SCAR-Net: Scalable ConvNet for Activity Recognition with multi-modal Sensor Data

Zabir Al Nazi

Abstract In recent times, context-aware activity recognition has made significant progress due to modern development of machine learning. Yet, it is challenging due to the major difference among each recognition tasks due to the change in sensor types, system design, bio-factors of the subjects, etc. In this work, a generalized solution has been presented for the Cooking Activity Recognition Challenge, where the model needs to predict the macro and micro activities from raw sensor data.
The proposed Convolutional Neural Network (SCAR-Net [1,2])[1] is an end2end, multi-head model which can work with sensor data from multiple sensors at once without truncating or padding the time-series data. SCAR-Net also performs exceptionally well without any kind of pre-processing to the data. Proposed method is tested with leave-one-subject-out cross-validation metrics (accuracy) to validate models' performance. SCAR-Net 1 achieves 54% average accuracy for the macro activity recognition task with 3 classes and SCAR-Net 2 achieves 27% average multi-label accuracy for the micro activity recognition task with 10 classes. The source-code for the complete project is available at `https://github.com/zabir-nabil/activity-recognition-abc`.

1 Introduction

Due to the rapid development of microelectronics and computer systems, many mobile devices has come to existence enabling sensors with complex characteristics. As the idea of Ubiquitous Sensing continued, devices with high computational power, low cost, small size has seen a steady growth. As a result, the research field focusing on the data acquired by pervasive sensors also expanded [1].

Zabir Al Nazi
Independent Researcher, e-mail: `zabiralnazi@yahoo.com`

[1] 1. SCAR-Net 1 for macro-activity recognition 2. SCAR-Net 2 for micro-activity recognition

Activity Recognition (AR) is a difficult and challenging task. The various methods presented in the literature vary largely in terms of the underlying sensing technologies, the models of machine learning and the realism of the world in which knowledge about behavior was collected. Regardless of the sensing technology and the machine learning model, literature is abundant with AR techniques which work extremely well on scripted or pre-segmented activity sequences [2].

In [3], authors proposed a ConvNet to perform efficient HAR using smartphone sensors by exploiting the inherent characteristics of activities and 1D time-series signals which achieved 94.79% accuracy on the test set.

A generic deep framework for activity recognition based on convolutional and LSTM recurrent units was proposed by [4] for multimodal wearable sensors, their Deep ConvLSTM outperformed other methods up to 9%.

In [10], authors proposed the multilabel classification technique with the Label Combination (LC) approach utilizing the role of sensor placements for recognizing various types of physical activities using the accelerometer sensor embedded in the smartphone. Their approach outperformed the traditional multi-class classification methods with minimized model build time.

Human Activity Recognition using Hierarchical Hidden Markov Models was proposed by [11] for streaming data. A novel online application of Hierarchical Hidden Markov Model is considered to detect the current activity on the live streaming of sensor events.

A similar approach but with Hierarchical Deep LSTM Networks by [12] also showed promising result when compared to state-of-the art methods. They used smoothing and denoising; then, the feature were selected and extracted from time–frequency-domain, the approach reached 99.15% accuracy.

In [13], authors proposed a data-driven approach for HAR with an ensemble of neural networks. Four base models were generated and evaluated using a support function fusion method. The process involved computing an output decision score for each base classifier for homogenous ensemble neural network.

An unsupervised retraining of automatic feature extraction layers and supervised fine-tuning of classification layers through a novel active learning model was proposed by [14] for recognizing activities of daily living (ADL).

In some recent works, computer vision and adversarial techniques are being applied on time-series data for classification [5, 15, 16, 17].

In the Cooking Activity Recognition Challenge, the task is to recognize both the macro activity (recipe) and the micro activities taking place during a 30 second window based on motion data collected with accelerometer and motion capture sensors. Even though the macro activity recognition is same as the usual activity recognition problem which is modeled as a multi-class classification problem, complex activities such as cooking are usually made up of several smaller activities which can have several advantages in understanding the activity in more details. The micro activity recognition task can be modeled as a multi-label classification problem [6, 7].

2 Methodology and Result Analysis

The most challenging part of the activity recognition task is the scalability. As the change in sensor organization and the activity changes the paradigm dramatically, a generalized model is needed to work across different tasks, subjects, and sensor set.

The proposed SCAR-Net is extremely generic, the same network can be used for both micro activity and macro activity recognition with only minor changes in the last layer. This feature makes SCAR-Net highly usable in broader set of tasks.

