HAL open science

Generative Resident Separation and Multi-label Classification for Multi-person Activity Recognition

Xi Chen, Julien Cumin, Fano Ramparany, Dominique Vaufreydaz

To cite this version:

Xi Chen, Julien Cumin, Fano Ramparany, Dominique Vaufreydaz. Generative Resident Separation and Multi-label Classification for Multi-person Activity Recognition. Context and Activity Modeling and Recognition (CoMoReA) Workshop at IEEE International Conference on Pervasive Computing and Communications (PerCom 2024), Mar 2024, Biarritz, France. hal-04538267

HAL Id: hal-04538267

https://hal.science/hal-04538267v1

Submitted on 9 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

GENERATIVE RESIDENT SEPARATION AND MULTI-LABEL CLASSIFICATION FOR MULTI-PERSON ACTIVITY RECOGNITION

AUTHOR VERSION

Xi Chen[1,2], Julien Cumin[1], Fano Ramparany[1], Dominique Vaufreydaz[2,]

[1] Orange Innovation

[2] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

ABSTRACT

This paper presents two models to address the problem of multi-person activity recognition using ambient sensors in a home. The first model, Seq2Res, uses a sequence generation approach to separate sensor events from different residents. The second model, BiGRU+Q2L, uses a Query2Label multi-label classifier to predict multiple activities simultaneously. Performances of these models are compared to a state-of-the-art model in different experimental scenarios, using a state-of-the-art dataset of two residents in a home instrumented with ambient sensors. These results lead to a discussion on the advantages and drawbacks of resident separation and multi-label classification for multi-person activity recognition.

1 Introduction

Ambient-based activity recognition has garnered growing interest due to its non-intrusive, privacy-friendly, and cost-effective properties. This technology leverages ambient sensors strategically placed in the environment (such as a home) to capture changes and interactions in their proximity. These recorded changes are referred to as sensor events. By analysing sequences formed by these sensor events, residents' activities in the environment can be identified. However, in real-home scenarios, there are often multiple residents, and sensor events captured in these situations correspond to potentially multiple and intertwined activities. Activity recognition in such situations is referred to as **multi-person activity recognition**.

The primary challenge in multi-person activity recognition is to separate activity information for each person. Existing methods can generally be categorized into 2 classes based on when this separation occurs: **resident separation** and **multi-label classification**. On one hand, resident separation aims to distinguish sensor events triggered by different residents, and subsequently, perform individual activity recognition on each separated sensor event sequence. On the other hand, multi-label classification methods involve extracting global features from sensor event sequences and then using these features to recognize multiple activity classes associated with individuals, thereby recognizing multi-person activities.

This paper presents several approaches: one based on resident separation, called Seq2Res, and another based on multi-label classification, called BiGRU+Q2L. A third approach combines them into a two-stage model. Unlike previous separation approaches that assign sensor events to residents one by one, Seq2Res employs a Sequence-to-Sequence (Seq2Seq) [18] architecture. It models the entire sensor sequence and generates separated sequences based on the modeled context. On the other hand, BiGRU+Q2L uses attention mechanisms to establish correlations not only among activity labels but also between labels and features. This enables a more accurate and flexible multi-label classification. Finally, the two approaches are combined in a model that separates resident information while considering the correlation of residents' activities.

This paper is organized as follows: Section 2 provides a summary of related work on resident separation and multi-label classification. Section 3 presents the Seq2Res and BiGRU+Q2L models, as well as their combination in a two-stage model. Section 4 describes the experimental results of these models as well as state-of-the-art models, on a state-of-the-art dataset. Finally, a conclusion is given in Section 5.

2 Related Work

2.1 Resident Separation

Crandall and Cook [7] use a supervised Naïve Bayes model to assign each sensor event to specific residents of a home, a problem often called **data association** in the literature. This method is highly reliant on the timing of events and resident schedule habits for classification, without consideration for spatiotemporal relationships between sensor events. Riboni *et al.* [16] modeled these spatiotemporal relationships by conducting a statistical analysis of the co-occurrence frequency of two adjacent sensor events within a defined temporal window in single-person data. If two sensor events, with rare co-occurrences in single-person data, happen in multi-person data within the defined temporal window, it suggests they come from different residents. In this approach, training requires pre-separated single-person data. Arrotta *et al.* [2] presented MICAR, a knowledge-based approach for data association, where sensor events are assigned to corresponding residents

using ontological reasoning on context. While less affected by data scarcity, it relies on explicit information like the location of each user, which may not always be available.

