AttnSense: Multi-level Attention Mechanism For Multimodal Human Activity Recognition

Haojie Ma, Wenzhong Li, Xiao Zhang, Songcheng Gao and **Sanglu Lu**
State Key Laboratory for Novel Software Technology, Nanjing University
mhj1137633684@gmail.com, lwz@nju.edu.cn

Abstract

Sensor-based human activity recognition is a fundamental research problem in ubiquitous computing, which uses the rich sensing data from multimodal embedded sensors such as accelerometer and gyroscope to infer human activities. The existing activity recognition approaches either rely on domain knowledge or fail to address the spatial-temporal dependencies of the sensing signals. In this paper, we propose a novel attention-based multimodal neural network model called AttnSense for multimodal human activity recognition. AttnSense introduce the framework of combining attention mechanism with a convolutional neural network (CNN) and a Gated Recurrent Units (GRU) network to capture the dependencies of sensing signals in both spatial and temporal domains, which shows advantages in prioritized sensor selection and improves the comprehensibility. Extensive experiments based on three public datasets show that AttnSense achieves a competitive performance in activity recognition compared with several state-of-the-art methods.

1 Introduction

Using the information extracted from various sensor modalities (accelerometer, gyroscope, etc.), a HAR system can recognize various activities, such as running, walking, etc. HAR systems are used in a large number of context-aware applications, including but not limited to medication intake, health monitoring and fitness tracker [Khan *et al.*, 2013].

Most HAR systems split a sensor signal into fixed sized sequences by a sliding window and classify each sequence to one activity by a recognition algorithm. Early recognition algorithms were mainly based on shallow supervised machine learning methods such as support vector machine (SVM) and random forest (RF). Such shallow supervised machine learning algorithms heavily relied on handcrafted features, which were usually limited by human domain knowledge [Bengio, 2013] and could only capture shallow features [Yang *et al.*, 2015]. With the rapid development of deep learning [LeCun *et al.*, 2015], more and more researchers tried to apply Deep Neural Networks (DNN), especially Convolutional Neural networks (CNN) and Recurrent Neural Networks (RNNs), to HAR systems to achieve automatic feature extraction without human domain knowledge [Wang *et al.*, 2018].

Recently, a novel deep learning model CNN-RNN had been successfully employed in HAR and outperformed the simple CNN and RNN model [Yao *et al.*, 2017] [Ordóñez and Roggen, 2016]. This CNN-RNN model used a CNN subnet to form a feature vector of the input signal at each timestep and feeds all generated feature vectors in a time window to an RNN subnet. To deal with multimodal sensing signals, some works [Radu *et al.*, 2017] [Yao *et al.*, 2017] adopted a *modality-specific architecture*, where separate CNN was built for each modality to first learn modality-specific information and then merged them to unified feature representation. The sequence of unified feature representation in a time window was further connected to the RNN subnet.

Still and all, two issues remain in HAR with the CNN-RNN model. Firstly, different sensor modalities come from different domains, and merge them without considering their difference may limit the model's ability. For example, the accelerometer features may be more significant in distinguishing the "walking" and "biking" activities; while the gyroscope features may be more significant in distinguishing the "turning-left" and "turning-right" activities. Treating them the same without distinction may degrade the performance of activity classification. Secondly, for a sensing signal time series, not all timesteps contribute equally to the activity recognition task. For example, the features on some timesteps may show more salient pattern than the others in distinguishing the "walking" and "running" activities. Therefore the model should consider the temporal dependencies for activity recognition.

To address these issues, we propose the usage of attention

This work was partially supported by the National Key R&D Program of China (Grant No. 2017YFB1001801), the Key R&D Program of Jiangsu Province, China (Grant No. BE2018116), the National Natural Science Foundation of China (Grant Nos. 61672278, 61832008, 61832005), the science and technology project from State Grid Corporation of China (Contract No. SGSNXT00YJJS1800031), the Collaborative Innovation Center of Novel Software Technology and Industrialization, and the Sino-German Institutes of Social Computing. Xiao Zhang was also supported by the program A for outstanding PhD candidate of Nanjing University. The corresponding authors are Wenzhong Li (lwz@nju.edu.cn) and Sanglu Lu (sanglu@nju.edu.cn).

mechanism for activity recognition. Attention model has been proved to perform well in the areas of speech and natural language processing [Kim and Lane, 2017], which can be viewed as weighted averaging of a series of input vectors. The attention weights represent the relative importance of the corresponding input, which can be learnt by neural networks. With such properties, the attention mechanism is suitable to fuse multi-modal sensing data in temporal-spacial domains.