2.1 Contribution

SCAR-Net is an end2end, multi-head model which can work with sensor data from multiple sensors at once without truncating or padding the time-series data. The main features which make SCAR-Net efficient is listed below:

- SCAR-Net is an end2end network which automatically extracts useful features without any additional feature extraction step.
- It is a multi-head model which can take inputs from multiple sensor modalities concurrently.
- It can handle variable length time-series data from multiple sensors without any padding or truncating.
- SCAR-Net can achieve good performance without any pre-processing on the data.
- It does not uses any recurrent units, so it is faster to train. Inference is also faster.
- The same network can be used for both micro and macro activity recognition.

2.2 Macro Activity Recognition

In the dataset, the first task is to classify each sample into one of the three classes **’cereal’, ’fruitsalad’, ’sandwich’**. The data has been collected using two smartphones (right arm and left hip), two smartwatches (both wrists) and one motion capture system with 29 markers. There were 4 subjects who prepared 3 recipes (sandwich, fruit salad, cereal) 5 times each. The data has been separated into training data and test data. Training data contains data from 3 subjects and test data contains the fourth subject’s data. Each recording has been segmented into 30-second segments.

For generalization and simplicity, only the data from smartphones and smartwatches were used for this task.

2.2.1 Segmentation, Feature Extraction, Classification: A baseline[2]

To create a baseline, a simpler approach has been investigated to extract features. The steps of this method is documented below:

1. Each time-series is segmented based on the neighbouring value. If $abs(x[t] - x[t-1]) > threshold$ where x is a signal, is satisfied then the signal is cut at index t into two signals.
2. For each of the segmented signals, twelve features are calculated. The features are described below.
3. For the feature set from each segmented signals, a final feature set is calculated by taking a weighted sum of the features calculated from segmented signals. The weight is given based on the length of each sub-signals.
4. Another weighted sum is calculated based on the sensor channel. Each sensor channel is assigned equal weight.
5. The final feature set of length 12 is stored.

The features used in this experiment are: absolute energy, spectral moment 2, LOG, WL, autocorrelation, binned entropy, C3, AAC, MSDC (mean second derivative central), ZC (zero/mean crossing), time reversal asymmetry statistic, variance.

Features from both time and frequency domain were used for the network. 12 features were calculated from each timeseries. List of features [8, 9] with mathematical definitions are listed below:
Absolute Energy is the sum of squared values.

$$E = \sum_{i=1}^{n} x_i^2 \tag{1}$$

Spectral Moments(SM2) is a statistical approach to extract power spectrum of ECG singal and it is defined as:

$$SM2 = \sum_{i=1}^{n} P_i f_i^2 \tag{2}$$

Waveform Length (WL) is used to measure the complexity of ECG signal and is defined as:

$$WL = \sum_{i=1}^{n-1} |x_{i+1} - x_i| \tag{3}$$

Binned Entropy(BE) is calculated as:

$$BE = - \sum_{k=0}^{min(max\_bins, len(x))} P_k log(p_k).where, p_k > 0 \tag{4}$$

[2] baseline method

Average amplitude change (AAC) is formulated as

$$ACC = \frac{1}{N} \sum_{i=1}^{N-1} |x_{i+1} - x_i| \tag{5}$$

Variance is the measure of how far a random variable is spread out and Time Reversal Asymmetry Statistic (TRAS) is

$$TRAS = \frac{1}{n - 2lag} \sum_{i=0}^{n-2lag} x_{i+2lag}^2 . x_{i+lag} - x_{i+lag} . x^2 \tag{6}$$

The features are used to train a random forest classifier (estimators = 50, max depth = 1), and a support vector machine to set a baseline accuracy for the task.

2.2.2 SCAR-Net 1: Network Architecture

SCAR-Net is a multi-input ConvNet, each input layer corresponds to a sensor channel. In this experiment, only four channels are considered **('right_arm', 'right_wrist', 'left_hip', 'left_wrist')**. The most challenging part of the dataset is the variance of the length. The major difference in the timeseries length makes it harder to decide a fixed length for the input. So, the SARS-Net is designed to work with any temporal dimension. To achieve it, SCAR-Net was trained with batch size 1. It was trained on the data of two subjects and then tested on the third one for selecting the optimal model. For the final training after model selection, the dataset was shuffled and trained with a very small learning rate (0.00005) for two epochs, and after each epoch the learning rate was divided by 1.5.

