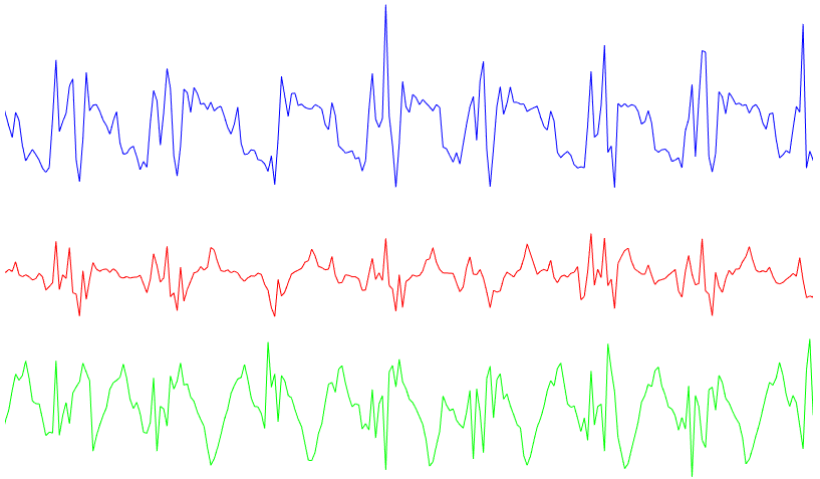
Challenges in Sensor-based Human Activity Recognition

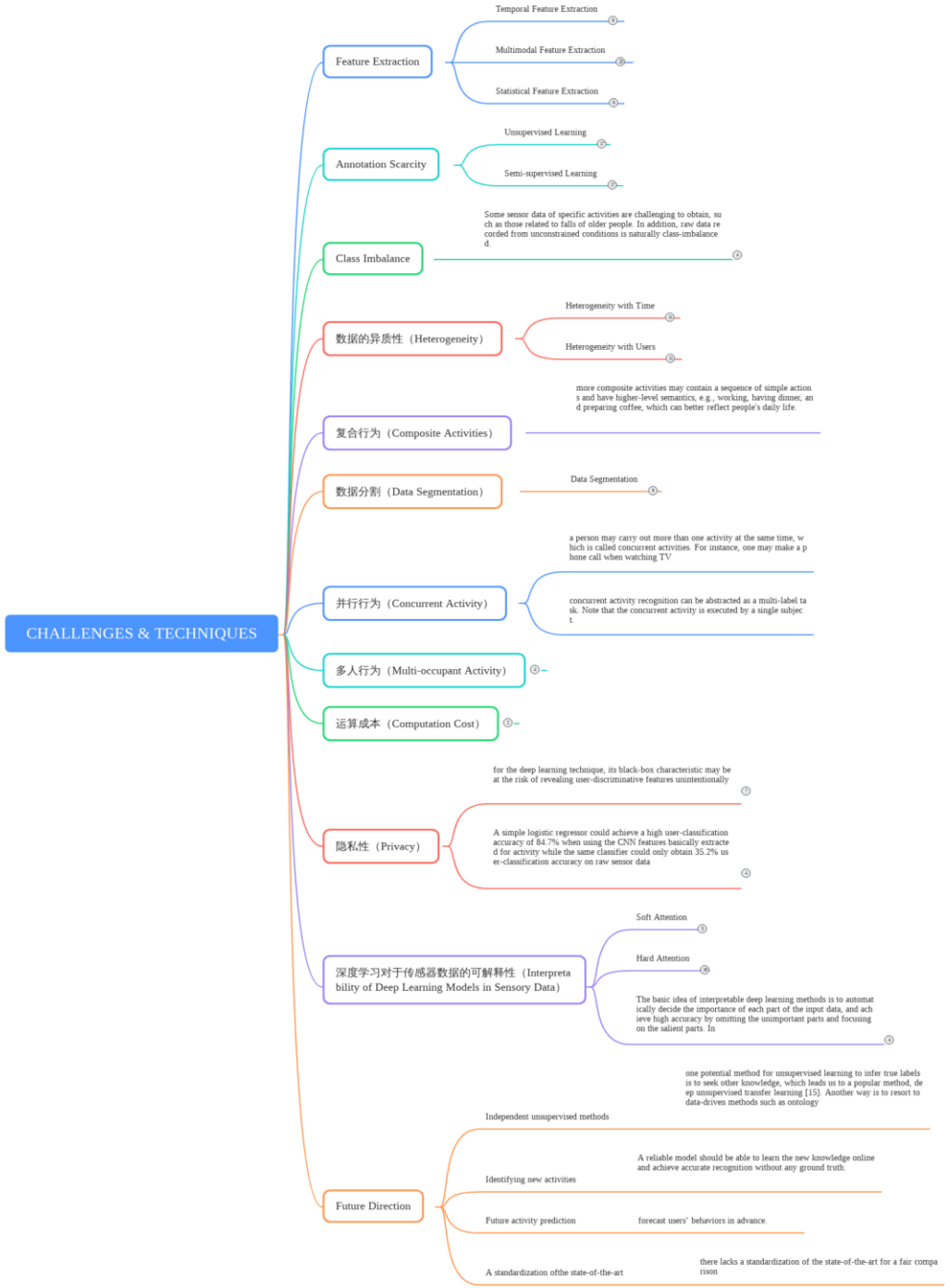


IMU (Inertial measurement unit)

3-axis accelerometer,

3-axis gyroscope,

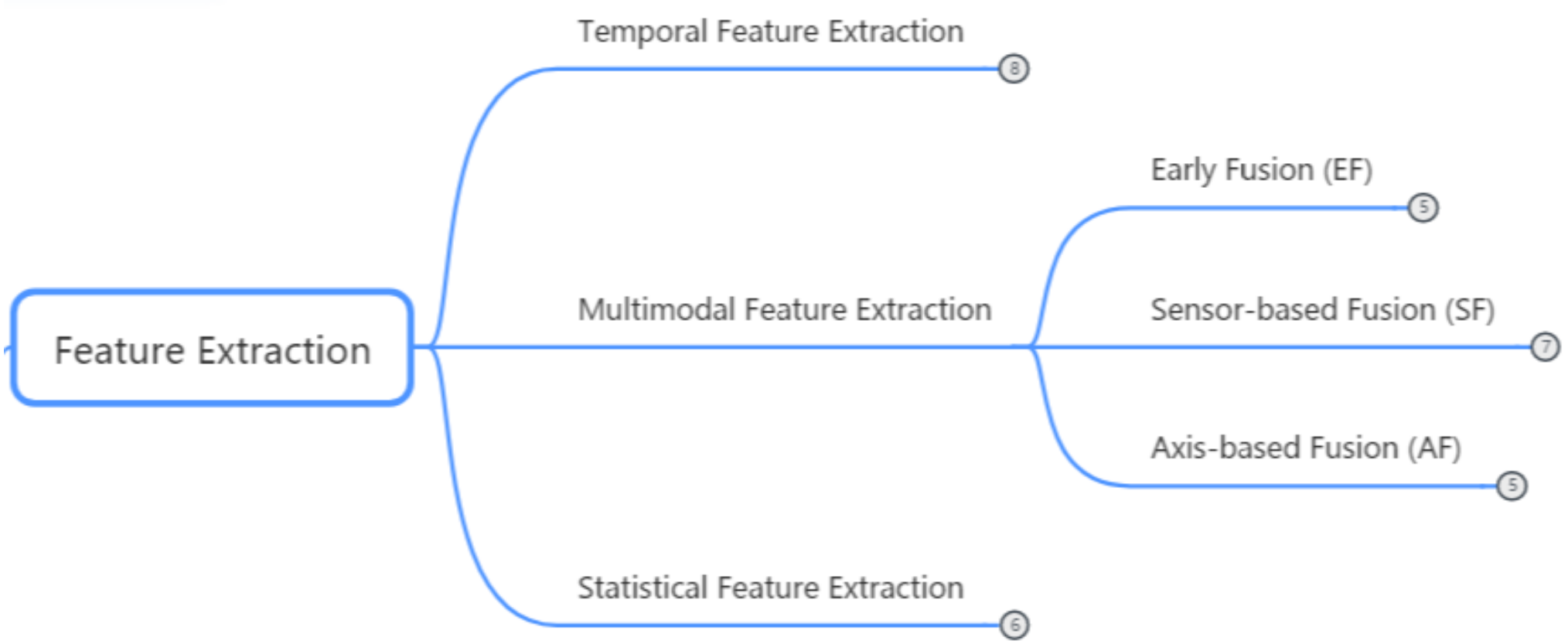
3-axis magnetometer



Challenges

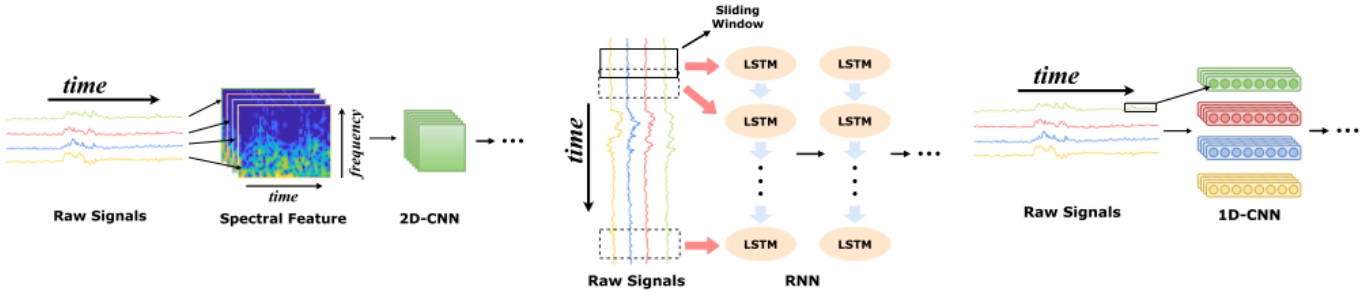
- Feature Extraction
- Annotation Scarcity (注释稀缺性)
- Class Imbalance
- Heterogeneity (数据异质性)
- Interpretability of Deep Learning Models in Sensory Data
- Data Segmentation
- Composite Activities（复合行为）
- Concurrent Activity (并行行为)
- Multi-occupant Activity (多人行为)
- Computation Cost （运算成本）
- Privacy

- **Feature Extraction**



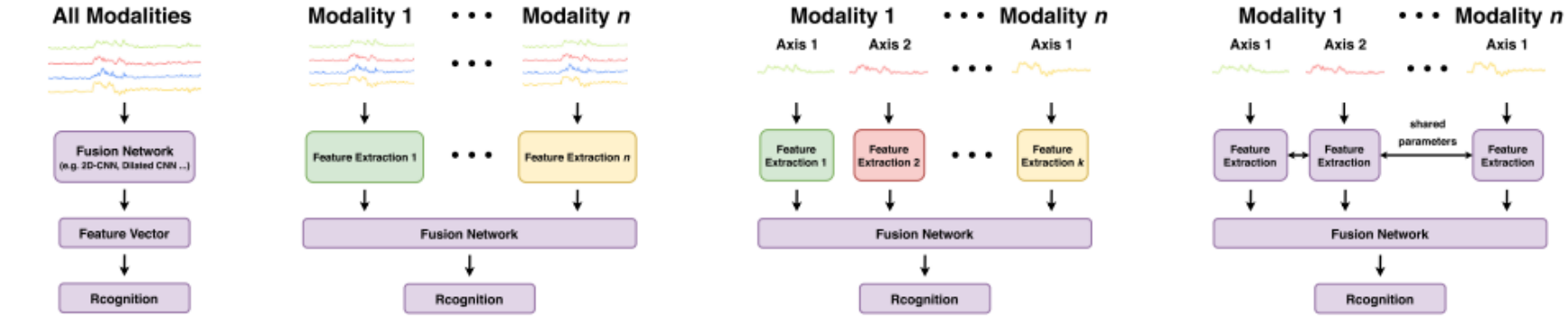
- **Feature Extraction**

  - Temporal Feature



(a) Time-Frequency-Spectral (b) RNN (c) CNN

  - Multimodal extraction



(a) Early Fusion (b) Sensor-based Fusion (c) Axis-based Fusion (d) Shared-filter Fusion

29. Rui Xi, Mengshu Hou, Mingsheng Fu, Hong Qu, and Daibo Liu. 2018. Deep dilated convolution on multimodality time series for human activity recognition. In 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 1–8

Feature Extraction Temporal Feature Extraction

- Expand the temporal features. Muti-kernel consume more computational resources. Applied a deep dilated CNN to time series for solving the issues. The dilated CNN uses dilated convolution kernels instead of the standard convolutional kernels to expand the convolution receptive field (i.e., time length) with no loss of resolution.
- Because the dilated kernel only adds empty elements between the elements of the conventional convolution kernel, it does not require an extra computational cost.

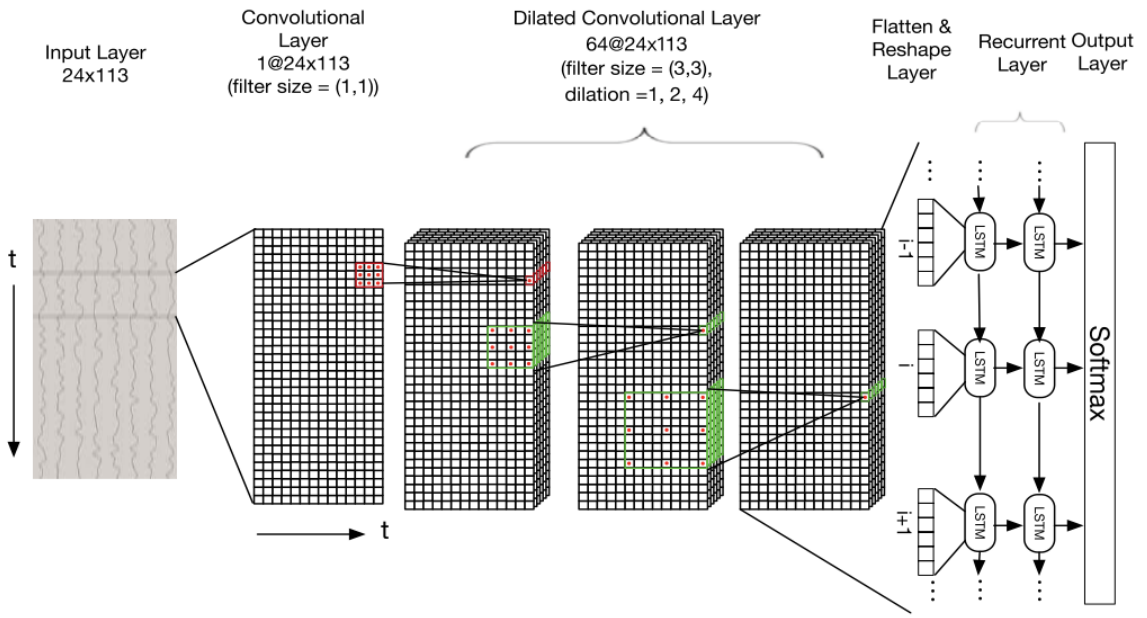


Fig. 3: Illustration of our proposed architecture for HAR on OPPORTUNITY dataset. The numbers before and after "@" refer to the number of feature maps and the dimension of a feature map in this layer. Note that we successively set dilation to 1, 2, 4 in three dilated convolutional layers. For simplicity, ReLU and normalization layer are not mentioned.

8. Sojeong Ha and Seungjin Choi. 2016. Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors. In 2016 International Joint Conference on Neural Networks (IJCNN). IEEE, 381–388.

Feature Extraction Temporal Feature Extraction Sensor-based Fusion (SF)

- Proposed to create a vector of different modalities at the early stage as well and to extract the common characteristics across modalities along with the sensor-specific characteristics; then both kinds of features are fused at the later part of the model.
- However for multi-modal data (accelerometer and gyroscope), [14] share weights for whole input signals in convolutional layer (full weights sharing), and extract same features without distinction of modalities, which might cause interferences between characteristics produced by accelerometers and gyroscopes for capturing modality-specific features.

