Semantically Encoding Activity Labels for Context-Aware Human Activity Recognition

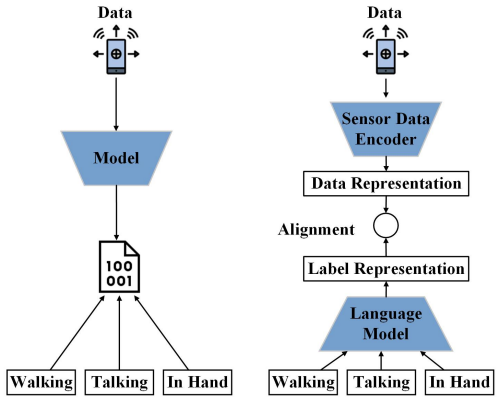
Wen Ge[§], Guanyi Mou[§], Emmanuel O. Agu, Kyumin Lee
Computer Science Department, *Worcester Polytechnic Institute* Worcester, MA, USA
{wge, gmou, emmanuel, kmlee}@wpi.edu

***Abstract*—Prior work has primarily formulated Context-Aware Human Activity Recognition (CA-HAR) as a multi-label classification problem, where model inputs are time-series sensor data and target labels are binary encodings representing whether a given activity or context occurs. These CA-HAR methods either predicted each label independently or manually imposed relationships using graphs. However, both strategies often neglect an essential aspect: activity labels have rich semantic relationships. For instance, walking, jogging, and running activities share similar movement patterns but differ in pace and intensity, indicating that they are semantically related. Consequently, prior CA-HAR methods often struggled to accurately capture these inherent and nuanced relationships, particularly on datasets with noisy labels typically used for CA-HAR or situations where the ideal sensor type is unavailable (e.g., recognizing speech without audio sensors). To address this limitation, we propose Semantically Encoding Activity Labels (SEAL), which leverage Language Models (LMs) to encode CA-HAR activity labels to capture semantic relationships. LMs generate vector embeddings that preserve rich semantic information from natural language. Our SEAL approach encodes input-time series sensor data from smart devices and their associated activity and context labels (text) as vector embeddings. During training, SEAL aligns the sensor data representations with their corresponding activity/context label embeddings in a shared embedding space. At inference time, SEAL performs a similarity search, returning the CA-HAR label with the embedding representation closest to the input data. Although LMs have been widely explored in other domains, surprisingly, their potential in CA-HAR has been underexplored, making our approach a novel contribution to the field. Our SEAL approach has been rigorously evaluated on three real-world datasets, demonstrating its superior performance. It consistently outperforms state-of-the-art methods by 7.8% to 22.6% in MCC and 3.9% to 8.4% in Macro-F1. Furthermore, SEAL performance is agnostic to the data encoding framework utilized, enhancing performance by 4.7% to 73.3% in MCC and 2.6% to 30.5% in Macro-F1 across different data encoding models. This robust performance opens up new possibilities for integrating more advanced LMs into CA-HAR tasks. We share our code and supplement material at https://github.com/GMouYes/SEAL.**

***Index Terms*—Context-Aware Human Activity Recognition, Language Modeling, Semantic Encoding, Machine Learning**

I. Introduction

Context-Aware Human Activity Recognition (CA-HAR) is a critical task that involves detecting human activities and their contexts using sensor data from devices such as smartphones and smartwatches [1], [2]. Traditionally, CA-HAR has been formulated as a multi-label classification task, where the inputs are time-series sensor data, and the target labels are binary 0/1 encodings that indicate the absence/presence of specific activities and contexts. Previous research (Fig. 1a) has focused on creating models that map sensor data to binary labels [3]–[5]. More recent work has explored modeling label co-occurrences using manually defined, learned graphical representations to capture some relationships between labels [6]–[8].



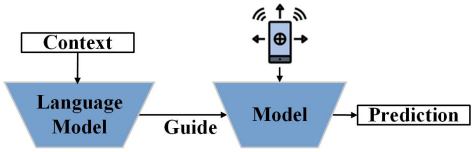
(a) Traditional approach. (b) Our approach.

Figure 1. Comparison of traditional Machine Learning approach v.s. our multi-modality alignment approach. Traditional approaches directly map labels into binary values. In contrast, our approach leverages the language model to encode the semantic relationship between context and activities within high-dimensional vector representations and leads to better performance.

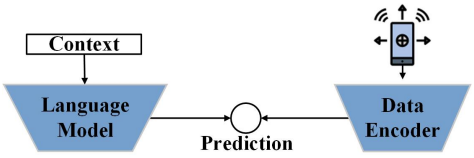
However, these approaches overlook a key consideration: activity labels have intrinsic, nuanced semantic relationships that are not captured when labels are assigned independent binary values, or missed by manually defined graphical relationships. For example, activities such as *Walking*, *Jogging*, and *Running* share similar movement patterns but differ in pace and intensity, suggesting that they are semantically close. On the other hand, misclassifying *Trembling* as *Standing* due to subtle signal changes could lead to missed early signs of medical conditions such as Parkinson's disease [9]. Current models struggle to represent these subtle relationships, leading to suboptimal CA-HAR performance.

To bridge the gap, we propose a novel framework called Semantically Encoded Activity Labels (SEAL), a unique approach that leverages Language Models (LMs) [10] and their associated rich embedding spaces to encode the rich semantics inherent in activity labels (Fig. 1b). This novel approach is a departure from traditional methods and is likely to pique the interest of researchers and practitioners in the field. LMs have

[§]Equal contribution



(a) Other HAR models using Language Models (LMs).



