Karlsruher Institut für Technologie

Institut für Telematik
Pervasive Computing Systems / TECO
Prof. Dr.-Ing. Michael Beigl

Improving Robustness in Human Activity Recognition: Shared-Specific Feature Modeling for Missing Modalities

Bachelor's Thesis

by

Lukas Probst

Department of Informatics

Responsible Supervisor:	Prof. Dr. Michael Beigl
Supervising Staff:	Yexu Zhou and Tobias King
Project Period:	28.11.2023 - 28.03.2024

Karlsruher Institut für Technologie

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Abschlussarbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt habe, die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht und weiterhin die Richtlinien des KIT zur Sicherung guter wissenschaftlicher Praxis beachtet habe.

Karlsruhe, den _____

_____

Unterschrift

Kurzfassung

Praktische Anwendungen von Modellen zur Erkennung menschlicher Aktivitäten (HAR) leiden oft unter fehlenden Sensorkanälen oder Geräten. Wenn Modalitäten fehlen, leiden aktuelle HAR-Modelle in der Regel unter einem erheblichen Leistungs- und Robustheitsabfall, was eine breitere Anwendung von HAR-Modellen in verschiedenen und dynamischen realen Umgebungen verhindert. Dies macht die Verbesserung der Fähigkeit von HAR-Modellen, menschliche Aktivitäten unter Bedingungen unvollständiger Sensordaten genau zu erkennen und zu interpretieren, zu einem aktiven Forschungsgebiet im weiteren Kontext des multimodalen Lernens.

In dieser Arbeit wird Shared-Specific Feature Modeling (ShaSpec) als neuartiger Ansatz für den Umgang mit fehlenden Modalitäten für diverse HAR-Daten vorgeschlagen. Durch die Verfeinerung von ShaSpec speziell für HAR stellt diese Arbeit einen methodischen Fortschritt dar, der die Auswirkungen fehlender Modalitäten auf die Modellleistung deutlich abschwächt. Spezielle Kodierer- und Dekodierer-Architekturen wurden entwickelt, um die Verarbeitung multimodaler Sensordaten zu optimieren und effektives Lernen sowohl aus vollständigen als auch aus unvollständigen Datensätzen zu gewährleisten. Rigorose Experimente mit Benchmark-Datensätzen, einschließlich DSADS und REALDISP, zeigen die verbesserte Fähigkeit des Modells, eine hohe Genauigkeit und Robustheit bei der Aktivitätserkennung beizubehalten, selbst bei erheblichen Datenlücken. Eine Ablationsstudie veranschaulicht darüber hinaus den Wert der einzelnen Komponenten des angepassten ShaSpec-Modells und hebt die wesentliche Rolle der Mechanismen zur Erzeugung fehlender Modalitätsmerkmale und des gemeinsamen Kodierers bei der Erzielung überlegener Leistungen im Vergleich zu bestehenden Methoden hervor.

Diese Arbeit trägt nicht nur zum akademischen Diskurs über HAR bei, sondern schafft auch einen grundlegenden Rahmen für künftige Innovationen bei der Entwicklung belastbarerer und anpassungsfähigerer multimodaler Lernmodelle.

Abstract

Practical applications of Human Activity Recognition (HAR) models often suffer from missing sensor channels or devices. When facing missing modalities, current HAR models typically suffer from a significant drop in performance and robustness, preventing the wider adoption of HAR models in diverse and dynamic real-world environments. This makes improving the ability of HAR models to accurately recognize and interpret human activities under conditions of incomplete sensor data an active research area in the wider context of multimodal learning.

This thesis proposes Shared-Specific feature modeling (ShaSpec) as a novel approach to dealing with missing modalities for diverse HAR data. By refining ShaSpec specifically for HAR, this thesis represents a methodological advancement that significantly mitigates the impact of missing modalities on model performance. Specialized encoder and decoder architectures are designed to optimize the processing of multimodal sensor data, ensuring effective learning from both complete and incomplete datasets. Rigorous experimentation on benchmark datasets, including DSADS and REALDISP, demonstrates the model's enhanced capability to maintain high accuracy and robustness in activity recognition, even with substantial data missing. An ablation study further illustrates the value of each component in the adapted ShaSpec model, highlighting the essential roles of missing modality feature generation mechanisms and shared encoder in achieving superior performance against existing methodologies.

This work not only contributes to the academic discourse on HAR but also lays a foundational framework for future innovations in creating more resilient and adaptable multimodal learning models.

Keywords: Deep Learning, Human Activity Recognition (HAR), Multimodal Learning, Missing Sensor Data, Shared-Specific Feature Modeling (ShaSpec)

Contents

List of Figures

List of Tables

1. Introduction

1.1 Motivation

Human Activity Recognition (HAR) leverages data from various sensors and devices to detect and estimate human activities. Providing accurate information on people's activities and behaviors is one of the most important tasks in Pervasive Computing [8]. Its applications span healthcare, fitness, elderly care, smart homes, and more, offering insights into human behavior and facilitating automated assistance or monitoring. Data is collected by phones and wearables like smartwatches or earbuds. The more sensors and devices record data, the greater the diversity of the data and therefore the chance to correctly record an activity. Integrated into the Internet of Things (IoT), HAR underpins advancements in smart environments and personalized technology interactions, enabling improved operational efficiency and better quality of life.

The complexity of sensor-based data is rapidly increasing as technology evolves. Real-world scenarios, however, often suffer from faulty sensors, empty batteries, or non-complying users, leading to gaps in the collected data. HAR models not only need to adapt to diverse data but also provide accurate and reliable predictions in the face of missing or incomplete data.

This advancement underscores the necessity for HAR models that not only adapt to diverse data types, but also remain robust in the face of missing or incomplete data. Real-world scenarios often present challenges such as sensor malfunctions, battery depletion or user non-compliance, leading to gaps in data collection. Addressing these challenges is crucial for the development of HAR systems that can provide accurate and reliable activity recognition, even under less-than-ideal conditions.

This challenge becomes even bigger in multimodal HAR systems, where the integration of data from various sources is key to achieving high accuracy. Usually, a modality is defined as a set of sensor channels of a certain device located at a specific body position, like the wrist (smartwatch), head (earbuds or smart glasses), lower body (phone in the pocket), chest (heart rate monitor), and more. Developing methodologies that can intelligently handle missing data becomes crucial for the advancement of HAR technologies.

By exploring innovative approaches to model adaptation and enhancement, this work tries to pave the way for HAR systems that are not only technically proficient but also practically applicable in diverse and unpredictable real-world environments.

1.2 Contribution

In response to the challenges outlined in section 1.1, this thesis makes several contributions to the field of HAR, particularly focusing on the robustness of models in the face of missing sensor data. This research focuses on the strategic adaptation and improvement of a multimodal learning with missing modality approach, called Shared-Specific feature modeling (ShaSpec) [18], which is specifically tailored to the unique requirements of HAR.

The main contributions of this work are as follows.

First, the adaption of the ShaSpec framework for HAR. By customizing the ShaSpec framework to suit the nature of HAR data, this thesis demonstrates the model's enhanced capability to extract and leverage both shared and specific features from available sensor data, ensuring robust performance even when certain sensor channels are unavailable. This demonstrates the potential of ShaSpec to improve the accuracy of activity recognition in the face of missing sensor modalities.

Second, the development of specialized Encoders and Decoder architectures. Recognizing the need for a model architecture that can handle the complexities of multimodal sensor data, this work presents the design and implementation of a Shared Encoder, a Specific Encoder and Decoder components. These components are designed to optimize the processing of HAR data, facilitating effective learning from both complete and incomplete sensor datasets. The custom architectures play a crucial role in enabling the adapted ShaSpec model to efficiently interpret sensor data and make accurate activity recognition predictions.

Third, the extensive evaluation and benchmarking of the ShaSpec framework for HAR. Through rigorous experimentation on benchmark datasets, including DSADS and REALDISP, this thesis provides a comprehensive evaluation of the adapted ShaSpec model's performance. The research not only benchmarks the model against existing HAR methodologies, such as the AALH framework [6], but also investigates its resilience to various levels of missing data. These evaluations underscore the model's superiority in handling incomplete sensor data, setting new robustness standards for HAR systems.

Fourth, an ablation study further investigates the impact of individual components of the ShaSpec model on its overall performance. By systematically removing and analyzing the effects of the shared encoder and missing modality feature generation mechanism, the study offers insights into their contributions towards the model's robustness to missing data. This analysis reinforces the significance of each component and validates the model's design choices.

The contributions of this thesis extend beyond theoretical advancements, offering practical insights and methodologies that can be directly applied to enhance the robustness and reliability of HAR systems. By addressing the critical challenge of missing modalities, this research moves the field of HAR one step closer to realizing its full potential in real-world applications, promising more resilient, accurate, and practical activity recognition solutions.

1.3 Outline

This thesis is structured to provide a clear, comprehensive exploration of enhancing the robustness of HAR models against missing modalities, with a focus on the adaptation of the ShaSpec model.

It is organized into the following chapters.

Chapter 2 delves into the theoretical foundations and state-of-the-art research in Deep Learning for HAR, emphasizing the challenges posed by missing modalities. It reviews existing methodologies, including deep learning approaches and multimodal learning frameworks, highlighting their strengths and limitations. The discussion provides context for the research, establishing a baseline for the innovations introduced in this thesis.

Chapter 3 describes the adapted ShaSpec framework in detail, focusing on the development of a Specific Encoder, a Shared Encoder, and a Decoder architecture tailored for HAR. It elaborates on the model's design principles, intended to enhance robustness to missing data, and outlines the methodologies for feature extraction, fusion and decoding.

Chapter 4 covers the practical aspects of translating the theoretical model into a functional HAR system. It discusses the technological environment, chosen libraries, and the step-by-step process of model implementation, including the configuration of Shared and Specific Encoders and Decoder components.

Chapter 5 presents the experimental setup, datasets used, evaluation metrics, and the benchmarking process against existing HAR models. It includes an extensive evaluation of the adapted ShaSpec model's performance on benchmark datasets, with a detailed analysis of its robustness to missing modalities. The chapter also features an ablation study to assess the impact of individual model components on overall performance.

Chapter 6 concludes the key findings on research, reflecting on the significance of the contributions and their implications for the field of HAR. It also discusses potential avenues for future research, suggesting ways to further improve HAR model robustness, and how to harness ShaSpec's characteristics in already existing HAR models.

