics for Health and Social Care*. (Cited on pages 4 and 54.)

Gu, F., Khoshelham, K., Valaee, S., Shang, J. and Zhang, R., 2018. Locomotion activity recognition using stacked denoising autoencoders. *IEEE Internet of Things Journal*. (Cited on page 55.)

Guan, Y. and Plötz, T., 2017. Ensembles of deep lstm learners for activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*. (Cited on pages 29, 38, 39, 41, 44, 55, 114, and 123.)

Günther, L.C., Kärcher, S. and Bauernhansl, T., 2019. Activity recognition in manual manufacturing: Detecting screwing processes from sensor data. *Procedia CIRP*. Elsevier. (Cited on page 2.)

Guo, X., Gao, L., Liu, X. and Yin, J., 2017a. Improved deep embedded clustering with local structure preservation. *The International Joint Conference on Artificial Intelligence*. (Cited on pages 115, 125, 127, 129, and 136.)

Guo, X., Liu, X., Zhu, E. and Yin, J., 2017b. Deep clustering with convolutional autoencoders. *International Conference on Neural Information Processing*. (Cited on pages 115 and 136.)

Guo, Y. and Gu, S., 2011. Multi-label classification using conditional dependency networks. *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*. (Cited on page 82.)

Hammerla, N.Y., Fisher, J., Andras, P., Rochester, L., Walker, R. and Plötz, T., 2015. Pd disease state assessment in naturalistic environments using deep learning. *AAAI Conference on Artificial Intelligence*. (Cited on page 28.)

Hammerla, N.Y., Halloran, S. and Plötz, T., 2016. Deep, convolutional, and recurrent models for human activity recognition using wearables. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. (Cited on pages 3, 29, 38, 39, 41, 44, 55, 63, 74, 75, 82, and 114.)

Haresamudram, H., Anderson, D. and Plötz, T., 2019. On the role of features in human activity recognition. *Proceedings of International Symposium on Wearable Computers*. (Cited on pages 38, 90, 100, 101, 102, 112, and 114.)

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (Cited on page 30.)

Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. *Neural Computation*. MIT Press. (Cited on page 19.)

Hofmann, T., Schölkopf, B. and Smola, A.J., 2008. Kernel methods in machine learning. *The Annals of Statistics*. JSTOR. (Cited on page 115.)

Hossain, H.S., Al Haiz Khan, M.A. and Roy, N., 2018. Deactive: scaling activity recognition with active deep learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*. ACM New York, NY, USA. (Cited on page 114.)

Hu, W., Miyato, T., Tokui, S., Matsumoto, E. and Sugiyama, M., 2017. Learning discrete representations via information maximizing self-augmented training. *Proceedings of the 34th International Conference on Machine Learning*. (Cited on pages 115, 119, 125, 127, and 136.)

Huynh, T. and Schiele, B., 2005. Analyzing features for activity recognition. *Proceedings Conference on Smart Objects and Ambient Intelligence: Innovative Context-aware Services: Usages and Technologies*. (Cited on page 2.)

Inoue, M., Inoue, S. and Nishida, T., 2018. Deep recurrent neural network for mobile human activity recognition with high throughput. *Artificial Life and Robotics*. Springer. (Cited on page 29.)

Jain, A.K., Murty, M.N. and Flynn, P.J., 1999. Data clustering: a review. *ACM Computing Surveys*. Acm. (Cited on page 125.)

Jayatilaka, A., Dang, Q.H., Chen, S.J., Visvanathan, R., Fumeaux, C. and Ranasinghe, D.C., 2019. Designing batteryless wearables for hospitalized older people. *Proceedings of International Symposium on Wearable Computers*. (Cited on pages 4 and 51.)

Jiang, W. and Yin, Z., 2015. Human activity recognition using wearable sensors by deep convolutional neural networks. *Proceedings of the 23rd ACM International Conference on Multimedia*. ACM. (Cited on page 114.)

Jordao, A., Nazare Jr, A.C., Sena, J. and Schwartz, W.R., 2018. Human activity recognition based on wearable sensor data: A standardization of the state-of-the-art. *arxiv preprint arxiv:1806.05226*. (Cited on page 42.)

Karras, T., Aila, T., Laine, S. and Lehtinen, J., 2018. Progressive growing of gans for improved quality, stability, and variation. *The International Conference on Learning Representations*. (Cited on page 90.)

Khan, M.A.A.H., Roy, N. and Misra, A., 2018. Scaling human activity recognition via deep learning-based domain adaptation. *The IEEE International Conference on Pervasive Computing and Communications*. (Cited on page 114.)