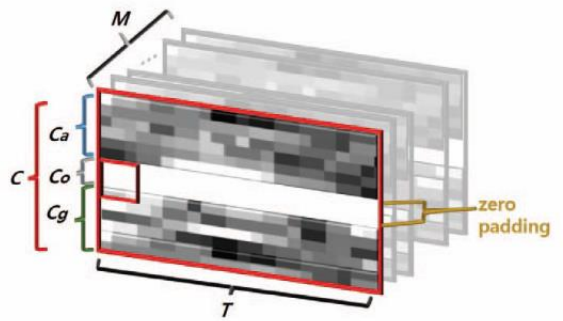
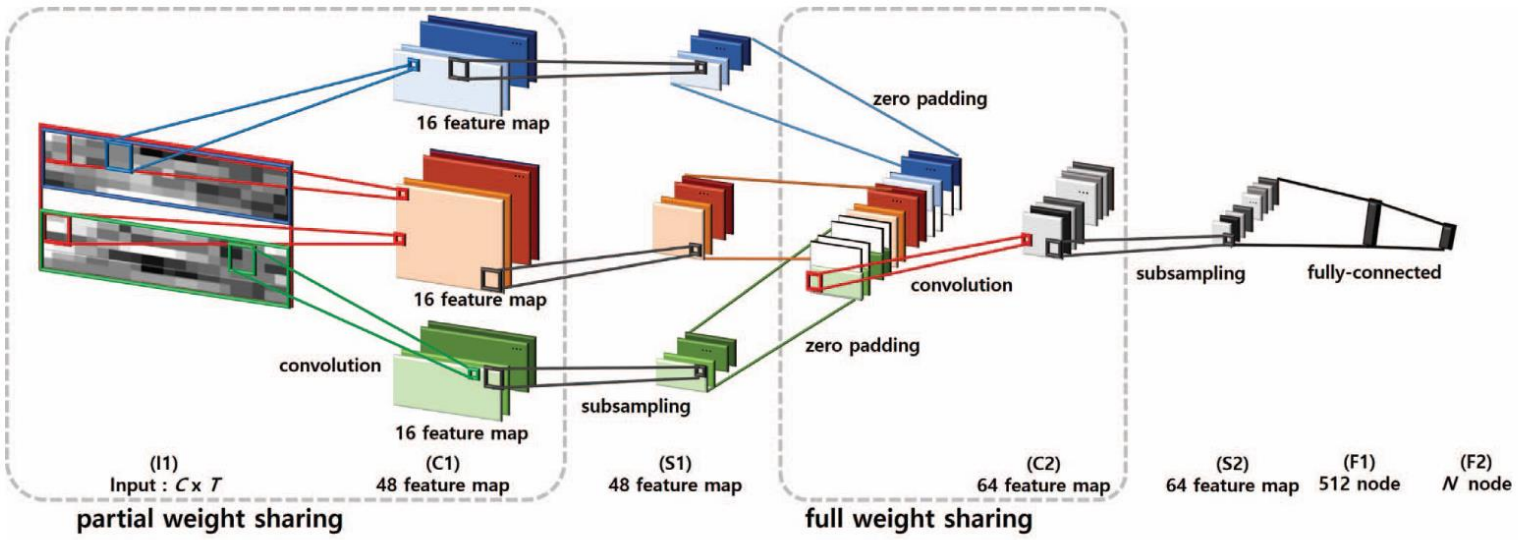


Fig. 5. Input batch of size $C \times T$ is formed by segmenting sensor signals via a sliding window.



14. Gierad Laput and Chris Harrison. 2019. Sensing Fine-Grained Hand Activity with Smartwatches. In Proceedings ofthe 2019 CHIConference on Human Factors in Computing Systems. ACM, 338.

Feature Extraction Temporal Feature Extraction Axis-based Fusion (AF).

- Developed a fine-grained hand activity sensing system through the combination of the time-frequency-spectral features and CNNs.
- They demonstrated 95.2% classification accuracy over 25 atomic hand activities of 12 people.
- The spectral features can not only be used for the wearable sensor activity recognition but also be used for the device-free activity recognition.

..\paper\14.pdf

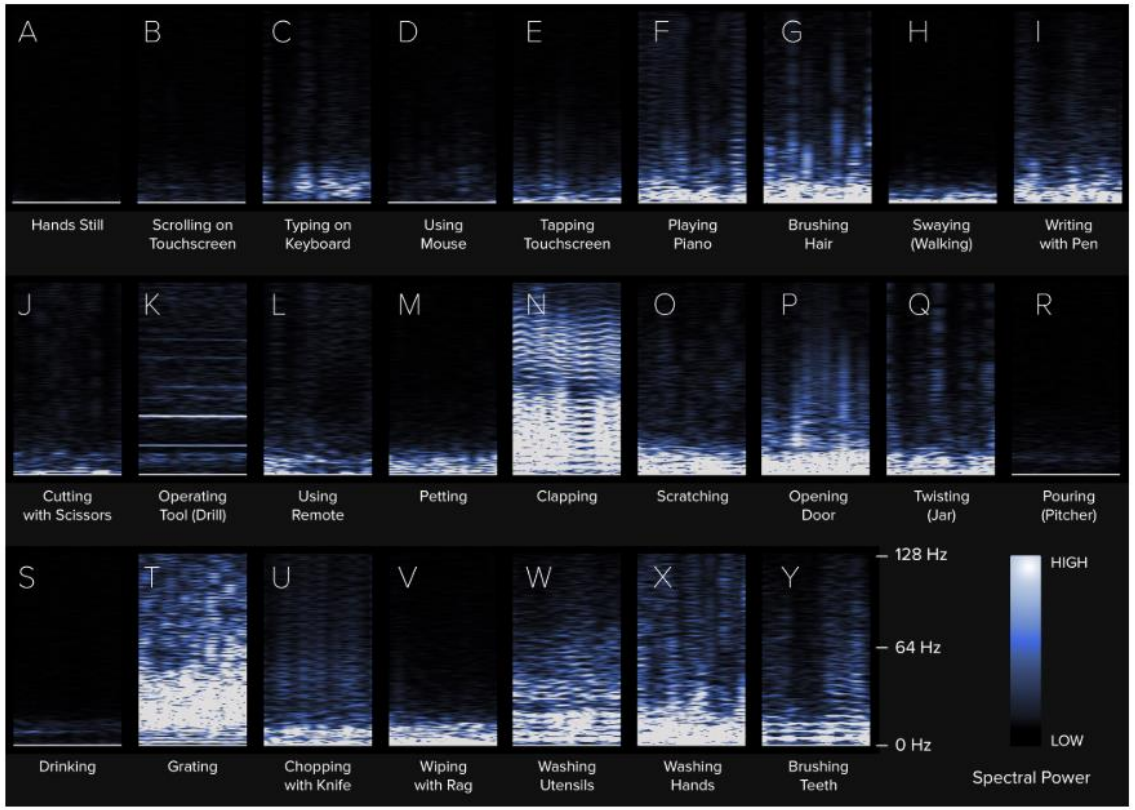


Figure 3. Example spectrograms of the 25 hand activities used in our obstacle course study (max of accelerometer axes shown). Y-axis is spectral power from 0 to 128 Hz. X-axis is time (3 seconds). Photos of these hand activities are shown in Figure 1, while Table 1 offers a rough estimation of how frequent these activities occur.

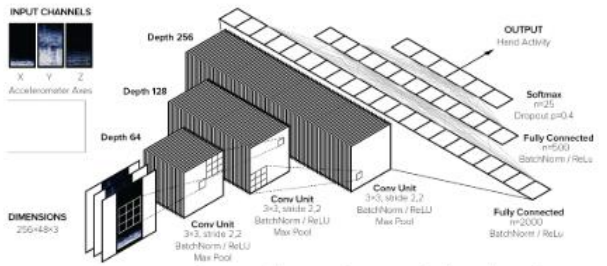


Figure 4. Our convolutional neural network (CNN) architecture, comprised of several convolutional units (three shown here), two fully connected layers, a dropout layer, and a softmax. We also apply batch normalization between non-linear layers (*i.e.*, activations).

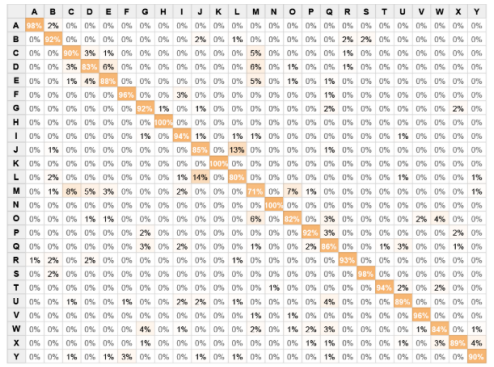


Figure 7. Across-user performance confusion matrix. Mean accuracy is 90.7% across our 25 activities and 12 users.

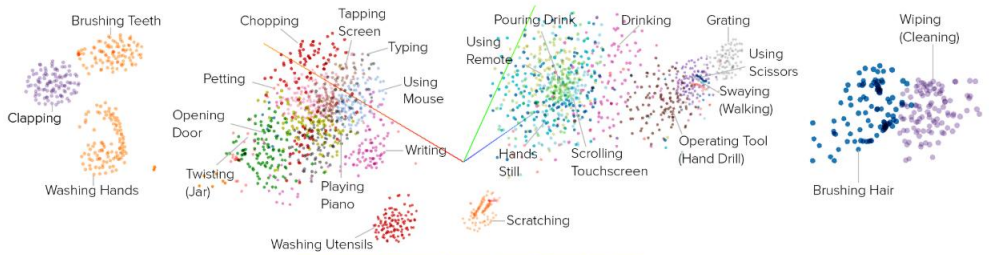


Figure 10. Clustering results from a t-SNE nonlinear dimensionality reduction (perplexity=40, 5000 iterations, projector.tensorflow.org) on a random subset of our 25 hand activities. Note how less intense activities (*e.g.*, pouring drink, scrolling touch screen) cluster together, while more vigorous hand activities (*e.g.*, clapping, scratching and wiping) emerge as distinct groups.

11. Chihiro Ito, Xin Cao, Masaki Shuzo, and Eisaku Maeda. 2018. Application of CNN for human activity recognition with FFT spectrogram of acceleration and gyro sensors. In Proceedings ofthe 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing andWearable Computers. ACM, 1503–1510

Feature Extraction Temporal Feature Extraction Axis-based Fusion (AF).

- Temporal features of acceleration and gyro signals are first represented by FFT spectrogram images and then vertically combined into a larger image for the following DCNN to learn inter-modality features.

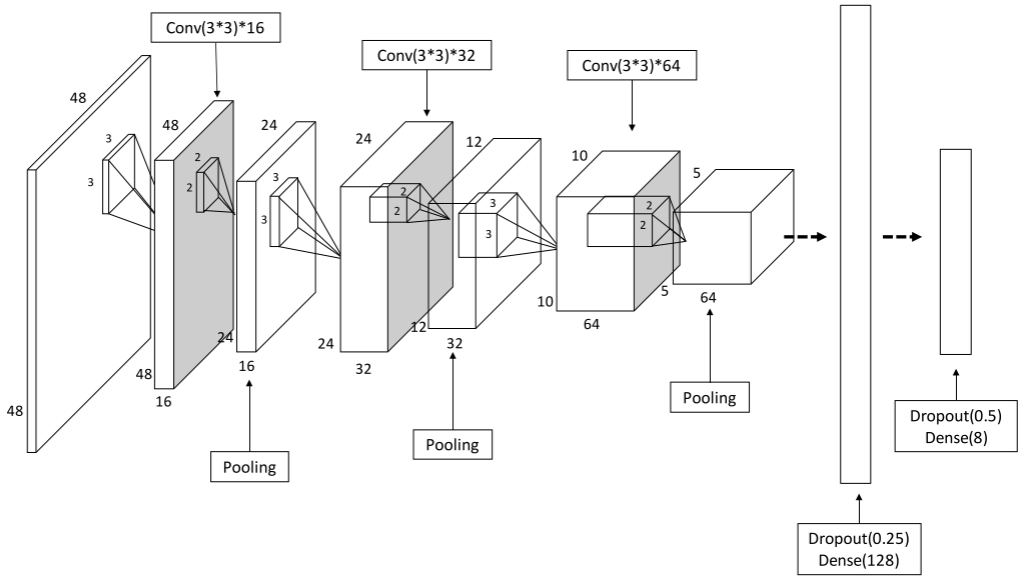


Figure 1: Proposed CNN model for human activity recognition using FFT spectrogram images.

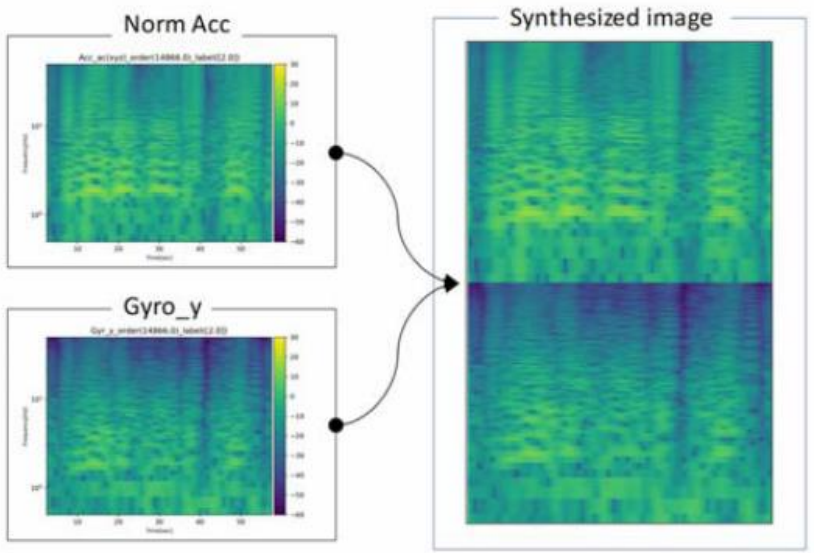
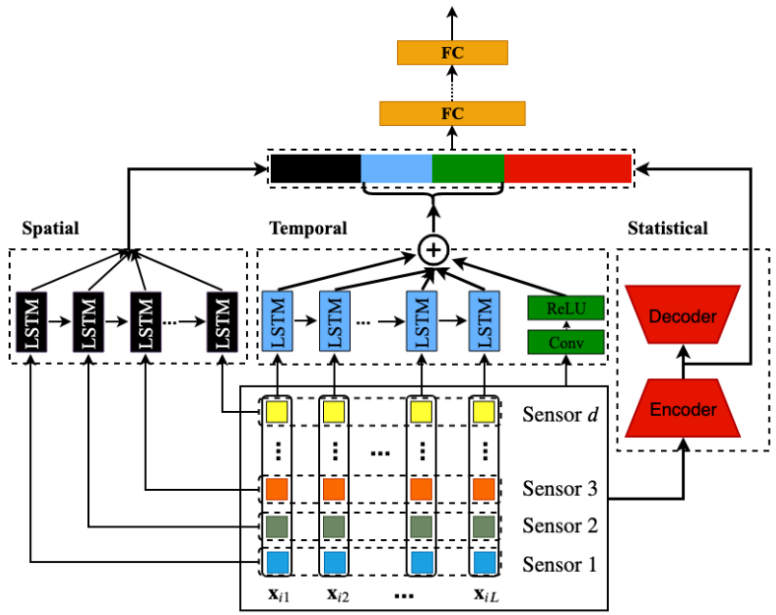


Figure 3: Example of synthesized image (right). FFT spectrogram images of norm Acc (left-above) and Gyr_y (left-below) were vertically arranged.

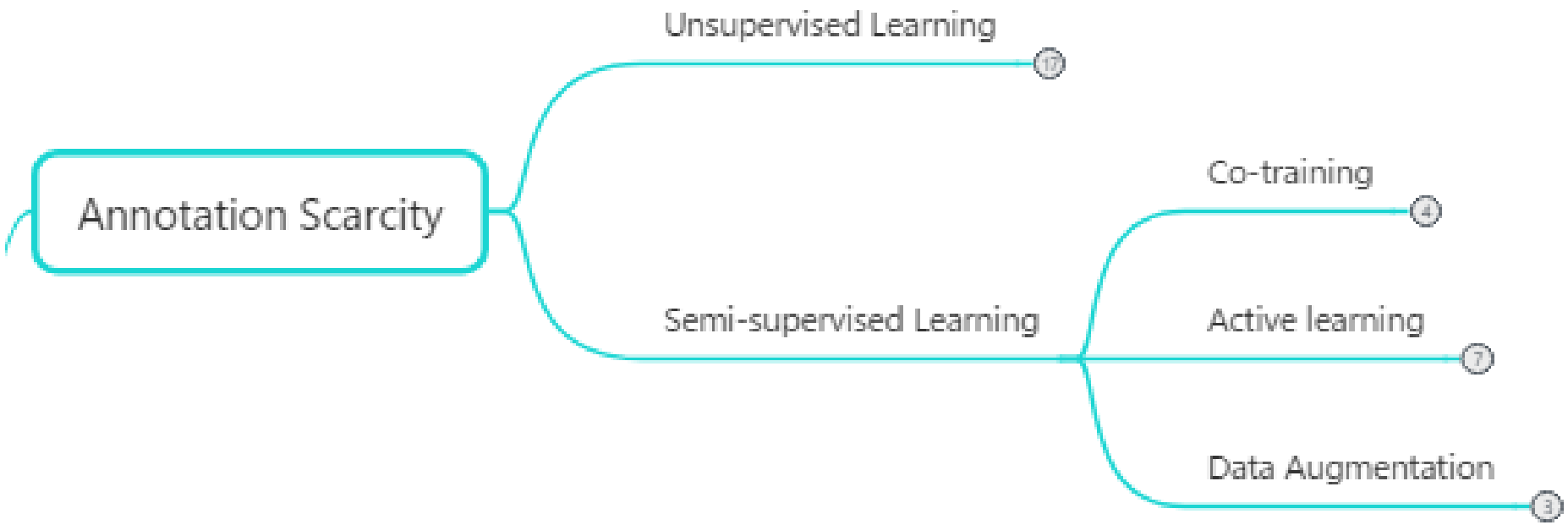
21. Hangwei Qian, Sinno Jialin Pan, Bingshui Da, and Chunyan Miao. 2019. A Novel Distribution-Embedded Neural Network for Sensor-Based Activity Recognition. In Proceedings ofthe Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019. 5614–5620
Feature Extraction Statistical Feature Extraction.