2. Background & Related Work

This chapter delves into the foundational research and recent advancements in HAR using Deep Learning (DL), focusing particularly on the challenges and solutions related to missing modalities in sensor data.

2.1 Deep Learning for HAR

2.1.1 Overview

DL has significantly advanced the capability of sensor-based activity recognition, a field central to developing smart environments and health monitoring systems. DL methods, particularly Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), have become integral in interpreting complex sensor data, due to their ability to extract hierarchical features and capture temporal dependencies. These architectures facilitate the automatic learning of discriminative features directly from raw data, a step change from traditional pattern recognition methods, which are highly dependent on human-crafted feature extraction.

The advent of Deep Convolutional Neural Networks (DCNN) [19] showcased the capabilities of CNNs for HAR tasks, underscoring their excellent potential in extracting local dependencies. However, to capture long-term temporal dependencies, CNNs usually need to be stacked very deep to obtain a larger receptive field.

RNNs face challenges with long-range dependencies due to the vanishing gradient problem [4], where gradients become too small to make significant changes in the model's parameter during backpropagation. This impedes the training of deep RNNs. Long Short-Term Memory (LSTM) networks [5] were introduced as a solution, utilizing gating mechanisms to control the flow of information and enabling them to learn long-term dependencies more effectively than traditional RNNs.

By combining CNNs' spatial feature extraction ability with RNNs' ability to handle sequential data, a series of CNN-RNN hybrid-models [14, 12, 21] were proposed, thereby creating a more efficient mechanism for capturing long-term dependencies.

In recent years, models based on self-attention [17] have risen to prominence, outperforming RNN-based models in their ability to grasp long-term temporal dependencies. The application of attention mechanisms in the realm of HAR has also been investigated [11, 10].

The integration of CNNs, RNNs, LSTMs, and attention-based models illustrates the field's dynamic progression, leading to advancements in accuracy and robustness, especially in dealing with high-dimensional and time-series sensor data.

2.1.2 Challenges

Despite the exceptional progress made by DL in HAR, several challenges persist, complicating the development of robust HAR systems. These challenges broadly encompass data diversity, model generalization and especially missing data (sensor channels or devices), in a context where multiple devices record sensor data at the same time. This thesis deals in particular with the latter challenge.

2.1.3 Case Study: Attend and Discriminate HAR Framework

The "Attend and Discriminate" HAR framework [1] marks a significant advancement in the use of DL for HAR using wearable sensors.

Architecture

The architecture of this model integrates convolutional layers at its base to extract primary features from sensor data. These features are then processed through attention layers, which selectively prioritize segments of data that are more indicative of certain activities. The final stage consists of classification layers that interpret these focused features to classify the activity, demonstrating a sophisticated approach to enhancing the accuracy of HAR systems.

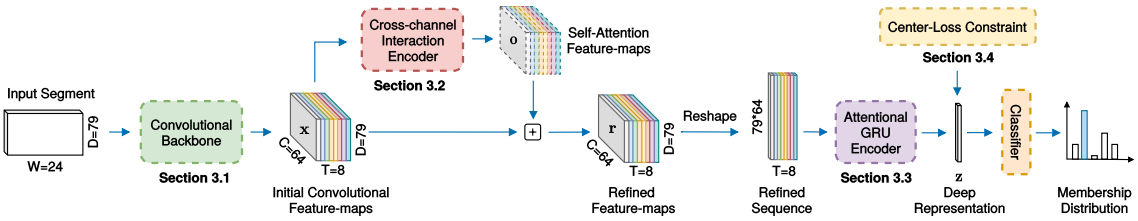

Figure 2.1: Overview of the Attend and Discriminate HAR framework.

Limitations

The model encounters limitations, when faced with missing data. It heavily depends on consistent and complete input from all sensors, and its performance can significantly diminish when faced with incomplete data. This presents a challenge in dynamic environments where sensor data might be missing or compromised, indicating a need for models that can robustly handle such inconsistencies in sensor inputs. Despite its strengths, the limitations of this framework in handling missing data highlight a critical area for further research and development. Addressing this limitation is essential for the advancement of HAR systems, ensuring their robustness and reliability in real world scenarios where sensor data may be incomplete. This underscores the need for models that can maintain high performance even under varied and less-than-ideal data conditions.

2.2 Multimodal Learning

Multimodal learning, a concept extensively explored in fields like computer vision and medical image analysis, involves integrating data from multiple sources or sensors to improve the accuracy and robustness of learning models. In these fields, multimodal learning leverages diverse data types like images, text, and audio, providing comprehensive insights that a single modality might miss. Most multimodal learning models assume that all modalities are completely represented. However, this is often not the case in practical scenarios.

In the context of HAR, a ”modality” typically refers to a set of sensors or wearable devices that collect data. These devices are designed for the purpose of recognizing, tracking, or analyzing human activities. Common modalities provide tri-directional acceleration, gyroscope and magnetic field measurements as well as orientation estimates in quaternion format (4D). Each modality yields a unique perspective and form of information pertinent to human movements or behaviors, playing a crucial role in the multimodal learning framework.

Multimodal learning within HAR capitalizes on the assortment of these data types to significantly enhance the accuracy and reliability of activity recognition systems. This enhancement is achieved through the comprehensive integration and analysis of information from a variety of sensor types or sources. Consequently, multimodal learning affords a more thorough understanding of human activities, surpassing the insights gained from systems reliant on single-modal data.

The integration of data from multiple modalities, particularly in configurations involving multiple sensors, substantially enriches the dataset available for HAR systems, facilitating more precise and reliable recognition of activities. By leveraging the complementary strengths of diverse sensor types and modalities, HAR systems are equipped to deliver a more complete depiction of human activities. This integrative approach markedly improves the system's capability in recognizing and differentiating between complex activities, thereby elevating the precision of activity recognition.

Approaches for Missing Modalities

In addressing the challenge of missing modalities in multimodal learning, various solutions have been explored in recent research. While considerable research has been directed towards accommodating modalities missing during testing, like Tsai et al. [16], the issue of incomplete training modalities remained less explored until recently.

2.2.1 AALH

A paper by Kang, Huang, and Zhang [6] presents a framework for HAR in scenarios where sensor data is incomplete. This framework, termed as Augmented Adversarial Learning framework for HAR (AALH), aims to create a generalizable model that can effectively work with various combinations of sensor placements and subject discrepancies.

Figure 2.2 illustrates the AALH framework along with its principal components designed to enhance HAR robustness in the face of missing sensor data. The key elements of this framework include:

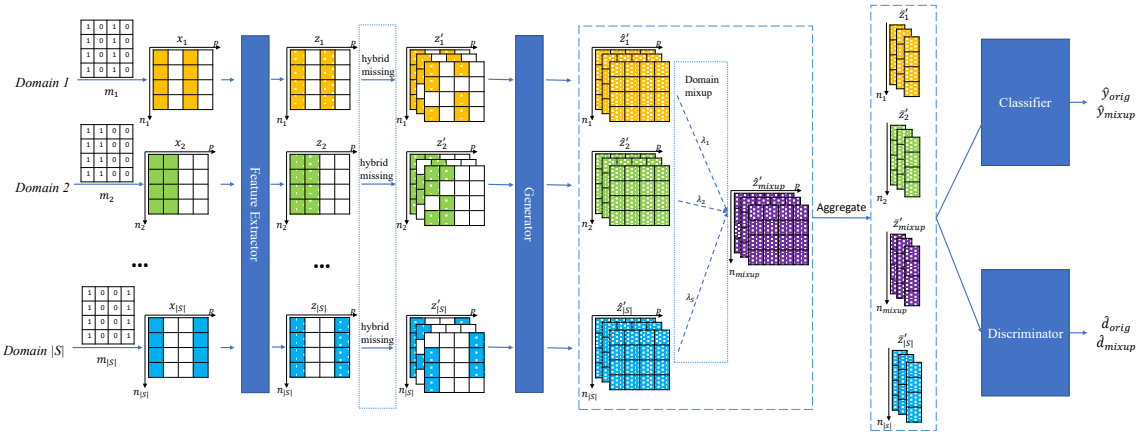

Figure 2.2: Overview of the AALH framework.

- Common Feature Extractor: Centralizes feature learning from different sensor domains.
- Hybrid Missing Strategy: Simulates missing sensor data by ignoring certain features, preparing the model for diverse data combinations.
- Augmented Adversarial Learning: Utilizes adversarial networks to generalize from artificial representations of data to handle missing sensors.
- Domain Mixup: Integrates data from different domains to enrich the model's exposure to various sensor configurations.
- Discriminator: A discriminator network assists in aligning features from different domains to a common latent space, encouraging domain invariance and enhancing class discrimination.

While AALH's methodical approach to HAR through adversarial learning is innovative, it raises concerns about the practicality and interpretability of the model. The complexity introduced by the adversarial learning components and the domain mixup might lead to difficulties in model convergence and scalability to larger datasets. Moreover, the framework's robustness in face of real-world sensor noise and the computational efficiency for deployment in on-device scenarios may be questioned. Although AALH successfully generates artificial data to address missing information, the method is not straightforward.

2.2.2 SMIL

The Severely Missing Modality (SMIL) model, introduced by Ma et al. [9], focuses on addressing challenges in multimodal learning when a significant portion of training data lacks modalities.

It is a method based on Bayesian meta-learning to achieve two objectives:

- *Flexibility*: Handling missing modalities during training, testing, or both.
- *Efficiency*: Learning from data, where up to 90% of training samples contain incomplete modalities.

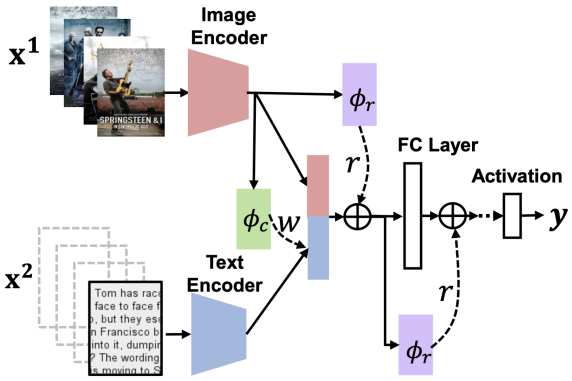

Figure 2.3: Overview of the SMIL training with severely missing modality.