Bouchabou *et al.* [5] drew inspiration from language models used in natural language processing. Each sensor event is considered as a token and modeled using a word embedding skip-gram model [12]. The advantage of this approach is its ability to establish more flexible and richer event correlations and can be directly applied in multi-person data. Similarly, SMRT [20] and GAMUT [21] also adopted skip-gram models to map sensors into a latent space. In addition, these methods used a linear Gaussian dynamic model and a Gaussian Mixture Probability Hypothesis Density (GM-PHD) filter to track residents' states in the latent space. While these probabilistic models enable unsupervised resident separation, their use of the linear dynamic model results in uneven tracking capabilities for residents or pets with diverse mobility profiles, rendering the model sensitive to anomalous sensor events.

2.2 Multi-label classification

To identify the resident associated with a specific activity without resident separation, multi-label classification methods typically combine each activity class with a corresponding resident identifier. Alternatively, these methods may opt for anonymous activity classification, without predicting the individual responsible for each activity.

The most straightforward multi-label classification method is binary relevance [1, 8, 10], where each activity label is predicted by an individual binary classifier. The problem of this approach is its inability to consider the interdependence between activity classes. For instance, in real-life scenarios, activities like "*User A using the toilet*" and "*User B using the toilet*" are less likely to occur simultaneously. Binary relevance struggles to learn such patterns because predictions for two labels are independent. Extending binary relevance, the classifier chain method [9, 13] introduces dependencies between binary classifiers by using the output of one classifier as a feature for the next classifier. However, the performances of such methods are highly dependent on the ordering of classifiers. Another extension of the binary relevance method is the work of Liu *et al.*, who propose Query2Label (Q2L) [11]. This method embedded labels as vectors and used Transformer decoders [19] to model inter-label relationships and then queried the label-related features from the feature space.

A number of works used a label combination method [4, 6, 14]. This method defines combinations of activities that are performed simultaneously by different persons as new labels and uses single-label classifiers to predict these combinations. As such, dependencies between the initial activity labels are hard-coded as the new labels, reducing the training complexity and often resulting in higher performances. For example, Chen *et al.* [6] achieve state-of-the-art performance with TransBi-GRU, a combination of Transformer [19] and Bidirectional Gated Recurrent Units (BiGRU) for feature extraction, with a label combination classifier at the end. Label combination has three main drawbacks: complexity grows exponentially with the number of persons and activity classes; class imbalance is exacerbated; the trained model cannot predict combinations of classes that do not exist in the training set.

