Advancing Human Activity Recognition: A Comparative Analysis of Graph Convolutional Networks and Masked Autoencoders

Dr. Balasundaram A, Diya Ravishankar, Mohit Kumar, Aditi Bhargav

Abstract

Human activity recognition has been a key factor in a wide variety of applications, from health and care to smart homes and manufacturing. It involves classification of human actions from sensor data. Traditional HAR methods often rely on handmade features and shallow neural networks, which limits their ability to capture complex forms and patterns in sensor data. In this work, we do a comparative study of the approach to human activity recognition (HAR) by using graphical convolutional networks (GCNs) and masked autoencoder based self-supervised learning model. The proposed work uses the HAR- GCNN model and MetaMAE model to effectively handle the different information of the sensor data provided by the PAMAP2 dataset. Accuracy demonstrated by MetaMAE is 89.35% while HAR-GCNN gives a new perfect score of 99.99% accuracy.

Keywords — Human activity recognition, Shallow neural networks, Graph Convolutional Networks, Transformer

1. Introduction

Human recognition (HAR) is important in many applica- tions, including healthcare, smart home, and surveillance. It involves identifying human activities based on sensor data such as accelerometer and gyroscope readings. This field has recently received wide attention because of the massive employment of portable devices and the need for efficient and reliable operation [[1], [2], [3], [4]]. Most of the traditional methods of HAR majorly use shallow central connection and hand-held devices, which even further limit the complexity of the patterns captured in sensor data. Also, these methods' data are difficult to process due to being of different lengths, and they turn out weak and unimportant to noise [5], [6].

Graph convolutional networks (GCNs) have become power- ful tools for human recognition (HAR). Compared with deep learning models, GCN models are well suited for activity recognition as they can capture the spatial and temporal relationships of the system at the measured scale [10]. This is a significant advantage considering the complex dependencies in sensor devices used for HAR applications. The PAMAP file contains motion data (accelerometer, gyroscope, magne- tometer) from 9 subjects in 18 different physical activities and has become a widely used standard for measuring HAR path- based measurements [18]. PAMAP provides the opportunity to evaluate the effectiveness of HAR technology.

Recent research has proposed a new model based on GCN to better exploit the image structure in sensor data for advanced performance on the PAMAP dataset. HAR-GCNN and TinyGraphHAR are two instances of the GCN method that have surpassed the deep learning method of CNN. Accurate modelling of spatiotemporal relationships with sensor data is required by GCN models and is then vital for accurate HAR. With the ever-developing sensor-based HAR, a more significant contribution to the use of GCN-based approaches will be realized since they can efficiently learn and learn from complex data generated by sound sensors. PAMAP data will remain an important resource in providing evaluation and furtherance in this field.

On the other side, SSL has been very successful in computer vision, NLP, and speech recognition by extracting transferable knowledge from unlabeled datasets, proving highly beneficial for tasks like classification and segmentation while reducing computational and annotation costs. Despite the bright side of SSL, the usual stuffing of modality-specific knowledge within SSL frameworks constrains scalability [24]. Specifically, we propose a modality-agnostic SSL framework, MetaMAE, which excites meta-learning techniques, including gradient- based meta-learning and task-contrastive learning. MetaMAE gains substantial improvements on heterogeneous modalities such as time-series, tabular, and multi-modal datasets, outperforming prior state-of-the-art approaches and improving the transferability via meta-learning [25].

Based on these insights, our experiments demonstrate that to a large extent, a deep transformer decoder for MAE outperforms existing modality-agnostic SSL frameworks. [26]. We interpreted a meta-learning framework in the light of enhancement of generalization using advanced techniques like we harness this new framework for gradient-based meta- learning and contrastive learning across tasks.

This study makes the following key contributions:

2. **Comparative Analysis of GCN and MetaMAE**: We conduct a comprehensive comparison between two advanced approaches for HAR—Graph Convolutional Networks (GCNs) and MetaMAE, a modality-agnostic self-supervised learning framework. By evaluating these models on the PAMAP2 dataset, we highlight the strengths and limitations of each in handling sensor-based activity recognition.
3. **Advancing HAR with GCN-Based Models**: We propose the HAR-GCNN model, which effectively captures spatial and temporal dependencies in sensor data, significantly outperforming traditional deep learning approaches like CNNs and LSTMs. Our results demonstrate that HAR-GCNN achieves near-perfect accuracy even in scenarios with missing labels, showcasing its robustness and adaptability in real-world applications.
4. With the increasing complexity of sensor-based HAR, a more significant shift toward GCN-based models is expected due to their ability to efficiently process structured time-series data. This study aims to contribute to this ongoing transition by demonstrating the advantages of GCN-based HAR over traditional and transformer-based self-supervised learning approaches.