(b) Our design.

Figure 2. Comparison of other HAR models using Language Models v.s. our design. While other approaches mainly use LMs as auxiliary components that provide guidance without directly participating in the decision-making process, our approach integrates LM as a primary contributor, allowing its active involvement in the activity recognition task.

become foundational in Natural Language Processing (NLP) due to their ability to transform natural language into vector representations that preserve semantics.

From a machine learning perspective, CA-HAR can be re-envisioned as an alignment task between two modalities: time-series data (sensor signals) and text (activity and context labels). During training, SEAL aligns the vector embeddings of sensor data with the embeddings of their corresponding activity/context labels. At inference time, it performs a similarity search to determine the label whose encoding is close to the input sensor data. By leveraging LMs and their associated embedding space with strong semantics, SEAL preserves semantic relationships between activity labels, capturing nuanced relationships between activities and contexts. Our work aligns with trending research in other domains [11], [12] where text-image alignments gained increasing momentum and enabled various real-world applications [13], [14].

Although prior work has explored incorporating LMs into CA-HAR [15], [16], these efforts primarily leveraged the reasoning capabilities of LMs by using them as auxiliary modules that refine the outputs of CA-HAR models but did not integrate them as a core module in an end-to-end framework (see Fig. 2). In contrast, our SEAL framework integrates LMs as a central component, serving as the primary encoder of activity labels. This innovative SEAL framework addresses the critical issue of semantic information loss inherent in binary label encodings, a significant limitation of prior CA-HAR methods. Furthermore, our SEAL approach achieves robust performance improvements and generalizes effectively, even on the relatively noisy datasets commonly used in CA-HAR. In this paper, we define context-aware human activity the as the <activity, phone placement> tuple.

In summary, our contributions are as follows:

- We identify the loss of semantic relationship information caused by binary label encodings widely used in prior CA-HAR frameworks. We reframe CA-HAR as a time-series-to-text alignment task in a shared vector embedding space, a perspective that should enlighten our audience about the nuanced relationships between activity labels.
- We introduce SEAL, a novel framework that, for the first time, leverages language models to semantically encode activity labels and align these textual encodings with sensor data representations, improving activity recognition.
- We extensively evaluate SEAL, comparing it to five state-of-the-art baselines on three real-world CA-HAR datasets. Our detailed evaluations provide robust evidence of SEAL's performance, demonstrating that it consistently outperforms baseline approaches, with MCC/Macro-F1 improvements ranging from 7.8%/3.9% to 22.6%/8.4%. This comprehensive evaluation should provide reassurance to our audience about the effectiveness of our approach.
- We provide rigorous analysis and statistical and visual evidence demonstrating our approach's effectiveness and contributions to the CA-HAR field.

II. Related Work

A. Traditional HAR

Early CA-HAR work trains separate machine learning models for each label independently [1], [17], overlooking the complex relationships between context and activities. This approach proved inadequate as the number of activities encountered and recognized grew, leading to scalability issues and suboptimal performance. In recent years, deep learning techniques, such as MLP [3], CNN [18], RNN [19], [20], and Hybrid Networks [4], [5], have gained popularity in HAR as they implicitly model relationships between activity labels by training a unified model. Despite their success, these models lack explicit mechanisms to capture the relationships between context and activity labels, often resulting in a limited understanding of the context's influence on activity recognition.

B. Graph-based HAR

To overcome the limitations of traditional HAR models, recent studies have delved into graph-based methods that explicitly encode relationships between context and activity labels. Martin *et al.* [21] proposed a GCN approach for HAR by modeling personalized mobility graphs from GPS trajectory. HAR-GCNN [22] leveraged chronological correlations between sequential activities to predict missing labels. HHGNN [6] and DHC-HGL [7] proposed modeling the activity and context labels as nodes in a heterogeneous hypergraph. However, the manual definition of these graphs may restrict the expressiveness and adaptability of the learned representations. Traditional and graph-based HAR works overlooked a crucial aspect of the problem: activity labels carry intrinsic semantic meanings. Employing independent and binary label encodings without capturing their semantics misses the nuanced, inherent relationships, similarities, and distinctions beyond simple co-occurrence and leads to sub-optimal performance. For example, while Walking, Jogging, and Running are semantically similar in movement patterns, they differ in pace and intensity.

C. Using Language Model for HAR

Recent advancements in LMs have demonstrated their capabilities in understanding and processing natural language, inspiring many applications in various other domains, including

computer vision (CV) [23] and health applications [24]. For instance, CLIP [11] introduces an innovative method for learning visual representations using natural language supervision. Their model learns a multi-modal embedding space by jointly training an image and text encoder, optimizing for high cosine similarity between correct image-text pairs while minimizing it for incorrect pairs. Similarly, ALIGN [12] developed a scalable model that processes noisy image-alt-text pairs using a dual-encoder architecture, aligning image and text representations in a shared latent space through a contrastive loss. TS2ACT [25] introduced a cross-modal contrastive learning framework for HAR, aligning time-series sensor data with web-sourced images. However, they relied on image-based augmentation rather than the inherent semantics of labels. Unlike SEAL, TS2ACT oversimplifies real-world scenarios by focusing on multi-class activity classification while ignoring co-occurring activities and contextual influences. These developments have encouraged researchers to explore the potential of language models for human activity recognition (HAR) tasks. ContextGPT [16] translated high-level contextual information into natural language using prompt engineering and retrieved the most plausible activities from pre-trained Large Language Models (LLMs) with the given context. The retrieved information is then infused into a HAR model to enhance recognition performance. Their approach used LM as an auxiliary component to help guide the model to focus more on plausible activities. Unlike their approach, SEAL integrates LM as a core module into an end-to-end model to capture the relationship between context and activities. It can be used directly for decision-making. Besides that, their reliance on LLMs for generating knowledge might lead to hallucinations and result in inconsistent activity predictions.