In this paper, we propose an attention-based deep neural network model called AttnSense for human activity recognition. The proposed model consists of four parts: (1) an individual convolutional subnet for each sensor to extract modality-specific features;; (2) an attention-fusion subnet that considers the relative importance of each sensor modality to fuse modality-specific features; (3) an attention-based Gated Recurrent Units (GRU) subnet that extract the importance of different timestep and fuse the hidden state of the GRU; (4) an output layer that uses a softmax function to obtain the probabilities for activity recognition.

The contributions of our paper are summarized as follows.

- We propose a novel attention-based deep neural network called AttnSense for multimodal human activity recognition which reasonably applies attention mechanism to multimodal sensor data fusion and enhancement of GRU.
- By integrating attention layers with both a CNN subnet and a GRU subnet to recognize multimodal sensing data, the proposed attention mechanism can capture spatial and temporal dependencies of the multimodal sensor signal, which amplifies the more important and informative modalities and timesteps during classification.
- Extensive experiments are conducted based on three HAR datasets, which verify the effectiveness and efficiency of AttnSense. Visualized analysis of attention weights is provided to improve the model's comprehensibility. Furthermore, we also study the impact of some hyper-parameters like the structure of CNN and the width of the sliding window.

2 Related Work

Researches on sensor-based human activity recognition can be summarized as two categories: shallow machine learning approaches and deep learning approaches.

Shallow machine learning approaches relied on handcrafted features, such as mean, variance, maximum, differences and etc [Figo *et al.*, 2010]. These extracted features are fed into some shallow supervised machine learning models such as support vector machine [Bulling *et al.*, 2011], random forest [Stisen *et al.*,] and Hidden Markov Model (HMM) [van Kasteren *et al.*, 2008] for activity recognition. However, handcrafted features is limited by human domain knowledge.

Deep learning approaches applied the deep neural network (DNN) framework to perform automatic feature extraction and classification, which provided promising results in HAR domain [Yang *et al.*, 2015]. Early works primarily targeted to feature representation learning aspect. Deep Belief Network (DBN) is used for feature extraction from sensing signals and show some interesting results [Plötz *et al.*, 2011]. Deep Convolutional Neural Networks (CNN) was also applied to human activity recognition, several works took different sensor modalities as an image and fed it to a 2D CNN for feature extraction, which had been shown to capture salient features in the spatial dimension and outperforms shallow machine learning approaches [Hammerla *et al.*, 2016][Yang *et al.*, 2015]. In the meantime, the RNN with long short-term memory (LSTM) was proposed and successfully applied in HAR [Hammerla *et al.*, 2016][Guan and Plötz, 2017], which can capture the long-term information in time series.

Moreover, a hybrid CNN-RNN model that combined CNN and RNN had shown promising results in activity recognition performance [Ordóñez and Roggen, 2016] [Yao *et al.*, 2017]. However, a potential issue with these models is that a neural network needs to be able to compress all the necessary information of a input sequence, but the input sequence usually involves irrelevant parts in the spatial and temporal dimension. In this situation, we propose a novel attention-based deep neural network model called AttnSense for multimodal human activity recognition.

3 Problem Definition

In this paper, we assume that there are K different sensors that are attached to the human body and are synchronized to emit data. For example, a smartwatch or an inertial measurement unit (IMU) is usually equipped with the accelerometer, gyroscope and magnetometer, where each sensor could generate a signal vector at a time (e.g. accelerometer generates a signal along the x-axis, y-axis and z-axis). For those sensors, the sensing data along time can be represented by a multidimensional time series $\boldsymbol{\mathcal{S}}$

$$\boldsymbol{\mathcal{S}} = [\ S_1, \cdots, S_t, \cdots\ ] \tag{1}$$

where $S_t = [\ s_{t1}, \cdots, s_{tk}, \cdots, s_{tK}\ ]^T$, and s_{tk} is the sensing signal of the k-th sensor at time t. In a real deployment, the sensor signals can contain noises, and s_{tk} can be represented by

$$s_{tk} \quad = \quad {s_{tk}}^* + n_t \tag{2}$$

where ${s_{tk}}^*$ represents the noiseless sensing signal of k-th sensor at time t, and n_t is a noise of independent, zero-mean Gaussian random variables with variance σ^2.

The *activity recognition problem* can be described as follows. Given the sensing signal time series $\boldsymbol{\mathcal{S}}$, detect a series of activities (e.g., sitting, standing, and walking) that infers the human behaviours in a duration.

4 Method

We proposed the method of activity recognition based on attention neural networks, which framework is illustrated in Fig. 1. The details are explained below.

4.1 Data Preprocessing

To capture the noise pattern and frequency features in the multidimensional sensing signals, we propose a comprehensive data preprocessing technique that consists of data augment, fast Fourier transform, and data segmentation.