Its training architecture is adept at managing severely missing modalities and uniformly achieving both objectives.

It utilizes a feature reconstruction network that leverages available modalities to approximate the missing ones, ensuring comprehensive feature space coverage for more effective learning. While SMIL represents a significant advance in tackling the challenge of incomplete training data, it is important to note that its specific design caters to particular tasks, which might not readily transfer to different domains that require a different balance of modality-specific and shared information.

SMIL ensures that the learning process can proceed even with severely incomplete modalities by generating a unified representation from available data. However, by not explicitly distinguishing between shared and unique features across modalities, it might not fully optimize the learning process for scenarios where recognizing and leveraging the distinctions between modality-robust and modality-specific features are crucial for performance [18]. This oversight can potentially limit the model's effectiveness in tasks where the interplay between these features significantly influences the outcome.

2.2.3 ShaSpec

Contrasting SMIL, the Shared-Specific feature modeling (ShaSpec) framework, proposed by Wang et al. [18], presents a simpler yet profoundly more effective approach. ShaSpec's ease of adaptability to a variety of tasks, including classification and segmentation, sets it apart from models like SMIL.

ShaSpec takes advantage of all available input modalities during training and evaluation by learning both features shared between modalities and features specific to each modality, learning a robust yet expressive representation. This dual-feature approach is key to ShaSpec's efficacy, particularly in HAR contexts, where understanding the interplay between shared and specific features is vital.

It is noteworthy that the original conceptualization of ShaSpec was not specifically designed for HAR data. Moreover, the ShaSpec paper lacks detailed insights into its encoder architectures, presenting an opportunity for exploration and customization.

As of the time of this thesis, the release of ShaSpec's code is restricted due to intellectual property considerations associated with the Imagendo project. Against this backdrop, the primary objective of this thesis is the strategic adaptation of

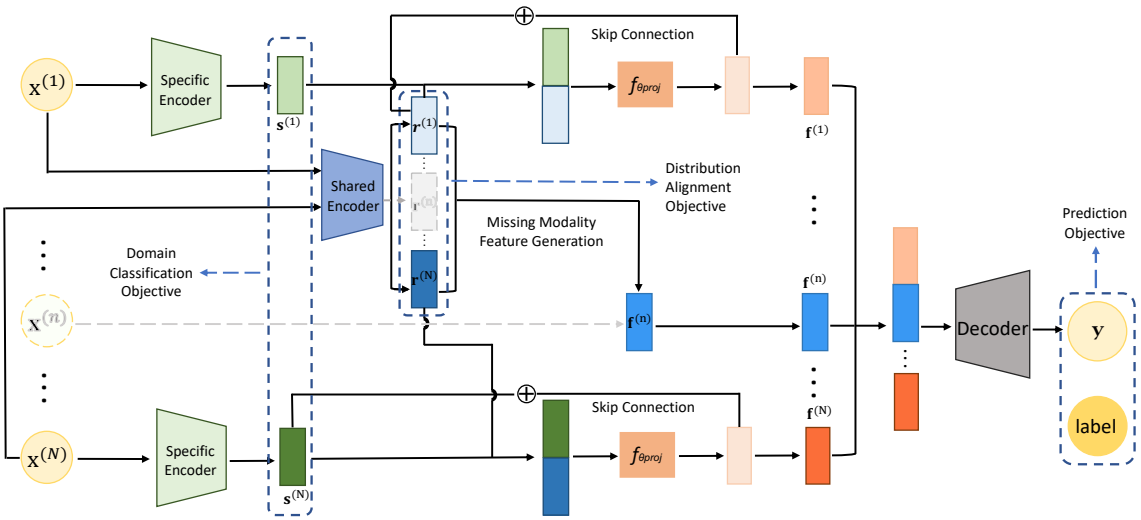

Figure 2.4: Overview of ShaSpec's architecture with missing modality feature generation.

ShaSpec for HAR, including the development of specialized encoder and decoder architectures to harness its full potential.

Comparing ShaSpec with AALH (see subsection 2.2.1), illuminates the evolution in multimodal learning strategies for managing missing modalities. The insights gained from analyzing AALH's adversarial learning approach enrich the foundation for developing a streamlined, efficient ShaSpec model tailored for HAR's dynamic requirements.

3. Methodology

This chapter describes the architecture of the ShaSpec components, which were specifically designed for HAR. While the overall structure of the proposed model is strongly based on the ShaSpec model [18] (see Figure 2.4), the details of encoder and decoder architectures constitute a novel contribution of this thesis.

3.1 Encoder

All available modalities are passed through one shared encoder and individual specific encoders to produce the shared and specific features. Both types of encoders share the same architecture, but the input and output shapes as well as the processing of weights differ significantly to accommodate their distinct functions. The shared encoder, on the one hand, focuses on extracting commonalities across modalities to form a unified representation, employing weights to emphasize features that are consistently present. The specific encoders, on the one hand, are tailored to highlight unique aspects of each modality, manipulating weights to accentuate modality-specific characteristics, ensuring that the combined feature space comprehensively captures both shared and specific elements of the input data.

3.1.1 Specific Encoder

The specific encoder takes raw input data $\mathrm{X}^{(i)} \in \mathbb{R}^{C \times T \times F}$, where (i) represents the ith modality. C denotes the number of sensor channels, T the temporal sliding window size, and F the number of filters. $F = 1$ for the raw data input, which has not undergone any processing. All available modalities are processed separately in this encoder.

Individual Convolutional Subnet

Four convolutional layers are applied for local context extraction. Each convolutional layer is activated by a ReLU or Tanh function. All individual convolutional layers have the same number of filters. Kernels only have a 1-dimensional structure along the temporal axis and kernel size is 5×1. The stride in each layer is set to 2 in order

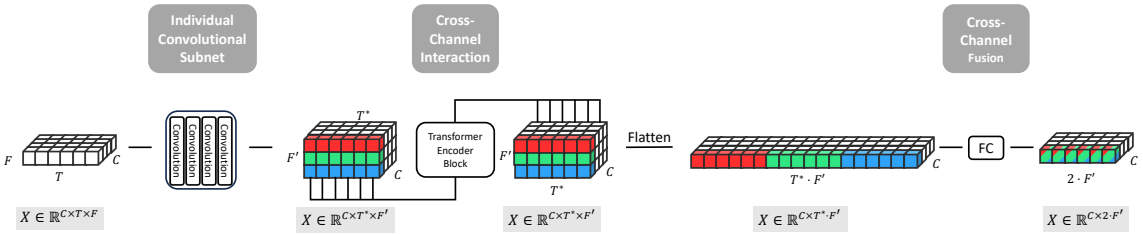

Figure 3.1: Overview of the specific encoder architecture.

to reduce the temporal dimension, i.e. halving the length with each convolutional layer. As a result, the output shape of this convolutional subnet is $\mathbb{R}^{C \times T^* \times F'}$, where T^* denotes the reduced temporal length and F' is the number of filters. All four convolutional layers have the same number of filters.

Transformer Encoder for Cross-Channel Interaction

A transformer encoder block [17] is utilized to learn the interaction across multiple sensor channels over time. It is specifically designed to interpret and encode the interactions occurring across the sensor channel dimension at each temporal step. Inspiration was drawn from Abedin et al. [1], who demonstrated the effectiveness of the self-attention mechanism in learning the collaboration between different sensor channels.

The transformer encoder block is composed of a scaled dot-product self-attention layer. For each time step in a sequence, the encoder extracts a slice of data corresponding to the time step. Then, it applies self-attention to this slice and refines the features by considering the interaction of this time step with all other time steps in the sequence. This iterative enhancement ensures each time step is contextually informed by the entire sequence's data, leading to a richer, more informed feature set.

Fully Connected Layer for Cross-Channel Fusion

To facilitate the fusion of learned features from various sensor channels, the tensor X with dimensions $\mathbb{R}^{C \times T^* \times F'}$ undergoes a flattening process, reshaping it to $\mathbb{R}^{C \times (T^* \cdot F')}$. Then, a fully connected (FC) layer is employed to perform a weighted summation across these features. The FC layer assigns distinct weights to different features within the same sensor channel, allowing for a nuanced integration of information.

As the final step, a specific feature is provided, denoted as $\mathrm{s}^{(i)}$ which has the shape $\mathbb{R}^{C \times (2 \cdot F')}$, effectively doubling the feature dimensionality.

All available modalities $i = 1, \cdots, N$ are passed through individual specific encoders to produce N specific features.

3.1.2 Shared Encoder

The shared encoder is responsible for extracting shared features among all available modalities. Unlike the specific encoder, which processes each modality independently, the shared encoder operates on a combined representation of all modalities.

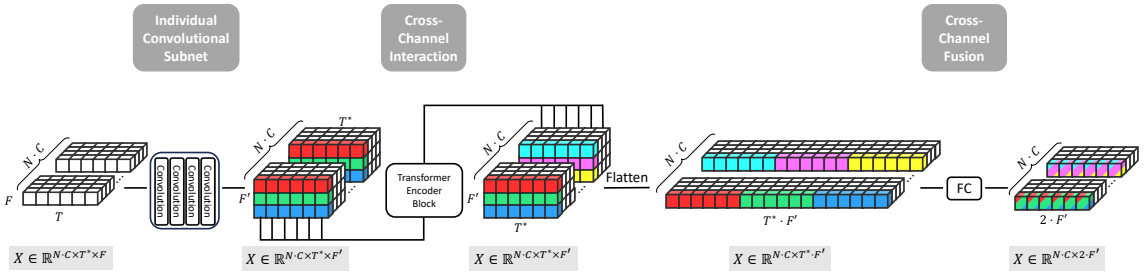

Figure 3.2: Overview of the shared encoder architecture.

All available modalities are concatenated along the C dimension, denoted by $\mathrm{X} \in \mathbb{R}^{N \cdot C \times T \times F}$, where N is the number of available modalities. This concatenated tensor is then processed through the same layers as the specific encoder. The key difference lies in the input dimensionality, which is expanded to accommodate all modalities.

As the final step, the output is then split back into N separate shared features, $\mathrm{r}^{(1)}, \cdots, \mathrm{r}^{(N)}$, each representing a common feature extracted across all available modalities.