Kim, E., Helal, S. and Cook, D., 2010. Human activity recognition and pattern discovery. *IEEE Pervasive Computing*. (Cited on page 72.)

Kingma, D.P. and Ba, J., 2015. Adam: A method for stochastic optimization. *The International Conference on Learning Representations*. (Cited on pages 39, 102, and 124.)

Kingma, D.P. and Welling, M., 2014. Auto-encoding variational bayes. *The International Conference on Learning Representations*. (Cited on page 92.)

Kunze, K., Barry, M., Heinz, E.A., Lukowicz, P., Majoe, D. and Gutknecht, J., 2006. Towards recognizing tai chi - an initial experiment using wearable sensors. *The 3rd International Forum on Applied Wearable Computing 2006*. (Cited on page 2.)

Kwapisz, J.R., Weiss, G.M. and Moore, S.A., 2011. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*. (Cited on pages 17, 62, 66, 80, and 81.)

Ladha, C., Hammerla, N.Y., Olivier, P. and Plötz, T., 2013. Climbax: Skill assessment for climbing enthusiasts. *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. Association for Computing Machinery. (Cited on page 2.)

Lara, O.D., Pérez, A.J., Labrador, M.A. and Posada, J.D., 2012. Centinela: A human activity recognition system based on acceleration and vital sign data. *Pervasive and Mobile Computing*. (Cited on page 3.)

LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. *Nature*. (Cited on pages 3 and 20.)

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P., 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*. Ieee. (Cited on page 100.)

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z. et al., 2017. Photo-realistic single image super-resolution using a generative adversarial network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (Cited on page 90.)

Lemey, S., Agneessens, S., Van Torre, P., Baes, K., Vanfleteren, J. and Rogier, H., 2016. Wearable flexible lightweight modular rfid tag with integrated energy harvester. *IEEE Transactions on Microwave Theory and Techniques*. (Cited on pages 4 and 54.)

Li, F., Qiao, H. and Zhang, B., 2018. Discriminatively boosted image clustering with fully convolutional auto-encoders. *Pattern Recognition*. Elsevier. (Cited on pages 115 and 136.)

Li, X., Zhang, Y., Zhang, J., Chen, S., Marsic, I., Farneth, R.A. and Burd, R.S., 2017. Concurrent activity recognition with multimodal cnn-lstm structure. *arxiv preprint arxiv:1702.01638*. (Cited on page 135.)

Liu, M.Y. and Tuzel, O., 2016. Coupled generative adversarial networks. *Advances in Neural Information Processing Systems*. (Cited on page 90.)

Maaten, L.v.d. and Hinton, G., 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*. (Cited on pages 68, 107, and 126.)

Mathur, A., Zhang, T., Bhattacharya, S., Veličković, P., Joffe, L., Lane, N.D., Kawsar, F. and Lió, P., 2018. Using deep data augmentation training to address software and hardware heterogeneities in wearable and smartphone sensing devices. *Proceedings of the 17th ACM/IEEE International Conference on Information Processing in Sensor Networks*. (Cited on page 30.)

Min, E., Guo, X., Liu, Q., Zhang, G., Cui, J. and Long, J., 2018. A survey of clustering with deep learning: From the perspective of network architecture. *IEEE Access*. (Cited on page 115.)

Miyato, T., Kataoka, T., Koyama, M. and Yoshida, Y., 2018a. Spectral normalization for generative adversarial networks. *The International Conference on Learning Representations*. (Cited on page 90.)

Miyato, T., Maeda, S.i., Koyama, M. and Ishii, S., 2018b. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE. (Cited on page 122.)

Mogren, O., 2016. C-rnn-gan: Continuous recurrent neural networks with adversarial training. *arxiv preprint arxiv:1611.09904*. (Cited on pages 90 and 92.)

Moshiri, P.F., Navidan, H., Shahbazian, R., Ghorashi, S.A. and Windridge, D., 2020. Using gan to enhance the accuracy of indoor human activity recognition. *arxiv preprint arxiv:2004.11228*. (Cited on page 90.)

Mukherjee, S., Asnani, H., Lin, E. and Kannan, S., 2019. Clustergan: Latent space clustering in generative adversarial networks. *Proceedings of the AAAI Conference on Artificial Intelligence*. (Cited on page 135.)

Murahari, V.S. and Plötz, T., 2018. On attention models for human activity recognition. *Proceedings of International Symposium on Wearable Computers*. (Cited on pages 3, 27, 29, 33, 38, 39, 41, 43, 44, 45, 48, and 114.)