(1) Data augment: In order to adapt to different noise patterns, we augment the training dataset by adding Gaussian

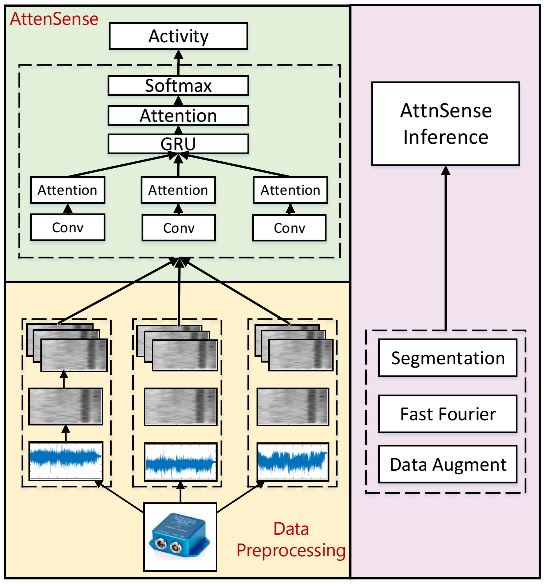


Figure 1: Overview of the proposed method.

random noise to each normalized instance in the training set, and each normalized instance will generate some noisy instances, we combine the original training dataset with the generated noisy instances to form a new training dataset.

(2) Fast Fourier transform: A spectrogram of a signal is a function of time and frequency that represents the changes in the energy content of a signal [Alsheikh *et al.*, 2016], and it can capture the intensity differences among the nearest data points. In order to calculate the spectrogram, we first split the signal $\boldsymbol{\mathcal{S}}$ into small size intervals with width m (m = 0.25s in our experiment), and apply Fast Fourier transform (FFT) to each interval to generate frequency features. After applying FFT, the raw signal $\boldsymbol{\mathcal{S}}$ can be transformed into the spectrogram represented as $\boldsymbol{\mathcal{F}}$.

(3) Data segmentation: A sliding window with width $\boldsymbol{N}$ and 25% overlap is used to split each spectrogram $\boldsymbol{\mathcal{F}}$ into same size sequences. We represent each sequence as $\boldsymbol{\mathcal{X}}$ = $[X_1, \cdots, X_t, \cdots, X_N]$, where X_t = $[x_{t1}, \cdots, x_{tK}]$ is the frequency representation of K sensors at timestep t. After data segmentation, each sequence $\boldsymbol{\mathcal{X}}$ is used as the input of the AttnSense model to detect the possible activities.

4.2 The AttnSense Model

In this section, we introduce AttnSense, an attention-based neural network model for activity sensing. As shown in Fig. 2, AttnSense consists of an individual convolutional subnet, an attention-fusion subnet, an attention-based GRU subnet, and an output layer, which are explained below.

Individual Convolutional Subnet

Convolutional neural networks had shown great potential in identifying the various salient patterns of sensing signals for HAR [Yang *et al.*, 2015]. Here we use an individual convolutional subnet to extract features of each sensor. The individual convolutional subnet consists of several stacked convolutional layers and pooling layers. The suitable number of convolutional layers will be discussed in section 5. In addition, a batch normalization layer is applied at each layer to reduce internal covariate shift.

As shown in Fig. 2, the frequency representation of the k-th sensor at time t, x_{tk}, is fed into the convolutional subnet, which outputs a feature vector v_{tk} for each sensor. Then all K generated feature vectors are used as the input to the attention-fusion subnet.

Attention-fusion Subnet

In the multimodal activity recognition problem, not all modalities are equally contributed to the classification task. In order to prioritize the important modalities, we introduce a self-attention network, which takes the sensors' feature vectors $[v_{t1}, \cdots, v_{tk}, \cdots, v_{tK}]$ as input and outputs an attention weight for each modality. And the attention weights represent the importance of different sensors in the HAR task. Then those feature vectors of all sensors are fused by using their attention scores as weights to form a uniform feature representation vector c_t. The self-attention structure can be formalized as follows:

$$\mu_{tk} = tanh(W_1 v_{tk} + b_1) \tag{3}$$

$$\alpha_{tk} = \frac{exp((\mu_{tk})^T w_1)}{\sum_k exp((\mu_{tk})^T w_1)} \tag{4}$$

$$c_t = \sum_k \alpha_{tk} v_{tk} \tag{5}$$

Here, we compute the hidden representation of v_{tk} through a one-layer MLP to get μ_{tk}, then we measure the weight of the k-th sensor as the similarity of μ_{tk} with a sensor-level context vector w_1 and get a normalized weight α_{tk} through a softmax function. c_t is the uniform representation of all K sensors which is computed by the sum of all sensors' feature vectors weighted by their attention weights. $\{W_1, b_1, w_1\}$ are parameters of the attention subnet which are randomly initialized and jointly learned during the training process.