III. Proposed Framework

A. Overall Architecture

Our primary goal is to align the time-series input with its corresponding textual labels while preserving the underlying semantic relationships and accounting for potential label co-occurrence. This formulation contrasts with traditional CA-HAR frameworks [3], [4], [6], [7], which use arbitrary binary encodings for labels and treat the task as a single-modal data-to-label mapping, $f : X \rightarrow Y_b$. In these approaches, f is a learnable mapping function, and Y_b represents the binary label format. However, binary encodings discard the semantic meanings of the original labels and fail to capture the inherent similarities, differences, and relationships between activities. In contrast, our approach uses an LM to encode the original text labels, retaining rich semantic relationships and ensuring that the labels' meaning preserved in the process. This enables a more accurate alignment between sensor data, activities and contexts. Fig. 3 shows our architecture in detail, which consists of three sub-modules:

- *Sensor Data Encoder (M_{data}):* Encodes input time-series sensor data into vector embedding representations.

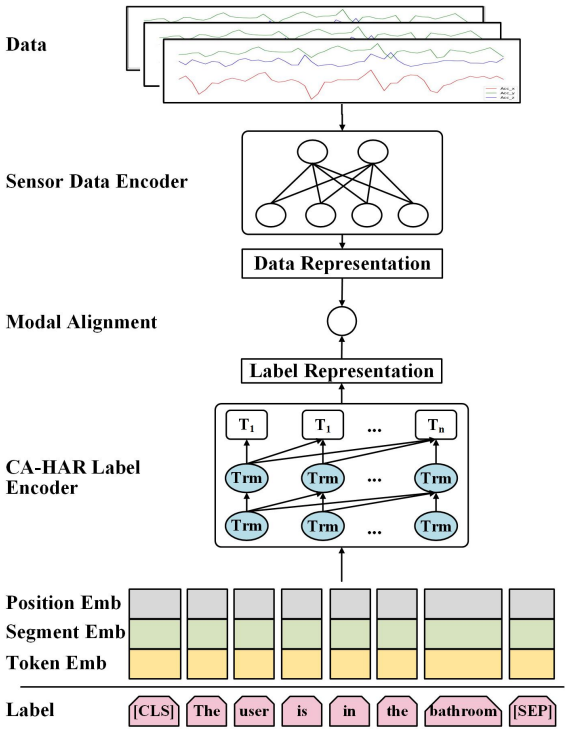


Figure 3. The SEAL framework consists of three main components: a Sensor Data Encoder, a CA-HAR Label Encoder, and a Modal Alignment. The Sensor Data Encoder transforms input sensor data into vector embedding representations, while the Label Encoder generates semantic label vector embedding representations from tokenized label sentences. Finally, the Modal Alignment component aligns the sensor data and CA-HAR label representations by maximizing their similarity, enabling SEAL to make accurate predictions. "Trm" are transformers modules within language models.

- *CA-HAR Label Encoder (M_{label}):* Encodes the activity labels in textual form into vector embedding representations.
- *Modal Alignment (M_{align}):* uses a custom objective function to aligns the vector embeddings generated by the sensor data and CA-HAR label encoders.

Our key contributions lie mostly in the CA-HAR label encoder and modal alignment components.

B. Sensor Data Encoder

The sensor data encoder (M_{data}) transforms input time-series sensor data into vector representations. Existing methods follow two main approaches:

Automatically learn feature representations from raw time series data using a CNN and/or Bi-LSTM as in [19], [26]: $M_{data}(X) = V_{data} \in \mathbb{R}^{n \times h_1}$.

Extract handcrafted features, followed by machine learning models to obtain refined feature representations as in [3]:

$$M_{data}(T(X)) = V_{data} \in \mathbb{R}^{n \times h_1}, T : \mathbb{R}^{n \times s \times d} \rightarrow \mathbb{R}^{n \times h'} \quad (1)$$

where h' and h_1 are hidden dimensions. While we utilized a two-layer MLP with activation and dropout to derive data representation from handcrafted features, it is instructive to note that our framework is highly adaptable to other data encoding methods. In future, with minimal modification, researchers can explore additional modalities such as audio, image, and video, as well as alternative data encoders [26], [27] to further

Table I
AVERAGE ACTIVITY RECOGNITION PERFORMANCE ON THE *WASH Scripted*, *WASH Unscripted*, AND *Extrasensory* DATASETS, EACH CELL CONTAINS MCC/MACRO-F1 SCORES. BEST RESULT: **BOLDED**. SECOND BEST RESULT: UNDERLINED.