- Managed to develop a Distribution-Embedded Deep Neural Network (DDNN) to integrate the statistical feature extraction process into an end-to-end for activity recognition.
- It encoded the idea of kernel embedding of distributions into a deep architecture, such that all orders of statistical moments could be extracted as features to represent each segment of sensor readings, and further used for activity classification in an end-to-end training manner.



Methods	DG		OPPOR		UCIHAR		PAMAP2	
	miF	maF	miF	maF	miF	maF	miF	maF
DDNN	**92.59**	**91.61**	**83.66**	86.01	**90.53**	**90.58**	**93.23**	**93.38**
DDNN$-f_1$	91.38	90.67	81.27	84.51	89.96	89.93	87.49	86.84
DDNN$-f_2$	89.67	88.97	77.96	82.27	88.60	88.58	89.37	89.43
CNN_Yang	87.96	86.65	9.98	2.95	88.12	88.11	70.17	70.46
DeepConvLSTM	87.21	84.28	75.47	78.92	89.05	89.07	84.31	82.73
DNN	88.91	86.47	77.05	80.25	87.65	87.72	80.31	79.82
CNN	89.23	88.85	10.66	3.56	86.66	86.77	89.75	89.72
LSTM	88.34	86.93	63.17	69.92	74.52	74.75	90.38	90.29
LSTM-f∗	67.3	-	67.2	90.8	-	-	92.9	-
LSTM-S∗	76.0	-	69.8	91.2	-	-	88.2	-
b-LSTM-S∗	74.1	-	74.5	**92.7**	-	-	86.8	-

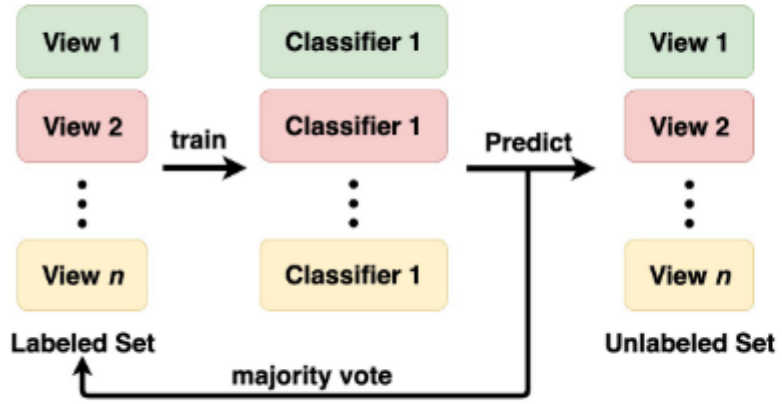
- **Annotation Scarcity**



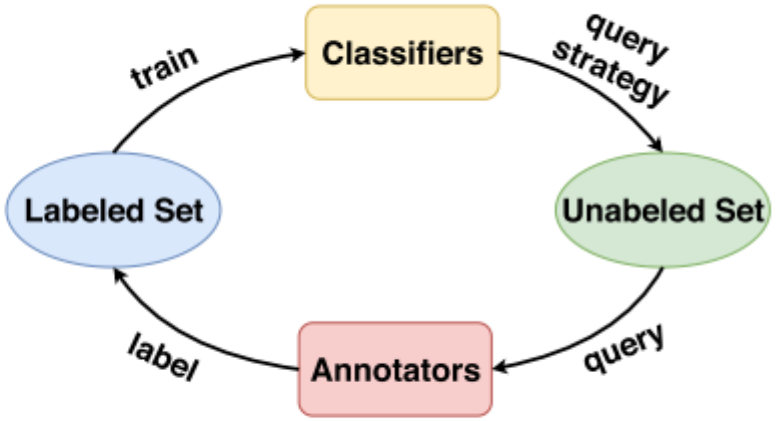
- **Annotation Scarcity**

  - Unsupervised Learning
    - Deep generative models such as Deep Belief Networks (DBNs) and autoencoders
    - They are useful in extracting features and finding patterns in massive data.

  - semi-supervised Learning
    - Co-training : extension of self-learning
    - Active Learning: requires annotators who are usually experts or users to label the data manually.



(a) Co-training



(b) Active Learning

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

Annotation Scarcity Unsupervised Learning

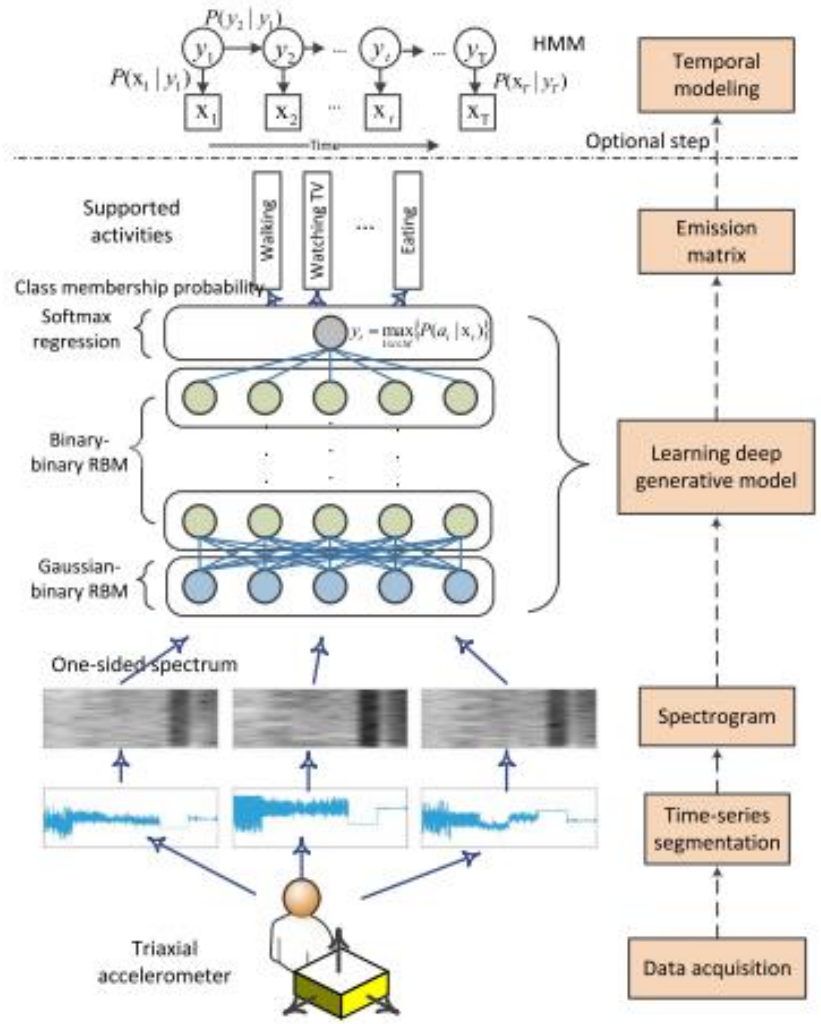
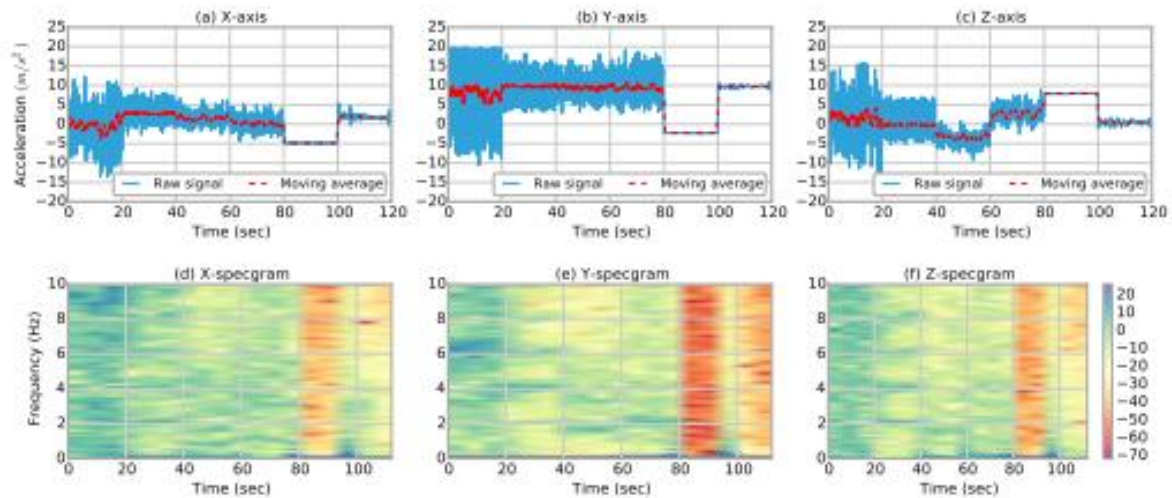


Figure 1: **Activity recognition using deep activity recognition model**. Our system automatically (1) takes triaxial acceleration time series, (2) extracts the spectrogram of windowed excerpts, (3) computes intrinsic features using a deep generative model, and then (4) recognizes the underlying human activities by finding the posterior probability distribution $\{P(a_i|\mathbf{x}_t)\}_{i=1}^{M}$. This deep architecture outperforms existing methods for human activity recognition using accelerometers as shown by the experimental analysis on real world datasets. Furthermore, an optional step involves using the emission probabilities out of the deep model to train a hidden Markov model (HMM) for modeling temporal patterns in activities.



2. Lu Bai, Chris Yeung, Christos Efstratiou, and Moyra Chikomo. 2019. Motion2Vector: unsupervised learning in human activity recognition using wrist-sensing data. In Proceedings ofthe 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019ACM International Symposium onWearable Computers. ACM, 537–542.

Annotation Scarcity Unsupervised Learning

- Proposed a method called Motion2Vector to convert a time period of activity data into a movement vector embedding within a multidimensional space.
- To fit with the context of activity recognition, they use a bidirectional LSTM to encode the input blocks of the temporal wrist-sensing data.
- Two hidden states generated are concatenated to form the embedded vectors which can be considered as an appropriate representation of the input movement.
- Classifiers such as C4.5, K nearest neighbor, and random forest are trained later for classification.
- The experiments showed that this method can achieve accuracy of higher than 87% when tested on public datasets

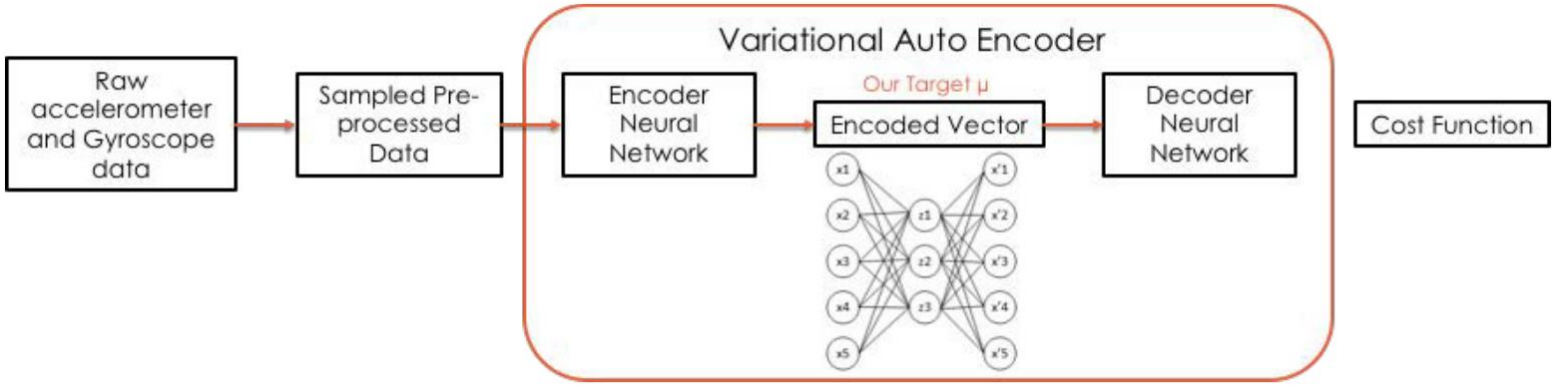


Figure 1: Deep learning architecture

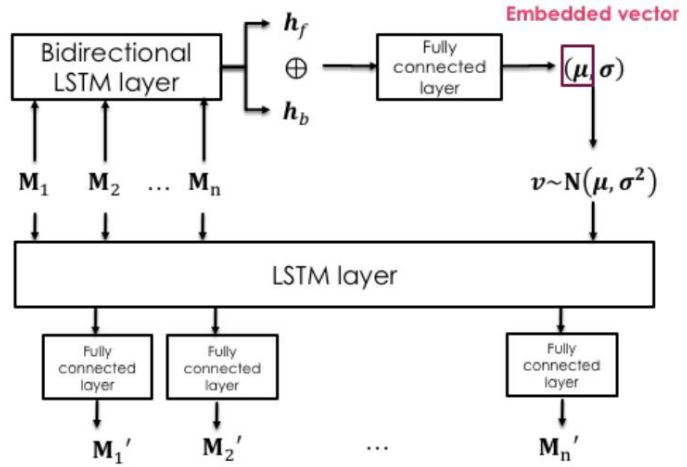
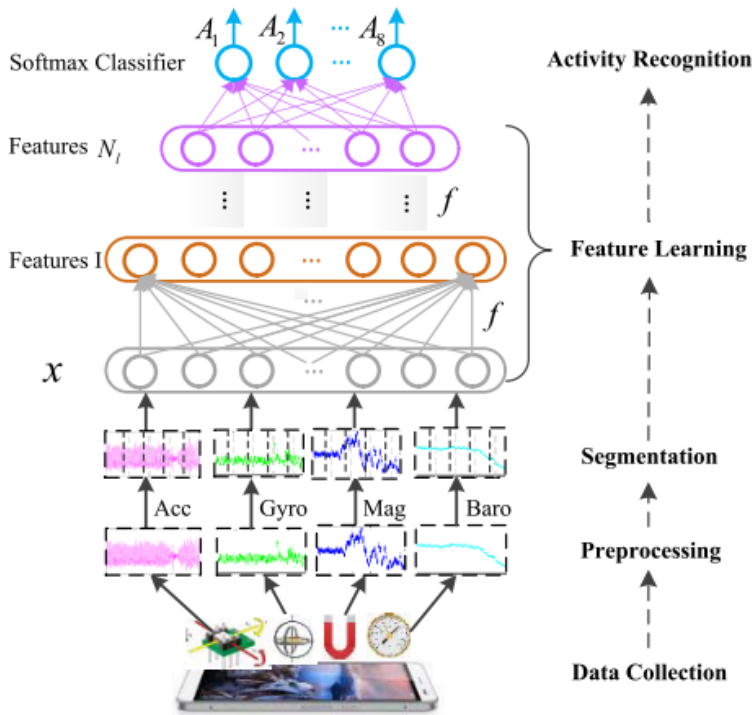


Figure 2: bidirectional LSTM

7. Fuqiang Gu, Kourosh Khoshelham, Shahrokh Valaee, Jianga Shang, and Rui Zhang. 2018. Locomotion activity recognition using stacked denoising autoencoders. IEEE Internet ofThings Journal 5, 3 (2018), 2085–2093.