Nguyen, D.T., Cohen, E., Pourhomayoun, M. and Alshurafa, N., 2017. Swallownet: Recurrent neural network detects and characterizes eating patterns. *The IEEE International Conference on Pervasive Computing and Communications Workshops*. (Cited on page 114.)

Okita, T. and Inoue, S., 2017. Recognition of multiple overlapping activities using compositional cnn-lstm model. *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. Association for Computing Machinery. (Cited on page 135.)

Ordóñez, F. and Roggen, D., 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*. (Cited on pages 3, 5, 29, 32, 41, 42, 45, 55, 63, 72, 74, 75, 76, 77, 83, and 114.)

Park, T., Liu, M.Y., Wang, T.C. and Zhu, J.Y., 2019. Semantic image synthesis with spatially-adaptive normalization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (Cited on page 90.)

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. and Lerer, A., 2017. Automatic differentiation in pytorch. *NIPS Autodiff Workshop*. (Cited on pages 39, 62, 82, 102, and 124.)

Perarnau, G., Weijer, J. van de, Raducanu, B. and Álvarez, J.M., 2016. Invertible Conditional GANs for image editing. *NIPS Workshop on Adversarial Training*. (Cited on page 92.)

Philipose, M., Smith, J.R., Jiang, B., Mamishev, A., Sumit Roy and Sundara-Rajan, K., 2005. Battery-free wireless identification and sensing. *IEEE Pervasive Computing*. (Cited on page 4.)

Plötz, T., Hammerla, N.Y. and Olivier, P.L., 2011. Feature learning for activity recognition in ubiquitous computing. *Twenty-second International Joint Conference on Artificial Intelligence*. (Cited on page 28.)

Qi, C.R., Su, H., Mo, K. and Guibas, L.J., 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (Cited on pages 56 and 60.)

Radford, A., Metz, L. and Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arxiv preprint arxiv:1511.06434*. (Cited on pages 90, 92, and 94.)

Ranasinghe, D., Torres, R.S., Hill, K. and Visvanathan, R., 2014. Low cost and batteryless sensor-enabled radio frequency identification tag based approaches to identify patient bed entry and exit posture transitions. *Gait and Posture*. Elsevier. (Cited on page 54.)

Ravi, N., Dandekar, N., Mysore, P. and Littman, M.L., 2005. Activity recognition from accelerometer data. *Proceedings of the 17th Conference on Innovative Applications of Artificial Intelligence*. (Cited on page 2.)

Reiss, A. and Stricker, D., 2012. Introducing a new benchmarked dataset for activity monitoring. *Proceedings of International Symposium on Wearable Computers*. (Cited on pages 18, 36, and 38.)

Reynolds, D., 2015. Gaussian mixture models. *Encyclopedia of Biometrics*. Springer. (Cited on page 115.)

Rezatofighi, S.H., Milan, A., Shi, Q., Dick, A.R. and Reid, I.D., 2018. Joint learning of set cardinality and state distribution. *The AAAI Conference on Artificial Intelligence*. (Cited on pages 73, 79, and 80.)

Ronao, C.A. and Cho, S.B., 2015. Deep convolutional neural networks for human activity recognition with smartphone sensors. *International Conference on Neural Information Processing*. (Cited on pages 29 and 114.)

Rumelhart, D., Hinton, G. and Williams, R., 1986. Learning internal representations by error propagation, parallel distributed processing, vol. 1. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press. (Cited on page 19.)

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A. and Chen, X., 2016. Improved techniques for training gans. *Advances in Neural Information Processing Systems*. (Cited on page 90.)

Shah, S.A. and Koltun, V., 2018. Deep continuous clustering. *arxiv preprint arxiv:1803.01449*. (Cited on page 115.)

Shinmoto Torres, R.L., Shi, Q., Hengel, A. van den and Ranasinghe, D.C., 2017a. A hierarchical model for recognizing alarming states in a batteryless sensor alarm intervention for preventing falls in older people. *Pervasive Mobile Computing*. (Cited on page 3.)

Shinmoto Torres, R.L., Visvanathan, R., Abbott, D., Hill, K.D. and Ranasinghe, D.C., 2017b. A battery-less and wireless wearable sensor system for identifying bed and chair exits in a pilot trial in hospitalized older people. *PloS ONE*. (Cited on page 54.)

Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arxiv preprint arxiv:1409.1556*. (Cited on page 30.)