Dataset	ExtraMLP	LightGBM	CRUFT	HHGNN	DHC-HGL	SEAL	Improv.(%)
WASH Scripted	0.387 / 0.637	0.411 / 0.657	0.409 / 0.652	0.389 / 0.640	0.411 / 0.655	**0.504 / 0.712**	22.6 / 8.4
WASH Uncripted	0.550 / 0.735	0.667 / 0.810	0.557 / 0.742	0.693 / 0.827	0.808 / 0.897	**0.921 / 0.959**	14.0 / 6.9
Extrasensory	0.636 / 0.787	0.710 / 0.837	0.769 / 0.871	0.808 / 0.896	0.855 / 0.923	**0.922 / 0.959**	7.8 / 3.9

Table II
EXAMPLES OF LABEL TRANSFORMATION FOR LANGUAGE MODEL INPUT. FOR SHORT-TERM ACTIONS, THE TRANSFORMED SENTENCES EXPLICITLY INCLUDED 'NOW' TO EMPHASIZE THE ACTION AS A TRANSITIONAL STATE OCCURRING IN THE PRESENT MOMENT.

Original Label	Transformed Sentence
In Pocket	The user has a phone in their pocket.
Bathroom	The user is in the bathroom .
Talking on Phone	The user is talking on the phone.
Sitting down (action)	The user is sitting down now.

improve model performance. As demonstrated in Section V, our SEAL framework imposes minimal restrictions on the data encoder and improves performance across different backbone models (e.g., MLPs [3], RNNs [28], and GNNs [7]).

C. CA-HAR Label Encoder

Unlike time-series input data, the CA-HAR labels are represented in textual form. Inspired by recent advances [11], we employ LMs to encode the labels into vector representations that preserve rich CA-HAR label semantics and relationships.

$$M_{label}(Y_t) = V_{label}, Str^{n\times C} \rightarrow \mathbb{R}^{n\times C\times h_2} \tag{2}$$

In this paper, we finetune a pre-trained BERT-based model to generate a CA-HAR label representation. Labels are first rewritten into complete sentences using templates as shown in Table II, then tokenized and fed into BERT to extract representations from the classification head:

$$\begin{aligned} Y_t' &= Tokenize(Rewrite(Y_t)) \\ V_{label} &= BERT(Y_t') < cls > \end{aligned} \tag{3}$$

where $< cls >$ refers to the classification head. While we recognize the availability of larger and more powerful language models [29]–[33], the novelty of our approach is that it is LM-agnostic. We focus on demonstrating that incorporating an LM as a CA-HAR label encoder is effective. Future work can explore larger models towards incremental performance gains or efficient models [34] to reduce computational costs, and facilitate deployment in real-world applications.

D. Modal Alignment

The modal alignment module aligns the data encoder and label encoder representations in a shared vector space, where similarity is measured using a dot product.

$$\begin{aligned} V_{data}' &= MLP_d(V_{data}) \in \mathbb{R}^{n\times h} \\ V_{label}' &= MLP_l(V_{label}) \in \mathbb{R}^{n\times C\times h} \\ \hat{Y}_b &= dot(V_{data}', V_{label}') \in \mathbb{R}^{n\times C} \end{aligned} \tag{4}$$

During SEAL training, the modal alignment module aligns the data and corresponding label vector embedding encodings by maximizing their similarity. During inference time, the model returns the labels with the LM encoding that are similar (close) to a given data encoding w.r.t thresholds for the multi-label problem. To optimize this alignment, we employ two loss functions depending on the nature of the labels. For mutually exclusive context labels such as phone placements, the Cross-Entropy (CE) loss is utilized.

$$L_{ce} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{n,c} log(\frac{exp(\hat{y}_{n,c})}{\sum_{i=1}^{C} exp(\hat{y}_{n,i})}) \tag{5}$$

For CA-HAR activities that may co-occur (e.g. *Walking* while *Talking On Phone*), Binary Cross-Entropy (BCE) is applied:

$$L_{bce} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} \omega_{n,c}\Big[y_{n,c}\log(\hat{y}_{n,c}) + (1-y_{n,c})\log(1-\hat{y}_{n,c})\Big] \tag{6}$$

In Eq. 5 and Eq. 6, N is the number of instances, C is the number of classes, and $y_{n,c}$ and $\hat{y}_{n,c}$ are the ground truth and predicted values, respectively.

IV. EVALUATION

We rigorously evaluated SEAL's performance on three CA-HAR datasets, namely WASH Scripted, WASH Unscripted, and Extrasensory. For more details about our dataset and experiment settings, please refer to our repository. Average activity recognition performance is presented in Table I, and detailed experimental results on *WASH Scripted* and *Extrasensory* are listed in Table III and Table IV. SEAL uses MLP for data encoder in this section, while we discuss different data encoders' impact on the model's performance in Sec. V.

1) Human Activity Recognition Performance: As shown in Table I, SEAL consistently outperformed all baselines across all datasets. It achieved 0.504 / 0.712, 0.921 / 0.959, and 0.922 / 0.959 MCC/Macro-F1 scores on *WASH Scripted*, *WASH Unscripted*, and *Extrasensory* dataset, respectively, thus improving on the performance of the best baselines by 7.8% / 3.9% to 22.6% / 8.4% on MCC/Macro-F1 scores. The average improvement on activity labels is higher than the average improvement on context labels (see Table V), especially in the WASH datasets. This suggests that activity labels, which are inherently more varied and nuanced, benefit more from LM's added semantic richness.

The best performing baseline, *DHC-HGL*, initializes context and activity nodes by aggregating handcrafted features associated with each node. As information propagates through the GNN, it learns predictive label representations that capture co-occurrence correlations among activities and contexts. On the other hand, our model takes a different approach to capture these label relationships. It introduces semantic label information through LM encoding in an innovative way. As a result, our model consistently outperforms DHC-HGL, the best baseline on all datasets, inspiring confidence in its

Table III
ACTIVITY RECOGNITION PERFORMANCE ON THE *WASH Scripted* DATASET, EACH CELL CONTAINS MCC/MACRO-F1 SCORES.