Attention-based GRU subnet

After attention-fusion layer, the output $[c_1, ..., c_N]$ is fed to a stacked GRU structure (two layers). GRU [chu, 2014] is a type of RNN. Similar to LSTM [Greff *et al.*, 2015], GRU can model long-term dependencies in a sequence and solves the vanishing gradient problem of conventional RNN. Since GRU has lower computational complexity than LSTM, we choose it to construct the recurrent layer. The stacked GRU structure transforms the input into the hidden layer output by various gate units worked as follows.

$$z_t = \sigma\big(c_t U^{(z)} + h_{t-1} W^{(z)}\big) \tag{6}$$

$$r_t = \sigma\big(c_t U^{(r)} + h_{t-1} W^{(r)}\big) \tag{7}$$

$$\tilde{h}_t = \tanh\big(c_t U^{(h)} + (r_t * h_{t-1}) W^{(h)}\big) \tag{8}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \tag{9}$$

Here r_t is a reset gate, and z_t is an update gate. Intuitively, the reset gate determines how to combine the new input with the previous memory, and the update gate defines how much of the previous memory to keep around.

Standard GRU generates hidden state at each timestep [chu, 2014], but only use hidden state at the last timestep as a single representation for the whole input sequence, which leads

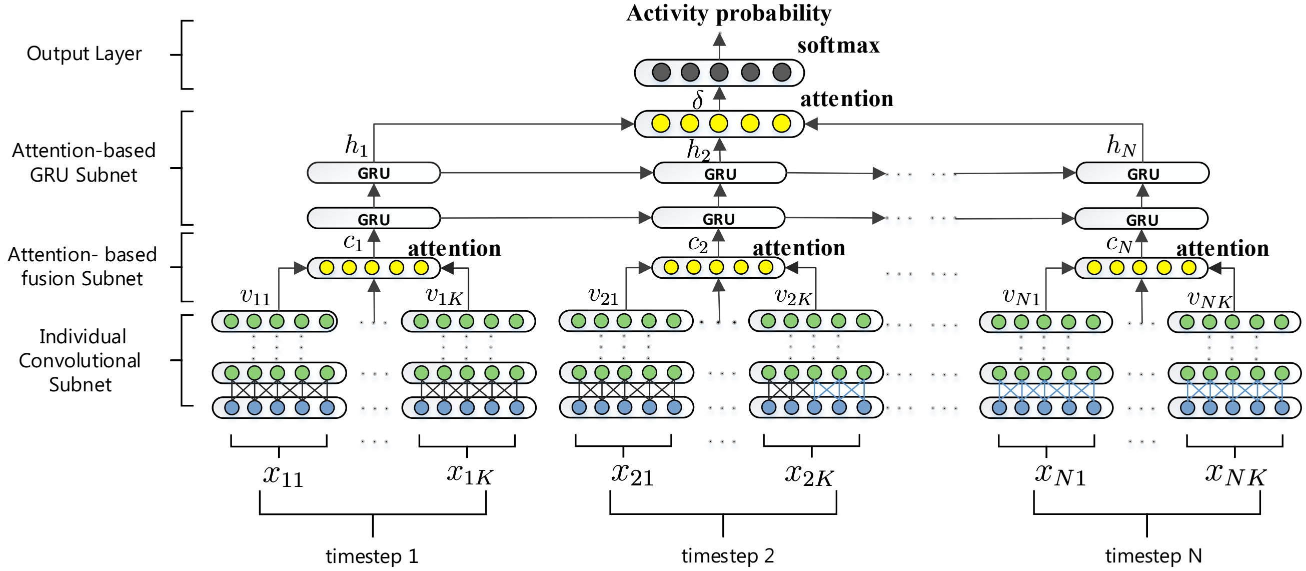


Figure 2: The AttnSense model.

to less impact on the classification for the front part of the sequence. A better method is to calculate the average of all hidden states, but for HAR problem, input sequence usually contain irrelevant information and not all timesteps contribute equally to the activity recognition task, so we again use the self-attention mechanism to calculate the weighted average sum of all hidden states.

We represent all GRU's hidden states as **H** = $[h_1, h_2, \cdots, h_t, \cdots, h_N]$, where h_t represents the GRU's hidden state at timestep t. The self-attention for GRU can be formalized as:

$$\gamma_t = tanh(W_2 h_t + b_2) \tag{10}$$

$$\beta_t = \frac{exp((\gamma_t)^T w_2)}{\sum_t exp((\gamma_t)^T w_2)} \tag{11}$$

$$\delta = \sum_t \beta_t h_t \tag{12}$$

Similarly, w_2 is a time-level context vector, β_t is a normalized weight through a softmax function and δ is the uniform representation of the whole sequence which is computed by the sum of all hidden state weighted by their attention weights. $\{W_2, b_2, w_2\}$ are parameters of the attention-based GRU subnet which are randomly initialized and jointly learned during the training process.