SCAR-Net architecture is shown in Figure 1. It uses same number of filters in each input branch and as the batch size is 1, the temporal data can be concatenated after few layers. Instead of using Fully Connected layers, global max pooling is used to flatten the features. This is used to avoid overfitting alongside L1, L2 regularization in the convolutional layers.

2.2.3 Unit Batch Training with Sigmoid at the last layer

Figure 2 illustrates the distribution of the length of the signals in the dataset. It is evident that the length has a really high variance, so choosing a fixed length model may not be optimal. As a result, SCAR-Net is designed to handle variable length signals without any pre-processing. To compensate that, the model is trained with batch size = 1. As it is ideal to train with higher batch size, this leads to convergence issues. It is observed from the experimentation that, under these situations it is easier to train the model with sigmoid activation in the last layer. So, SCAR-Net was trained with binary cross-entropy loss and Adam optimizer.

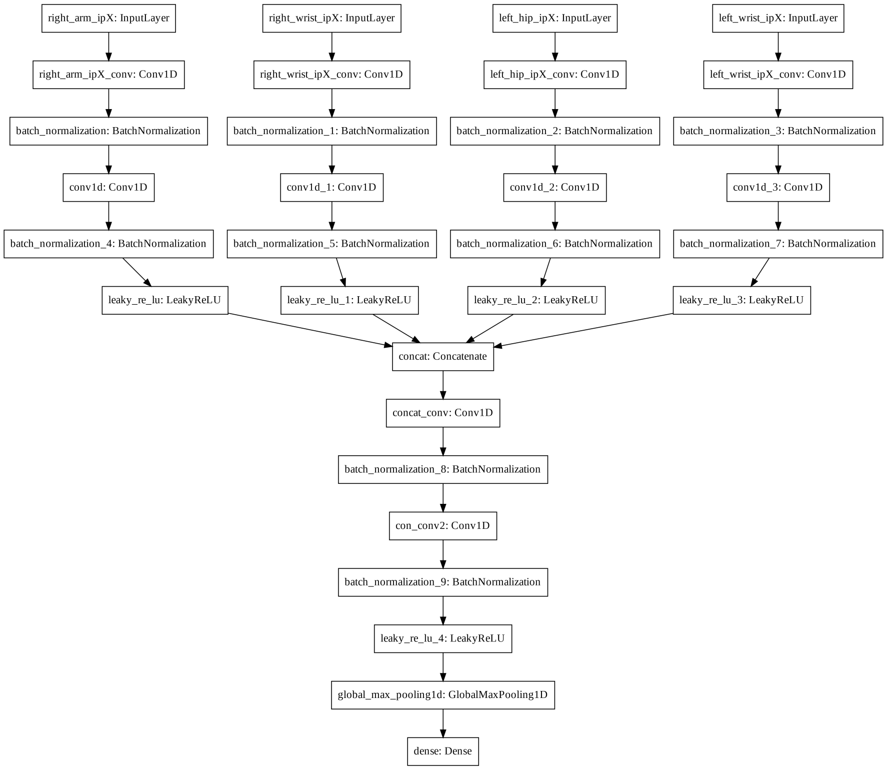
right_arm_ipX: InputLayer
right_wrist_ipX: InputLayer
left_hip_ipX: InputLayer
left_wrist_ipX: InputLayer
right_arm_ipX_conv: Conv1D
right_wrist_ipX_conv: Conv1D
left_hip_ipX_conv: Conv1D
left_wrist_ipX_conv: Conv1D
batch_normalization: BatchNormalization
batch_normalization_1: BatchNormalization
batch_normalization_2: BatchNormalization
batch_normalization_3: BatchNormalization
conv1d: Conv1D
conv1d_1: Conv1D
conv1d_2: Conv1D
conv1d_3: Conv1D
batch_normalization_4: BatchNormalization
batch_normalization_5: BatchNormalization
batch_normalization_6: BatchNormalization
batch_normalization_7: BatchNormalization
leaky_re_lu: LeakyReLU
leaky_re_lu_1: LeakyReLU
leaky_re_lu_2: LeakyReLU
leaky_re_lu_3: LeakyReLU
concat: Concatenate
concat_conv: Conv1D
batch_normalization_8: BatchNormalization
con_conv2: Conv1D
batch_normalization_9: BatchNormalization
leaky_re_lu_4: LeakyReLU
global_max_pooling1d: GlobalMaxPooling1D
dense: Dense

Fig. 1 SCAR-Net architecture

2.2.4 Result Analysis

Table 1 Result Comparison (task 1)

Method	Accuracy
SVM (baseline)	0.30
Random Forest (baseline)	0.33
SCAR-Net 1	0.54

For task 1 (macro activity), first the model is trained on the data of three subjects separately for 3 epochs each with binary cross-entropy loss. After that the model is again trained for 2 epochs on a shuffled set which contains data for all the subjects with a low initial learning rate (0.00005).