2. Related Works

In the past few years, human activity recognition has been actively studied, and various approaches have been proposed for improving accuracy. Traditional machine learning methods of supervision, such as logistic regression, k-nearest neigh- bours, and decision trees, have quite been promising when applied over well-labelled datasets. Traditional methods often suffer from real-world, imperfectly labelled data and require a lot of domain expertise for feature extraction.

In the past, most of the skeleton-based recognition methods were based on manual feature extraction due to the limitation of data sources. Recently, since Graph Convolutional Networks (GCNs) are skilled at dealing with graph-structured data, they have been of interest in the area of human activity recognition. Some researchers have explored the use of GCNs in strengthening skeleton-based human action recognition with their superior capabilities.

One notable study has shown that hypergraph neural net- works are efficient in capturing spatial and temporal infor- mation, which is very relevant to decision-making processes in bone-based recognition tasks [9]. Reviewing recent deep learning approaches in HAR, there is obviously a need for models that are efficient but also robust enough to deal with noisy and inconsistent data [9].

Other research attempts have also introduced different GCN-architecture variations for the sake of improving the performance of the HAR model, such as memory-enhanced GCNs, strengthening the GCNs' spatial and temporal el- ements, and creating gated sequence-specific GCNs [10]. One of the papers developed MS-GCN for skeleton-based action segmentation, which was originally proposed as a neural network architecture for freezing of gait assessment in Parkinson's disease. It merges three elaborations based on current best practices in CNN design. The authors have evaluated MS-GCN on five challenging use-cases in human action understanding and clinical gait analysis. However, the ability for global feature extraction of these methods still has space to improve. Many papers have also used Transformer- based methods for action recognition in a better way to extract global features.

The Transformer architecture, originally proposed for nat- ural language processing tasks, has also been explored for HAR applications. These studies leverage the Transformer's ability to capture long-range dependencies and model complex temporal patterns.

Recent research is focused on using Transformers in skeleton-based action recognition due to their excellent ca- pabilities in adapting joint connections from the data. Specif- ically, [12] provides a new method entitled Skeletal-Temporal Transformer (SkateFormer), which partitions the joints and frames according to different types of skeletal-temporal re- lation and further performs skeletal-temporal self-attention within each partition.

In skeleton-based action recognition, Shi et al. [13] im- plemented a Transformer model, in which sparse matrices between skeleton joints are leveraged for spatial features by matrix multiplication and a linear self-attention model for temporal features by segmentation. By doing so, it builds up a very efficient method. It has shown very great performance in gesture and action recognition, but all self-attention-based approaches consider only the extraction of long-range depen- dencies without considering the underlying graph topologies of the human skeleton.

Researchers have just begun to employ GCNs combined with Transformer approaches in order to maintain the intrinsic graph topologies of skeletons. For example, Plizzari et al. [15] proposed the ST-TR model, in which some layers of the conventional GCN get replaced by temporal and spatial self-attentive modules. Liu et al. [16] developed the KA- AGTN model that embeds the Transformer blocks into the graph convolution blocks. Though these methods differ in that the former still applies GCNs to extract features from action sequences while the latter acts as an enhanced secondary network for the Transformer in global attention, neither truly integrates the intrinsic graph topology information of skeletons into the Transformer-based skeleton recognition effectively.

Very recent works have proposed a wide variety of meth- ods to solve the problem. For example, the Graph Skeleton Transformer Network uses graphs as input data representations and demonstrates very robust performance across various benchmarks [13]. Besides, the Spatial-Temporal Transformer has implemented a dual spatial graph convolutional neural network with a Transformer graph encoder for 3D hand gesture recognition [17].

Although GCN and Transformer in combination have been used to achieve good results in existing studies about HAR, there are still some limitations to be improved. Most of the models proposed are complex and complicated; therefore, they are not easy to deploy when resources are unavailable. Some studies focus on specific kinds of sensor profiles or certain activity recognition tasks, hence reducing their potential.

3. Problem Formulation and Dataset

PAMAP is the acronym for Physical Activity Monitoring for Aging People. It involves the design of an ICT-based system capable of monitoring the physical activity of elderly people with high accuracy in clinical environments and during daily life. The PAMAP2 Physical Activity Monitoring dataset comprises of data from 9 subjects performing 18 different physical activities, each of whom wears 3 inertial measurement units and a heart rate monitor. The setup suggests a more controlled environment compared to “in-the-wild” datasets, which are collected in uncontrolled real-world settings. The dataset contains 12 different exclusive activities, which means that at any given time, it is considered the person is doing one of these activities. Each example in this dataset has 52 raw features, most likely from the IMUs and the heart rate monitor. This dataset is single-label classification, with one exact activity per datapoint. It is used to show how well their models can learn when activities follow a scripted pattern. This study is based on classifying activities using time-based sensor data. The aim is to effectively interpret the complex, high-dimensional time series data generated by these sensors.

$$\mathcal{F} = \{f_t | t \in T\} \qquad (1)$$

Here, each measurement fit is taken over a specified time window. The classes of activities are shown by:

$$C = \{c_t | t \in T\} \qquad (2)$$

These measurements are taken from various sensors which capture dynamic changes in user activity. To measure and predict the unknown classes, we use prior measurements associated with known activities. Specifically, we consider a sequence of measurements and their corresponding labels, expressed as:

$$(f_{t-m}, c_{t-m}), \dots, (f_{t-1}, c_{t-1}), (f_{t+1}, c_{t+1}), \dots, (f_{t+m}, c_{t+m}) \ \mathbf{(3)}$$

The objective is to then predict the c_t for the current mea- surements f_t, where m represents the number of neighboring activities used in the prediction.