Activity	ExtraMLP	LightGBM	CRUFT	HHGNN	DHC-HGL	SEAL	Improv.(%)
Lying Down	0.183 / 0.513	0.185 / 0.520	0.202 / 0.526	0.221 / 0.540	0.227 / 0.544	**0.260 / 0.563**	14.5 / 3.5
Sitting	0.530 / 0.718	0.550 / 0.739	0.492 / 0.690	0.522 / 0.716	0.509 / 0.709	**0.646 / 0.796**	17.5 / 7.7
Walking	0.885 / 0.942	0.869 / 0.934	0.871 / 0.935	0.874 / 0.937	0.899 / 0.949	**0.909 / 0.954**	1.1 / 0.5
Sleeping	0.505 / 0.701	**0.596 / 0.763**	0.519 / 0.712	0.477 / 0.684	0.503 / 0.701	0.570 / 0.745	-4.4 / -2.4
Talking On Phone	0.393 / 0.632	0.528 / 0.723	0.601 / 0.768	0.493 / 0.698	0.406 / 0.640	**0.616 / 0.778**	2.5 / 1.3
Bathroom	0.316 / 0.576	0.348 / 0.598	0.362 / 0.614	0.313 / 0.577	0.344 / 0.602	**0.528 / 0.725**	45.9 / 18.1
Standing	0.356 / 0.598	0.345 / 0.597	0.380 / 0.622	0.342 / 0.591	0.401 / 0.642	**0.454 / 0.669**	13.2 / 4.2
Jogging	0.632 / 0.786	0.826 / 0.907	0.671 / 0.812	0.653 / 0.799	0.649 / 0.797	**0.837 / 0.913**	1.3 / 0.7
Jumping	0.429 / 0.656	0.576 / 0.754	0.545 / 0.731	0.526 / 0.721	0.614 / 0.779	**0.707 / 0.837**	15.1 / 7.4
Running	0.715 / 0.840	0.713 / 0.841	0.592 / 0.761	0.683 / 0.821	0.637 / 0.791	**0.808 / 0.896**	13.0 / 6.5
Stairs-Going Down	0.460 / 0.680	0.351 / 0.608	0.371 / 0.618	0.419 / 0.652	0.440 / 0.666	**0.557 / 0.742**	21.1 / 9.1
Stairs-Going Up	0.316 / 0.589	0.274 / 0.567	0.280 / 0.568	0.308 / 0.589	0.377 / 0.636	**0.408 / 0.644**	8.2 / 1.3
Typing	0.612 / 0.772	**0.770 / 0.873**	0.744 / 0.859	0.558 / 0.737	0.572 / 0.747	0.745 / 0.857	-3.2 / -1.8
Coughing	0.195 / 0.532	0.155 / 0.521	0.199 / 0.533	0.182 / 0.528	0.250 / 0.560	**0.300 / 0.589**	20.0 / 5.2
Sneezing	0.170 / 0.520	0.153 / 0.522	0.200 / 0.536	0.181 / 0.527	0.242 / 0.561	**0.302 / 0.591**	24.8 / 5.3
Trembling	0.415 / 0.647	0.384 / 0.632	0.453 / 0.673	0.377 / 0.624	0.390 / 0.635	**0.607 / 0.771**	34.0 / 14.6
Laying Down (a)	0.190 / 0.524	0.186 / 0.526	0.223 / 0.540	0.202 / 0.531	0.229 / 0.550	**0.247 / 0.558**	7.9 / 1.5
Sitting Down (a)	0.131 / 0.502	0.119 / 0.505	0.136 / 0.509	0.127 / 0.504	0.166 / 0.525	**0.178 / 0.532**	7.2 / 1.3
Sitting Up (a)	0.149 / 0.507	0.140 / 0.511	0.171 / 0.517	0.163 / 0.515	0.187 / 0.537	**0.204 / 0.538**	9.1 / 0.2
Standing up (a)	0.153 / 0.507	0.142 / 0.509	0.163 / 0.512	0.151 / 0.512	0.173 / 0.528	**0.191 / 0.535**	10.4 / 1.3

Table IV
ACTIVITY RECOGNITION PERFORMANCE (MCC/MACRO-F1 SCORES) ON THE *Extrasensory* DATASET.