Sønderby, C.K., Caballero, J., Theis, L., Shi, W. and Huszár, F., 2017. Amortised map inference for image super-resolution. *The International Conference on Learning Representations*. (Cited on page 90.)

Srivastava, N., Mansimov, E. and Salakhudinov, R., 2015. Unsupervised learning of video representations using lstms. *International Conference on Machine Learning*. (Cited on page 118.)

Stiefmeier, T., Roggen, D., Ogris, G., Lukowicz, P. and Tröster, G., 2008. Wearable activity tracking in car manufacturing. *IEEE Pervasive Computing*. (Cited on pages 17, 38, and 123.)

Subasi, A., Radhwan, M., Kurdi, R. and Khateeb, K., 2018. Iot based mobile healthcare system for human activity recognition. *The 15th Learning and Technology Conference*. (Cited on page 2.)

Taigman, Y., Polyak, A. and Wolf, L., 2017. Unsupervised cross-domain image generation. *The International Conference on Learning Representations*. (Cited on page 90.)

Tieleman, T. and Hinton, G., 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*. (Cited on page 62.)

Torres, R.L.S., Ranasinghe, D.C., Shi, Q. and Sample, A.P., 2013. Sensor enabled wearable rfid technology for mitigating the risk of falls near beds. *IEEE International Conference on RFID*. (Cited on pages 18, 54, and 61.)

Torres, R.L.S., Visvanathan, R., Abbott, D., Hill, K.D. and Ranasinghe, D.C., 2017. A battery-less and wireless wearable sensor system for identifying bed and chair exits in a pilot trial in hospitalized older people. *PloS ONE*. (Cited on pages 4 and 54.)

Torres-Huitzil, C. and Alvarez-Landero, A., 2015. Accelerometer-based human activity recognition in smartphones for healthcare services. *Mobile Health*. Springer. (Cited on page 2.)

Um, T.T., Pfister, F.M.J., Pichler, D., Endo, S., Lang, M., Hirche, S., Fietzek, U. and Kulić, D., 2017. Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. Association for Computing Machinery. (Cited on pages 26, 30, 47, and 48.)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in Neural Information Processing Systems*. (Cited on page 32.)

Vepakomma, P., De, D., Das, S.K. and Bhansali, S., 2015. A-wristocracy: Deep learning on wrist-worn sensing for recognition of user complex activities. *The IEEE 12th*

International Conference on Wearable and Implantable Body Sensor Networks. (Cited on page 19.)

Walse, K.H., Dharaskar, R.V. and Thakare, V.M., 2016. Pca based optimal ann classifiers for human activity recognition using mobile sensors data. *Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 1*. (Cited on page 19.)

Wang, J., Chen, Y., Gu, Y., Xiao, Y. and Pan, H., 2018a. Sensorygans: an effective generative adversarial framework for sensor-based human activity recognition. *The International Joint Conference on Neural Networks*. (Cited on pages 90 and 93.)

Wang, J., Chen, Y., Hao, S., Peng, X. and Hu, L., 2019. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*. (Cited on pages 55 and 114.)

Wang, L., Gu, T., Tao, X., Chen, H. and Lu, J., 2011. Recognizing multi-user activities using wearable sensors in a smart home. *Pervasive and Mobile Computing*. Elsevier. (Cited on page 2.)

Wang, X., Girshick, R., Gupta, A. and He, K., 2018b. Non-local neural networks. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. (Cited on page 32.)

Ward, J.A., Lukowicz, P. and Gellersen, H.W., 2011. Performance metrics for activity recognition. *ACM Transactions on Intelligent Systems and Technology*. (Cited on page 43.)

Wen, Y., Zhang, K., Li, Z. and Qiao, Y., 2016. A discriminative feature learning approach for deep face recognition. *European Conference on Computer Vision*. (Cited on pages 31 and 34.)

Wickramasinghe, A. and Ranasinghe, D., 2015. Recognising activities in real time using body worn passive sensors with sparse data streams: To interpolate or not to interpolate? *Proceedings of the 12th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. (Cited on pages 55, 63, and 65.)

Wickramasinghe, A., Ranasinghe, D.C., Fumeaux, C., Hill, K.D. and Visvanathan, R., 2017. Sequence learning with passive rfid sensors for real-time bed-egress

recognition in older people. *IEEE Journal of Biomedical and Health Informatics*. (Cited on page 2.)

Wold, S., Esbensen, K. and Geladi, P., 1987. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*. Elsevier. (Cited on page 115.)

Xie, J., Girshick, R. and Farhadi, A., 2016. Unsupervised deep embedding for clustering analysis. *International Conference on Machine Learning*. (Cited on pages 115, 119, 125, 127, 129, and 136.)