Table 1 shows the performance metrics after cross-validation. The model was trained with data of 2 subjects, and tested on the other subject.

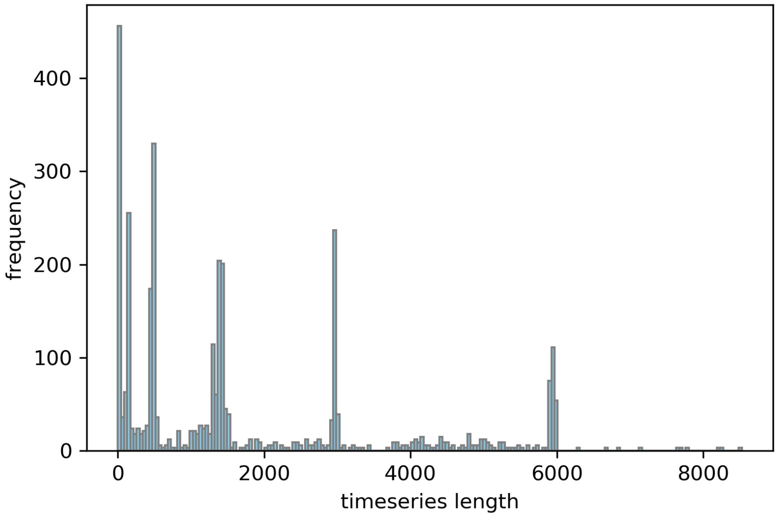

Fig. 2 Histogram plot of the timeseries length

Here, SCAR-Net outperforms the baseline methods by a big margin while maintaining all the metrics comparable. Note that, accuracy is the subject-wise cross validation accuracy, SCAR-Net was trained without any data pre-processing.

2.3 Micro Activity Recognition

The same sensor data is used for this task too, but here there can be multiple labels for a task. There are 10 classes in the dataset which are **’Add’, ’Cut’, ’Mix’, ’Open’, ’Peel’, ’Pour’, ’Put’, ’Take’, ’Wash’, ’other’**.

2.3.1 SCAR-Net 2: Network Architecture

SCAR-Net 2 has the same architecture as SCAR-Net 1. But, as there are 10 classes in this task, in the last layer 10 neurons are used. It is also trained with binary cross-entropy loss and Adam optimizer. Learning rate scheduling was used to reduce the learning rate after each epoch with a factor of 1.5.

Table 2 SCAR-Net 2 performance metrics (task 2)

accuracy	0.272
f1-score	0.289
mean absolute error	0.269

2.3.2 Result Analysis

For task 2 (micro activity), first the model is trained on the data of three subjects separately for 10 epochs each with binary cross-entropy loss. After that the model is again trained for 5 epochs on a shuffled set which contains data for all the subjects with a low initial learning rate (0.00001).

Table 2 shows the performance metrics after cross-validation. The model was trained with data of 2 subjects, and tested on the other subject.

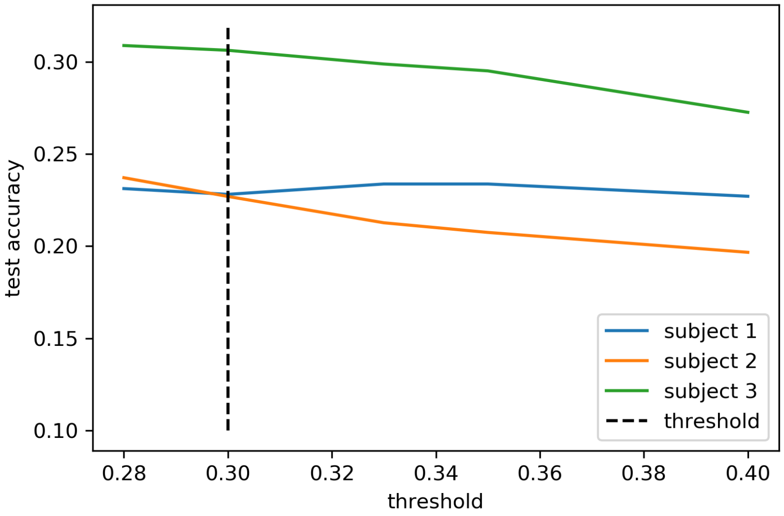

Fig. 3 Threshold selection for multi-label classification

Multi-label classification requires to set a threshold to the sigmoid output from the model to generate hard labels. To select a threshold, the test accuracy for each subject was considered for a range of threshold. The data is shown in figure 3.