Activity	ExtraMLP	LightGBM	CRUFT	HHGNN	DHC-HGL	SEAL	Improv.(%)
Lying Down	0.936 / 0.968	0.911 / 0.955	0.968 / 0.984	0.962 / 0.981	0.977 / 0.989	**0.991 / 0.996**	1.4 / 0.7
Sitting	0.797 / 0.898	0.755 / 0.876	0.897 / 0.948	0.870 / 0.935	0.907 / 0.953	**0.964 / 0.982**	6.3 / 3.0
Walking	0.554 / 0.739	0.560 / 0.746	0.709 / 0.838	0.718 / 0.846	0.772 / 0.878	**0.895 / 0.945**	15.9 / 7.6
Sleeping	0.950 / 0.975	0.934 / 0.967	0.977 / 0.988	0.979 / 0.989	0.985 / 0.993	**0.994 / 0.997**	0.9 / 0.4
Talking	0.681 / 0.822	0.697 / 0.833	0.834 / 0.912	0.858 / 0.927	0.895 / 0.946	**0.945 / 0.972**	5.6 / 2.7
Bath-Shower	0.496 / 0.695	0.729 / 0.848	0.654 / 0.801	0.780 / 0.880	0.827 / 0.908	**0.857 / 0.924**	3.6 / 1.8
Toilet	0.429 / 0.652	0.524 / 0.716	0.599 / 0.767	0.670 / 0.813	0.743 / 0.859	**0.846 / 0.918**	13.9 / 6.9
Standing	0.621 / 0.787	0.573 / 0.760	0.797 / 0.891	0.766 / 0.875	0.828 / 0.911	**0.920 / 0.959**	11.1 / 5.3
Running	0.616 / 0.774	0.656 / 0.802	0.733 / 0.850	0.866 / 0.930	0.883 / 0.938	**0.945 / 0.972**	7.0 / 3.6
Stairs-Going Down	0.441 / 0.661	0.791 / 0.886	0.618 / 0.780	0.695 / 0.831	0.807 / 0.898	**0.870 / 0.932**	7.8 / 3.8
Stairs-Going Up	0.446 / 0.664	0.654 / 0.801	0.651 / 0.801	0.706 / 0.837	0.770 / 0.876	**0.886 / 0.940**	15.1 / 7.3
Exercising	0.665 / 0.808	0.735 / 0.852	0.792 / 0.887	0.830 / 0.909	0.868 / 0.930	**0.945 / 0.972**	8.9 / 4.5

effectiveness. By incorporating semantic label representations, our model excels in recognizing nuanced, confusing, and ill-posed activities even in unconstrained real-world scenarios.

2) Detailed Activities: We provide detailed results for the *WASH Scripted* and *Extrasensory* datasets. Due to its scripted data collection protocol, the *WASH Scripted* dataset contains high-fidelity data. The *Extrasensory* dataset is publicly available, enabling other researchers to attempt reproducing our results and reflecting more complex, real-world, unscripted scenarios. Due to space limitations, the SEAL model's performance on *WASH Unscripted* was omitted, while it also follows a consistent pattern, with similar observations and conclusions.

Overall, SEAL performed well on activities such as *Talking (On Phone)*, *Bathroom(-Shower)*, and *Toilet*, which could be attributed to those activities' more distinct patterns and precise semantic meanings that our model can capture. The controlled lab environment in the *WASH Scripted* data collection allowed the inclusion of some rare activities, such as *Trembling*, *Coughing*, and *Sneezing*, along with short-term actions such as *Laying Down (action)*, *Sitting Down (action)*, *Sitting Up (action)*, and *Standing Up (action)*. SEAL exhibited exceptional recognition of these activities, achieving average MCC/Macro-F1 improvements of 26.3% / 8.4% on rare activities and 8.7% / 1.1% on short-term actions. For activities such as *Trembling*, *Coughing*, and *Sneezing*, sensor signals may not exhibit significant changes, especially in the absence of audio sensors. The semantic information from encoding labels helped the model detect these activities reliably. Similarly, for short-term actions, the label sentences indicated their brief nature, guiding the model to concentrate on the abrupt changes in sensor signals. This approach allowed the SEAL to more accurately identify the activities based on the enriched label descriptions, even when the raw sensor data was less distinctive.

For labels that are often misclassified between each other, such as *Jogging* vs. *Running* [35] and *Stairs Going Up* vs. *Stairs Going Down* [36], SEAL achieved even more noticeable improvements. Additionally, we observed impressive improvements on activities with hierarchical structures, such as *Exercising* (which encompasses *Walking*, *Running*, *Jogging*, and *Jumping*). SEAL delivered outstanding recognition performance, particularly in distinguishing activities that other CA-HAR models have historically confused (e.g., differentiating between Going Up and Going Down Stairs). Furthermore, it performed remarkably well on activities that are difficult to classify based solely on sensor data, such as *Talking On Phone*, where audio data is ideal for accurate recognition.

V. ANALYSIS

A. SEAL*'s performance on context recognition*

The average context recognition performance across all datasets is listed in Table V. SEAL achieved MCC/Macro-F1 scores of 0.796/0.890, 0.960/0.980, and 0.977/0.988 on the three datasets, respectively. Notably, SEAL significantly improved context recognition performance on the two unscripted datasets, where how users held their phones while

Table V
AVERAGE CONTEXT RECOGNITION PERFORMANCE ON THE *WASH Scripted*, *WASH Unscripted*, AND *Extrasensory* DATASETS.

Dataset	ExtraMLP	LightGBM	CRUFT	HHGNN	DHC-HGL	SEAL	Improv.(%)
WASH Scripted	0.719 / 0.845	0.705 / 0.839	0.707 / 0.836	0.765 / 0.875	**0.807 / 0.897**	0.796 / 0.890	-1.4 / -0.8
WASH Unscripted	0.686 / 0.823	0.687 / 0.827	0.704 / 0.835	0.813 / 0.901	0.902 / 0.950	**0.960 / 0.980**	6.4 / 3.2
Extrasensory	0.784 / 0.883	0.835 / 0.913	0.897 / 0.946	0.953 / 0.976	0.950 / 0.975	**0.977 / 0.988**	2.5 / 1.2