Output Layer

The output of attention-based GRU subnet is further connected to an output layer.

$$label = \operatorname*{argmax}_{a \in \mathcal{A}}(softmax(W_3 \cdot \delta + b_3)) \tag{13}$$

Here $\mathcal{A}$ is the set of all activities. We use a fully-connected layer and a softmax function to transform δ to the probability of each activity, and derive the predicted label by searching the activity with maximum probability.

5 Performance Evaluation

In order to evaluate the effectiveness of the proposed model, we conduct extensive experiments based on three HAR datasets. In what follows, we will first describe the experimental setup and the numerical results.

5.1 Experimental Setup

We build our model using TensorFlow and train it on a GPU GTX 1070ti. The batch size is set to 64, and the network is optimized using Rmsprop with learning rate 0.0001. The parameters in optimizers are initialized by the default setting.

We use F-measure (F1) as the performance metric in the evaluation. Since the traditional F1 score is used to measure the performance of binary classification, we extended it to a weighted F1 score, F_w, by weighting classes according to their sample proportion. Furthermore, we ran 20 repetitions of the experiments and report averaged F_w as the final measure of a model's performance.

$$F_w = \frac{1}{C} \cdot \sum_{i=1}^{C} \frac{2 \cdot \text{Precision}_\text{i} \cdot \text{Recall}_\text{i}}{\text{Precision}_\text{i} + \text{Recall}_\text{i}} \tag{14}$$

where for a given class i, $\text{Precision}_\text{i} = \frac{\text{TP}_\text{i}}{\text{TP}_\text{i}+\text{FP}_\text{i}}$; $\text{Recall}_\text{i} = \frac{\text{TP}_\text{i}}{\text{TP}_\text{i}+\text{FN}_\text{i}}$; $\text{TP}_\text{i}, \text{FP}_\text{i}$ represents the number of true and false positive respectively; and FN_i is the number of false negatives.

5.2 Dataset Description

We evaluate AttnSense on three public HAR datasets. These datasets are recorded in different contexts by either worn or embedded into objects that subjects manipulated. The statistics of the three datasets are depicted in Table 1.

The first dataset is Heterogeneous [Stisen *et al.*,]. It contained sensing data of accelerometer and gyroscope collected from 9 users performing 6 activities. An important fact of the dataset is that users perform these activities with 12 different smartphones and smartwatches, which increases the complexity of the task and can test the model's robustness. We preprocess the dataset as described and use the whole data from participant 1 for testing, and the rest of the dataset for training.

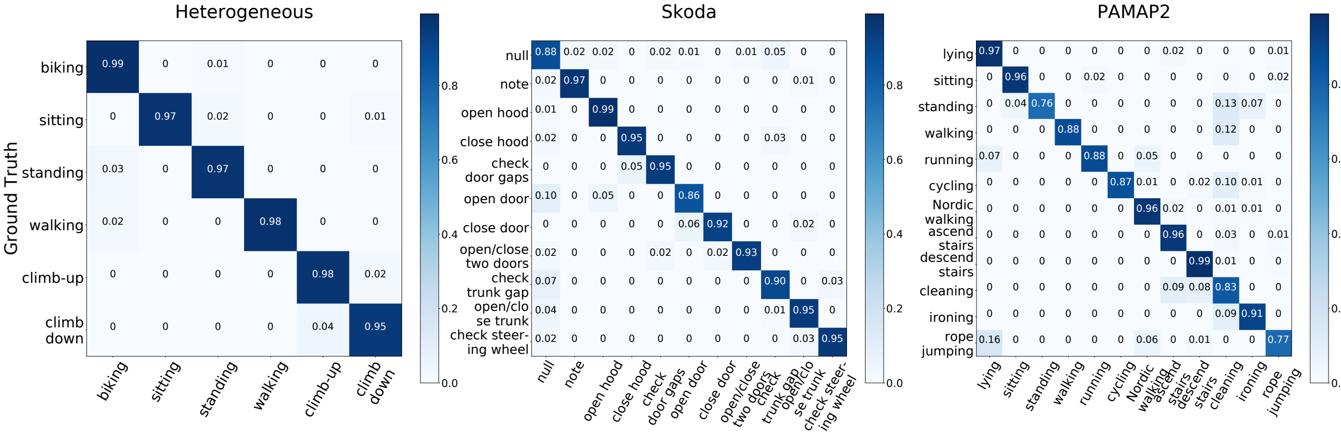


Figure 3: Confusion matrix of three datasets.