Annotation Scarcity Unsupervised Learning

- we propose a deep learning method for LAR, which consists of stacked denoising autoencoders
- Moreover, the proposed method can make use of unlabeled data for model fitting in an unsupervised pretraining phase, which is especially useful when labeled data are scarce



No.	Activity	Definition
A1	Still	The user remains still and does not use the phone.
A2	Walking	The user walks naturally with a phone.
A3	False Motion	The user remains still while using the phone for texting, calling, playing games, etc.
A4	Running	The user runs with the phone swinging in hand naturally.
A5	Upstairs	Going up stairs.
A6	Downstairs	Going down stairs.
A7	UpElevator	Taking an elevator upward.
A8	DownElevator	Taking an elevator downward.

Fig. 1. Architecture of the proposed method.

6. Kaixuan Chen, Lina Yao, Dalin Zhang, Xianzhi Wang, Xiaojun Chang, and Feiping Nie. 2019. A semisupervised recurrent convolutional attention model for human activity recognition. IEEE transactions on neural networks and learning systems (2019).

Annotation Scarcity Semi-supervised Learning Co-training

- Applied co-training with multiple classifiers on different modalities of the data.
- Three classifiers are trained on acceleration, angular velocity, and magnetism, respectively.
- The learned classifiers are used for predicting the unlabeled data after each training round.
- If most of the classifiers reach an agreement on predicting an unlabeled sample, this sample is labeled and moved to the labeled set for the next training round.
- The training flow is repeated until no confident samples can be labeled, or the unlabeled set is empty. Then a new classifier is trained on the final labeled set with all modalities

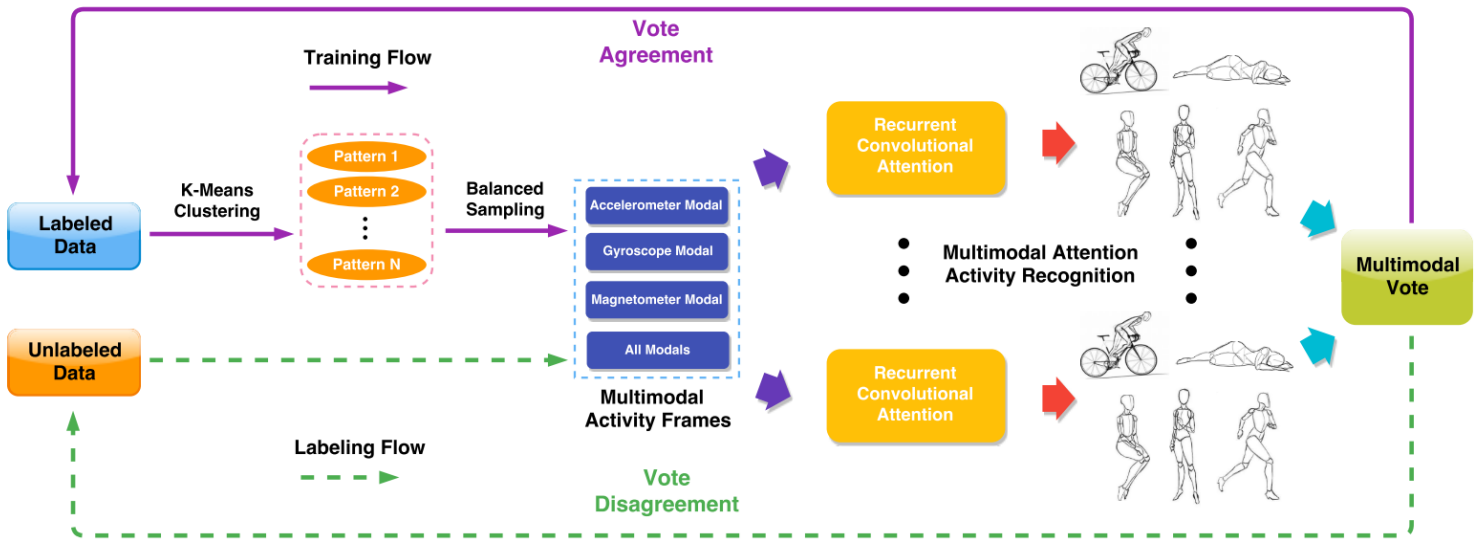
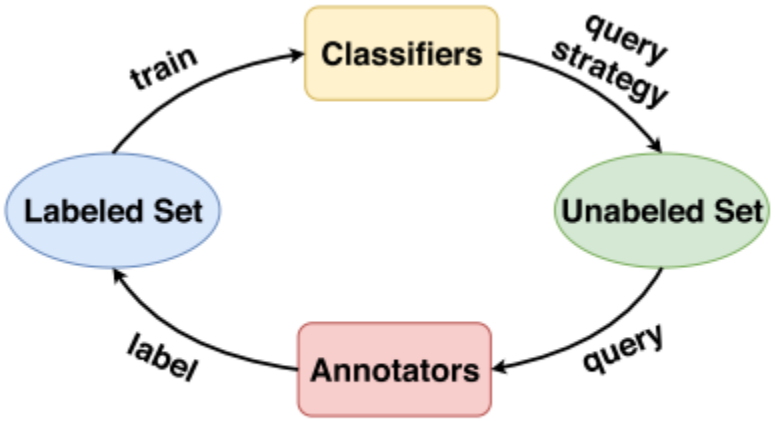


Fig. 1. Workflow of the proposed pattern-balanced co-training framework. The framework contains two flows. 1) Training flow (solid lines): labeled data are categorized into N patterns via k-means clustering and data of patterns are sampled evenly to train multimodal classifiers. 2) Labeling flow (dashed lines): predict the unlabeled data with trained models. If most of the classifiers reach an agreement on predicting a sample, this sample is labeled; otherwise, keep it unlabeled.

10. HM Hossain and Nirmalya Roy. 2019. Active Deep Learning for Activity Recognition with Context Aware Annotator Selection. In Proceedings ofthe 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 1862–1870.

Annotation Scarcity Semi-supervised Learning Active learning

- The first problem is that outliers can be easily mistaken for important samples. When entropy is calculated for selection, apart from informativeness, larger entropy may also mean outliers because outliers belong to none of the classes. Therefore, a joint loss function was proposed
- The second problem considered in this work is how to reduce the workload of annotators as annotators are required to master domain knowledge for accurate labels. Multiple annotators are employed in this work. They are selected from the intimate people of users. The annotator selection is made by the reinforcement learning algorithm according to the heterogeneity and the relations of users. The contextual similarity is used to measure the relations among users and annotators. The experimental results show that this work has an 8% improvement in accuracy and has a higher convergence rate.



(b) Active Learning

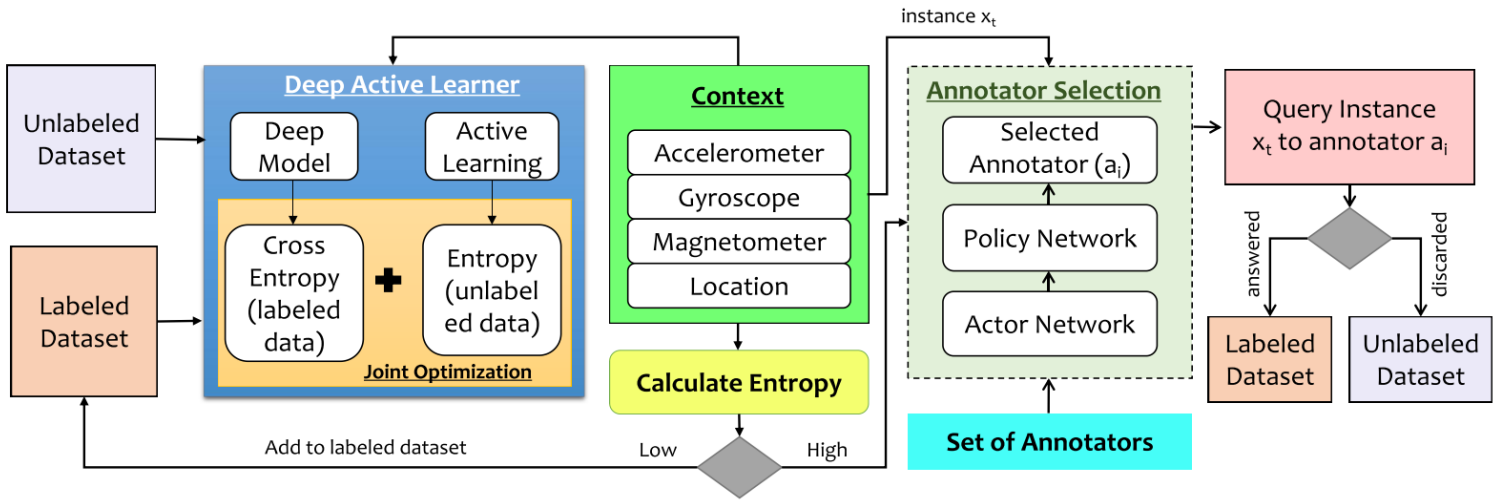


Figure 1: A High level architecture of our proposed model. The left side of the figure illustrates our active learning enabled deep model and in the right side of the figure our annotator selection pipeline is shown.

27. Jiwei Wang, Yiqiang Chen, Yang Gu, Yunlong Xiao, and Haonan Pan. 2018. SensoryGANs: An Effective Generative Adversarial Framework for Sensor-based Human Activity Recognition. In 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 1–8

Annotation Scarcity Data augmentation

- As sensory data is heterogeneous, a unified GAN may not be enough to depict the complex distribution of different activity,
- Wang et al. employed three activity-specific GANs for three activities. After generation, the synthetic data are fed into classifiers for prediction with original data.
- We should note that although this work uses deep generative networks, the generation process depends on labels so the process is not unsupervised.

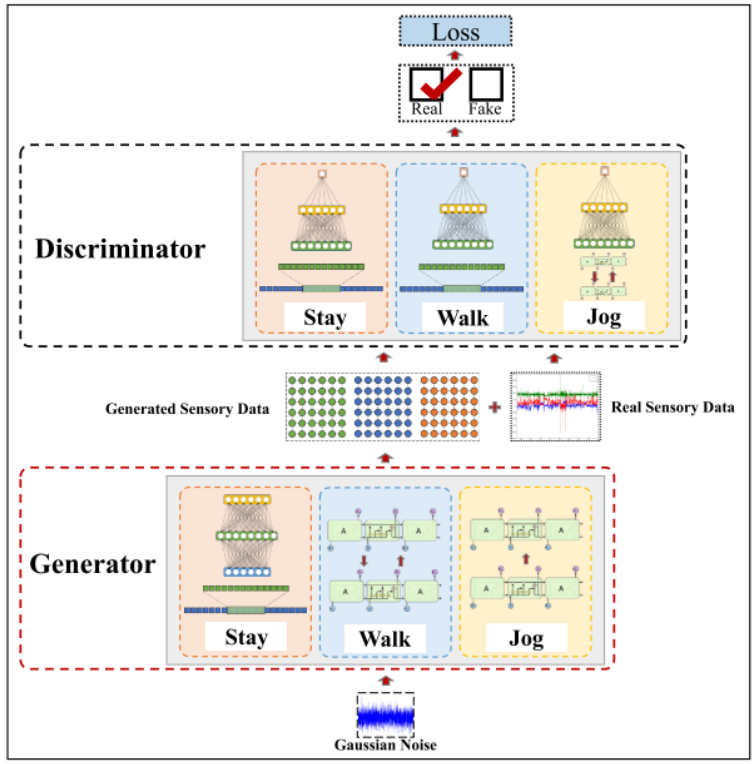


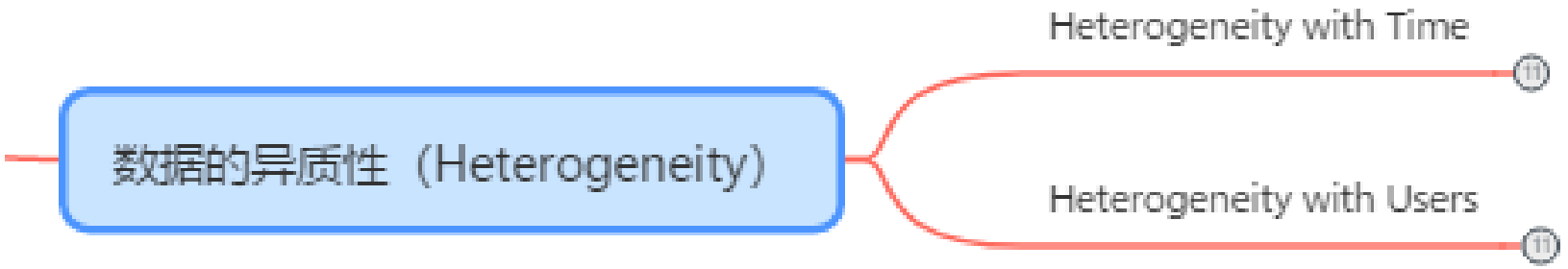
Fig. 1: The framework of SensoryGANs models

Algorithm 1 Pseudocode of SensoryGANs

Input: (1) random noises z; (2) real sensor data x
Output: synthetic sensor data $\tilde{x}$

The training and generating processes for three different activites are same
for the number of iterations **do**
  for the number of batches **do**
    Sample a batch of real sensor data $\{x_1, x_2, ..., x_n\}$
    Sample a batch of synthetic sensor data $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$
    Update the discriminator with real and synthetic sensor data by Equation 3
    for k steps **do**
      Sample random noises from the Gaussian distribution, $z \sim N(\mu, \sigma^2)$
      Use random noises and gradients from the fixed discriminator to update the generator by Equation 2
for the number of sensor samples that need to be generated **do**
  Sample random noises from the Gaussian distribution, $z \sim N(\mu, \sigma^2)$
  Use the trained generator to output synthetic sensor data
return synthetic sensor data

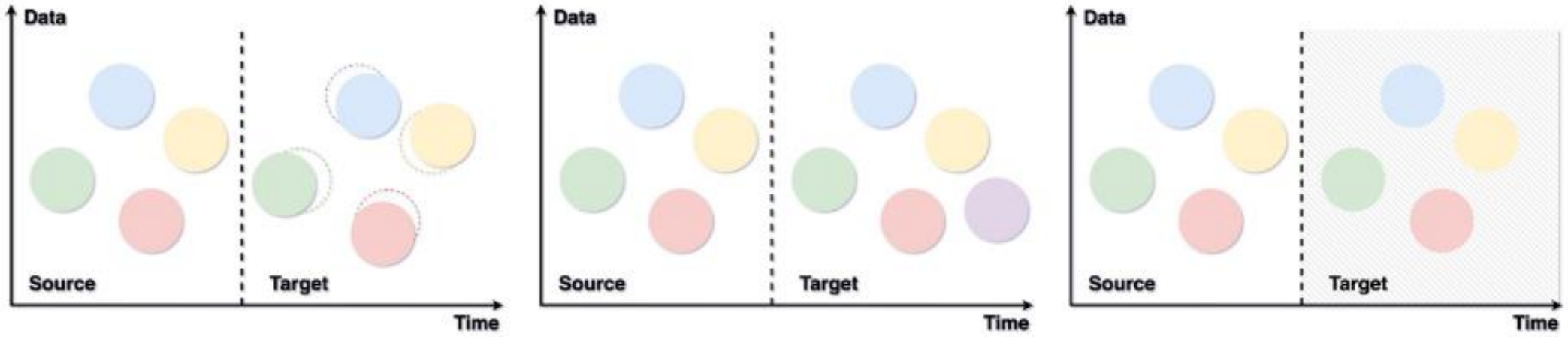
- **Heterogeneity**



- **Heterogeneity**

  - Users:
    - Owing to biological and environmental factors, the same activity can be performed differently by different individuals.

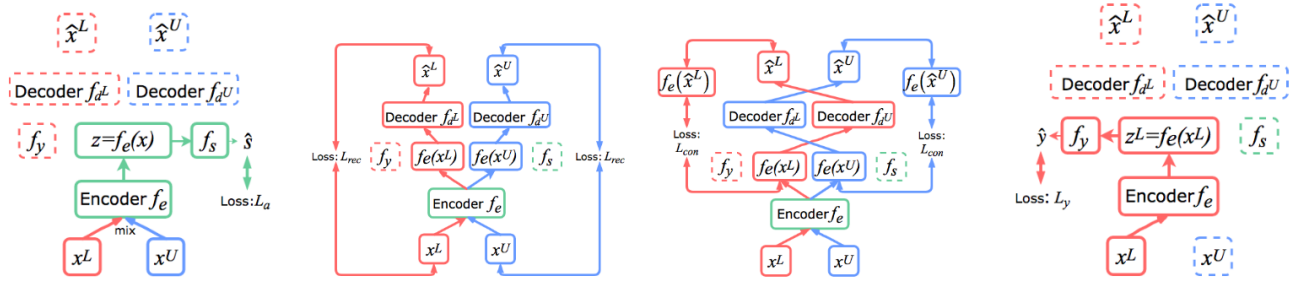
  - Time:
    - the streaming sensory data changes over time.