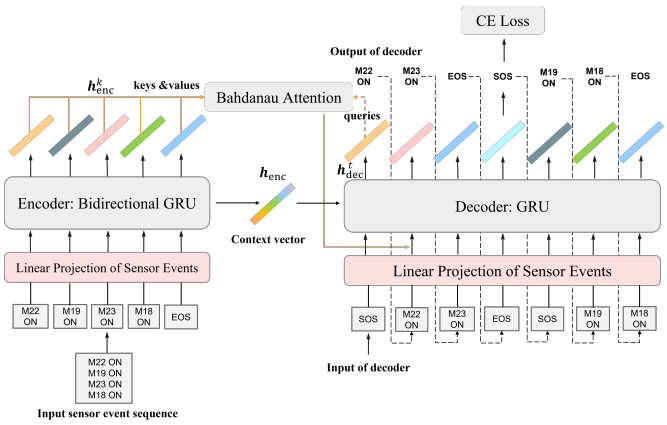

Figure 1: *Framework of the proposed Seq2Res model for resident separation.*

3 Proposed models

3.1 Seq2Res resident separation model

In this research, it is assumed that, as in [6, 16], there are two residents living in a smart home equipped with ambient sensors. Theoretically, this research can be extended to scenarios involving more than two residents, but this is left as future work. Given a sequence of sensor events $\{e_k\}$, the objective is to assign each e_k to one of two sets $\{e^1_k\}$ and $\{e^2_k\}$ where each set represents an event sequence of a resident. Existing resident separation methods generally determine the belonging of the next sensor event based only on the resident's state at the previous time step. These methods not only overlook longer-term contextual information but also can lead to error accumulation. To address these issues, the present proposal attempts to relax the constraints of the previous algorithms. In contrast with previous approaches, a generative method is used to "translate" a sensor event sequence triggered by multiple individuals into separate event sequences for each resident. As a result, the separated sequences $\{e^1_k\}$ and $\{e^2_k\}$ are no longer constructed by partitioning sensor events one by one but are generated based on the overall context of the input sequence. This means that the two sequences no longer guarantee that $\{e^1_k\} \cup \{e^2_k\} = \{e_k\}$. An attention-based Sequence to Sequence (Seq2Seq) architecture is used, based on the work of Bahdanau *et al.* [3]. Figure 1 illustrates the proposed model, which is called **Seq2Res** (Sequence to Residents).

3.1.1 Input sequence encoding

This step is illustrated on the left side of Figure 1. As in [5], a sensor event is represented as a numerical token. To better capture spatiotemporal relationships between sensors, the input token sequence is mapped into an embedding space of dimension D. Then, a bidirectional GRU is used to encode the bidirectional temporal characteristics of the event sequence,

resulting in output vectors $\{\mathbf{h}_{\text{enc}}^k\}$ and a context vector $\mathbf{h}_{\text{enc}}$, with a dimension of $2D$.

3.1.2 Output sequence generation

Designing decoders capable of generating separated sequences presents a significant challenge, particularly in generating one resident's event sequence while taking into account the generation of the other resident's sequence. A straightforward but effective approach involves employing a single decoder for the sequential generation of two sequences. This requires passing the hidden state from the first sequence generation to the subsequent sequence. In this work, the decoder is designed to produce a unified sequence where e_k^1 and e_k^2 are continuously generated. The initial hidden state for generating $\{e_k^2\}$ is thus the hidden state of $\{e_k^1\}$, and the initial hidden state for generating $\{e_k^1\}$ is $\mathbf{h}_{\text{enc}}$.

As illustrated in the right part of Figure 1, taking a "Start of Sequence" token SOS as input and $\mathbf{h}_{\text{enc}}$ as hidden state, a GRU-based decoder is applied to give an output vector and a new hidden state vector. The output is then mapped into the probability vector of events using a fully connected network and a softmax function. The event with the highest probability will be considered as the generated event in this step. Generated events will serve as input for the next step, prompting the decoder to generate based on the context and the existing sequence. After the first sequence is generated, the model is trained to generate an "End of Sequence" token EOS, followed by an SOS to prompt the generation of the second sequence until a second EOS is finally generated.

Since the generation of the first sequence depends on the first event, we set the resident who triggers the first event in the input sequence as resident 1. To train the model, labels of separated sequence are of the form $\{\{e_k^1\}, \text{EOS}, \text{SOS}, \{e_k^2\}, \text{EOS}\}$, where $\{e_k^i\}$ is the event sequence of resident i. Cross Entropy Loss (CE Loss) is used as the loss function.

3.1.3 Bahdanau Attention

Due to the inherent limitations of encoding the entire input with a single context vector, an attention mechanism is added to enhance the decoder's aligning capability. This allows the decoder to focus on different parts of the encoder's output at each step of the decoding process.

Specifically, the Bahdanau attention mechanism [3] is applied. In each decoding step t, the decoder's hidden state at the previous time step $\mathbf{h}_{\text{dec}}^{t-1}$ functions as the query. The encoder's outputs $\{\mathbf{h}_{\text{enc}}^k\}$ serve as both keys and values. Following the computation of Bahdanau attention scores, a weighted average is computed across the encoder's output vectors $\{\mathbf{h}_{\text{enc}}^k\}$, yielding a context vector $\mathbf{c}_t$. This context vector $\mathbf{c}_t$ is then concatenated with the embedding vector of the decoder's input at time t and subsequently fed into the GRU. The Bahdanau attention score $s_{t,k}$ between the query $\mathbf{h}_{\text{dec}}^{t-1}$ and the key $\mathbf{h}_{\text{enc}}^k$ is computed as

$$s_{t,k} = \mathbf{v}^\top \tanh(\mathbf{W}\mathbf{h}_{\text{dec}}^{t-1} + \mathbf{U}\mathbf{h}_{\text{enc}}^k + \mathbf{b}),$$

where $\mathbf{v}$, $\mathbf{W}$, $\mathbf{U}$, and $\mathbf{b}$ are learnable parameters.