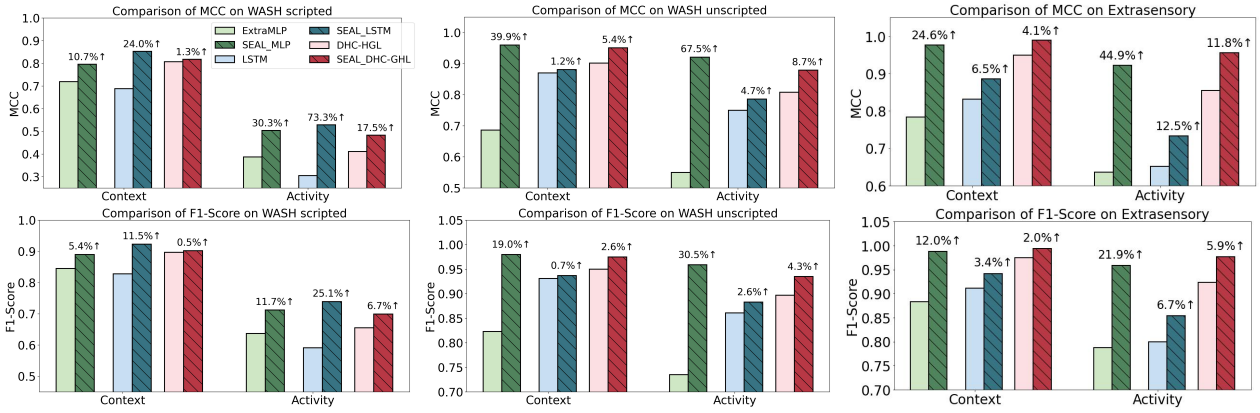


Figure 4. Result of SEAL using different backbones. We observe that SEAL can show improvement with all backbones across all datasets.

performing various activities was not constrained. This reflects the model's strength in handling complex, real-world scenarios with more diverse and unpredictable conditions. The performance dropped slightly for context recognition in *WASH Scripted*. This might be due to the nature of this dataset, where activities and contexts were performed under specific instructions, leading to less natural variability. Additionally, as pointed out in the dataset section, the *WASH Scripted* dataset has fewer instances, which could make it more susceptible to noise, causing the SEAL to misinterpret the context.

B. SEAL *with different backbones*

We hypothesize that semantic encoding could improve activity recognition in two ways: 1) directly for inference and 2) as a source of supplementary information to enhance other label encoding methods, such as Graph Neural Networks (GNNs). To evaluate the merits of these two semantic encoding approaches, we explored three types of deep learning backbones, spanning models that only leveraged handcrafted features or raw signals to models that obtained label representation using GNNs: *ExtraMLP*, *LSTM*, and *DHC-HGL*. The results are shown in Fig. 4 where LM label encoding provides backbone-agnostic improvements with consistent improvements in both the MCC and Macro-F1 scores across all datasets. Notably, SEAL achieved the most significant improvements on *ExtraMLP* and *LSTM*. Their MCC / Macro-F1 scores increased from 1.2% / 0.7% to 39.9% / 19.0% for context labels and from 4.7% / 2.6% to 73.3% / 30.5% for activity labels.

Even though DHC-HGL inherently captures label relationships, it still benefits from the language model's semantic encoding, achieving MCC/Macro-F1 improvements ranging from 1.3% / 0.5% to 5.4% / 2.6% for context labels and from 8.7% / 4.3% to 17.5% / 6.7% for activity labels. Although the degree of improvement for DHC-HGL is less than that of *ExtraMLP* and *LSTM*, language modeling can further enrich label information even for models already adept at capturing inherent relationships. Additionally, while context and activity labels exhibited gains from incorporating the Language Model encodings, the improvements were more substantial for activity labels across all datasets. This suggests that SEAL is particularly effective in making accurate predictions in scenarios with varied activity patterns, which are likely more intricate than contextual information. Interestingly, SEAL outperformed SEAL_DHC-HGL on the two WASH datasets (Scripted and Unscripted), further validating the effectiveness of leveraging the Language Model's semantic capabilities.

Aside from these analyses, we also provide visualizations of the learned label embeddings across all datasets along with an in-depth analysis in the repository.

VI. CONCLUSION

In this paper, we proposed SEAL, a novel framework that leverages language models for CA-HAR. Unlike traditional HAR models that overlook semantic relationships among labels, SEAL directly integrates LMs into the decision-making process, capturing the semantic relationships for improved activity recognition. Experimental results demonstrated that SEAL outperforms SOTA CA-HAR models, achieving 0.504/0.712, 0.921/0.959, and 0.922/0.959 MCC/Macro-F1 scores across three datasets. The results further highlight its ability to distinguish semantically similar activities with subtle differences or abrupt changes. These findings demonstrate the effectiveness of incorporating LMs into CA-HAR tasks and suggest promising directions for extending this approach to multi-modal data incorporating additional modalities, including image and hierarchical label structure.

ACKNOWLEDGMENT

This research is supported by NSF grant IOS-2430277 and DARPA grant HR00111780032-WASH-FP-031.

REFERENCES

[1] Y. Vaizman, K. Ellis, and G. Lanckriet, "Recognizing detailed human context in the wild from smartphones and smartwatches," *IEEE Pervasive Computing*, vol. 16, no. 4, pp. 62–74, 2017.

[2] C. Bettini, G. Civitarese, and R. Presotto, "Caviar: Context-driven active and incremental activity recognition," *Knowledge-Based Systems*, vol. 196, p. 105816, 2020.

[3] Y. Vaizman, N. Weibel, and G. Lanckriet, "Context recognition in-the-wild: Unified model for multi-modal sensors and multi-label classification," *ACM IMWUT*, vol. 1, no. 4, pp. 1–22, 2018.