Yang, B., Fu, X., Sidiropoulos, N.D. and Hong, M., 2017. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. *Proceedings of the 34th International Conference on Machine Learning*. (Cited on pages 115 and 136.)

Yang, J., Parikh, D. and Batra, D., 2016. Joint unsupervised learning of deep representations and image clusters. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (Cited on page 115.)

Yang, J.B., Nguyen, M.N., San, P.P., Li, X.L. and Krishnaswamy, S., 2015. Deep convolutional neural networks on multichannel time series for human activity recognition. *Proceedings of the 24th International Conference on Artificial Intelligence*. (Cited on pages 5, 29, 55, 72, 74, 75, and 114.)

Yao, R., Lin, G., Shi, Q. and Ranasinghe, D.C., 2018. Efficient dense labelling of human activity sequences from wearables using fully convolutional networks. *Pattern Recognition*. (Cited on pages 3, 5, 18, 29, 38, 41, 44, 72, 74, 75, and 83.)

Yao, S., Hu, S., Zhao, Y., Zhang, A. and Abdelzaher, T., 2017. Deepsense: A unified deep learning framework for time-series mobile sensing data processing. *Proceedings of the 26th International Conference on World Wide Web*. (Cited on page 29.)

Yoon, J., Jarrett, D. and Schaar, M. van der, 2019. Time-series generative adversarial networks. *Advances in Neural Information Processing Systems*. (Cited on page 90.)

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R.R. and Smola, A.J., 2017. Deep sets. *Advances in Neural Information Processing Systems*. (Cited on page 56.)

Zeng, M., Nguyen, L.T., Yu, B., Mengshoel, O.J., Zhu, J., Wu, P. and Zhang, J., 2014. Convolutional neural networks for human activity recognition using mobile sensors.

6th International Conference on Mobile Computing, Applications and Services. (Cited on pages 29, 55, 74, 75, and 114.)

Zhang, H., Cisse, M., Dauphin, Y.N. and Lopez-Paz, D., 2018a. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*. (Cited on page 35.)

Zhang, H., Goodfellow, I., Metaxas, D. and Odena, A., 2019. Self-attention generative adversarial networks. *International Conference on Machine Learning*. (Cited on page 32.)

Zhang, J., Dang, H., Lee, H.K. and Chang, E.C., 2018b. Flipped-adversarial autoencoders. *arxiv preprint arxiv:1802.04504*. (Cited on pages 91, 92, 94, 95, 97, 101, 102, and 107.)

Zhang, L., Wu, X. and Luo, D., 2015. Human activity recognition with hmm-dnn model. *The IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing*. (Cited on page 28.)

Zhang, M. and Sawchuk, A.A., 2012a. Motion primitive-based human activity recognition using a bag-of-features approach. *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. (Cited on page 2.)

Zhang, M. and Sawchuk, A.A., 2012b. Usc-had: a daily activity dataset for ubiquitous activity recognition using wearable sensors. *Proceedings of the ACM Conference on Ubiquitous Computing*. (Cited on pages 18 and 100.)

Zhang, R., Isola, P. and Efros, A.A., 2016. Colorful image colorization. *European Conference on Computer Vision*. (Cited on page 101.)

Zhang, Y., Li, X., Zhang, J., Chen, S., Zhou, M., Farneth, R.A., Marsic, I. and Burd, R.S., 2017. Poster abstract: Car - a deep learning structure for concurrent activity recognition. *The 16th ACM/IEEE International Conference on Information Processing in Sensor Networks*. (Cited on page 135.)

Zheng, H., Wang, H. and Black, N., 2008. Human activity detection in smart home environment with self-adaptive neural networks. *The IEEE International Conference on Networking, Sensing and Control*. (Cited on page 2.)

Zheng, N., Wen, J., Liu, R., Long, L., Dai, J. and Gong, Z., 2018. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. *Thirty-second aaai conference on artificial intelligence*. (Cited on page 27.)

Zhong, Z., Zheng, L., Kang, G., Li, S. and Yang, Y., 2020. Random erasing data augmentation. *The Association for the Advancement of Artificial Intelligence*. (Cited on page 30.)

Zhu, J.Y., Park, T., Isola, P. and Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE International Conference on Computer Vision*. (Cited on page 90.)

Zhuang, Z. and Xue, Y., 2019. Sport-related human activity detection and recognition using a smartwatch. *Sensors*. Multidisciplinary Digital Publishing Institute. (Cited on page 2.)