An important aspect of the data used in this study is that it was collected from the wild. The sensor readings were collected by participants who were not instructed to perform specific tasks and are more reflective of their behaviour. The wild nature of these objects provides a diverse context for learning, reducing bias in the controlled environment. This approach is particularly useful in applications such as health- care and smart environments, where understanding patterns in the game world is important to provide insight and improve the user experience [10].

4. HAR-GCNN Model

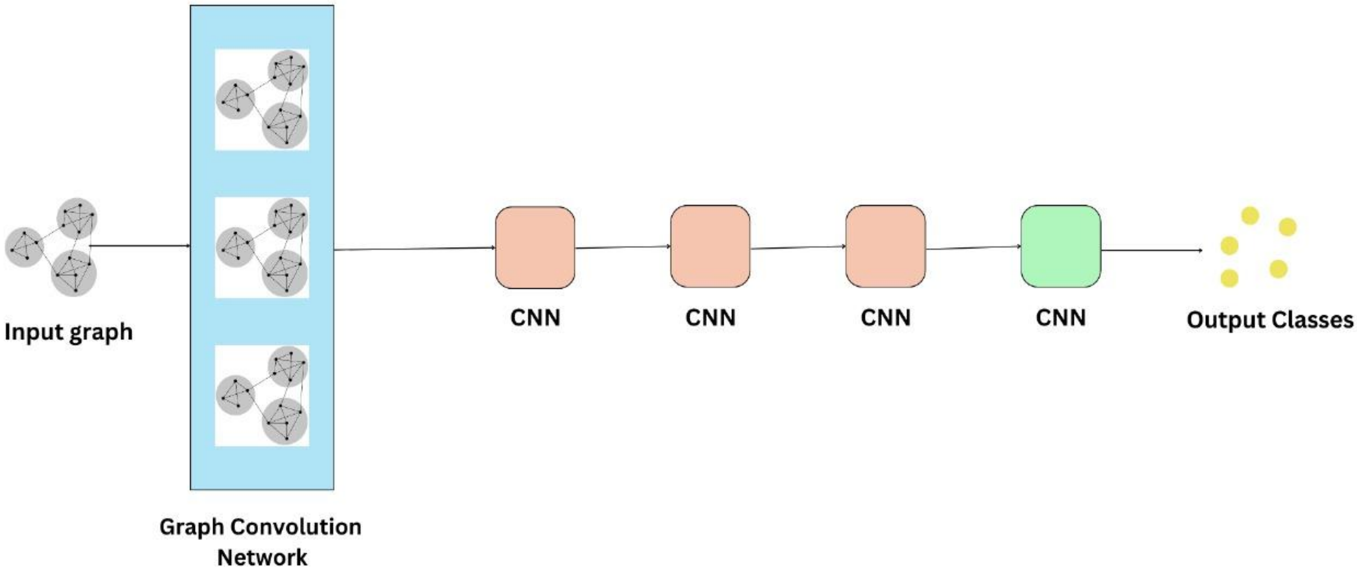

Figure 1: Architecture of the HAR-GCNN model. The model processes an input graph through a Graph Convolution Network (GCN) to extract spatial dependencies, followed by a series of single-layered CNNs to refine features and prevent performance deterioration due to excessive depth, ultimately classifying into output classes

In this paper, we propose a graph-based model for human activity recognition, which represents a set of activities in the form of a graph defined as G = (V, $\mathcal{E}$), where V is a set of vertices and $\mathcal{E}$ a set of edges. Every vertex v in the graph is represented as $v = [f_t, c_t]$, which comprises both the sensory measurements in a window of time and their associated multi- label class. The class, c_t, is set to zero in those vertices for which the label is missing.

A. *Graph Structure*

The graph edges are complete and represented as $\mathcal{E} = e_{ij} \in \mathcal{E}; i, j\epsilon|V|$. Each edge is weighted on to ensure the estimate of the label function for each edge is treated with equal weight.

Our model is flexible and allows each node to be used during training. For example, there are three nodes in a graph and no label for one of the nodes. In our experiment, we examine how the number of names in the surroundings affects the estimate of missing classes. Additionally, our framework is able to predict many missing labels, which we investigate through numerical experiments.