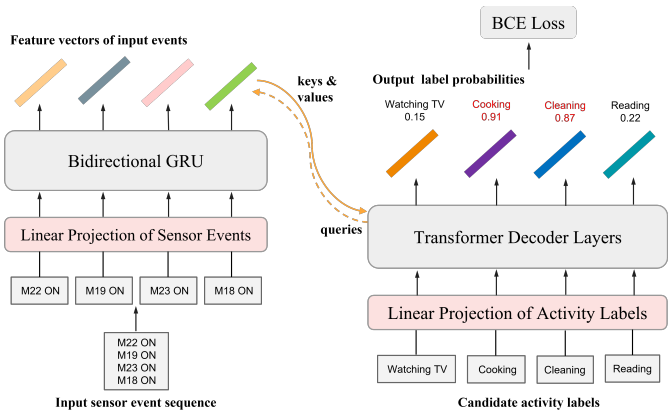

Figure 2: *Framework of the proposed BiGRU+Q2L model, with feature extraction from sensor events depicted on the left, and Query2Label transformer decoder for multi-label classification on the right.*

3.2 BiGRU+Query2Label multi-label classification model

In the following, we introduce a multi-label classifier based on an attention mechanism, while employing a BiGRU model as a sequence feature extractor. Figure 2 illustrates the framework of this method.

3.2.1 BiGRU-based feature extractor

The feature extractor of the proposed multi-label classification model is presented on the left side of Figure 2. The input event sequence, whether separated or not, is linearly mapped to the embedding space and then processed by a BiGRU to extract bidirectional temporal information. In the TransBiGRU model [6], 6 composite layers of Transformer encoder coupled with BiGRU are used. Following preliminary experiments on the same dataset as in [6], it appears that BiGRU is the important component, and Transformer layers play a lesser role. As such, only BiGRU layers are used for the encoder in this work. Comparative results between the two models can be found in Section 4.5.

3.2.2 Query2Label (Q2L) multi-label classifier

Given an input event sequence $\{e_k\}$ and L candidate activity labels, multi-label classification is to predict $\{y_l\}_{1\leq l\leq L}$, where $y_l \in \{0, 1\}$ is a binary indicator to describe whether the class l is present in the sequence. A straightforward multi-label classification method is Binary Relevance (BN): the features extracted by the BiGRU are averaged and then fed into L independent binary fully connected classifiers to predict $\{y_l\}_{1\leq l\leq L}$. This model is denoted as BiGRU+BN. To enhance the model's ability to extract label correlations and pay attention to important features of the sequence, we further propose the BiGRU+Q2L model, depicted in Figure 2, which utilizes the Query2Label (Q2L) [11] model as the multi-label classifier.

Query2Label embeds candidate labels into a label embedding vector and then feeds them into a Transformer decoder, each layer consisting of a self-attention module, a cross-attention module, and a position-wise feed-forward network. In the self-attention module, query, key, and value are all the label embedding vectors. Unlike binary relevance, the correlation between labels can be learned in this module. In the cross-attention module, the queries are label embeddings, whereas keys and values are the temporal features extracted by the encoder. This module allows each label to be associated with its desired features and pool them by linear combinations. The queried feature vectors are then fed to the position-wise feed-forward networks for further non-linear transformations.

Therefore, the output vectors at each label position in the Transformer decoder are a fusion and transformation of the label-related features of the input sequence. We apply a linear transformation to each output, followed by the sigmoid function to frame it into the range $[0, 1]$. This numerical value, denoted as p_l, represents the probability of presence of the label l. Empirically, labels with probabilities greater than 0.7 are finally considered as output predictions. Binary Cross Entropy Loss (BCE Loss) is used as the loss function.

3.3 Multi-label classification with resident separation

As mentioned earlier, the distinction between resident separation and multi-label classification lies in the timing of information separation: they are not mutually exclusive. The proposed Seq2Res and BiGRU+Q2L/BN models can be combined into a two-stage model, where Seq2Res is first used to perform resident separation on the mixed sequence, and then the output separated sequence is treated as a whole input for activity recognition in BiGRU+Q2L/BN. Compared to directly conducting individual activity recognition after resident separation, this two-stage approach allows for co-consideration of sensor events from both individuals before classification. The input of BiGRU+Q2L is the sequence of softmax probability vectors from the output of the fully connected network of Seq2Res, rather than the exact most probable events, to retain the most information.