The second dataset is Skoda [Stiefmeier *et al.*, 2008]. It described the activities of assembly-line workers in a car production environment, where each worker wore a number of accelerometers on both arms while performing 10 activities, including checking the boot, opening/closing engine bonnet, etc. In addition, it contained an extra null activity, which means the subject did nothing in that time. We preprocess the dataset as described and use 10% of the data in each class for testing, and the rest 90% data for training.

The third dataset is PAMAP2 [Reiss and Stricker, 2012]. It recorded signals from three inertial measurement units (IMUs) located on the hand, chest, and ankle. Each IMU consisted of an accelerometer, gyroscope, magnetometer, temperature and heart rate sensor. We select accelerometer, gyroscope, and magnetometer as input to our model. The dataset was collected from 9 participants, which includes 12 activities ("walking", "lying down", "standing", etc) and 6 optional activities ("watching TV", "folding laundry", etc). Since those optional activities are rarely performed by the participants, we excluded them from the analysis. We preprocess the dataset as described and use the whole data from participant 6 for testing, and the rest of the dataset for training.

Moreover, we perform 4-fold cross-validation for Skoda dataset and leave-one-subject-out validation for Heterogeneous and PAMAP2 dataset to obtain the best model configuration.

5.3 Compared Algorithms

We compare AttnSense with the following algorithms:

- **Random forest (RF) [Liaw and Wiener, 2002]**: The random forests are a classical ensemble model which construt a multitude of decision trees at training time.
- **SVM [Hearst, 1998]**: A simple support vector machine (SVM) with radial basis function (RBF) kernel.
- **CNN [LeCun *et al.*, 2015]**: A single CNN model with three convolutional layers, a pooling layer, and a fully connected layer.
- **LSTM [Hochreiter and Schmidhuber, 1997]**: A single layer LSTM model.
- **DeepConvLSTM [Ordóñez and Roggen, 2016]**: This model uses a deep convolutional neural network to extract feature and a recurrent neural network to learn time dependencies.
- **DeepSense [Yao *et al.*, 2017]**: This model is the state-of-the-art model in Heterogeneous dataset, which used a CNN network to extract feature of each sensor and combined them by another merge convolutional layer, then it used a LSTM network to learn time dependencies.

Name	Subject	S. Rate	Activity	Sample	Sensor
Heterogeneous	9	100 Hz	6	43,930,257	A, G
Skoda	1	98 Hz	11	22,000	A
PAMAP2	9	100 Hz	12	2,844,868	A, G, M

Table 1: Description of datasets (A=accelerometer, G=gyroscope, M=magnetometer).

In addition, for shallow models, we extract all time-domain features mentioned the literature [Figo *et al.*, 2010], including mean, std, median, maximum, and etc.

To verify the contributions of different components in our model, we consider two variants of our model as follows:

- **AS-noAF**: This model removes the attention fusion layer and uses naive concatenation to fuse feature vectors of different sensors.
- **AS-noAG**: This model removes the attention mechanism for GRU and uses the average of GRU's hidden state.

5.4 Numerical Results

We compare AttnSense and its variants with the baseline algorithms on those three datasets. The results are shown in Table 2, and the normalized confusion matrixs are illustrated in Fig. 3. In addition, the F1 scores of DeepSense and DeepConvLSTM are from those literatures [Yao *et al.*, 2017], [Ordóñez and Roggen, 2016], and [Guan and Plötz, 2017] accordingly.

As shown in the table, deep models outperform shallow such as SVM and RF. The hybrid neural network models such as DeepConvLSTM and DeepSense performs better than simple CNN and LSTM, but DeepConvLSTM shows poor performance on the PAMAP2 dataset. AttnSense performs

Model	Heterogeneous	Skoda	PAMAP2
AttnSense	**0.965 ± 0.010**	**0.931 ± 0.022**	**0.893 ± 0.013**
AS-noAF	0.949 ±0.023	0.919 ± 0.036	0.854 ± 0.011
AS-noAG	0.945 ± 0.018	0.921 ± 0.015	0.867 ± 0.007
DeepSense	0.931	–	–
DeepConvLSTM	–	0.912	0.748
LSTM	0.812 ± 0.016	0.893 ± 0.041	0.751 ± 0.036
CNN	0.808 ± 0.032	0.845 ± 0.028	0.817 ± 0.041
RF	0.743 ± 0.009	0.827 ± 0.033	0.742 ± 0.022
SVM	0.756 ± 0.017	0.816 ± 0.019	0.706 ± 0.013