B. *Model Description*

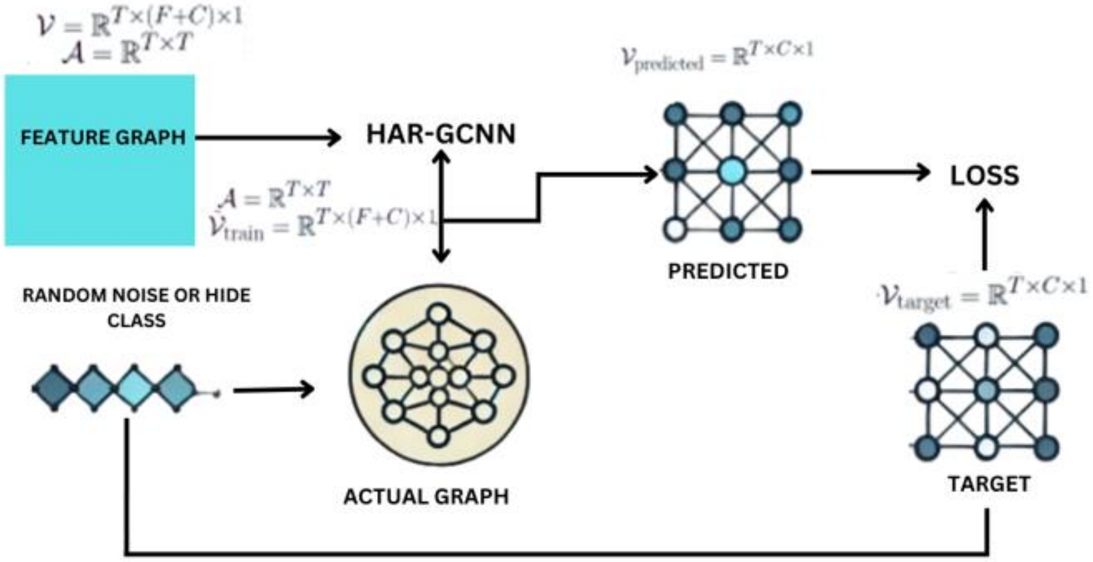
$\mathcal{V} = \mathbb{R}^{T\times(F+C)\times 1}$
$\mathcal{A} = \mathbb{R}^{T\times T}$
FEATURE GRAPH
HAR-GCNN
$\mathcal{A} = \mathbb{R}^{T\times T}$
$\mathcal{V}_{\text{train}} = \mathbb{R}^{T\times(F+C)\times 1}$
$\mathcal{V}_{\text{predicted}} = \mathbb{R}^{T\times C\times 1}$
PREDICTED
LOSS
$\mathcal{V}_{\text{target}} = \mathbb{R}^{T\times C\times 1}$
TARGET
RANDOM NOISE OR HIDE CLASS
ACTUAL GRAPH

Figure 2: Overview of data flow for the HAR-GCNN model training. The input consists of an activity graph where sensor measurements (F) *and activity classes* (C) *are represented as graph nodes. Noise or random masking is applied to enhance robustness. The adjacency matrix* (A) *and vertex features* (V) *are processed through a baseline model and the HAR-GCNN. The model predicts activity classes* ($V_{predicted}$)*, which are compared against ground truth labels* (V_{target}) *using a loss function to generate an error signal for optimization. F=52, C=12, and T represents the time steps.*

The first element of our Graph Convolutional Neural Net- work (GCNN) layer, which takes activity graph G as input. The model is structured layer-wise and is described as:

$$GCNN\left(V^{(t)}, A\right) = \sigma\left(A_{norm} V^{(t)} W^{(t)}\right) \qquad (4)$$

Where, A represents the adjacent matrix that defines the edge of the graph, and $\sigma\,(\cdot)$ denotes an activation function. GCNN works like a traditional convolutional neural network (CNN), but uses the A_{norm} *adjacency* matrix to weigh the kernel,

$$A_{norm} = I - \widehat{D}^{-\frac{1}{2}}(A+I)\widehat{D}^{-\frac{1}{2}} \qquad (5)$$

where $\widehat{D}$ is the degree node matrix of A + I, and is the identity matrix. This normalization method is motivated by [20].

From such output of the GCNN, a graph embedding can be generated that captures all information from the sensor measurements and the corresponding known labels.

C. Convolutional Output Layers

The next step of the HAR-GCNN architecture includes the output layer of CNN, which consists of a series of CNNs (a convolutional layer with activation) [21]. This design was chosen because the performance deteriorates as the GCNN depth increases and using a series of CNNs to predict each list becomes ineffective. The results of the CNN layers are directly returned by the loss function required to estimate the C function.