3.2 Residual Block

The residual block fuses each specific feature $\mathrm{s}^{(i)} \in \mathbb{R}^{C \times 2 \cdot F'}$ with each shared feature $\mathrm{r}^{(i)} \in \mathbb{R}^{C \times 2 \cdot F'}$ by linearly projecting their concatenation $\mathrm{s}^{(i)}\mathrm{r}^{(i)} \in \mathbb{R}^{C \times 4 \cdot F'}$. Then, a skip connection adds the ith shared feature $\mathrm{r}^{(i)}$ to the output of the linear projection, aiding in better gradient propagation and network training to get the fused feature:

$$\mathrm{f}^{(i)} = f_{\theta_{\text{proj}}}(\mathrm{s}^{(i)}, \mathrm{r}^{(i)}) + \mathrm{r}^{(i)} \tag{3.1}$$

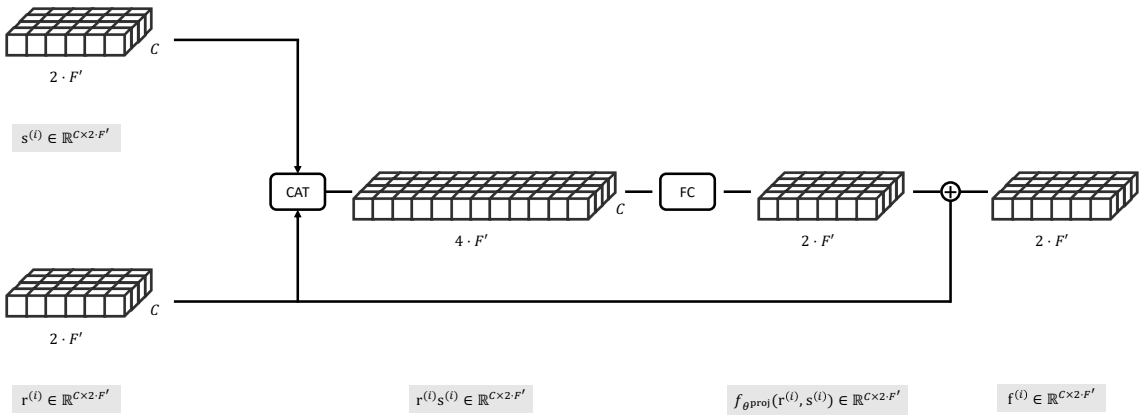

Figure 3.3: Overview of the residual block architecture.

3.3 Missing Modality Feature Generation

As described by Wang et al. [18], missing modalities are generated by taking the average of all other extracted shared features.

$$\mathrm{f}^{(n)} = \frac{1}{N-1} \sum_{i=1, i \neq n}^{N} \mathrm{r}^{(i)} \tag{3.2}$$

This results in a good representation for missing modalities. However, the more modalities are missing and need to be generated, the more homogeneous the fused features get.

3.4 Decoder

In the final step of the model, the prediction for each class is produced with a fully connected layer (FC layer) as the decoder.

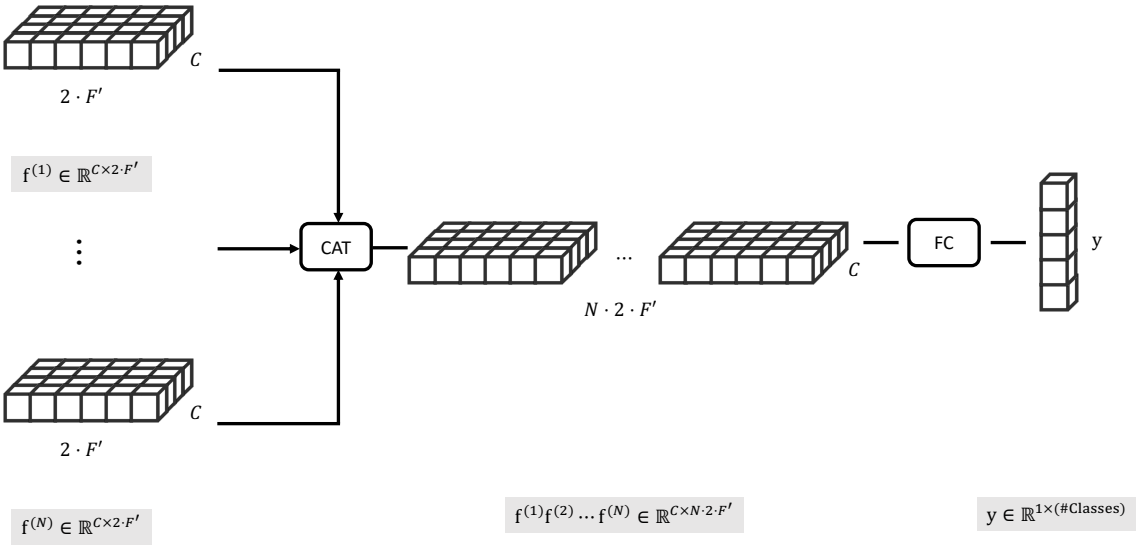

Figure 3.4: Overview of the decoder with FC layer.

In preparation for the decoding, the fused features are concatenated along the $2 \cdot F'$ dimension, resulting in a $\mathbb{R}^{C \times N \cdot 2 \cdot F'}$ shape. Before applying the FC layer, the concatenated fused features are flattened to $\mathbb{R}^{C \cdot N \cdot 2 \cdot F'}$.

Then, the FC layer projects the flattened feature vector to the desired output dimension $\mathbb{R}^{1 \times (\#Classes)}$, suitable for class labels.

4. Implementation

This chapter delves into the practical aspects of translating the theoretical methods discussed in chapter 3 into a functional model with Python. It outlines the technological choices made during development, the setup of the development environment, and the step-by-step implementation of the model.

4.1 Technological Environment and Setup

To facilitate the development and training of the model, the PyTorch framework was chosen for its comprehensive support for deep learning models and its ease of use [15]. Development began on a local environment using Python 3.10.12. and PyTorch 2.0.1. However, the final training sessions were conducted on a more powerful server, utilizing Python 3.9.13 and PyTorch 1.9.1 to ensure compatibility and performance optimization. This dual-environment approach allowed for rapid prototyping and debugging in a familiar development setup, while still being able to harness powerful computational resources for model training.

4.2 Model Implementation

The implementation of the model can be broken down into several key components, each contributing to the model's ability to process and learn from multimodal data effectively.

For clarity and simplicity in the graphical representations within chapter 3, the batch size dimension, denoted as B, has been implicitly set to 1. This decision was made to focus on the structural and functional aspects of the model's architecture. In practice, the model was trained with $B = 128$.

4.2.1 Base Encoder

The BaseEncoder class serves as the foundation for feature extraction, utilizing a sequence of convolutional layers for channel-wise feature extraction. It incorporates a self-attention mechanism to enhance cross-channel interaction and a fully connected layer for cross-channel fusion.

The self-attention implementation is heavily borrowed from the fastai deep learning library (https://github.com/fastai/fastai/blob/5c51f9eabf76853a89a9bc5741804d2ed4407e49/fastai/layers.py) and its used notation from Zhang et al. [20]. This encoder is critical for extracting meaningful features from its input. It also prevents duplicate code, since the Shared Encoder has the same architecture as the Specific Encoder.

4.2.2 Shared and Specific Encoders

Two specialized encoders, Shared Encoder and Specific Encoder, build upon the BaseEncoder. The Shared Encoder extracts common features across all available modalities, while the Specific Encoder is tailored to capture unique characteristics of each available modality.

This dual approach enables the model to learn both shared and specific representations, crucial for handling multimodal data effectively.

4.2.3 Residual Block and Feature Generation

The Residual Block allows the integration of specific and shared features, facilitating a richer representation by combining the strengths of both encodings. For scenarios with missing modalities, the Missing Modality Feature Generation class innovatively generates features, ensuring the model's robustness to incomplete data.

4.2.4 Decoder

Finally, the Decoder class is responsible for translating the fused feature representations into predictions. It plays a crucial role in the model's ability to make sense of the processed data and output meaningful results.

4.3 Code Availability

To ensure transparency and facilitate further research, the entire codebase, including the detailed implementation of the ShaSpec model for HAR, is made publicly available on GitHub at https://github.com/probstlukas/shaspec-har.

5. Experiment

This chapter describes the evaluation methods, the baseline model, benchmark HAR datasets, model training, results and ablation studies.

5.1 Setup

5.1.1 Evaluation Rules and Metrics

The evaluation of the model's performance, particularly in terms of intersubject generalization, was conducted using Leave-One-Subject-Out (LOSO) Cross-Validation (CV) on two distinct datasets. This approach involves excluding data from certain subjects at a time from the training set and using it for testing, thereby ensuring that the model's ability to generalize across different subjects can be accurately assessed. To quantify the model's performance, two evaluation metrics were employed: accuracy and the weighted F1-score. These metrics were chosen because they provide a comprehensive view of the model's effectiveness, taking into account both the precision and recall of predictions across various classes, and the overall percentage of correct predictions, respectively.

To prevent the model from memorizing fixed patterns of missing modalities, they are randomly omitted based on a predefined miss rate for each experiment. This approach introduces a dynamic element to the data, simulating a variety of missing data scenarios and significantly enhancing representation diversity. By incorporating this randomness, the model is trained to adapt to various combinations of available data, mirroring real-world situations where data may be missing unpredictably. This method aligns with the evaluation's goal to rigorously test the model's ability to generalize across different subjects under varying conditions of data completeness.

The CV experiments were run twice, using different pseudo-random number seeds (1 and 42) for each experiment. This strategy helps to mitigate the potential impact of random variation in the selection of training and testing sets on the model's performance evaluation. For each run, the accuracy and weighted F1-score were taken, and then averaged across both runs.

The average scores are reported in Table 5.2 as the final performance measures of the model, providing a reliable estimate of its ability to generalize across subjects.

5.1.2 Baseline

The results of the AALH framework [6] by Kang, Huang, and Zhang (see subsection 2.2.1) will serve as the baseline. They evaluated their model on two datasets: the Daily and Sports Activities Data Set (*DSADS*) [3] and the REAListic sensor DISPlacement activity recognition dataset (*REALDISP*) [2], and reported both the accuracy and weighted F1 score.

To measure the performance with missing modalities, they used sensor miss rates of 0.0, 0.2, 0.4, and 0.6 for DSADS, and 0.1, 0.3, 0.5, and 0.7 for REALDISP.