(a) Concept Drift (b) Concept Evolution (c) Open-Set

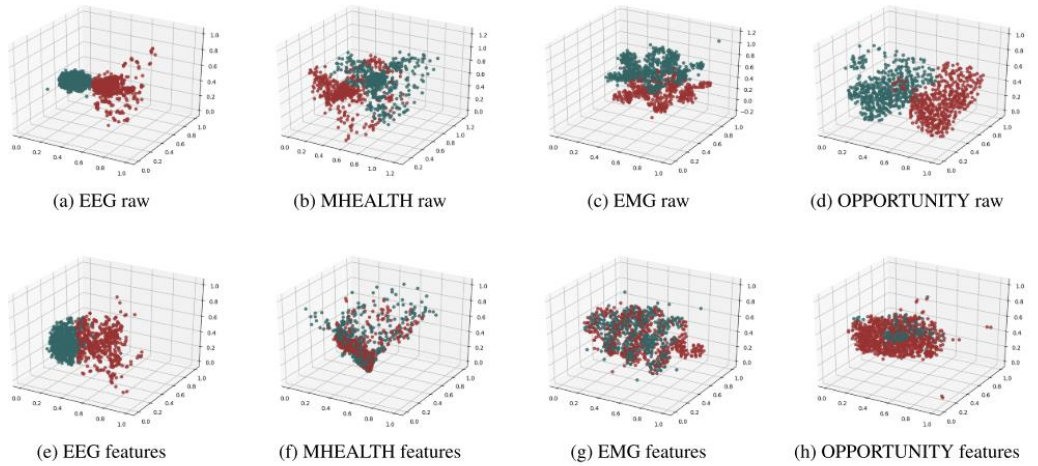
4. Kaixuan Chen, Lina Yao, Dalin Zhang, Xiaojun Chang, Guodong Long, and Sen Wang. 2019. Distributionally Robust Semi-Supervised Learning for People-Centric Sensing. In The Thirty-Third AAAIConference on Artificial Intelligence, AAAI, Honolulu, Hawaii USA, January 27 February 1, 2019. 3321–3328.

- Defined person-specific discrepancy and task-specific consistency for people-centric sensing applications.
- Person-specific discrepancy means the distribution divergence of data collected from different people, and task-specific consistency denotes the inherent similarity of the same activity.
- Their learned features not only reduce person-specific discrepancy but also preserve task-specific consistency, guaranteeing the recognition accuracy after transferring.
- ..\paper\4.pdf.pdf



(a) Person-Specific Discrepancy (b) Task-Specific Consistency (c) Prediction

Figure 1: The overview of the proposed model. We define three components of the training procedure: (a) person-specific discrepancy, (b) task-specific consistency, (c) prediction. Four losses are proposed for the objective: the adversarial loss L_a to reduce person-specific discrepancy, the reconstruction loss L_{rec} and the latent consistency loss L_{con} to preserve task-specific consistency and the prediction loss L_y. When minimizing the losses, only the activated parts are trained (indicated as solid lines) while the rest remain fixed (indicated as dashed lines). Red, blue and green denote the training procedures that are associated with the labeled samples, unlabeled samples, and a mixture of all the training samples, respectively.



(a) EEG raw (b) MHEALTH raw (c) EMG raw (d) OPPORTUNITY raw

(e) EEG features (f) MHEALTH features (g) EMG features (h) OPPORTUNITY features

Figure 3: Visualization of Latent Features. Green points correspond to the labeled data or features, while the red points correspond to the unlabeled data or features. In all cases, our model is effective in reducing distribution discrepancy.

25. Elnaz Soleimani and Ehsan Nazerfard. 2019. Cross-Subject Transfer Learning in Human Activity Recognition Systems using Generative Adversarial Networks. arXiv preprint arXiv:1903.12489 (2019).

Heterogeneity User

- The authors generated data of the target domain directly from the source domain with GANs to enhance the training of the classifier.
- This paper presents a novel method of adversarial knowledge transfer named SA-GAN stands for Subject Adaptor GAN which utilizes Generative Adversarial Network frame- work to perform cross-subject transfer learning in the domain of wearable sensor-based Human Activity Recognition.

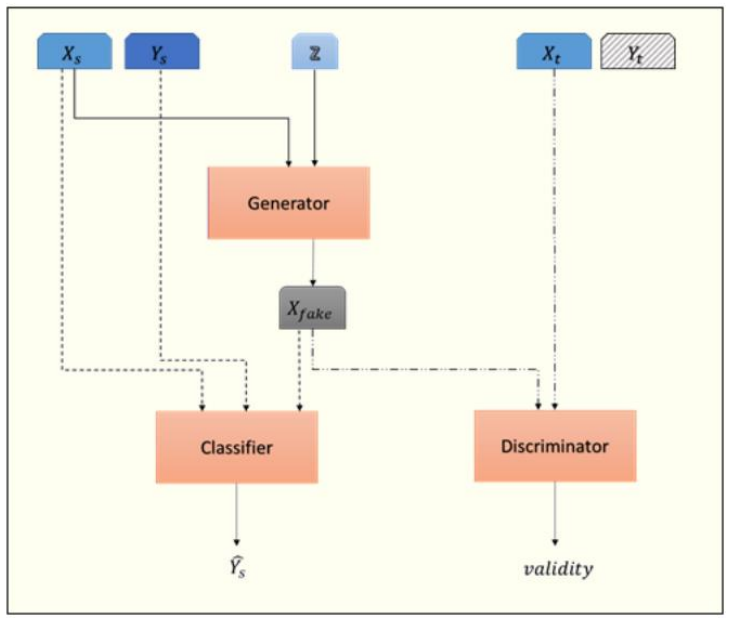


Figure 1: Abstract structure of SA-GAN. Dotted, dashed and solid lines depict input data flow to the Discriminator, Classifier, and Generator respectively.

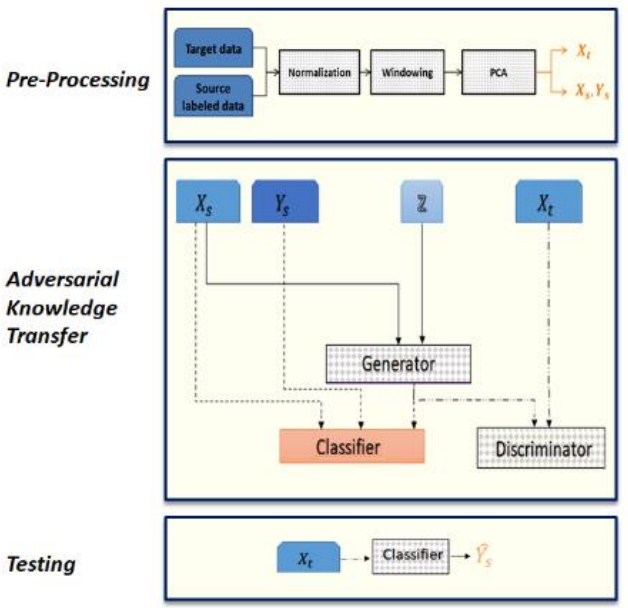


Figure 3: Overview of applying the proposed approach. The output of each step is colored in orange.

22. Seyed Ali Rokni, Marjan Nourollahi, and Hassan Ghasemzadeh. 2018. Personalized Human Activity Recognition Using Convolutional Neural Networks. In Thirty-Second AAAIConference on Artificial Intelligence

- Proposed to personalize their models with transfer learning.
- In the training phase, CNN is firstly trained with data collected from a few participants (source domain).
- In the test phase, only the top layers of the CNN are fine-tuned with a small amount of data for the target users (target domain).
- Annotation for target users is required. GAN is also serviceable for addressing heterogeneity with users.

38. Gautham Krishna Gudur, Prahalathan Sundaramoorthy, and Venkatesh Umaashankar. 2019. ActiveHARNet: Towards On-Device Deep Bayesian Active Learning for Human Activity Recognition. arXiv preprint arXiv:1906.00108 (2019)

Heterogeneity with Time

- Active learning is a special type of incremental learning. In streaming data systems, active learning is able to query ground truth for some samples when change is detected in the data streams. It encourages to select the most efficient samples to update the models for the new concepts.
- proposed a deep Bayesian CNN with dropout to obtain the uncertainties of the model and select the most informative data points to be queried according to the uncertainty query strategy. Owing to the active learning, the model supports updating continuously and capturing the changes of data over time.

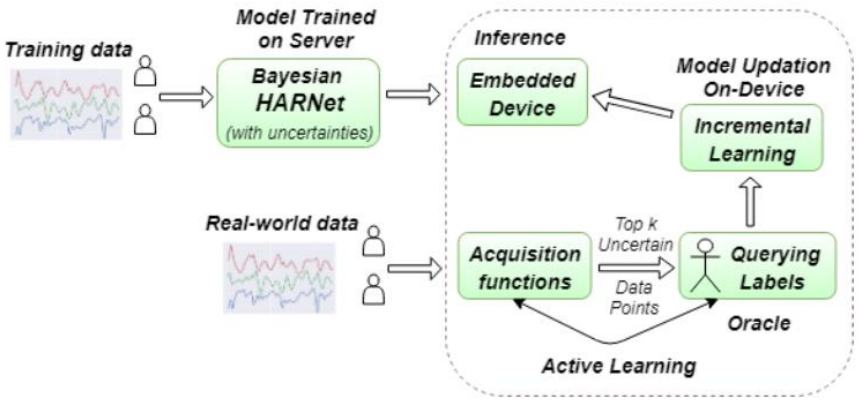


Figure 1: *ActiveHARNet* Architecture

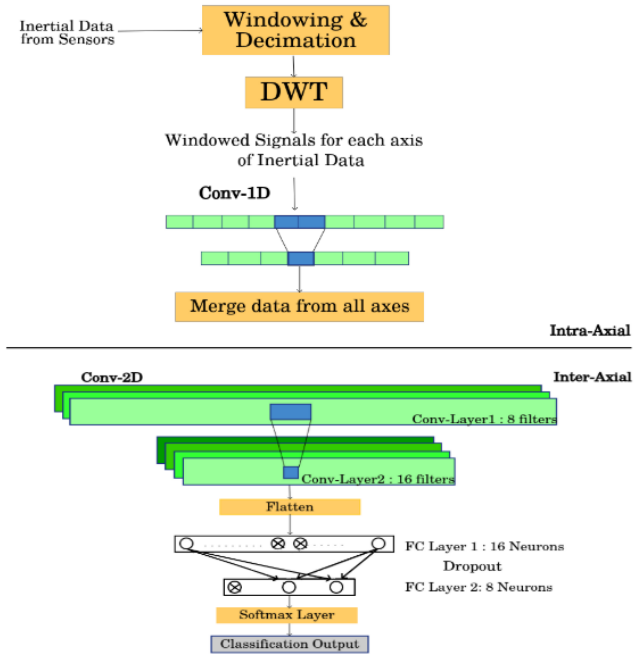
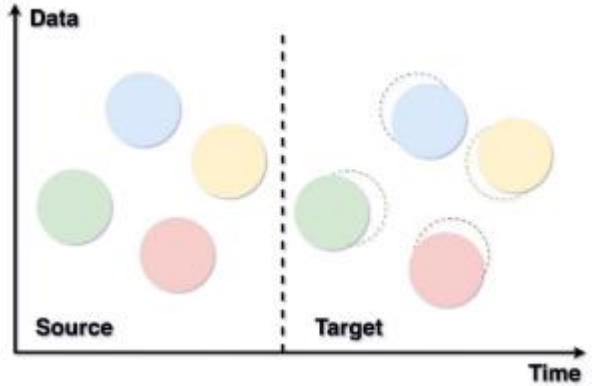
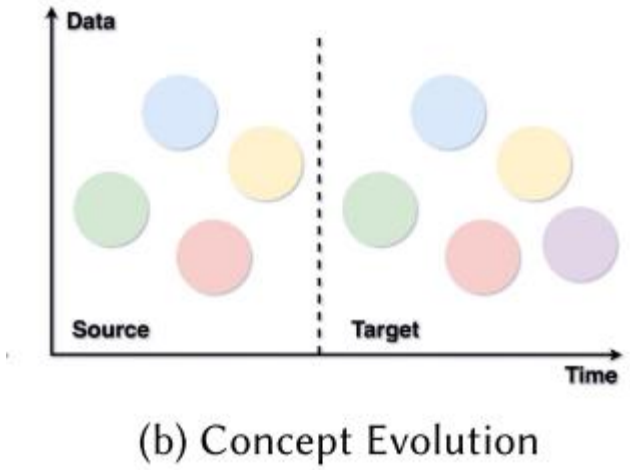


Figure 2: *HARNet Architecture*



16. Harideep Nair, Cathy Tan, Ming Zeng, Ole J Mengshoel, and John Paul Shen. 2019. AttriNet: learning mid-level features for human activity recognition with deep belief networks. In Proceedings ofthe 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings ofthe 2019 ACM International Symposium on Wearable Computers. ACM, 510–517.

- In the application phase, the concepts of the new activities still need to be learned.
- It is essential to study activity recognition systems which can recognize new activities in the streaming data settings.
- It is difficult due to the restricted access to annotated data in the application phase. One approach is to decompose activities into mid-level features such as arm up, arm down, leg up, and leg down. This method demands experts to define the mid-level attributes for further training, and the capability is limited when new activities composed of new attributes appear



(b) Concept Evolution

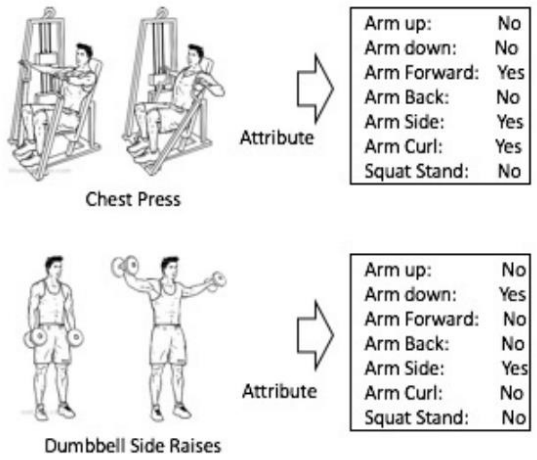


Figure 1: High-level activities can be represented by a set of mid-level attributes.

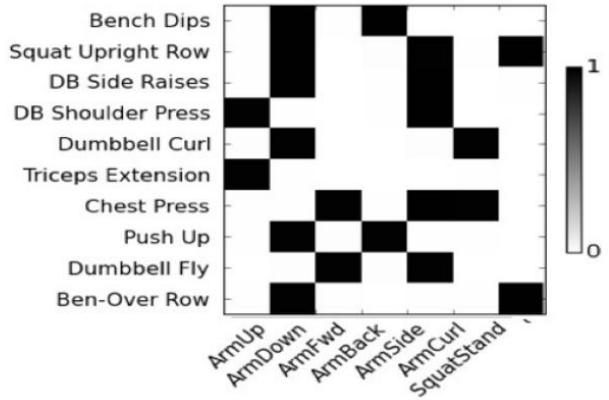


Figure 3: Correlation between human activities (in rows) and attributes (in columns) for the Exercise Activity dataset

- **Interpretability of Deep Learning Models in Sensory Data**

• Interpretability of Deep Learning Models in Sensory Data

- Soft:
  - “soft” means differentiable. Soft attention assigns weight from 0 to 1 to each element of the inputs.

- Hard:
  - whether to attend to a part of inputs or not

17. Mark Nutter, Catherine H Crawford, and Jorge Ortiz. 2018. Design of Novel Deep Learning Models for Real-time Human Activity Recognition with Mobile Phones. In 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 1–8.

Interpretability of Deep Learning Models in Sensory Data Traditional

- Present a detailed study of the traditional hand-crafted features used for shallow/statistical models that consist of a over 561 manually chosen set of dimensions.
- We show, through (PCA) and (SVM) pipeline, that the number of features can be significantly reduced – less than 100 features that give the same performance.
- In addition, we show that features derived from frequency-domain transformations do not contribute to the accuracy of these models.
- Finally, we provide details of our learning technique which creates 2D signal images from windowed samples of IMU data.