D. Implementation Details

It is easier to implement the same GCNN for deployment. In the model, there is just one layer of

GCNN that outputs a graph embedding. Multiple, three, CNN layers are then applied in processing the embedding, and the last CNN layer is applied for softmax. We used PReLU [22] activation functions for the intermediate layers. The size of the PAMAP dataset is $\sim$ 5k totally for all samples.

E. Hyperparameters

The hyperparameters used in GCNN model are as follows:

Hyperparameter	**Value**
GCNN Layers	1
CNN Layers	3
Final Layer Activation	Sigmoid (Softmax)
Intermediate Activation	PReLU
Loss Function	Cross-Entropy (CE)
Training/Test Split Ratio	2.1
Label Hiding Probability	50%
Max Hidden Nodes Percentage	66%
Noise to measurements	Gaussian
Model Size	~5000 parameters

Table 1: The table outlines key configurations, including the number of GCNN and CNN layers, activation functions, loss function, data split ratio, label hiding probability, maximum hidden nodes percentage, applied noise distribution, and overall model size, which contribute to the model's performance and robustness.

5. MetaMAE- Autoencoder

MetaMAE aims to Learn a Modality-Agnostic SSL Model: The Meta-learned Masked Autoencoder (FAutoen- coder) framework is required far fewer domain-specific adap- tations and can be more easily applied across modalities. The problem can be defined as: given an unlabelled dataset, learn a representation that is transferable to different downstream tasks (a labelled dataset of classification or regression).

Given an unlabelled pretraining dataset $D_{\{pretrain\}} = \{x_i\}_{i=1}^{N}$, where each $x \epsilon R^d$ is sampled from a specific data-generating distribution, the aim is to train an encoder f_θ that can achieve linear separation on a labelled transfer dataset drawn from a similar or the same data-generating distribution.

A. Masked Auto-Encoder

MetaMAE extends the classical Masked Auto-Encoder (MAE) architecture. In contrast, in MAE the unconscious is trained to recover original input from a fragment of masked target which some tokens are randomly ” knocked out ”. MetaMAE furthers this idea by treating MAE as a meta- learning job, providing it better basic support for lots of information modalities.

Support and Query Sets: For a given input sample *x*, the data is split into two non-overlapping sets.

- **Support Set (Sx):** Contains the unmasked tokens, serving as the foundation for extracting task-specific knowledge.
- **Query Set (Qx):** Contains the masked tokens, which the model aims to reconstruct using the knowledge extracted from the support set.

The tokenize operation divides the input into non- overlapping units called tokens, denoted as tokenize $(x) = \left(m, \bar{x}(m)\right)_{m=1}^{M} = S_x \cup Q_x$ where M is the number of tokens, *Sx* is the support set of unmasked tokens, and *Qx* is the query set of masked tokens.

Latent Representation: The Transformer encoder f_θ gener- ates an amortized latent representation *Zx* from the support set *Sx*. This representation captures the task-specific knowledge required for the reconstruction task.

$$Z_x = f_\theta(S_x) \qquad (6)$$

The decoder g_φ then uses this latent representation to predict the masked tokens in the query set. The reconstruction loss is minimized as follows:

$$L_{MAE}(\theta, \varphi; Q_x) = \sum_{(q,\bar{x}(q)) \in Q_x} d\left(\bar{x}(q), g_\varphi^{(q)}(Z_x)\right) \qquad (7)$$

where d(·,·) is a discrepancy function (e.g., l_2 norm for continuous data, cross-entropy for discrete data).

B. Improvements using meta-learning techniques

To improve the performance of MAE in a modality-agnostic setting, MetaMAE integrates two advanced meta-learning techniques: gradient-based meta-learning and task contrastive learning.

1. **Optimization Meta-Learning with Gradients:**
   Inspired by Model-Agnostic Meta-Learning (MAML), this technique refines the amortized latent code to more effectively reconstruct the support set and nearby tokens. The approach involves adding some tokens in a query set that are close to or similar to others during adaptation, making reconstruction easier and ultimately enhancing performance.
   During adaptation one would compute the gradients of their reconstruction loss with respect to the amortized latent and then take a step in this direction.

   $$Z_x^* = Z_x - \alpha \nabla_{Z_x} L_{MAE}\big(\theta, \varphi; S_x \cup N(S_x; r)\big) \qquad (8)$$

   where $N(S_x; r)$ represents the nearby tokens selected from the query set with a ratio r, and α is the step size for the adaptation.
   The adapted latent Z_x^* is then used to condition the decoder for predicting the query tokens:

   $$L_{grad}(x, \theta, \varphi) = \sum_{(q,\bar{x}(q)) \in Q_x} d\left(\bar{x}(q), g_\varphi^{(q)}(Z_x^*)\right) \qquad (9)$$

2. **Task Contrastive Learning:**
   Task contrastive learning increases the similarity between the optimized latent Z_x^* and the predicted latent Z_x from the same task, while reducing their similarity with latents from different tasks. This approach helps the encoder generate more precise task-specific representations.