4 Experimental Results

4.1 Dataset

The proposed approaches are evaluated on a real-world dataset known as the Multiresident ADL Activities (ADLMR) dataset[1], which was published by the Center for Advanced Studies in Adaptive Systems (CASAS) of the Washington State University [17]. This dataset comprises 26 subsets, each representing a single day, with each day containing sensor events triggered by 2 residents (a different pair each day) performing activities among a shared set of 15 classes. There is a total of 37 different ambient sensors in the home, such as motion and opening sensors. Each sensor event has been manually labeled with the identity of the resident triggering this event, along with the activity they were engaged in. Therefore, this dataset can be used for both resident separation as in [16] and multi-subject activity recognition as in [6].

[1] `http://casas.wsu.edu/datasets/adlmr.zip` (last seen on 11/2023)

4.2 Experimental setup

4.2.1 Evaluation method

For each experiment, we used 10-fold cross-validation. Data was partitioned using the scikit-learn library. Since each day is performed by different pairs of residents, one day can not be split into different folds to ensure cross-resident independence. As such, each fold contains entire days. In order to cover the whole 26 days, the test set in the first 6 folds contains data for 3 days, ranging from day 1 to day 18; in the subsequent 4 folds, each test set contains data for 2 days, covering days 18 to 26.

4.2.2 Data preparation

Due to typographical errors in the annotations of the original dataset, we corrected some of the labels. In addition, for events with labels of single-resident activity, we assigned the last performed activity of the other resident as a second label, so that each event is labeled with the activity of both residents. To reduce data redundancy, we excluded events in which motion sensors were automatically deactivated. By applying a sliding window approach, we segmented data so that each data instance contains 16 sensor events, with a step of 3 events between each instance. The activity label of an instance is the result of majority voting between the labels of the last 3 events.

4.2.3 Parameters

All parameters were set following preliminary experiments. For the encoder of Seq2Res, the embedding size and the hidden size of BiGRU are 128. The output size and the context vector size are then $2 \times 128 = 256$. A dropout rate of 0.1 is used. For the decoder, the embedding size and hidden size are both 256. Given that the input to the GRU is a concatenation of the embedded input vector and the attention-queried vector, the input size for GRU is $2 \times 256 = 512$. A dropout rate of 0.4 is used for the decoder. The initial learning rate for training is 0.001, with a halving schedule every 80 epochs. Seq2Res was trained for a total of 300 epochs.

For BiGRU+Q2L, the event embedding size and the hidden size of BiGRU are both 128, resulting in an output feature vector size of $2 \times 128 = 256$. The label embedding size of Q2L is consequently set to 256. A dropout rate of 0.3 is used. A learning rate of 1×10^{-4} is used for a 100 epochs training. The Adam optimizer is used, and a batch size of 100 is used across all training sessions.

4.3 Metrics

4.3.1 Resident separation

Former resident separation compute typically the accuracy of each event assignment through one-to-one comparisons between the prediction and the ground truth in terms of their

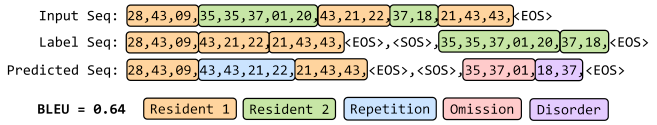

Figure 3: *Example of an input sequence of Seq2Res model with its label and prediction. Each number represents the token of a sensor event.*

Table 1: *Performance of resident separation with Seq2Res.*

BLEU		Seq2Res (ours)	SMRT [20]
Class	Fill medication dispenser	**0.7906**	0.6152
	Hang up clothes	**0.8397**	0.7225
	Move couch and table	**0.4844**	0.4326
	Read on couch (user B)	**0.4799**	0.4509
	Water plants	0.7174	**0.7411**
	Sweep kitchen floor	**0.6212**	0.4899
	Play checkers	**0.6389**	0.5016
	Set out dinner ingredients	**0.7017**	0.6387
	Set dinner table	**0.6993**	0.6538
	Read on couch (user A)	**0.5713**	0.5372
	Pay electric bill	**0.5619**	0.5445
	Prepare picnic basket	**0.6086**	0.6073
	Retrieve dishes	**0.5887**	0.5513
	Pack supplies in basket	**0.6290**	0.4708
	Pack food in basket	**0.6266**	0.4864
Overall BLEU		**0.6385**	0.5608

positions. Our separation method consists of sequence generation, making it challenging to establish a direct one-to-one alignment. Moreover, calculating accuracy on a one-to-one alignment fails to consider the coherence of the separated sequence, which are crucial for extracting temporal information from the sequence. Hence, we borrowed a metric commonly used in machine translation, Bilingual Evaluation Understudy (BLEU) [15], to assess our separation results. An illustrative example is given in Figure 3, in which the BLEU score between the prediction and the label is 0.64.