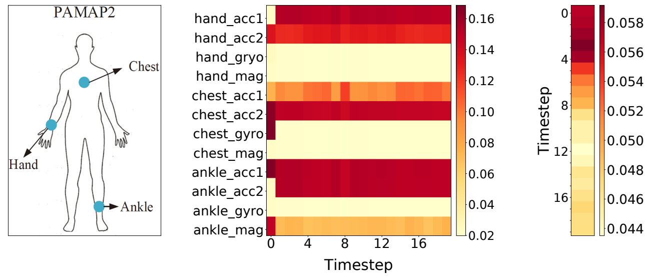
Table 2: F1 scores of different algorithms.

the best among all algorithms, which achieves performance improvement compared to the state-of-the-art. It verifies that AttnSense has greater capability to capture temporal-spacial patterns in multimodal sensing data for HAR. It is worth noting that AttnSense outperforms its variants AS-noAF and AS-noAG, which indicate that the attention mechanism plays an important role in our proposed model.

Visualizing Attention Weights

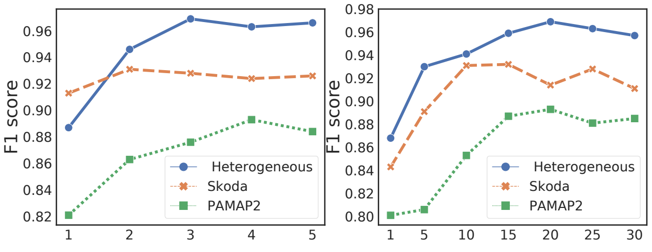
We provide visualized analysis to the attention weights, which can be used to evaluate the impact of different sensor modalities placed on different parts of the human body. Figure 4(a) shows the positions of IMUs for PAMAP2 dataset; Figure 4(b) shows the attention weights of different sensor modalities for the running activity; and Figure 4(c) shows the temporal attention weights of GRU subnet for the running activity.

According to Figure 4(b), the attention fusion layer puts a high emphasis on the acc1, acc2 of the hand, acc2 of the chest, and acc1, acc2 of the ankle, which is intuitively interpretable for running activity. Compared with simple the naive concatenation mechanism, which treats all the sensor modality equally, the proposed attention mechanism automatically learns the priority of different sensors for HAR task, which works better in feature selection and fusion, and achieves better performance in dealing with multimodal high-dimensional time series. According to Figure 4(c), the temporal attention model highlights the hidden state close to the beginning of the signal, which means the beginning of the signal sequence show more salient pattern of the running activity and our proposed model can capture the important part of the sequence to increase performance and comprehensibility.



(a) Position of sensors (b) Attention weight of different sensors (c) Temporal attention

Figure 4: Visualization of attention weights of running activity in PAMAP2 dataset.



(a) Number of convolutional layer (b) Windows_size

Figure 5: The performance of activity recognitions under different numbers of convolutional layers and different sliding windows size.

Number	Parameters of CNN
1	conv3-64 → pool → FC
2	conv3-32 → conv3-32 → pool → FC
3	conv3-32 → conv3-32 → pool → conv3-64 → pool → FC
4	conv3-32 → conv3-32 → pool → conv3-64 → conv3-64 → pool → FC
5	conv3-32 → conv3-32 → pool → conv3-64 → conv3-64 → conv3-64 → pool → FC

Table 3: Structure of convolutional layers (conv3 represents convolutional layer with 1 × 3 kernel, pool represents max pooling layer and FC represents fully connected layer).

Parameter Analysis

Here we study the influence of some parameters including the structure of convolutional subnet and the length of sliding window (each timestep correspondings to 0.25s). The five kinds of CNN structures are illustrated in Table 3 and the results are given in Figure 5. It can be found that increasing the number of convolutional layers tends to improve the performance of the model, but it approaches to be stable after a certain number. The best numbers of convolutional layers for Heterogeneous, Skoda and PAMAP2 dataset are 3, 2, and 4 accordingly. In addition, the length of sliding window also influence the model's performance, and a small sliding window usually results in poor recognition accuracy. We get the best performance when using 20, 15, and 20 width sliding window for Heterogeneous, Skoda and PAMAP2 accordingly.

6 Conclusion

Recognizing human activities from multimodal sensing data is a challenging task. In this paper, we proposed an attention-based deep neural network model called AttnSense and a comprehensive data preprocessing technique to solve the problem. AttnSense adopted a hybrid framework to combine the attention mechanism with CNN-RNN architecture to fuse multimodal sensor information and RNN hidden state, which has greater capability to capture temporal-spacial patterns in multimodal sensing data for HAR. The data preprocessing technique also help our model to capture the noise pattern and frequency features in the multidimensional sensing signals. As demonstrated in the experiments, the proposed method outperformed the state-of-the-art methods.