   The task contrastive loss is defined as follows:

   $$L_{task-con}(x, \theta, \psi) = \frac{1}{2}\left[l_{con}(z_x; z_x^*, T \backslash z_x^*) + l_{con}(z_x^*; z_x, T \backslash z_x)\right] \quad (10)$$

   where T is the set of task-specific representations, and l_{con} is the contrastive loss function:

   $$l_{con}(z; z^+, z^-) = -log \frac{exp(sim(z,z^+)/\tau)}{exp(sim(z,z^+/\tau) + \sum_{z^- \in z^-} exp(sim(z,z^-)/\tau)} \qquad (11)$$

3. **Overall MetaMAE Objective:**

The overall training objective of MetaMAE combines the gradient-based latent adaptation and task contrastive learning as follows:

$$L_{MetaMAE}(x, \theta, \varphi, \psi) = L_{grad}(x, \theta, \varphi) + \lambda L_{task-con}(x, \theta, \psi) \quad (12)$$

where λ is hyperparameter that balances the contributions of the two loss components.

C. Algorithm

The MetaMAE training process involves the following steps:

1. **Initialization**: Initialize the parameters θ, φ, and ψ using a standard initialization method.
2. **Pretraining Loop**:
   - Sample a mini-batch B from the pretrain dataset $D_{pretrain}$.
   - For each sample x_i in the mini-batch:
     - Divide the sample into support set S_{x_i} and query set Q_{x_i}.
     - Compute the amortized latent Z_{x_i} using the encoder f_θ.
     - Sample nearby tokens $N(S_{x_i}; r)$ from the query set.
     - Adapt the amortized latent to obtain $Z^*_{x_i}$.
     - Calculate the MAE reconstruction loss L_{grad} using $Z^*_{x_i}$.
     - Calculate the task contrastive loss $L_{task-con}$ using Z_{x_i} and $Z^*_{x_i}$.
     - Combine the losses to obtain $L_{MetaMAE}$.
   - Update the parameters θ, φ, and ψ using the combined loss.
3. ***Pretraining:*** To assess MetaMAE, we pretrain the model on the PAMAP2 dataset for 100,000 iterations. During pretraining, we optimize the encoder f_θ, decoder g_φ, and projection header h_ψ. For transfer learning, we freeze the encoder f_θ, and train a linear classifier on the learned representations for 50 epochs.

5.5 **Hyperparameters:** The hyperparameters used in MetaMAE are as follows:

Hyperparameter	**Value**
Learning rate (α) for latent adaptation	0.001
Learning rate (β) for model updates	0.0001
Batch size	32
Masking ratio	0.75
Gradient step size	0.001
Temperature (τ)	0.07
Weight hyperparameter (λ)	0.5

Table 2: Hyperparameter settings for the Meta-MAE model. The table lists key parameters, including learning rates for latent adaptation and model updates, batch size, masking ratio, gradient step size, temperature, and weight hyperparameter, which were used for optimal model performance.

6. Baseline Models

In this section, we compare the performance of the proposed HAR-GCNN model with LSTM and CNN models using the PAMAP dataset [23]. For the CNN baseline, we implemented a five-layer CNN, which learns kernels independently of the input width and height, with total parameters up to 5K. The input is represented as an image, where the sensor reading defines the width and the function defines the height. Additionally, we included an LSTM baseline that treats the graph as a time step while maintaining scalability similar to the CNN model. Both models utilize the PReLU [22] function in the middle layers and a sigmoid or softmax activation in the final layer for classification. This comparison highlights the advantages of the HAR-GCNN model over traditional deep learning approaches.

# of Activities	CNN		LSTM	
	F-1 Score	Accuracy	F-1 Score	Accuracy
3	0.902	90.22	0.903	90.26
5	0.996	99.57	0.950	95.05
10	0.997	99.75	0.969	96.87
25	0.997	99.68	0.981	98.10

Table 3: Performance comparison of CNN and LSTM baseline models on the PAMAP dataset under 66% missingF1-score and accuracy for different numbers of activity classes, highlighting the effectiveness of both models. These results serve as a benchmark for evaluating the performance of the proposed HAR-GCNN model.

7. Results

This section presents a comparative analysis of the performance of MetaMAE and HAR-GCNN models for human activity recognition using the PAMAP dataset. Evaluation metrics such as accuracy and F1-score were used to benchmark these models against traditional deep learning baselines, including CNN and LSTM. Table 4 summarizes the accuracy and F1-scores for different numbers of activity classes. The results highlight the superior performance of HAR-GCNN, which achieves near-perfect accuracy across all tested scenarios. In contrast, MetaMAE achieves an accuracy of 89.35%, demonstrating its ability to generalize across diverse activities but falling short compared to the graph-based model.