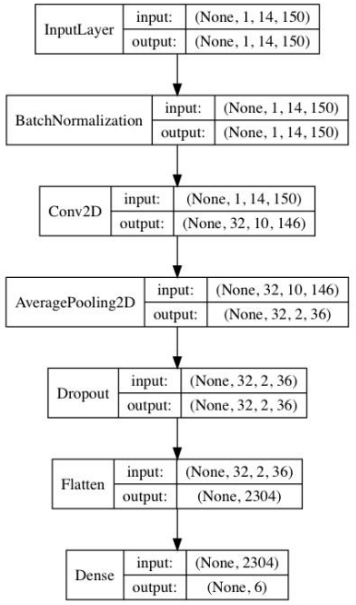


Fig. 1. Deep Neural Network pipeline for classifying 2D gray scale signal images which are created from windowed samples of IMU data as described

32. Ming Zeng, Haoxiang Gao, Tong Yu, Ole J Mengshoel, Helge Langseth, Ian Lane, and Xiaobing Liu. 2018. Understanding and improving recurrent networks for human activity recognition by continuous attention. In Proceedings ofthe 2018 ACM International Symposium on Wearable Computers. ACM, 56–63

Interpretability of Deep Learning Models in Sensory Data Soft Attention

- Developed attention mechanisms in two perspectives.
- They first propose sensor attention on the inputs to extract the salient sensory modalities and then apply temporal attention to an LSTM to filter out the inactive data segments. Spatial

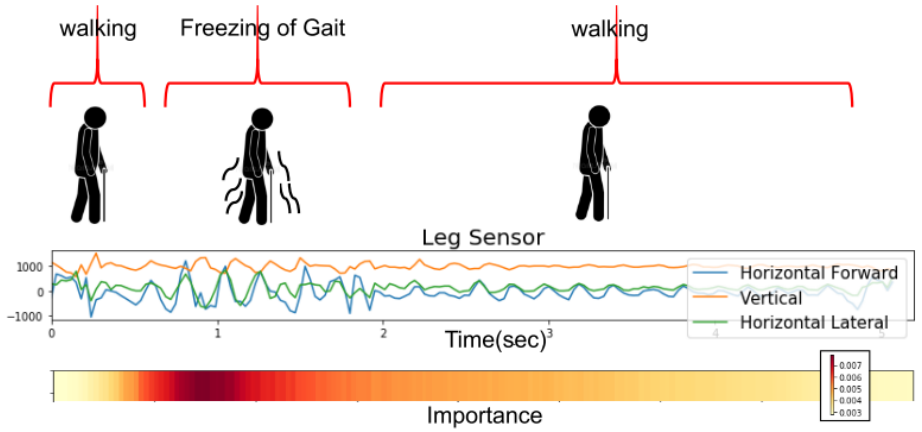
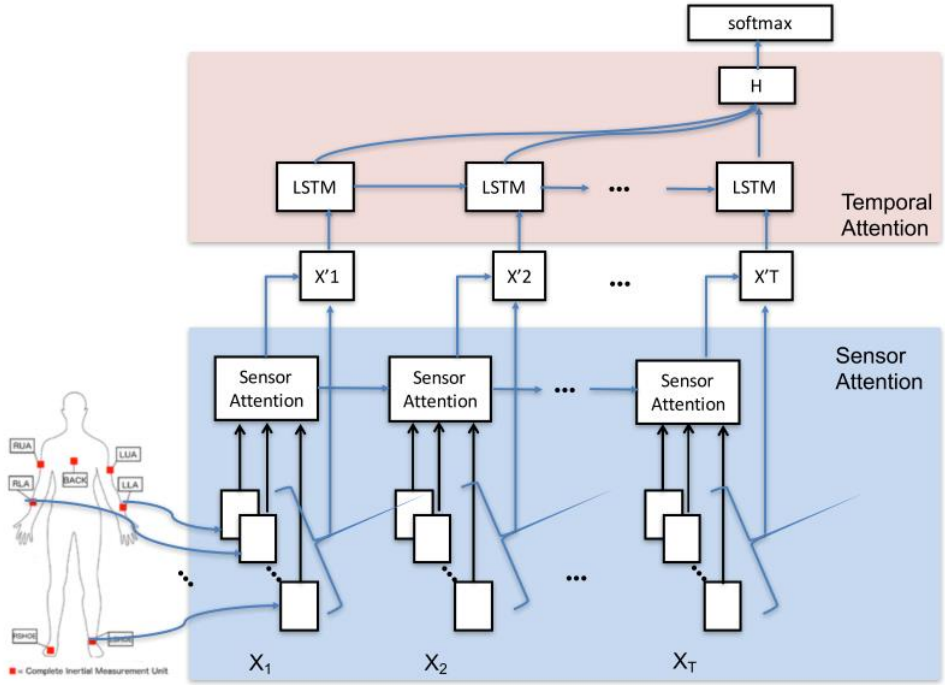


Figure 1. An example of a Freezing of Gait (FOG) detection for Parkinson disease from Daphnet Gait (DG) dataset [2]. The important acceleration signal components for FOG is shown in dark red.



24. Yu-Han Shen, Ke-Xin He, and Wei-Qiang Zhang. 2018. SAM-GCNN: A Gated Convolutional Neural Network with Segment-Level Attention Mechanism for Home Activity Monitoring. In 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). IEEE, 679–684.

Interpretability of Deep Learning Models in Sensory Data Soft Attention

- Considered the temporal context.
- They designed a segment-level attention approach to decide which time segment contains more information.
- Combined with gated CNN, the segment-level attention better extracts temporal dependencies.

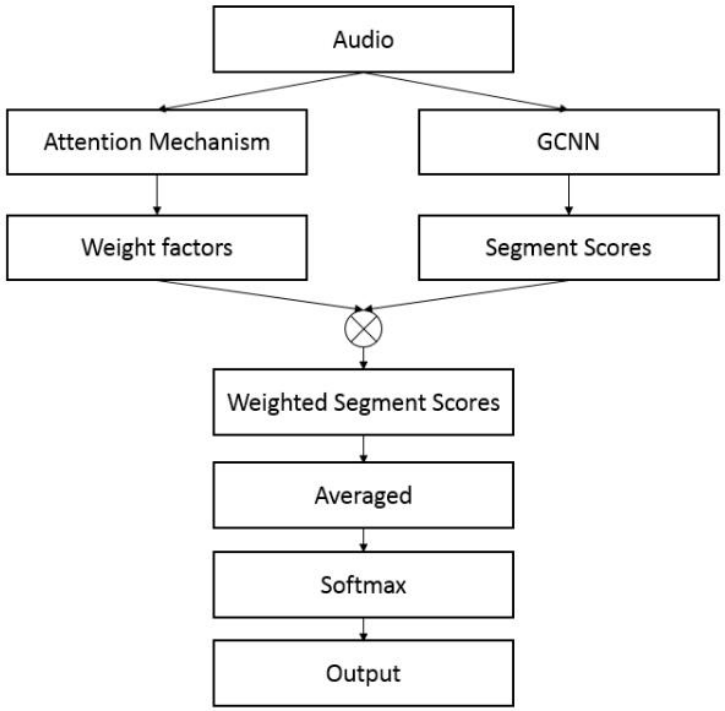


Fig. 1. Overall architecture of proposed system

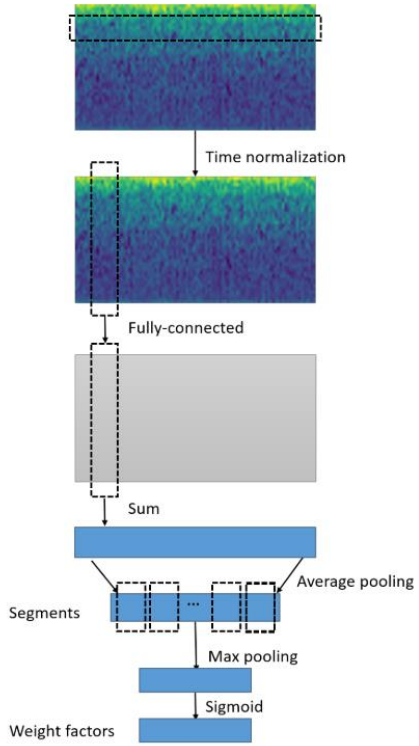
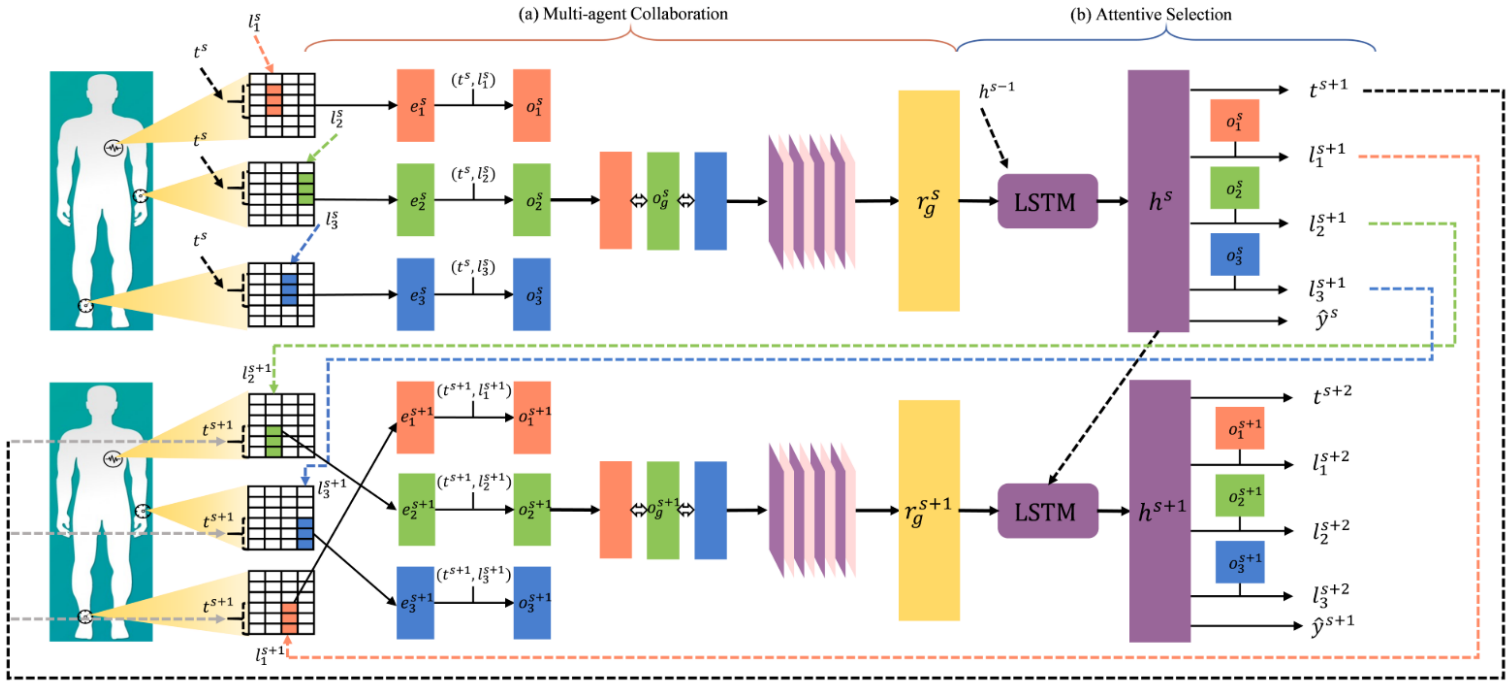


Fig. 4. Segment-Level Attention Mechanism.

5. Kaixuan Chen, Lina Yao, Dalin Zhang, Bin Guo, and Zhiwen Yu. 2019. Multi-agent Attentional Activity Recognition. In Proceedings ofthe Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI, Macao, China, August 10-16, 2019. 1344–1350.

Interpretability of Deep Learning Models in Sensory Data Hard Attention

- further considered the intrinsic relations between activities and sub- motions from human body parts.
- They employ multiple agents to concentrate on modalities that are related to sub-motions.
- Multiple agents coordinate to portray the activities.
- The visualization of the selected modalities and body parts validates that the attention mechanism provides insights into how sensory data elements affect the models' prediction of activities.



Algorithm 1 Training and Optimization

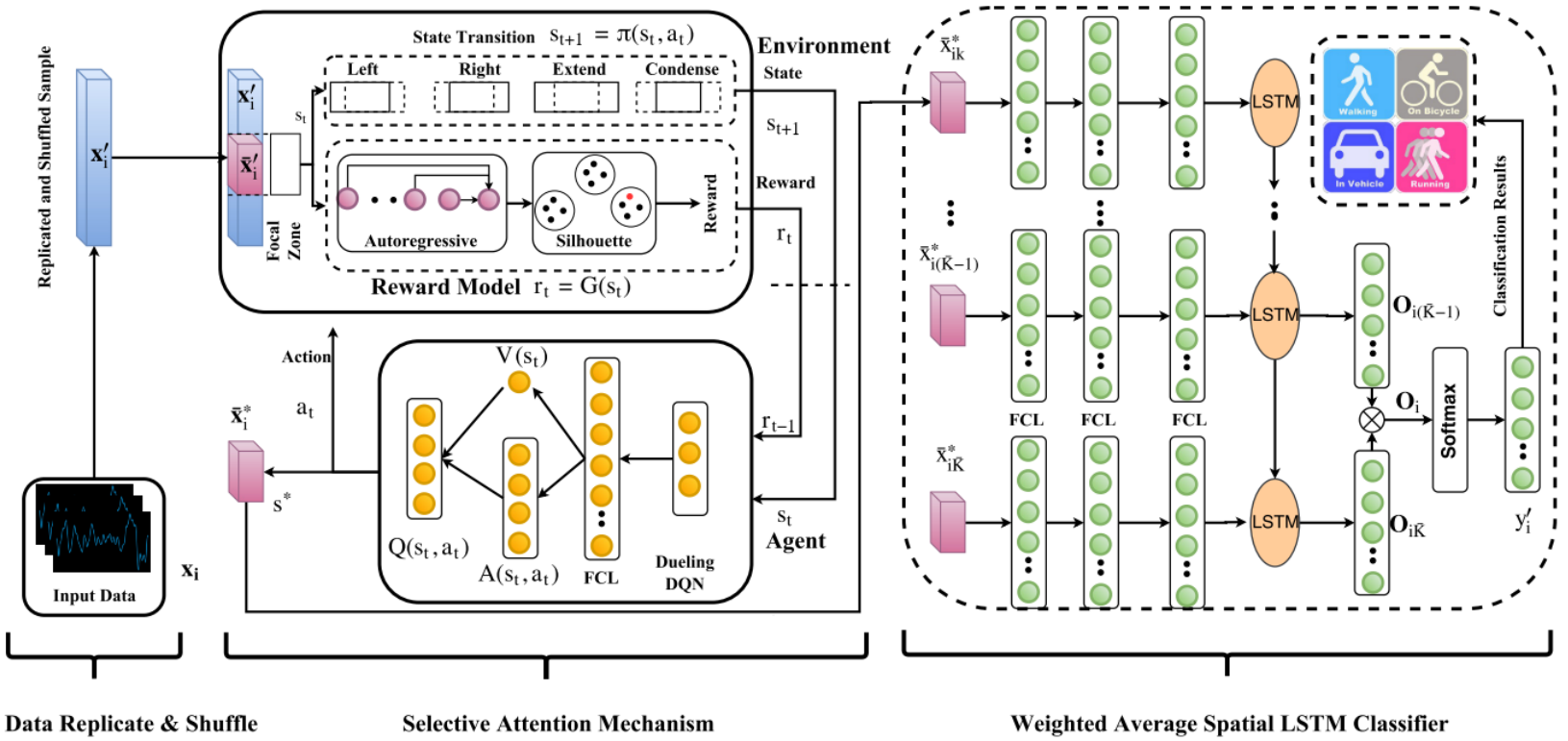
Require: sensory matrix $\mathbf{x}$, label y,
the length of episodes S,
the number of Monte Carlo samples M.
Ensure: parameters Θ.