References

[Alsheikh *et al.*, 2016] Mohammad Abu Alsheikh, Ahmed Selim, Dusit Niyato, Linda Doyle, Shaowei Lin, and Hwee-Pink Tan. Deep activity recognition models with triaxial accelerometers. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[Bengio, 2013] Yoshua Bengio. Deep learning of representations: Looking forward. In *International Conference on Statistical Language and Speech Processing(SLSP'13)*, pages 1–37. Springer, 2013.

[Bulling *et al.*, 2011] Andreas Bulling, Jamie A Ward, Hans Gellersen, and Gerhard Troster. Eye movement analysis for activity recognition using electrooculography. *IEEE transactions on pattern analysis and machine intelligence*, 33(4):741–753, 2011.

[chu, 2014] Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014.

[Figo *et al.*, 2010] Davide Figo, Pedro C. Diniz, Diogo R. Ferreira, and João M. Cardoso. Preprocessing techniques for context recognition from accelerometer data. *Personal Ubiquitous Comput.*, 14(7):645–662, October 2010.

[Greff *et al.*, 2015] Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber. LSTM: A search space odyssey. *CoRR*, abs/1503.04069, 2015.

[Guan and Plötz, 2017] Yu Guan and Thomas Plötz. Ensembles of deep lstm learners for activity recognition using wearables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(2):11:1–11:28, June 2017.

[Hammerla *et al.*, 2016] Nils Y. Hammerla, Shane Halloran, and Thomas Plötz. Deep, convolutional, and recurrent models for human activity recognition using wearables. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 1533–1540. AAAI Press, 2016.

[Hearst, 1998] Marti A. Hearst. Support vector machines. *IEEE Intelligent Systems*, 13(4):18–28, July 1998.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[Khan *et al.*, 2013] Wazir Zada Khan, Yang Xiang, Mohammed Y Aalsalem, and Quratulain Arshad. Mobile phone sensing systems: A survey. *IEEE Communications Surveys & Tutorials*, 15(1):402–427, 2013.

[Kim and Lane, 2017] Suyoun Kim and Ian Lane. End-to-end speech recognition with auditory attention for multi-microphone distance speech recognition. In *INTERSPEECH*, 2017.

[LeCun *et al.*, 2015] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 5 2015.

[Liaw and Wiener, 2002] Andy Liaw and Matthew Wiener. Classification and Regression by randomForest. *R News*, 2(3):18–22, 2002.

[Ordóñez and Roggen, 2016] Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1), 2016.

[Plötz *et al.*, 2011] Thomas Plötz, Nils Y. Hammerla, and Patrick Olivier. Feature learning for activity recognition in ubiquitous computing. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, IJCAI'11, pages 1729–1734. AAAI Press, 2011.

[Radu *et al.*, 2017] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D. Lane, Cecilia Mascolo, Mahesh K. Marina, and Fahim Kawsar. Multimodal deep learning for activity and context recognition. *IMWUT*, 1:157:1–157:27, 2017.

[Reiss and Stricker, 2012] Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *Proceedings of the 2012 16th Annual International Symposium on Wearable Computers*, ISWC '12, pages 108–109, Washington, DC, USA, 2012. IEEE Computer Society.

[Stiefmeier *et al.*, 2008] Thomas Stiefmeier, Daniel Roggen, Georg Ogris, Paul Lukowicz, and Gerhard Tröster. Wearable activity tracking in car manufacturing. *IEEE Pervasive Computing*, 7, 2008.

[Stisen *et al.*,] Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Mikkel Baun Prentow, Anind Dey, Tobias Sonne, and Mads Møller Jensen. Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems (SenSys'15)*, pages 127–140. ACM.

[van Kasteren *et al.*, 2008] Tim van Kasteren, Athanasios Noulas, Gwenn Englebienne, and Ben Kröse. Accurate activity recognition in a home setting. In *Proceedings of the 10th International Conference on Ubiquitous Computing*, UbiComp '08, pages 1–9, New York, NY, USA, 2008. ACM.

[Wang *et al.*, 2018] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 2018.

[Yang *et al.*, 2015] Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiaoli Li, and Shonali Krishnaswamy. Deep convolutional neural networks on multichannel time series for human activity recognition. In *Proceedings of the 24th International Conference on Artificial Intelligence (Ijcai'15)*, volume 15 of *IJCAI'15*, pages 3995–4001. AAAI Press, 2015.

[Yao *et al.*, 2017] Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek Abdelzaher. Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In *Proceedings of the 26th International Conference on World Wide Web (WWW'17)*, pages 351–360. International World Wide Web Conferences Steering Committee, 2017.