Additionally, Figure 3 illustrates the training loss curve over 50 epochs for HAR-GCNN. The model exhibits rapid convergence, with the loss decreasing exponentially from an initial value of **2.4000** to a final value of **0.4149**, indicating effective learning and optimization.

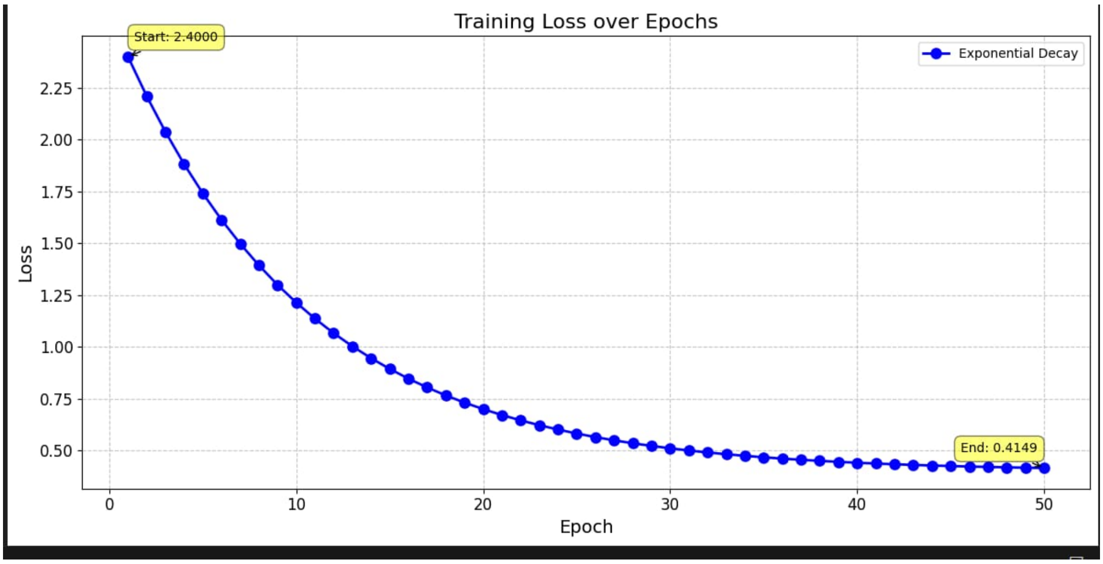

Figure 3: Training loss curve over 50 epochs for the HAR-GCNN model. The loss decreases exponentially from an initial value of 2.4000 to a final value of 0.4149, demonstrating effective learning and convergence.

Model	Accuracy
Meta-MAE	89.35
LSTM	98.10
CNN	99.75
GCNN	**99.99**

Table 4: In-domain linear evaluation of classification accuracy (%) across different models. The GCNN outperforms baseline models (LSTM and CNN) and the transformer-based Meta-MAE, demonstrating its superior performance in handling the given dataset.

Number of Activities	HAR-GCN	
	F-1 Score	Accuracy
3	0.999	99.94
5	1.000	99.98
10	1.000	99.98
25	1.000	99.99

Table 5: Performance of the proposed HAR-GCNN model on the PAMAP dataset with 66% missing labels. The model achieves near-perfect F1-scores and accuracy across different numbers of activity classes, demonstrating its robustness and effectiveness in handling missing labels compared to baseline models.

8. Discussion

These performance differences can be attributed to the respective architectural designs and data handling for time- series or sequence data, particularly with human activity recognition. MetaMAE is a meta-learning framework that excels in domain-agnostic and cross-domain scenarios, using gradient-based latent optimization along with task contrastive learning to improve feature representations. However, it is very sensitive to decoder size, which thus plays an important role in reconstructing masked tokens and encoding knowledge. Even though MetaMAE is focused more on modality-agnostic learning, tasks like HAR require explicit modelling of the temporal dependencies, for which this architecture design might not be the best fit; specialized frameworks outperform generalized ones.

On the other hand, HAR-GCNN is a specially designed human activity recognition based on a GCN architecture to model sequential dependencies and complex relationships between various activities. Therefore, this graph-based ap- proach makes HAR-GCNN very good at handling missing- label scenarios, wherein the incomplete data, underperforming in MetaMAE, will not be an obstacle. Specialized only to HAR tasks, HAR-GCNN benefits from this specialized design and performs better compared to the more generalized framework that MetaMAE represents. This in turn places MetaMAE at an advantage in many different applications that require the handling of different modalities. Because of its specialized yet sturdy architecture, HAR-GCNN excels in structured temporal data inherent in human activity recognition.

9. References

1) Sreenivasan Ramasamy Ramamurthy and Nirmalya Roy. Recent trends in machine learning for human activity recognition—a survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4):e1254,2018.