$\Theta = RandomInitialize()$
while training **do**
  duplicate $\mathbf{x}$ for M times
  for i from 1 to M **do**
    $l_1^{1(i)}, l_2^{1(i)}, l_3^{1(i)}, t^{1(i)} = RandomInitialize()$
    for s from 1 to S **do**
      extract $e_1^{s(i)}, e_2^{s(i)}, e_2^{s(i)}$
      $o_1^{s(i)}, o_2^{s(i)}, o_3^{s(i)} \leftarrow Eq.\ 1$
      $o_g^{s(i)}, r_g^{s(i)}, h^{s(i)} \leftarrow Eq.\ 2, Eq.\ 3, Eq.\ 4$
      $l_1^{s(i)}, l_2^{s(i)}, l_3^{s(i)}, t^{s(i)} \leftarrow Eq.\ 5, Eq.\ 6$
      $\hat{y}^{s(i)} \leftarrow Eq.\ 7$
      record $\tau_{1:s}^{(i)}$
    end for
    $R^{(i)} \leftarrow Eq.\ 8$
  end for
  $\hat{y} = \frac{1}{M}\sum_{i=1}^{M} \hat{y}^{S(i)}$
  $L_c, \nabla_\Theta \bar{R} \leftarrow Eq.\ 9, Eq.\ 13$
  $\Theta \leftarrow \Theta - \nabla_\Theta L_c + \nabla_\Theta \bar{R}$
end while
return Θ

36. Xiang Zhang, Lina Yao, Chaoran Huang, Sen Wang, Mingkui Tan, Guodong Long, and Can Wang. 2018. Multi- modality sensor data classification with selective attention. In Twenty-Seventh International Joint Conference on Artificial Intelligence.

Interpretability of Deep Learning Models in Sensory Data Hard Attention

- use dueling deep Q networks as a core of hard attention to focus on the salient parts of multimodal sensory data
- we introduce a selective attention mechanism into the reinforcement learning scheme to focus on the crucial dimensions of the data.



3. Kaixuan Chen, Lina Yao, XianzhiWang, Dalin Zhang, Tao Gu, Zhiwen Yu, and Zheng Yang. 2018. Interpretable parallel recurrent neural networks with convolutional attentions for multi-modality activity modeling. In 2018 International Joint

Interpretability of Deep Learning Models in Sensory Data Hard Attention

- Mined important modalities and elide undesirable features with policy gradient.
- The attention is embedded into an LSTM to make selections step by step because LSTM incrementally learns information in an episode.

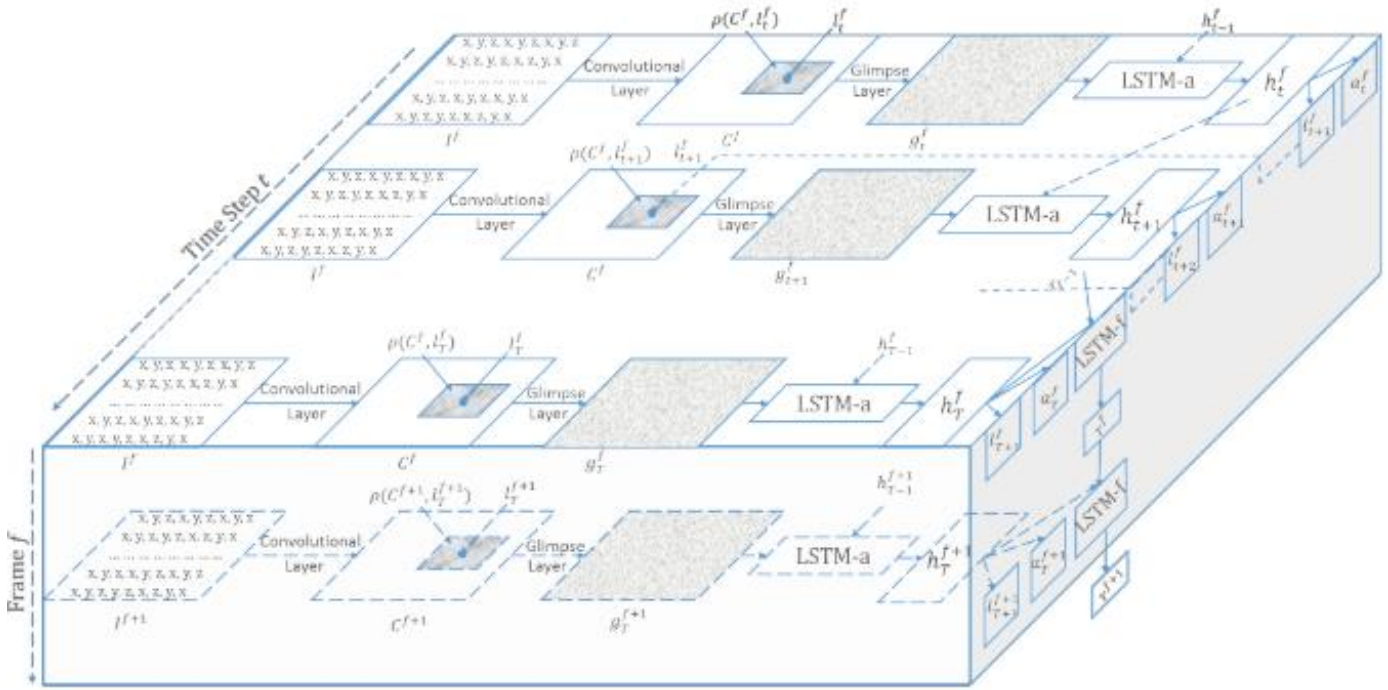


Fig. 1. Work-flow of the Proposed Approach. Dashed arrows indicate the time step t for attention based LSTM and the frame f for activity frame based LSTM, respectively. For each time step t, the input frame goes through a convolutional network to obtain a higher-level representation C^f. We extract a retina region $\rho(C^f, l_t^f)$ at location l_t^f, which is decided by the last time step $t-1$. $\rho(C^f, l_t^f)$ next goes through a glimpse layer to get the glimpse g_t^f as input of the attention based LSTM-a which decides the action a_t^f and the next location l_{t+1}^f. For the activity frame based LSTM-f takes the last action of each frame a_f^T as input and outputs the final prediction.

- **Data Segmentation**

  - Continuously streaming signals, a fixed-size window
  - Ideally, one partitioned data segment processes only one activity,
  - However, one window may not always share the same label

26. Alireza Abedin Varamin, Ehsan Abbasnejad, Qinfeng Shi, Damith C Ranasinghe, and Hamid Rezatofighi. 2018. Deep Auto-Set: A Deep Auto-Encoder-Set Network for Activity Recognition Using Wearables. In Proceedings ofthe 15th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services. ACM, 246–253

Data Segmentation

- Varamin et al. designed a multi-label architecture to simultaneously predict the number of ongoing activities and the occurring possibility of each alternative activity within a window.
- Using the optimal parameters learned from the training dataset, a Maximum A Posteriori (MAP) inference was adopted to output the most likely activity set by combining the multi-label outputs.

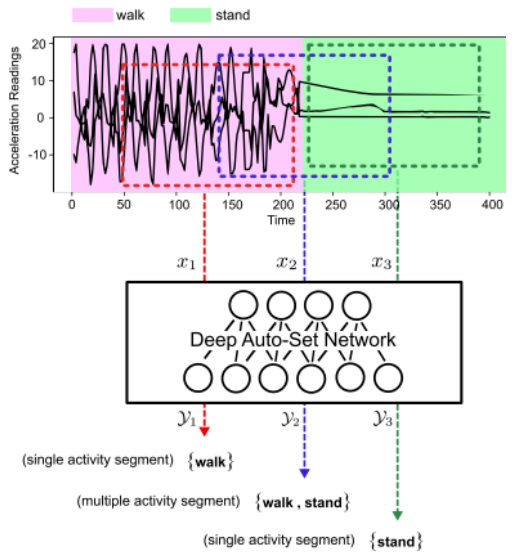
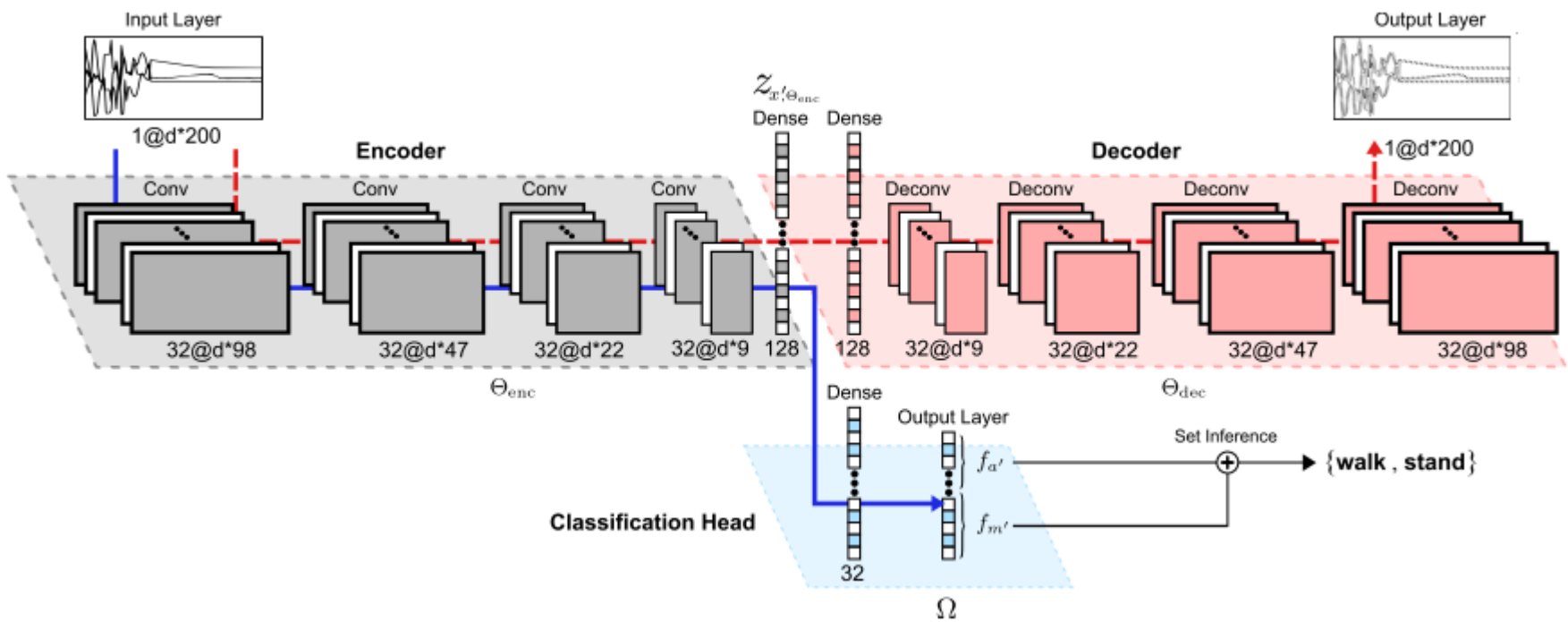


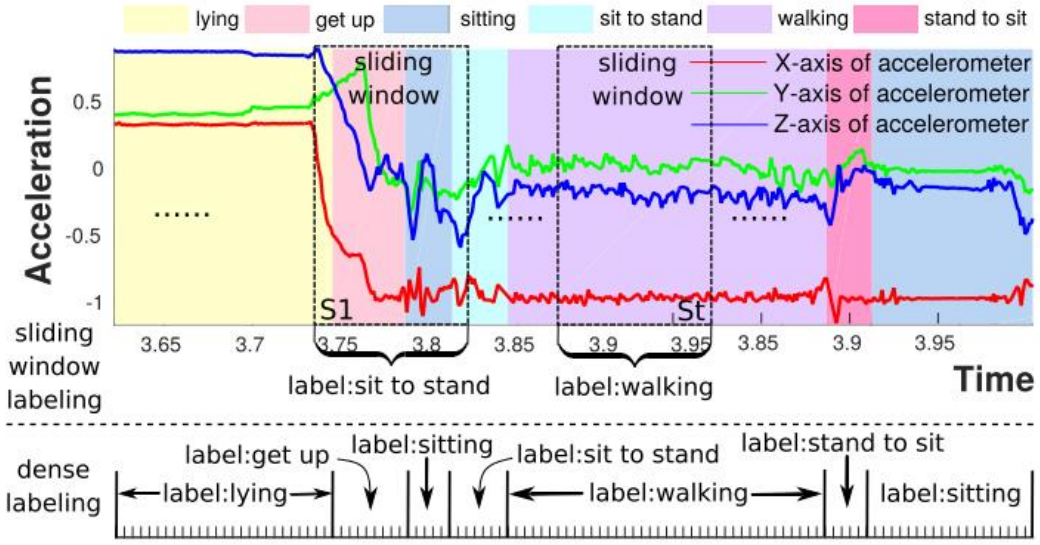
Figure 1: An illustration of our novel *Deep Auto-Set* n work to perform precise activity recognition from tin series data. Our network consumes windowed raw senso excerpts (x), automatically extracts distinctive features a outputs corresponding *sets of activities* ($\mathcal{Y}$) with various c dinalities.



31. Rui Yao, Guosheng Lin, Qinfeng Shi, and Damith C Ranasinghe. 2018. Efficient dense labelling of human activity sequences from wearables using fully convolutional networks. Pattern Recognition 78 (2018), 252–266.

Data Segmentation

- Fixed size sliding window has two key : i) the samples in one window may not always share the same label. Consequently, using one label for all samples within a window inevitably lead to loss of information; ii) the testing phase is constrained by the window size selected during training while the best window size is difficult to tune in practice.
- We propose an efficient algorithm that can predict the label of each sample, which we call dense labeling, in a sequence of human activities of arbitrary length using a fully convolutional network.



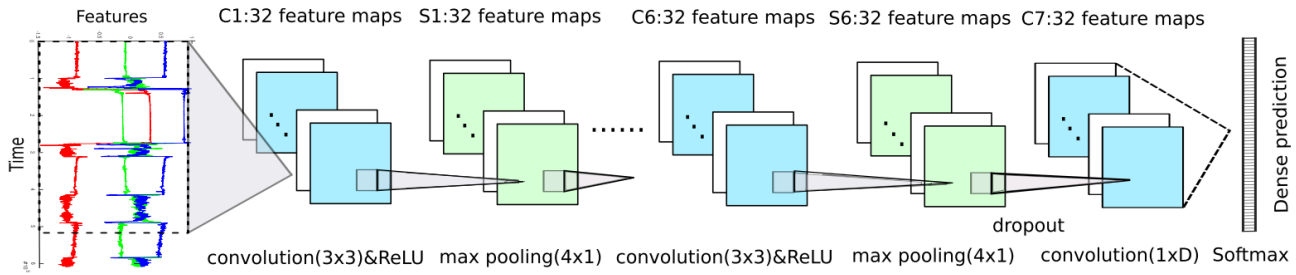


Figure 3: Detailed view of the proposed fully convolutional network architecture. The input 2D time sequence is from the wearable sensors, where one dimension is time and the other is the attributes of all sensors. The proposed model repeats the layers convolution, ReLU, and max pooling six times, where the kernel size is reported in brackets. One dropout layer then performs zeroing operation, followed by the 7-th convolutional layer, where D is the dimension of input data. The final softmax layer performs dense prediction. To ensure the output of each layer is the same length as the input sequence, we add padding in each convolutional and pooling layers.

- **Composite Activities**（复合行为）
  - More composite activities may contain a sequence of simple actions and have higher-level semantics

- **Concurrent Activity** (并行行为)
  - A person may carry out more than one activity at the same time

- **Multi-occupant Activity** (多人行为)
  - Living and working spaces are usually resided by multiple subjects

23. Silvia Rossi, Roberto Capasso, Giovanni Acampora, and Mariacarla Staffa. 2018. A Multimodal Deep Learning Network for Group Activity Recognition. In 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 1–6.

Multi-occupant Activity

- Both wearable and ambient sensors were used to recognize group activities of two occupants.
- The ambient sensors were leveraged for extracting context information which is represented by disparate functional indoor areas.
- The sensor data of different occupants was input into different RBMs separately and then merged into a sequential network, a DBN and an MLP for the inference of the group activity.
- Pretty high accuracy of nearly 100% was achieved. However, most targeting scenarios that two occupants performed the same activity together were constrained.