Unfortunately, the authors of the AALH framework did not make their code available for public use. As a result, direct comparison between the outcomes of this thesis and the AALH framework requires recreating the experimental setup as closely as possible to ensure accuracy and fairness in the comparison. Therefore, the comparison should still be approached with a degree of caution due to potential differences in setup and configuration details.

5.1.3 Datasets and Data Preprocessing

To rigorously assess the proposed model's performance, it underwent evaluation on two widely used HAR benchmark datasets [3, 2]. Key details of these datasets are concisely summarized in Table 5.1.

Dataset	Freq. [Hz]	#Subjects	#Activities	#Channels	#Modalities	#SW
DSADS	25	8	19	9 A, G, M	5	5 s
REALDISP	50	17	33	13 A, G, M, Q	9	3 s

Table 5.1: Overview of the datasets used in experiments. #Channels = number of sensor channels *per* modality, A = accelerometer, G = gyroscope, M = magnetometer, Q = orientation estimates in quaternion format (4D) and #SW = sliding window length.

DSADS

The dataset comprising eight subjects is divided into four groups, with each group holding two subjects. 90% of the data is used for training purposes, while the remaining 10% is reserved for validation.

REALDISP

The dataset covers nine body positions (modalities), 17 subjects, and 33 activities. For the activity column, zero denotes that no activity label is present. To reduce noise and irrelevant information, data which does not have an activity label was dropped. Only the first 16 subjects are utilized and divided into four groups, each containing four subjects. One group belongs to the test data, while the other three groups serve as the source domains. 90% of the data is used for training purposes, while the remaining 10% is reserved for validation. There are three sensor displacement scenarios in the original dataset: *ideal-placement*, *self-placement*, and *mutual-displacement*. To relieve training efforts, only the scenario for ideal-placement sensor data is being used. In the ideal-placement scenario all sensors are positioned by the instructor to predefined locations within each body part.

5.1.4 Training

For the model training, the Adam optimizer [7] is used, renowned for its effectiveness in handling sparse gradients and adaptive learning rates. The training starts with an initial learning rate of 10^{-3}, which is systematically decreased by 10% every 20 epochs to fine-tune the network's convergence. Training lasts for a maximum of 60 epochs, with an early stopping mechanism activated if there's no improvement in validation metrics for 15 consecutive epochs, ensuring efficiency by preventing overfitting and unnecessary computations. The batch size is set to 128.

At every epoch, the data is reshuffled. This shuffling is beneficial as it ensures that each batch seen by the network during training is different, preventing the model from memorizing the order of the training data and hence promoting a more generalized learning. By introducing randomness in the order of examples, it helps in breaking any correlations present in the sequence of training data, leading to more robust learning outcomes.

The loss function employed for both training and validation is the Cross-Entropy (CE) loss, a standard choice for classification problems. The CE loss effectively measures the difference between the predicted probability distributions and the actual class distributions and guides the model to make more accurate predictions.

For the encoder, detailed in section 3.1, the configuration includes setting the filter number to 64, which is the number of output filters in the convolution layer, and a filter size of 5, determining the extent of each filter's reach. The self-attention divisor is set to 8, optimizing the balance between computational efficiency and the capacity to model complex interactions in the data. The ReLU activation function [13] is selected for its effectiveness in introducing non-linearity, promoting sparse activations that contribute to the network's efficiency and interpretability.

All experiments were conducted on a cluster with NVIDIA Tesla V100 SXM2 32GB GPU.

5.2 Results

Following the AALH setup as closely as possible, the proposed model was trained on DSADS and REALDISP datasets with varying miss rates.

5.2.1 Performance Benchmarking Against AALH

Comparison with the AALH framework reveals distinct performance characteristics in response to varying levels of missing data. Specifically, AALH's performance deteriorates sharply with missing data while ShaSpec maintains a steadier level of performance under similar conditions, showing less sensitivity to missing data rates. This resilience makes ShaSpec a robust alternative for dealing with incomplete data, though it does not always match AALH's effectiveness when most data is present.

A direct comparison of the results with the AALH framework is shown in Table 5.2.

Figure 5.1 and Figure 5.2 show the comparison of accuracy, with a visible reduction in the performance gap to AALH at higher miss rates.

Dataset	Miss Rate	ShaSpec		AALH	
		Acc	F1	Acc	F1
DSADS	0.0	0.862(0.037)	0.857(0.038)	0.934(0.016)	0.933(0.017)
	0.2	0.867(0.028)	0.857(0.031)	0.905(0.042)	0.900(0.045)
	0.4	0.878(0.037)	0.870(0.037)	0.874(0.039)	0.866(0.044)
	0.6	0.851(0.023)	0.838(0.025)	0.762(0.084)	0.754(0.089)
	Average	0.865(0.031)	0.856(0.033)	0.869(0.034)	0.863(0.037)
REALDISP	0.0	0.934(0.027)	0.930(0.032)	0.979(0.006)	0.978(0.006)
	0.1	0.929(0.032)	0.925(0.036)	0.971(0.008)	0.971(0.008)
	0.3	0.934(0.024)	0.934(0.024)	0.955(0.016)	0.955(0.016)
	0.5	0.919(0.019)	0.918(0.020)	0.908(0.027)	0.906(0.028)
	0.7	0.829(0.035)	0.826(0.036)	0.753(0.040)	0.746(0.042)
	Average	0.909(0.027)	0.907(0.030)	0.913(0.012)	0.911(0.012)

Table 5.2: Comparison of ShaSpec and AALH framework performance on DSADS and REALDISP datasets. (Values in the cells represent "mean(std)"; F1 refers to the weighted F1-score.)

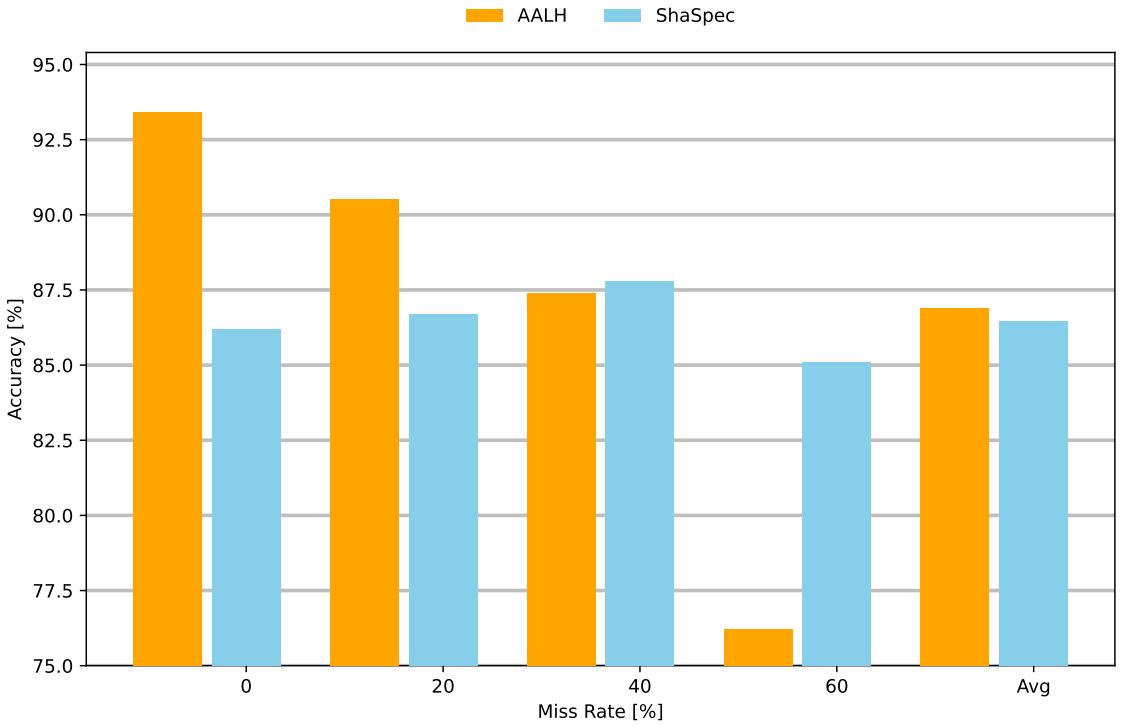

Figure 5.1: Model performance of ShaSpec and AALH with varying rates of missing modality data on DSADS.

DSADS

For DSADS, the performance of ShaSpec is not only sustained, but even improves until a 40% miss rate. This unexpected increase could be attributed to the model's ability to leverage the remaining modalities more effectively, possibly through enhanced feature extraction and representation learning that compensates for the missing data. With 60% of data missing, the model still achieves 98.72% of the perfor-

mance compared to when full data is available. It can maintain high accuracy even when a significant portion of data is absent.

REALDISP

The model demonstrates a smaller performance gap to AALH when evaluated on the REALDISP dataset compared to the DSADS dataset. This observation suggests several key insights into the model's behavior and its interaction with different types of datasets. For the REALDISP dataset, which encompasses a broader range of subjects, activities, sensor channels, modalities and overall more complex data types, the model's ability to maintain closer performance levels to the baseline, even as the miss rate increases, indicates its effective handling of richer information. It may facilitate more robust feature extraction and generalization capabilities.