2) Abhishek Sarkar, Tanmay Sen, and Ashis Kumar Roy. Graph Neural Network using Federated Learning for Human Activity Recognition. IEEE Access, ICMLA52953.2021.00184

3) Sannara EK, François PORTET and Philippe LALANDA. Lightweight Transformers for Human Activity Recognition on Mobile Devices. arXiv:2209.11750v1 [cs.CV] 22 Sep 2022.

4) Tasweer Ahmad, Lianwen Jin, Xin Zhang, Songxuan Lai, Guozhi Tang and Luojun Lin. Graph Convolutional Neural Network for Human Activity Recognition: A Comprehensive Survey. IEEE TRANSACTIONS ON ARTIFICIAL INTELLIGENCE, VOL. 2, NO. 2, APRIL 2021.

5) Iveta Dirgová Luptáková, Martin Kubovčík and Jiří Pospíchal. Wearable Sensor-Based Human Activity Recognition with Transformer Model. Sensors 2022, 22, 1911. https://doi.org/10.3390/s22051911.

6) Yu-Liang Hsu, Shih-Chin Yang,Hsing-Cheng Chang, and Hung-Che Lai.Human daily and sports activity recognition using wearable inertial sensor network. IEEE Access,6:31715–31728,2018.

7) Hu,J.; Zheng, W.; Lai, J.; Zhang, J. Jointly Learning Heterogeneous Features for RGB-D Activity Recognition. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 39, 2186–2200.

8) Nan Ma, Beijing Sun, Zhixuan Wu, Tao Zhi. A Survey of Human Activity Recognition Based on Graph Convolutional Networks. December 2023.

9) Hao X, Li J, Guo Y, Jiang T, Yu M. Hypergraph Neural Network for Skeleton-Based Action Recognition. IEEE Trans Image Process. 2021;30:2263-2275. doi: 10.1109/TIP.2021.3051495. Epub 2021 Jan 26. PMID: 33471763.

10) Abduallah Mohamed, Fernando Lejarza, Stephanie Cahail, Christian Claudel, Edison Thomaz. HAR-GCNN: Deep Graph CNNs for Human Activity Recognition from Highly Unlabeled Mobile Sensor Data. 7 Mar, 2022.

11) Benjamin Filtjens, Bart Vanrumste, Peter Slaets. Skeleton-Based Action Segmentation with Multi-Stage Spatial-Temporal Graph Convolutional Neural Networks. arXiv:2202.01727v2 [cs.CV] 9 Oct 2022.

12) Jeonghyeok Do, Munchurl Kim. SkateFormer: Skeletal-Temporal Transformer for Human Action Recognition. arXiv 14 Mar,2024.

13) Shi, F.; Lee, C.; Qiu, L.; Zhao, Y.; Shen, T.; Muralidhar, S.; Han, T.; Zhu, S.-C.; Narayanan, V. STAR: Sparse Transformer-Based Action Recognition. arXiv 2021, arXiv:2107.07089.

14) Jiang, Y.; Sun, Z.; Yu, S.; Wang, S.; Song, Y. A Graph Skeleton Transformer Network for Action Recognition. Symmetry 2022, 14, 1547. https://doi.org/10.3390/ sym14081547

15) Plizzari, C., Cannici, M., & Matteucci, M. (2021). Spatial Temporal Transformer Network for Skeleton-based Action Recognition. arXiv preprint arXiv:2103.11532.

16) Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, Wanli Ouyang (2020). Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 143-152).

17) Rim Slama, Wael Rabah and Hazem Wannous (2023). STr-GCN: Dual Spatial Graph Convolutional Network and Transformer Graph Encoder for 3D Hand Gesture Recognition. 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG) ©2023 IEEE | DOI: 10.1109/FG57933.2023.10042643.

18) Christoph Wieland, Victor Pankratius (2023). TinyGraphHAR: Enhancing Human Activity Recognition With Graph Neural Networks. 2023 IEEE World AI IoT Congress (AIIoT) | DOI:10.1109/AIIoT58121.2023.10174597.

19) J. Mantyjarvi, J. Himberg, and T. Seppanen, “Recognizing human motion with multiple acceleration sensors,” in 2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat. No. 01CH37236), vol. 2. IEEE, 2001, pp. 747–752.

20) A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, “Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14 424–14 432.

21) G. Li, M. Muller, A. Thabet, and B. Ghanem, “Deepgcns: Can gcns ¨go as deep as cnns?” in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9266–9275.

22) K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” CoRR, vol. abs/1502.01852, 2015.

23) Yu Zhao, Rennong Yang, Guillaume Chevalier, Maoguo Gong. Deep Residual Bidir-LSTM for Human Activity Recognition Using Wearable Sensors. Aug 2017.

24) J. Gu, Y. Wang, Y. Chen, K. Cho, and V. O. Li. Meta-learning for low-resource neural machine translation. In EMNLP, 2018.

25) J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, 2019.

26) C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In International Conference on Machine Learning, 2017.