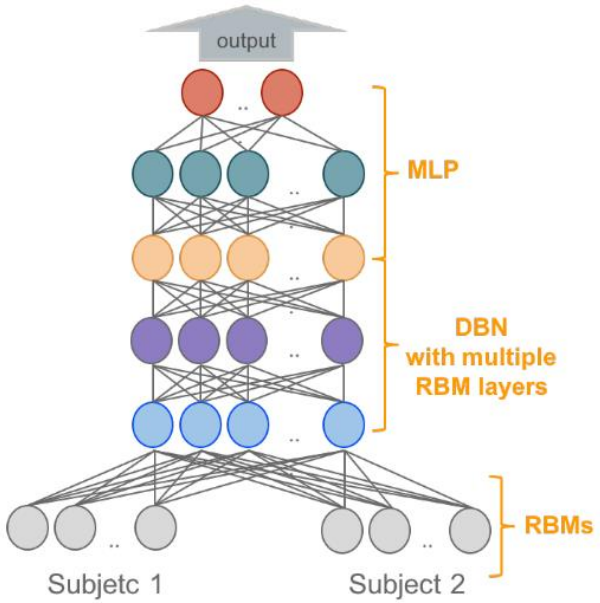


Fig. 1. Multimodal DBN

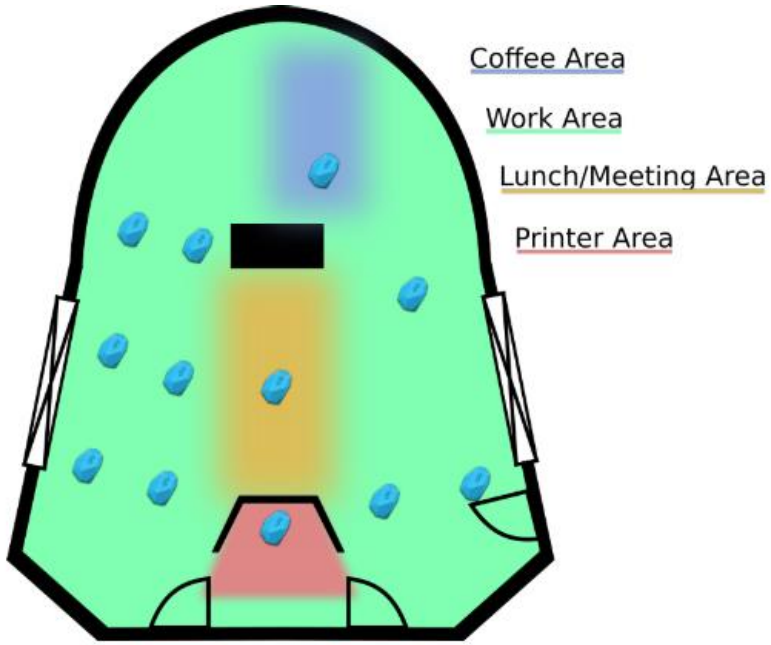


Fig. 2. PRISCA Lab map

Class	Exit	Meeting	Coffee-break	Working	Lunch	Single activity	Recall
Exit	1772	0	0	0	0	0	100%
Meeting	0	766	0	0	5	0	99.3%
Coffee-break	0	0	436	0	0	0	100%
Working	0	0	0	181	0	0	100%
Lunch	0	8	0	0	720	0	98.9%
Single activity	0	9	11	0	0	889	97.8%
Precision	100%	97.8%	97.5%	100%	99.3%	100%	

37. Yanyi Zhang, Xinyu Li, Jianyu Zhang, Shuhong Chen, Moliang Zhou, Richard A Farneth, Ivan Marsic, and Randall S Burd. 2017. Car-a deep learning structure for concurrent activity recognition. In 2017 16th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN). IEEE, 299–300.

Concurrent Activity

- Designed an individual fully-connected network for each candidate activity on top of shared multimodal fusion features.
- The final decision-make layer classified each activity independently by independent softmax layers.
- A key drawback of this kind of structure is that the computational cost would increase considerably with the number of activities rises.

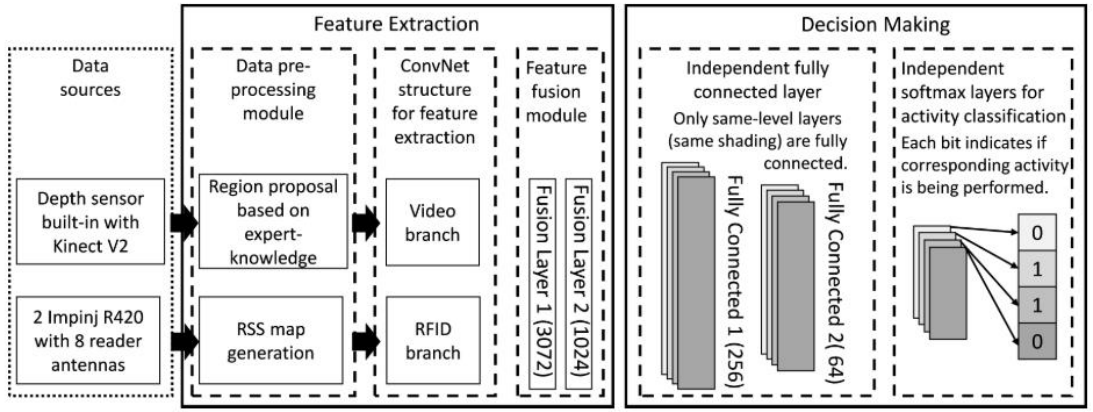


Figure 1: The overall structure of our system for concurrent activity recognition

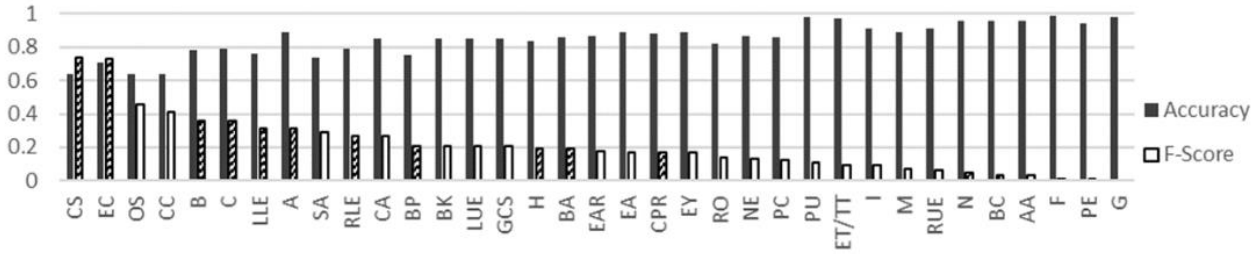


Figure 2: The recognition activity and F1-Score for 35 activities.

20. Ivan Miguel Pires, Nuno Pombo, Nuno M Garcia, and Francisco Flórez-Revuelta. 2018. Multi-Sensor Mobile Platform for the Recognition of Activities of Daily Living and their Environments based on Artificial Neural Networks.. In Twenty-Seventh International Joint Conference on Artificial Intelligence. 5850–5852

Computation Cost

- Demonstrates the combination of hand-crafted features and a neural network is a potential plan to achieve real-time activity recognition on a mobile device.

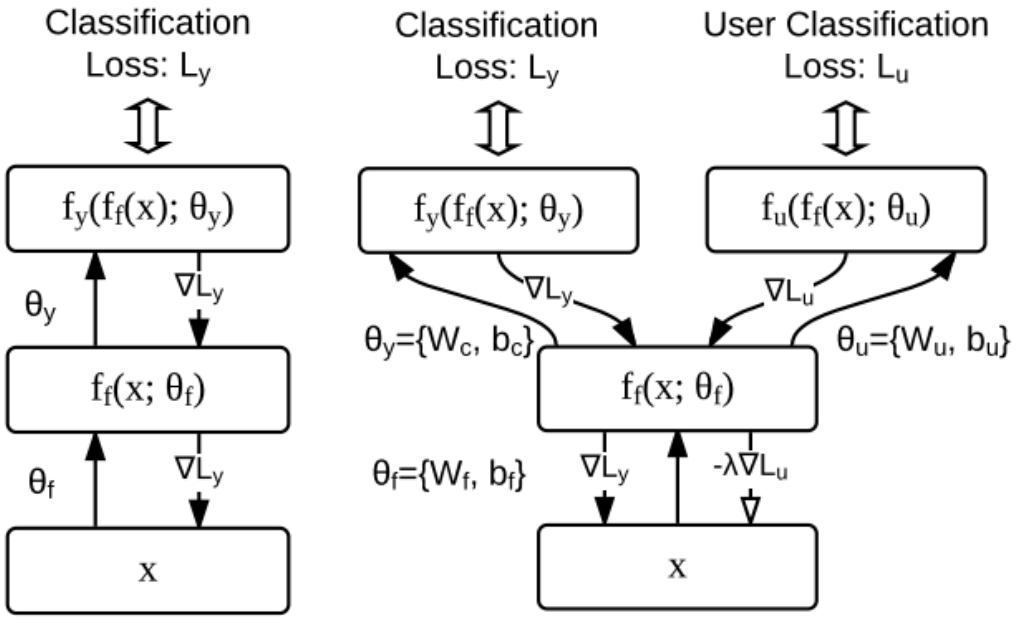
- This paper focuses on the demonstration of a mobile application that implements a framework, that forks their implementation in several modules, including data acquisition, data processing, data fusion and classification methods based on the sensors' data acquired from the accelerometer, gyro- scope, magnetometer, microphone and Global Positioning System (GPS) receiver.

- **Privacy**
  - Sensitive user information (age, weight, gender)
  - CNN features show powerful user-discriminative ability

12. Yusuke Iwasawa, Kotaro Nakayama, Ikuko Yairi, and Yutaka Matsuo. 2017. Privacy Issues Regarding the Application of DNNs to Activity-Recognition using Wearables and Its Countermeasures by Use of Adversarial Training.. In Twenty-Sixth International Joint Conference on Artificial Intelligence. 1930–1936

Privacy

- The authors investigated the privacy issue of using CNN features for human activity recognition.
- Their empirical studies revealed that although CNN is trained with a cross-entropy loss only targeting activity classification, the obtained CNN features still showed powerful user-discriminative ability.



(a) Feedforwad NN (b) User Adversarial NN

Figure 1: User-adversarial neural networks.

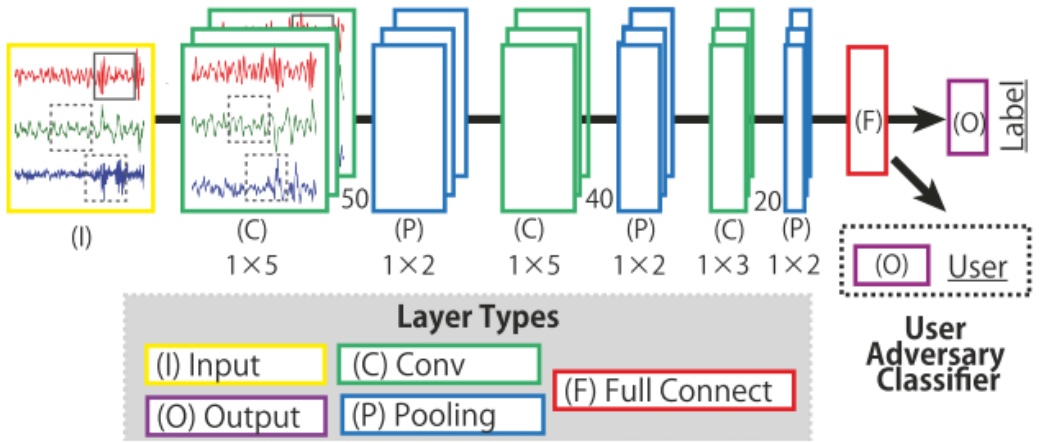


Figure 2: Network architecture used for the evaluations.

34. Dalin Zhang, Lina Yao, Kaixuan Chen, Guodong Long, and SenWang. 2019. Collective Protection: Preventing Sensitive Inferences via Integrative Transformation. In The 19th IEEE International Conference on Data Mining (ICDM). IEEE, 1–6

Privacy

- Borrowed the idea of image style transformation from the computer vision community to protect all private information at once.
- The authors creatively viewed raw sensor signals from two aspects: "style" aspect that describes how a user performs an activity and was influenced by user's identical information like age, weight, gender, height, et al.; s"content" aspect that describes what activity a user performs.
- They proposed to transform raw sensor data to have the "content" unchanged but the "style" is similar to random noises. Therefore, the method has the potential to protect all sensitive information at once.

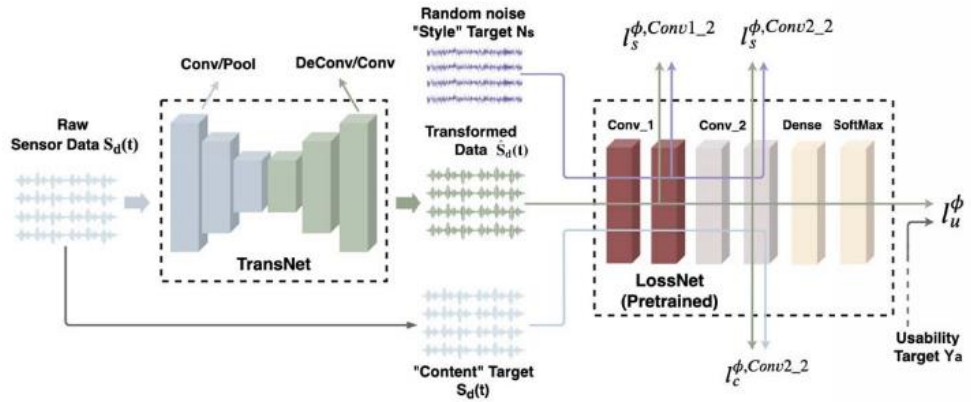


Figure 1. Framework Overview. We first pretrain the LossNet on raw sensor data for inferring desired information. Then the LossNet is fixed, and used to define the loss functions that measure "style" difference between transformed data and random noise and "content" difference between transformed data and raw data. We also define a usability loss to specifically keep the inference accuracy of the desired information. We train the TransNet through minimizing a weighted combination of the above loss functions to protect user sensitive information while simultaneously preserve the desired information.

19. NhatHai Phan, Yue Wang, Xintao Wu, and Dejing Dou. 2016. Differential privacy preservation for deep auto-encoders: an application of human behavior prediction. In Thirtieth AAAIConference on Artificial Intelligence

Privacy

- Proposed to perturb the objective functions of the traditional deep auto-encoder to enforce the ϵ-differential privacy.
- In addition to the privacy preservation in feature extraction layers, an ϵ-differential privacy preserving softmax layer was also developed for either classification or prediction.
- Different from the above approaches, this method provided theoretical privacy guarantees and error bounds.

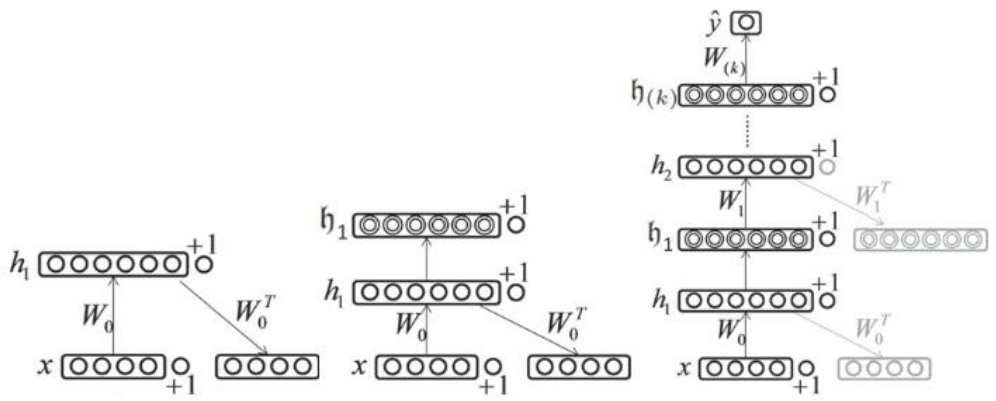


Figure 1: Examples of Auto-encoder, Private Auto-encoder, Deep Private Auto-encoder.

Algorithm 1: Pseudo Code of a dPA model

1) Derive polynomial approximation of data reconstruction function $RE(D, W)$ (Eq. 7), denoted as $\widehat{RE}(D, W)$
2) The function $\widehat{RE}(D, W)$ is perturbed by using *functional mechanism* (FM) (Zhang et al. 2012), the perturbed function is denoted as $\overline{RE}(D, W)$
3) Compute $\overline{W} = \arg\min_W \overline{RE}(D, W)$
4) Private Auto-encoder (PA) stacking
5) Derive and perturb the polynomial approximation of cross-entropy error $C(\theta)$ (Eq. 8), the perturbed function is denoted as $\overline{C}(\theta)$
6) Compute $\overline{\theta} = \arg\min_\theta \overline{C}(\theta)$; Return $\overline{\theta}$

15. Haojie Ma, Wenzhong Li, Xiao Zhang, Songcheng Gao, and Sanglu Lu. 2019. AttnSense: Multi-level Attention Mechanism For Multimodal Human Activity Recognition. In Proceedings ofthe Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019. 3109–3115.
Interpretability of Deep Learning Models in Sensory Data Soft Attention