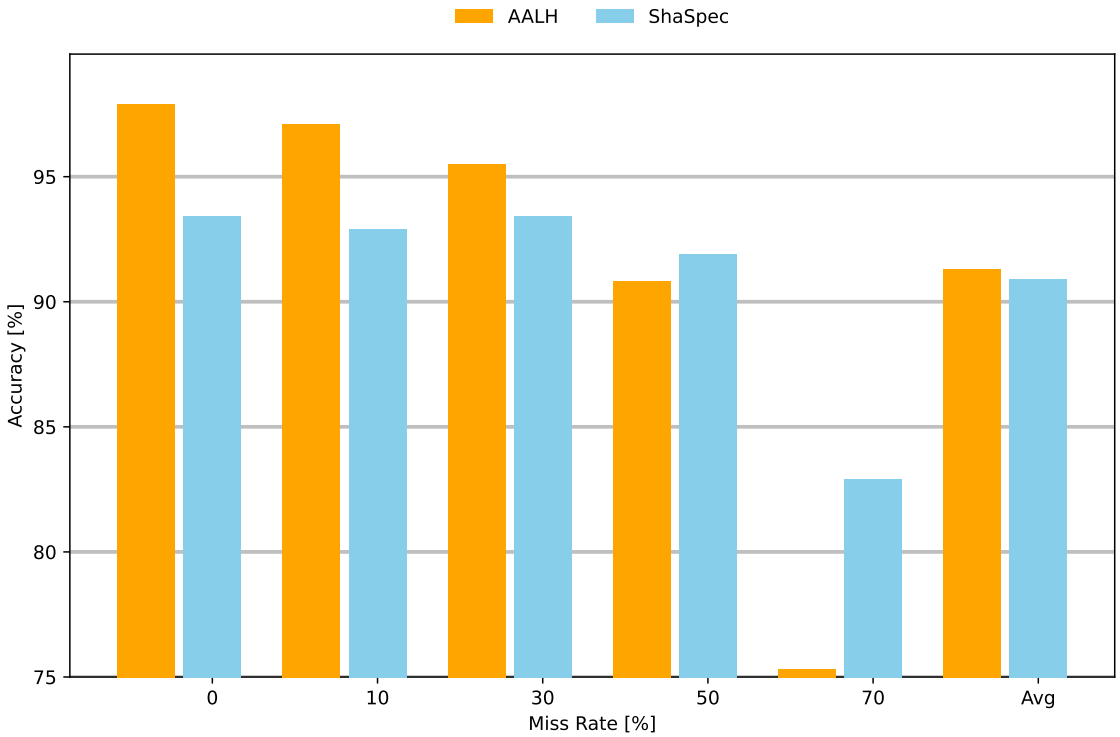

Figure 5.2: Model performance of ShaSpec and AALH with varying rates of missing modality data on REALDISP.

5.2.2 Robustness Index: A Relative Performance Metric

Assessing model performance through direct comparison of accuracy (see subsection 5.2.1) can be misleading, when experimental setups, model training or fine-tuning differ significantly. The impossibility of replicating the AALH setup precisely due to undisclosed details makes a nuanced approach that emphasizes relative changes in performance across varying miss rates necessary.

A fair and insightful comparison is obtained by considering a Robustness Index (RI), which quantifies the relative performance retention of a model at different miss rates compared to its performance with complete data (0% miss rate).

The Robustness Index RI_i for a given condition i is computed as:

$$\mathrm{RI}_i = \left(\frac{P_i}{P_{baseline}}\right) \cdot 100 \tag{5.1}$$

where P_i is the model's performance at the ith miss rate and $P_{baseline}$ is the performance with full data.

This metric illuminates the true superiority of ShaSpec, distinguishing it as the more stable and robust model, particularly in demanding conditions characterized by high rates of missing data.

The Robustness Index comparison for each dataset can be found in Figure 5.3 and Figure 5.4.

DSADS

ShaSpec consistently outperforms AALH, with an average robustness score that is 7.25% higher.

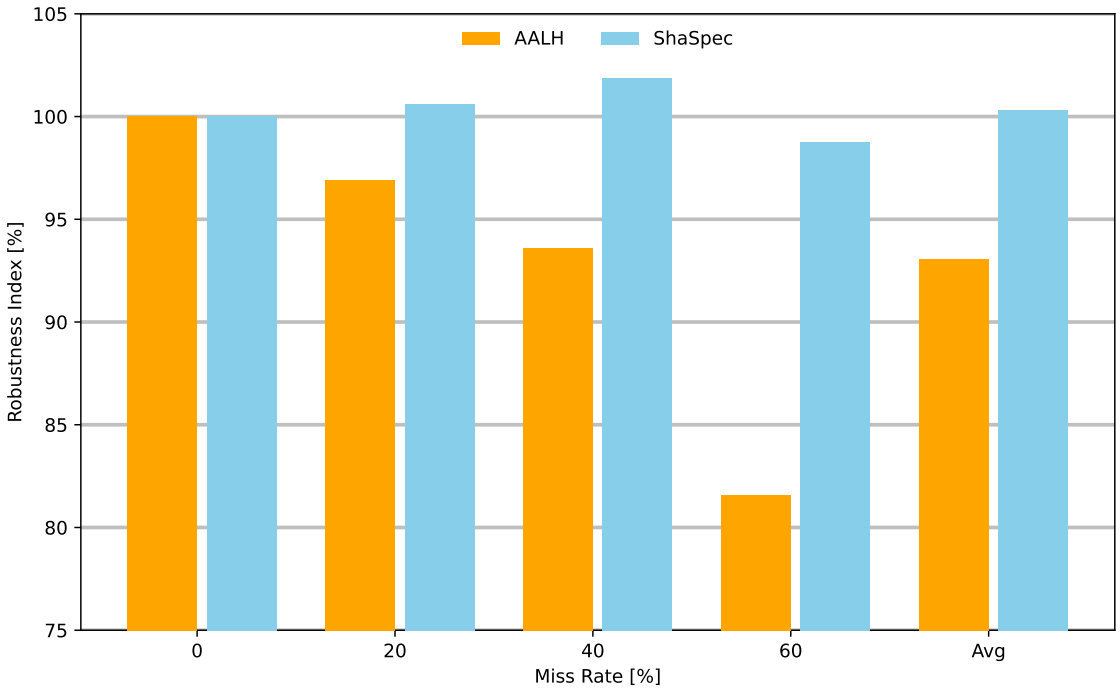

Figure 5.3: Robustness index comparison of ShaSpec with AALH for DSADS.

REALDISP

A similar trend can be observed for REALDISP, with ShaSpec surpassing AALH in each condition and achieving an average robustness score that is 4.06% higher.

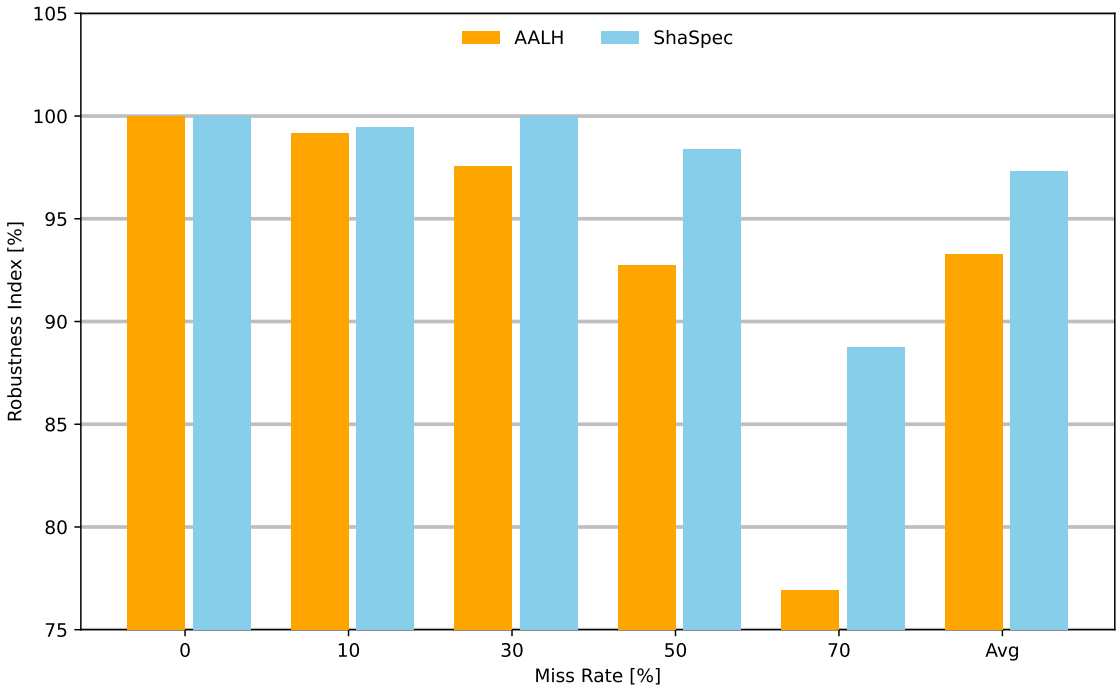

Figure 5.4: Robustness index comparison of ShaSpec with AALH for REALDISP.

5.3 Ablation Study

The purpose of the ablation study is to systematically observe the performance impact when selectively removing components from the model. This method allows for the isolation and understanding of the contribution of individual components to the overall model performance. To examine the impact of different components independently, two ablation studies were conducted.

In these studies, besides omitting a component, the missing modality features are filled with generated noise features. Specifically, noise generated from a normal (Gaussian) distribution is employed to compare randomness to the regular generation method, as described in section 3.3 (see Equation 3.2).

Without loss of generality, let the nth modality $\mathrm{x}^{(n)}$ be missing. The missing modality is then modeled to follow a standard normal distribution:

$$\mathrm{f}^{(n)} \sim \mathcal{N}(0, 1) \tag{5.2}$$

This introduces a level of randomness that is more challenging for the model to inadvertently learn or adapt to, ensuring that any performance change can be more directly attributed to the absence of the specific component under study.

A more naive approach might involve the use of zero features to indicate missing information. While intuitively appealing for their simplicity, they are not inherently understood by neural network models as ”missing” or ”lacking information.” Instead, models can learn to associate zero values with specific outcomes or states, inadvertently incorporating these values into their decision-making process. This would lead to misleading results in these ablation studies, as the model might adjust its weights

to exploit the presence of zeros artificially, rather than dealing with the absence of genuine information.

By utilizing random noise features, the integrity of the ablation study is preserved, ensuring that variations in the model's performance are attributable to the absence of the intended component rather than the model's adaptation to easily identifiable placeholders like zero features. This method allows for a more accurate assessment of the importance and contribution of each component to the model's overall effectiveness.

5.3.1 Performance Without Missing Modality Feature Generation

To assess what difference the feature generation for missing modalities makes when modalities are unavailable, the Missing Modality Feature Generation component was dropped.

Omitting this part should result in reduced performance when faced with missing data, demonstrating the importance of this component in handling incomplete data. When the model is trained with full data, the performance should stay the same.

To test this conjecture, the model was adapted to operate without the Missing Modality Feature Generation step. During training and testing, instead of generating features for missing modalities, the model is forced to make predictions with whatever information is available.