Given a generated sequence c and a reference sequence r, the BLEU metric considers the precision of N-grams, which is the proportion of N-gram phrases in the generated sequence c that appear in the reference sequence r, and penalizes shorter sequences. Like most studies, the final BLEU of this research is the average of scores corresponding to N from 1 to 4.

4.3.2 Activity Recognition

We use the standard metrics of accuracy, recall, precision, and F1 score.

4.4 Results on resident separation

We conduct experiments on resident separation using the Seq2Res model and reproduce the SMRT (Sensor-based Multi-resident Tracking) [20] under the same protocol for comparison. Table 1 reports, for both methods, the cross-validation average BLEU for each activity class and the overall average BLEU. We see that the overall performance of Seq2Res is higher than that of SMRT. This could be attributed to the fact that Seq2Rees, compared to SMRT, is better able to consider a longer context (by the encoder) while ensuring the correlation between two consecutive sensor events (by the decoder). The overall BLEU of Seq2Res reaches 0.6385. Three types of error are generally observed in generated sequences: repetition, omission and disorder, as illustrated in Figure 3. A significant variation across different classes is also observed. For example, "*Move couch and table*" and "*Read on couch (user B)*" have BLEU scores below 0.5. These two activities always occur in close proximity, and the trajectories of the two residents significantly overlap, making it more difficult to separate. Conversely, "*Fill medication dispense*" and "*Hang up clothes*" take place at opposite sides of the house, with minimal overlap in the trajectories of the residents. As a result, the BLEU scores for these two activities reach around 0.8. In general, the separation ability of the model is negatively correlated with the degree of overlap in the actions of the two residents, which conforms to intuition.

Table 2: *Performance of activity recognition models for 3 scenarios.*

Scenario	Model	Accuracy (%)	Macro-F1 (%)
No_Sep	BiGRU+BN	87.66 (0.31)	86.09 (0.37)
	TransBiGRU+BN [6]	86.85 (0.34)	85.13 (0.38)
	BiGRU+Q2L (ours)	**88.47 (0.23)**	**87.07 (0.25)**
S2S_Sep	BiGRU+BN	79.36 (0.54)	76.74 (0.70)
	TransBiGRU+BN Chen et al. [6]	73.38 (0.56)	70.12 (0.66)
	BiGRU+Q2L (ours)	**83.26 (0.39)**	**81.08 (0.47)**
GT_Sep	BiGRU+BN	88.70 (0.40)	87.15 (0.48)
	TransBiGRU+BN Chen et al. [6]	87.35 (0.38)	85.68 (0.44)
	BiGRU+Q2L (ours)	**90.87 (0.32)**	**89.57 (0.38)**

4.5 Results on activity recognition

In this section, 3 scenarios are used to compare the performance of multi-resident activity recognition models:

- No_Sep: Inputs are event sequences without separation.
- S2S_Sep: Inputs are event sequences generated by Seq2Res as introduced in Section 3.3.
- GT_Sep: Inputs are event sequences separated based on the ground truth labeled in the dataset.

Under these 3 scenarios, we first evaluate the recognition accuracy and macro-F1 score of 3 different models: TransBiGRU [6] using a binary relevance classifier, BiGRU+BN and BiGRU+Q2L (both presented in Section 3.2.2). For TransBiGRU, we used the same hyperparameters as in [6].