For the inference stage, a threshold of 0.3 is selected based on the test accuracy on a range of thresholds. The best model achieves an average of 0.27 cross-validation accuracy on the unseen subjects.

3 Conclusion

In this work, SCAR-Net is presented which can handle variable length data in an efficient way without any pre-processing. The model is faster to train, hard to overfit, mostly consists of convolutional blocks which makes it a suitable choice for activity recognition task. Even though the model shows promising outcomes the performance metrics is not high due to the complexity of the cross-subject features. The generalization across subject is a challenging task for data with such diversified distribution. In the future, the SCAR-Net can be augmented by adding an embedding layer at the end to design a siamese network and trained with triplet loss - which can further improve the accuracy, and exploit meta-learning characteristics.

Appendix

Table 3 Processing and Resources

Sensor modalities	right arm, right wrist, left hip, left wrist
Features used	Convolutional blocks (automatic)
Programming language and libraries used	Python: tensorflow 2, tsfresh, matplotlib
Window size and Post processing	None
Training and testing time	4.2 minutes, 1.3 minutes
Machine specification	RAM: 16 GB, CPU: i7, GPU: RTX 2060

References

1. Lara O, Labrador M (2012) A survey on human activity recognition using wearable sensors. Ieee Communications Surveys & Tutorials 15:1192-1209.
2. Krishnan N, Cook D (2014) Activity recognition on streaming sensor data. Pervasive And Mobile Computing 10:138-154.
3. Ronao C, Cho S (2016) Human activity recognition with smartphone sensors using deep learning neural networks. Expert Systems With Applications 59:235-244.
4. Ordóñez F, Roggen D (2016) Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors 16:115.
5. Qin Z, Zhang Y, Meng S, Qin Z, Choo K (2020) Imaging and fusing time series for wearable sensor-based human activity recognition. Information Fusion 53:80-87.
6. Cooking Activity Recognition Challenge `https://abc-research.github.io/cook2020/learn/`
7. Paula lago, shingo takeda, kohei adachi, sayeda shamma alia, moe matsuki, brahim benai, sozo inoue, francois charpillet. Cooking activity dataset with macro and micro activities. (IEEE Dataport,2020), doi: 10.21227/hyzg-9m49
8. Christ M, Braun N, Neuffer J, Kempa-liehr A (2018) Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). Neurocomputing 307:72-77.

9. Alnazi Z, Biswas A, Rayhan M, Abir T (2019) Classification of ECG signals by dot Residual LSTM Network with data augmentation for anomaly detection.
10. Mohamed, R. Multi-label classification for physical activity recognition from various accelerometer sensor positions. *Journal Of Information And Communication Technology*. **17**, 209–231 (2020)
11. Asghari, P., Soleimani, E. Nazerfard, E. Online human activity recognition employing hierarchical hidden Markov models. *Journal Of Ambient Intelligence And Humanized Computing*. **11**, 1141–1152 (2020)
12. Wang, L. Liu, R. Human activity recognition based on wearable sensor using hierarchical deep LSTM networks. *Circuits, Systems, And Signal Processing*. **39**, 837–856 (2020)
13. Irvine, N., Nugent, C., Zhang, S., Wang, H. Ng, W. Neural Network Ensembles for Sensor-Based Human Activity Recognition Within Smart Environments. *Sensors*. **20**, 216 (2020)
14. Akbari, A. Jafari, R. Personalizing Activity Recognition Models with Quantifying Different Types of Uncertainty Using Wearable Sensors.. *Ieee Transactions On Bio-medical Engineering*. (2020)
15. Soleimani, E. Nazerfard, E. Cross-subject transfer learning in human activity recognition systems using generative adversarial networks. *Arxiv Preprint Arxiv:1903.12489*. (2019)
16. Wang, J., Zhao, Y., Ma, X., Gao, Q., Pan, M. Wang, H. Cross-Scenario Device-Free Activity Recognition Based on Deep Adversarial Networks. *Ieee Transactions On Vehicular Technology*. (2020)
17. Ismailfawaz, H., Forestier, G., Weber, J., Idoumghar, L. Muller, P. Adversarial attacks on deep neural networks for time series classification. *Arxiv*. pp. arXiv–1903 (2019)