Dataset	**Miss Rate**	**ShaSpec**	
		Acc	**F1**
DSADS	0.0	0.862(0.037)	0.857(0.038)
	0.2	0.880(0.014)	0.871(0.013)
	0.4	0.855(0.016)	0.844(0.021)
	0.6	0.836(0.046)	0.830(0.047)
	Average	0.858(0.028)	0.851(0.030)
REALDISP	0.0	0.934(0.027)	0.930(0.032)
	0.1	0.930(0.031)	0.926(0.036)
	0.3	0.915(0.020)	0.912(0.023)
	0.5	0.879(0.031)	0.876(0.033)
	0.7	0.768(0.039)	0.766(0.040)
	Average	0.885(0.029)	0.882(0.033)

Table 5.3: Performance of ShaSpec without the missing modality feature generation on DSADS and REALDISP datasets. (Values in the cells represent "mean(std)"; F1 refers to the weighted F1-score.)

5.3.2 Performance Without Shared Encoder

To evaluate the contribution of the Shared Encoder to the model's ability to recognize activities from multimodal data, it is removed completely. This also means that the Residual Block and Missing Modality Feature Generation are omitted.

If the Shared Encoder significantly contributes to performance, one would expect that its removal should lead to a major decrease in performance as the model can no longer leverage shared representations across modalities.

For this ablation study, the architecture was modified to bypass the Shared Encoder and train the model using only the specific features for each modality. The results are shown in Table 5.4.

Dataset	Miss Rate	ShaSpec	
		Acc	F1
DSADS	0.0	0.883(0.023)	0.878(0.025)
	0.2	0.843(0.016)	0.831(0.022)
	0.4	0.837(0.036)	0.828(0.040)
	0.6	0.810(0.041)	0.800(0.044)
	Average	0.843(0.029)	0.834(0.033)
REALDISP	0.0	0.940(0.023)	0.939(0.025)
	0.1	0.936(0.025)	0.933(0.030)
	0.3	0.906(0.029)	0.903(0.031)
	0.5	0.881(0.035)	0.878(0.036)
	0.7	0.782(0.044)	0.780(0.044)
	Average	0.889(0.031)	0.887(0.033)

Table 5.4: Performance of ShaSpec without the shared encoder on DSADS and REALDISP datasets. (Values in the cells represent "mean(std)"; F1 refers to the weighted F1-score.)

A visual comparison of the performance for ShaSpec, ShaSpec without missing modality feature generation, and ShaSpec without shared encoder, is shown in Figure 5.5 and Figure 5.6.

DSADS

Figure 5.5 shows that, on average, there is a significant decrease in performance when integral components of the ShaSpec are removed. Ablating the Missing Modality Feature Generation leads to a modest performance decrease of 0.63% on average. Despite this, with a miss rate of 60%, it accounts for 1.5%, which is not insignificant. Training the model without the Shared Encoder results in an additional decrease of 1.49% on average, cumulating in a total decrease of 2.12% compared to the full model.

Interestingly, the model without the Shared Encoder surpasses the complete ShaSpec by 2.1% at a 0% miss rate. This might imply that for certain datasets or under specific conditions, the complexity introduced by the Shared Encoder could potentially be counterproductive, leading to overfitting or less optimal feature combinations for the task at hand. Without the Shared Encoder, the model relies on simpler, but possibly more relevant feature interactions that are more suited for the classification task with the full data available.

An unexpected finding is the performance of the model without the generation of missing modality features at a miss rate of 20%, with an increase of 1.3%. As the missing modalities are not generated for this model, but replaced by random noise, a drop in performance was to be expected, as it can be observed for the REALDISP dataset in Figure 5.6.

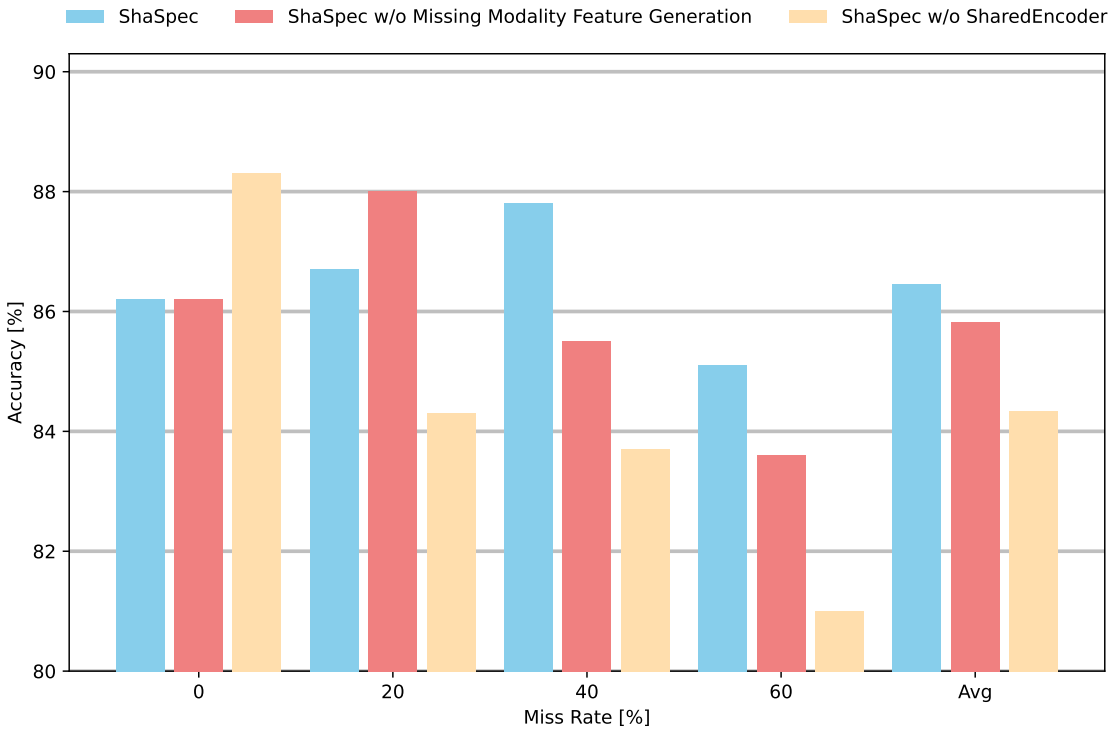

Figure 5.5: Comparison of the full ShaSpec model performance against its ablated variants on the DSADS dataset.

REALDISP

The ablation study on the REALDISP dataset, as depicted in Figure 5.6, reveals a decline in accuracy when the Missing Modality Feature Generation is omitted. Surprisingly, the removal of the Shared Encoder yields a marginally superior performance on average than the model lacking only the Missing Modality Feature Generation. This unexpected improvement may be linked to the model's ability to concentrate on the specific features for each modality, which seems to work better in this case.
Nonetheless, a holistic view of the results confirm that the ShaSpec model, with all its components operational, is optimally equipped to handle the complexities of varying miss rates.

As demonstrated by the ablation study, the complete ShaSpec model is the most capable, particularly under high miss rates. The significance of both the Missing Modality Feature Generation and the Shared Encoder is highlighted in their contribution to the model's overall accuracy.

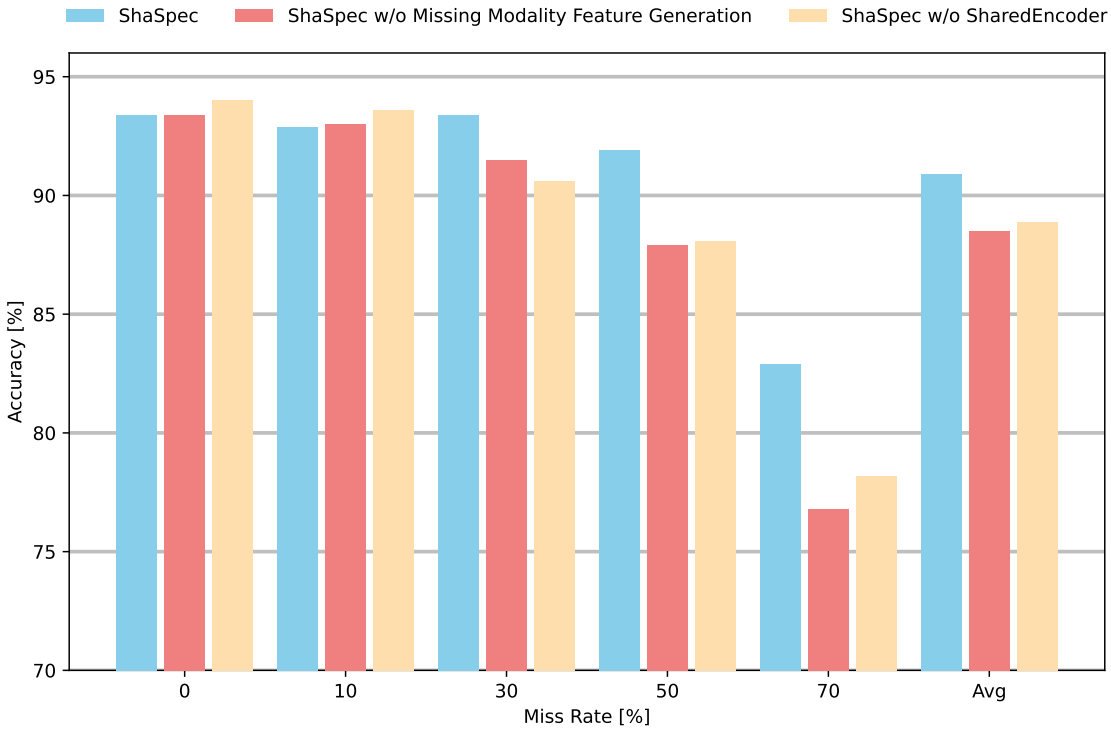

Figure 5.6: Comparison of the full ShaSpec model performance against its ablated variants on the REALDISP dataset.

6. Conclusion and Future Work

6.1 Conclusion

This thesis evaluates a shared-specific feature modeling (ShaSpec) method designed for HAR in respect of robustness for missing modalities. The proposed model is compared to an HAR-specific framework called AALH, which also generates artificial data for missing modalities. The proposed ShaSpec for HAR method outperforms AALH in all cases in terms of relative performance robustness. However, ShaSpec performs worse than AALH with full data and low miss rates, which could be attributed to a different setup or model training.