Table 2 reports the average cross-validation performance of these 3 models for each scenario, with standard deviations in parentheses. BiGRU+Q2L achieves the best performance for all 3 scenarios. Compared to the BiGRU+BN model, BiGRU+Q2L exhibits statistically significant improvement in performances, especially for macro-F1 compared to accuracy. This indicates that the Query2Label classifier can help address data imbalance between activity classes, which is a common problem in human activity recognition. The performance of TransBiGRU [6] is significantly lower than BiGRU+BN, a lighter model with fewer parameters, for all scenarios. For the

Table 3: *Performance of BiGRU+Q2L for each activity class, for 3 scenarios.*

Metric		Precision (%)			Recall (%)			F1-score (%)			Count
Scenario		No_Sep	S2S_Sep	GT_Sep	No_Sep	S2S_Sep	GT_Sep	No_Sep	S2S_Sep	GT_Sep	
Class	Fill medication dispenser	89.54	85.79	93.94	92.71	87.55	96.12	91.10	86.66	95.02	4770
	Hang up clothes	89.97	85.19	88.44	92.13	87.65	90.51	91.04	86.40	89.46	2890
	Move couch and table	86.00	75.92	87.80	84.01	79.79	86.63	84.99	77.81	87.21	2450
	Read on couch (user B)	84.15	77.43	90.58	84.22	80.92	90.92	84.18	79.13	90.75	2410
	Water plants	82.25	76.10	86.10	85.06	78.62	88.72	83.63	77.33	87.39	2000
	Sweep kitchen floor	92.58	87.81	93.90	89.26	85.61	93.04	90.89	86.70	93.47	4840
	Play checkers	93.64	90.64	95.53	90.94	87.79	93.15	92.27	89.19	94.33	5910
	Set out dinner ingredients	80.05	73.96	83.30	80.95	77.29	89.92	80.50	75.59	86.48	1970
	Set dinner table	83.71	79.77	86.41	83.56	80.51	86.33	83.63	80.14	86.37	3480
	Read on couch (user A)	81.08	73.77	83.70	79.34	75.16	81.86	80.20	74.46	82.77	2970
	Pay electric bill	84.72	79.06	86.22	82.39	77.99	83.06	83.54	78.52	84.61	3070
	Prepare picnic basket	91.05	91.12	94.83	91.42	85.96	93.62	91.23	88.46	94.22	5710
	Retrieve dishes	90.02	90.97	93.08	92.35	86.59	93.05	91.17	88.73	93.06	5750
	Pack supplies in basket	83.44	70.13	85.10	85.18	74.48	87.13	84.30	72.24	86.19	3080
	Pack food in basket	84.38	69.84	84.25	84.88	73.88	85.13	84.63	71.81	84.69	3220
Average		86.44	80.50	88.88	86.56	81.32	89.28	86.49	80.88	89.06	3635

S2S_Sep scenario, where the difference is significantly large, the excessively deep network of TransBiGRU is overfitting on the noise in the generated separated sequences, resulting in significant drops in performance.

Comparing the results across the three scenarios, we observe that GT_Sep (using ground truth resident labels) has an accuracy and macro-F1 score that are 2.4% and 2.5% higher, respectively, than No_Sep in the BiGRU+Q2L model (and similar gaps for the other 2 models). This indicates that resident separation does help multi-resident activity recognition when this separation is perfectly accurate. Performances for the S2S_Sep scenario reach the same orders of magnitude, with accuracies as high as 83.26% with BiGRU+Q2L. However, they are significantly lower than for both GT_Sep and No_Sep scenarios, for all models. This suggests that, although the separated sequences generated by the Seq2Res model are overall representative of the true separation, the errors introduced during the generation process have a significant impact on the final activity classification. In general, these results show that resident separation can improve activity recognition, but only if the separated sequences are very accurate, which remains a scientific challenge.

To further investigate the behavior of models for each scenario, we report in Table 3 the cross-validation average precision, recall, and F1 scores of the BiGRU+Q2L model, per activity class. We observe that models for the S2S_Sep scenario achieve better results for classes with larger numbers of instances (e.g. "*Play checkers*", "*Prepare picnic basket*", "*Retrieve dishes*"), but underperform for classes with a smaller number of instances (e.g. "*Water plants*", "*Set out dinner ingredients*"). This could be because the noise introduced by generative resident separation increases the variation of the classifier input, which requires more instances to learn. On the other hand, comparing GT_Sep and No_Sep highlights that accurate resident separation helps to recognize classes with a small number of instances (e.g. "*Set out dinner ingredients*", "*Read on couch (user B)*") because the temporal features of the accurately separated sequences may be easier to learn.