[4] W. Ge and E. Agu, "Cruft: Context recog. under uncertainty using fusion and temporal learning," in *Proc. ICMLA*, IEEE. IEEE, 2020, pp. 747–52.

[5] W. Ge and E. O. Agu, "Qcruft: Quaternion context recog. under uncertainty using fusion & temporal learning," in *Proc. ICSC*, IEEE. IEEE, 2022, pp. 41–50.

[6] W. Ge, G. Mou, E. O. Agu, and K. Lee, "Heterogeneous hyper-graph neural networks for context-aware human activity recognition," in *Proc PerCom Workshops*. IEEE, 2023, pp. 350–354.

[7] W. Ge, G. Mou, E. Agu, and K. Lee, "Deep heterogeneous contrastive hyper-graph learning for in-the-wild context-aware human activity recognition," *Proc. ACM IMWUT*, vol. 7, no. 4, pp. 1–23, 2024.

[8] R. Mondal, D. Mukherjee, P. K. Singh, V. Bhateja, and R. Sarkar, "A new framework for smartphone sensor-based human activity recognition using graph neural network," *IEEE Sensors Journal*, vol. 21, no. 10, pp. 11 461–11 468, 2020.

[9] E. de Carvalho Costa, F. B. Santinelli, G. F. Moretto, C. Figueiredo, A. E. von Ah Morano, J. A. Barela, and F. A. Barbieri, "A multiple domain postural control assessment in people with parkinson's disease: traditional, non-linear, and rambling and trembling trajectories analysis," *Gait & Posture*, vol. 97, pp. 130–136, 2022.

[10] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PMLR. PMLR, 2021, pp. 8748–8763.

[12] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*. PMLR, 2021, pp. 4904–4916.

[13] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo *et al.*, "Improving image generation with better captions," *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, vol. 2, no. 3, p. 8, 2023.

[14] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[15] Y. Zhou, J. Yang, H. Zou, and L. Xie, "Tent: Connect language models with iot sensors for zero-shot activity recognition," *arXiv preprint arXiv:2311.08245*, 2023.

[16] L. Arrotta, C. Bettini, G. Civitarese, and M. Fiori, "Contextgpt: Infusing llms knowledge into neuro-symbolic activity recognition models," *arXiv preprint arXiv:2403.06586*, 2024.

[17] X. Gao, H. Luo, Q. Wang, F. Zhao, L. Ye, and Y. Zhang, "A human activity recognition algorithm based on stacking denoising autoencoder and lightgbm," *Sensors*, vol. 19, no. 4, p. 947, 2019.

[18] S. Münzner, P. Schmidt, A. Reiss, M. Hanselmann, R. Stiefelhagen, and R. Dürichen, "Cnn-based sensor fusion techniques for multimodal human activity recognition," in *Proc. ACM Int'l Symp. wearable computers*. ACM, 2017, pp. 158–165.

[19] S. Mekruksavanich and A. Jitpattanakul, "Lstm networks using smartphone data for sensor-based human activity recognition in smart homes," *Sensors*, vol. 21, no. 5, p. 1636, 2021.

[20] Y. Zhao, R. Yang, G. Chevalier, X. Xu, and Z. Zhang, "Deep residual bidir-lstm for human activity recognition using wearable sensors," *Mathematical Problems in Engineering*, vol. 2018, pp. 1–13, 2018.

[21] H. Martin, D. Bucher, E. Suel, P. Zhao, F. Perez-Cruz, and M. Raubal, "Graph convolutional neural networks for human activity purpose imputation," in *Spatiotemporal workshop co-located with NIPS*, 2018.

[22] A. Mohamed, F. Lejarza, S. Cahail, C. Claudel, and E. Thomaz, "Hargcnn: Deep graph cnns for human activity recog. from highly unlabeled mobile sensor data," in *Proc. PerCom*. IEEE, 2022, pp. 335–40.

[23] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[24] X. Yang, A. Chen, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, A. B. Costa, M. G. Flores *et al.*, "A large language model for electronic health records," *NPJ digital medicine*, vol. 5, no. 1, p. 194, 2022.

[25] K. Xia, W. Li, S. Gan, and S. Lu, "Ts2act: Few-shot human activity sensing with cross-modal co-learning," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 4, pp. 1–22, 2024.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[27] H. Li, A. Shrestha, H. Heidari, J. Le Kernec, and F. Fioranelli, "Bi-lstm network for multimodal continuous human activity recognition and fall detection," *IEEE Sensors Journal*, vol. 20, no. 3, pp. 1191–1201, 2019.

[28] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE trans. Signal Proc.*, vol. 45, no. 11, pp. 2673–2681, 1997.

[29] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[30] Y. Liu, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[31] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, and S. Riedel, "Language models as knowledge bases?" *arXiv preprint arXiv:1909.01066*, 2019.

[32] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[33] A. Open, "Chatgpt (mar 14 version)[large language model]," 2023.

[34] V. Sanh, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[35] B. Chakraborty, O. Rudovic, and J. Gonzalez, "View-invariant human-body detection with extension to human action recognition using component-wise hmm of body parts," in *Int'l Conf. Automatic Face & Gesture Rec.* IEEE, 2008, pp. 1–6.

[36] E. Fridriksdottir and A. G. Bonomi, "Accelerometer-based human activity recognition for patient monitoring using a deep neural network," *Sensors*, vol. 20, no. 22, p. 6424, 2020.