Experimental results on two multi-sensor, multi-subject datasets demonstrate the superiority of ShaSpec for high miss rates. For the first dataset, it achieves an average accuracy of 85.1% when the miss rate is as high as 0.6, surpassing the baseline model by 8.9%. For the second dataset, it achieves an average accuracy of 82.9% when the miss rate is as high as 0.7, surpassing the baseline model by 7.6%. This significant improvement in accuracy underscores the effectiveness of ShaSpec in handling complex datasets.

6.2 Other Potential Research

6.2.1 Different Shared Encoder Type

In the current methodology, the Shared Encoder takes a concatenated tensor which consists of all available modalities as input. After the extraction of features, the output is then split back into N separate shared features, $\mathrm{r}^{(1)}, \ldots, \mathrm{r}^{(N)}$.

Another approach would be to apply a weighting mechanism to integrate features from different modalities, as depicted in Figure 6.1. Instead of directly concatenating inputs, this approach would compute a weighted sum of the modalities' features, giving different importance to each modality based on their relevance in capturing shared information. This weighted sum is then processed through the Shared Encoder layers to extract common features. The output remains the same, with N shared features.

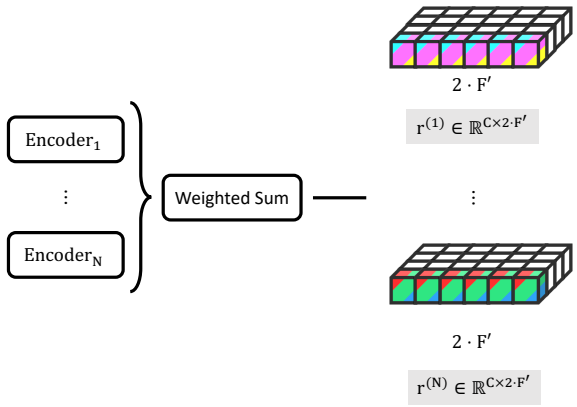

Figure 6.1: Overview of the shared encoder with weighted sum approach.

Due to time constraints, this method could not be tested. However, it can be expected that the Shared Encoder with concatenated modalities performs better than the variant with weighted sum, regardless. With concatenated modalities, the Shared Encoder takes information from all available modalities at the same time, so that they can have a more nuanced interaction with each other.

6.2.2 Improving Missing Modality Feature Generation

The process for generating features when a modality is missing involves averaging all other extracted shared features. While this method gives a good representation for missing modalities, the more modalities that are missing and need to be generated, the more homogeneous the fused features become, potentially leading to a loss in distinctiveness and performance degradation.

To improve and extend this method, a few strategies can be considered.

- **Weighted averaging**: Instead of a simple average, a weighted average of the features could be used, where the weights are learned during training. This can potentially give more importance to features that are more informative for the missing modality.
- **Incorporate uncertainty**: Uncertainty of the generated features could be modeled, potentially by using Bayesian methods or by learning a distribution over the possible values of the missing features.
- **Auxiliary loss functions**: Auxiliary loss functions that penalize homogeneity and encourage diversity in the generated features could be introduced.
- **Modality-specific generators**: Instead of a single process to generate missing features, multiple generators with each specializing in generating features for specific modalities could be used.
- **Feature attention mechanisms**: Attention mechanisms that can selectively focus on the most relevant features from the available modalities for generating the missing ones could be implemented.

Each of these methods or a combination of them could potentially improve the feature generation process for missing modalities.

When deciding whether to use these strategies, one must balance the potential benefits of improved model performance against the increased computational demands and complexity.

ShaSpec in its current form offers a simple but effective architecture, and the introduction of new components could negate these inherent advantages.

6.2.3 Utilization of Shared Features at Specific Miss Rates

The ablation studies discussed in section 5.3 highlight the potential for a model that relies solely on specific features before the miss rate reaches a certain threshold. Such a model could dynamically switch its operational mode from using exclusively specific features to both shared and specific features as the miss rate increases. This approach would capitalize on the strengths observed in the ablated version of ShaSpec that lacks a Shared Encoder, which demonstrated an unexpected increase in accuracy at a 0% miss rate.

This proposed model would employ a conditional architecture, where the reliance on shared features becomes predominant at predetermined miss rates, informed by empirical evidence from the ablation studies. The advantage of such a model lies in its adaptability; it would not only provide a robust solution under conditions of substantial data loss but also maintain high performance when full data is available.

Implementing this conditional model would involve developing a mechanism to detect the miss rate in real-time and adjust the model's feature utilization accordingly. This mechanism could be as simple as a rule-based switch or as complex as a learned function that predicts the optimal mode of operation based on the incoming data.

The introduction of this conditional model holds promise for practical applications where sensor data may frequently be compromised. It would prevent a drastic decrease in performance in response to missing modalities, providing a more reliable and robust tool for HAR in variable conditions.

6.2.4 Fine-Tuning

Fine-tuning is a critical process in DL, allowing a model to adjust its weights to better fit the specific nuances of a given dataset. In the case of ShaSpec, fine-tuning could involve the calibration of hyperparameters such as learning rate, number of epochs, weight decay, or batch size. In addition, more subtle adjustments in the Encoder architecture like the tuning of layer-specific parameters can be made.

Furthermore, fine-tuning can be extended to encompass the model's reaction to varying rates of missing data. By systematically adjusting the model at different miss rates, it is conceivable that performance could be further optimized for each scenario.

Fine-tuning ShaSpec would necessitate an iterative process, potentially involving a grid search or Bayesian optimization methods to explore the hyperparameter space efficiently. Given the complex nature of multimodal datasets and the nuances of missing data, this fine-tuning stage could significantly enhance the model's performance, especially in edge cases where standard training procedures might fall short.

6.2.5 Integration With Existing HAR Frameworks

A promising direction for future research lies in the integration of ShaSpec's robust architecture with pre-established models in the domain of HAR. Given the enhanced performance and stability of ShaSpec under varying miss rates, its components could be instrumental in reinforcing the resilience of existing HAR frameworks.

This hybridization strategy could take several forms.

- **Feature generation module**: Introducing ShaSpec's Missing Modality Feature Generation methodology to existing models could provide them with an improved ability to handle data sparsity.
- **Shared encoder mechanism**: Incorporating ShaSpec's Shared Encoder could enable models to better capitalize on commonalities across different modalities, potentially enhancing their feature extraction processes.
- **Dynamic feature integration**: Existing models could benefit from a dynamic feature integration approach that adapts the reliance on shared versus specific features based on real-time data completeness, inspired by ShaSpec's methodology.

The potential research would involve a systematic investigation into the compatibility of ShaSpec's components with other models, assessing how such integration affects overall performance. Additionally, it would require the development of adaptable interfaces within the existing frameworks to accommodate the integration seamlessly. This proposed symbiosis between ShaSpec's robust elements and established models could mark a significant advancement in the creation of versatile and reliable HAR systems, capable of maintaining high accuracy in the face of incomplete or missing sensor data.

The undertaking of this initiative could not only validate ShaSpec's contributions but also stimulate collaborative advancements in the HAR field, pushing the boundaries of what's achievable with current technology.

Bibliography

[1] Alireza Abedin et al. "Attend and Discriminate: Beyond the State-of-the-Art for Human Activity Recognition Using Wearable Sensors". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5.1 (2021).

[2] Oresti Banos, Mate Toth, and Oliver Amft. *REALDISP Activity Recognition Dataset*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5GP6D. 2014.

[3] Billur Barshan and Kerem Altun. *Daily and Sports Activities*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5C59F. 2013.

[4] Y. Bengio, P. Simard, and P. Frasconi. "Learning long-term dependencies with gradient descent is difficult". In: *IEEE Transactions on Neural Networks* 5.2 (1994), pp. 157–166.

[5] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Comput.* 9.8 (1997), 1735–1780. ISSN: 0899-7667.

[6] Hua Kang, Qianyi Huang, and Qian Zhang. "Augmented Adversarial Learning for Human Activity Recognition with Partial Sensor Sets". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6.3 (2022).

[7] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG].

[8] Oscar D. Lara and Miguel A. Labrador. "A Survey on Human Activity Recognition using Wearable Sensors". In: *IEEE Communications Surveys & Tutorials* 15.3 (2013), pp. 1192–1209.

[9] Mengmeng Ma et al. "SMIL: Multimodal Learning with Severely Missing Modality". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 3. 2021, pp. 2302–2310.

[10] Saif Mahmud et al. "Human Activity Recognition from Wearable Sensor Data Using Self-Attention". In: *CoRR* abs/2003.09018 (2020). arXiv: 2003.09018.

[11] Vishvak S. Murahari and Thomas Ploetz. "On Attention Models for Human Activity Recognition". In: *CoRR* abs/1805.07648 (2018). arXiv: 1805.07648.

[12] Ronald Mutegeki and Dong Seog Han. "A CNN-LSTM Approach to Human Activity Recognition". In: *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. 2020, pp. 362–366.

[13] Vinod Nair and Geoffrey E. Hinton. "Rectified linear units improve restricted boltzmann machines". In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML'10. Haifa, Israel: Omnipress, 2010, 807–814. ISBN: 9781605589077.

[14] Francisco Javier Ordóñez and Daniel Roggen. "Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition". In: *Sensors* 16.1 (2016). ISSN: 1424-8220.

[15] Adam Paszke et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. 2019. arXiv: 1912.01703 [cs.LG].

[16] Yao-Hung Hubert Tsai et al. *Learning Factorized Multimodal Representations*. 2019. arXiv: 1806.06176 [cs.LG].

[17] Ashish Vaswani et al. "Attention is all you need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, 6000–6010. ISBN: 9781510860964.

[18] Hu Wang et al. *Multi-modal Learning with Missing Modality via Shared-Specific Feature Modelling*. 2023. arXiv: 2307.14126 [cs.CV].

[19] Jian Bo Yang et al. "Deep convolutional neural networks on multichannel time series for human activity recognition". In: *Proceedings of the 24th International Conference on Artificial Intelligence*. IJCAI'15. Buenos Aires, Argentina: AAAI Press, 2015, 3995–4001. ISBN: 9781577357384.

[20] Han Zhang et al. *Self-Attention Generative Adversarial Networks*. 2019. arXiv: 1805.08318 [stat.ML].

[21] Yexu Zhou et al. "Automatic Remaining Useful Life Estimation Framework with Embedded Convolutional LSTM as the Backbone". In: Feb. 2021, pp. 461–477. ISBN: 978-3-030-67666-7.