5 Conclusion

In this paper, two models are presented: Seq2Res for resident separation, and BiGRU+Q2L for multi-resident activity recognition. On the CASAS ADLMR dataset, BiGRU+Q2L achieves better performance than another state-of-the-art model TransBiGRU, with a simpler architecture. While the Seq2Res model shows potential for the resident separation task, the quality of the generated sequences is still limited. As such, the combination of Seq2Res and BiGRU+Q2L does not yet reach the same performance as using only BiGRU+Q2L.

Experiments with ground truth separation have highlighted that using perfect resident separation before multi-resident activity recognition can significantly improve performance. Therefore, future work on improving resident separation must be conducted, such as post-processing methods to improve the sequences generated by Seq2Res.

References

[1] Alaa Alhamoud, Vaidehi Muradi, Doreen Böhnstedt, and Ralf Steinmetz. Activity recognition in multi-user environments using techniques of multi-label classification. In *Proceedings of the 6th International Conference on the Internet of Things*, pages 15–23, 2016.

[2] Luca Arrotta, Claudio Bettini, and Gabriele Civitarese. Micar: Multi-inhabitant context-aware activity recognition in home environments. *Distributed and Parallel Databases*, 41(4):571–602, 2023.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[4] Asma Benmansour, Abdelhamid Bouchachia, and Mohammed Feham. Modeling interaction in multi-resident activities. *Neurocomputing*, 230:133–142, 2017.

[5] Damien Bouchabou, Sao Mai Nguyen, Christophe Lohr, Benoit Leduc, and Ioannis Kanellos. Fully convolutional network bootstrapped by word encoding and embedding for activity recognition in smart homes. In *Deep Learning for Human Activity Recognition: Second International Workshop, DL-HAR 2020, Held in Conjunction with IJCAI-PRICAI 2020, Kyoto, Japan, January 8, 2021, Proceedings 2*, pages 111–125. Springer, 2021.

[6] Dong Chen, Sira Yongchareon, Edmund M-K Lai, Jian Yu, Quan Z Sheng, and Yafeng Li. Transformer with bidirectional gru for nonintrusive, sensor-based activity recognition in a multiresident environment. *IEEE Internet of Things Journal*, 9(23):23716–23727, 2022.

[7] Aaron S Crandall and Diane Cook. Attributing events to individuals in multi-inhabitant environments. In *2008 IET 4th International Conference on Intelligent Environments*, pages 1–8. IET, 2008.

[8] Manan Jethanandani, Thinagaran Perumal, Yuh-Ching Liaw, Jieh-Ren Chang, Abhishek Sharma, and Yipeng Bao. Binary relevance model for activity recognition in home environment using ambient sensors. In *2019 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, pages 1–2. IEEE, 2019.

[9] Manan Jethanandani, Abhishek Sharma, Thinagaran Perumal, and Jieh-Ren Chang. Multi-label classification based ensemble learning for human activity recognition in smart home. *Internet of Things*, 12:100324, 2020.

[10] Rahul Kumar, Imroj Qamar, Jaskaran Singh Virdi, and Narayanan Chatapuram Krishnan. Multi-label learning for activity recognition. In *2015 International Conference on Intelligent Environments*, pages 152–155. IEEE, 2015.

[11] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021.

[12] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[13] Raihani Mohamed, Thinagaran Perumal, Md Nasir Sulaiman, and Norwati Mustapha. Multi resident complex activity recognition in smart home: A literature review. *Int. J. Smart Home*, 11(6):21–32, 2017.

[14] Raihani Mohamed, Thinagaran Perumal, Md Nasir Sulaiman, Norwati Mustapha, and MN Zainudin. Modeling activity recognition of multi resident using label combination of multi label classification in smart home. In *AIP conference proceedings*, volume 1891. AIP Publishing, 2017.

[15] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[16] Daniele Riboni and Flavia Murru. Unsupervised recognition of multi-resident activities in smart-homes. *IEEE Access*, 8:201985–201994, 2020.

[17] Geetika Singla, Diane J Cook, and Maureen Schmitter-Edgecombe. Recognizing independent and joint activities among multiple residents in smart environments. *Journal of ambient intelligence and humanized computing*, 1:57–63, 2010.

[18] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[20] Tinghui Wang and Diane J Cook. smrt: Multi-resident tracking in smart homes with sensor vectorization. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2809–2821, 2020.

[21] Tinghui Wang and Diane J Cook. Multi-person activity recognition in continuously monitored smart homes. *IEEE transactions on emerging topics in computing*, 10 (2):1130–1141, 2021.