Deep Learning for Smartphone-Based Human Activity Recognition Using Multi-sensor Fusion

Charlene V. San Buenaventura, Nestor Michael C. Tiglao(✉), and Rowel O. Atienza

Electrical and Electronics Engineering Institute, University of the Philippines, Velasquez St. Diliman, 1101 Quezon City, Philippines
charlene.san_buenaventura@upd.edu.ph,
{nestor,rowel}@eee.upd.edu.ph

Abstract. In the field of ubiquitous computing, machines need to be aware of the present context to enable anticipatory communication with humans. This leads to human-centric applications that have the primary objective of improving the Quality-of-Life (QoL) of its users. One important type of context information for these applications is the current activity of the user, which can be derived from environmental and wearable sensors. Due to the processing capabilities and the number of sensors embedded in a smartphone, this device exhibits the most promise among other existing technologies in human activity recognition (HAR) research. While machine learning-based solutions have been successful in past HAR studies, several design struggles can be easily resolved with deep learning. In this paper, we investigated Convolutional Neural Networks and Long Short-Term Memory Networks in dealing with common challenges in smartphone-based HAR, such as device location and subject dependency, and manual feature extraction. We showed that the CNN model accomplished location- and subject-independent recognition with overall accuracy of 98.38% and 90.61%, respectively. The LSTM model also performed location-independent recognition with an accuracy of 97.17% but has a subject-independent recognition accuracy of only 80.02%. Finally, optimal performance of the network was achieved by performing Bayesian Optimization using Gaussian Processes in tuning the design hyperparameters.

Keywords: Deep learning · Human activity recognition · Sensor fusion
Hyperparameter optimization

1 Introduction

For any system that requires human-machine interaction (HMI), user and environmental context are necessary to enable machines to better serve humans and improve their quality of life. An important behavioral context for HMI is the current activity being performed by the user, which can be useful for anticipatory communications between machines and humans.

It is also becoming more essential to add some form of intelligence to systems that involve HMI. For human activity recognition (HAR), pattern recognition and machine learning strategies have been the most prevalent and widely implemented solutions.

J.-L. Chen et al. (Eds.): WiCON 2018, LNICST 264, pp. 65–75, 2019.
https://doi.org/10.1007/978-3-030-06158-6_7

Although classical machine learning strategies have made remarkable progress in the field of HAR, they require domain knowledge and thus prohibit generalization across multiple application domains. Furthermore, machine learning algorithms have inadequate capabilities in modelling input data dependencies and are limited to recognition of simple tasks.

Unlike traditional shallow learning based activity recognition, deep learning algorithms accomplish sensor fusion naturally and do not require combining multi-sensor signals prior to feeding it into the network. Moreover, features are automatically learned in a hierarchical manner to accurately perform recognition. One category of deep learning networks, called Convolutional Neural Network (CNN), has been shown to be more effective in classifying data that have inherent order in them, such as time series sensor data. Another type of deep learning, called Recurrent Neural Networks (RNN), is also commonly used for time series data since learning is done through time.

Although deep learning algorithms are generally better than shallow ones, their performance greatly depends on the hyperparameters set before training. Therefore, careful tuning of design hyperparameters is crucial in determining the success of these networks. While hyperparameters are usually set by the designer manually before training, finding the optimal hyperparameter settings can be done in a more automated manner that is based on the actual data to be processed.

In this study, we used CNN to model more complex dependencies in the raw sensor input and perform accurate activity recognition in smart phones. Moreover, since CNNs have been shown to capture local dependencies and extract scale-invariant features, this paper investigates the network's ability to perform subject and device location independent recognition. We also examined RNN in its capability to model the inter-temporal dependencies within time series data. Furthermore, the recognition performance, through the network's innate ability to automatically extract hierarchical features, was improved by leveraging on sensor fusion. Lastly, hyperparameter optimization was performed by employing Bayesian Optimization using Gaussian Processes.

The rest of the paper is structured as follows. Section 2 summarizes relevant work on human activity recognition in literature. Section 3 presents the methodology in building the system. In Sect. 4, experimental results are reported and analyzed to give further insights to the current HAR problem. Finally, conclusion is drawn in Sect. 5 along with recommendations and future work.

2 Human Activity Recognition

Human activity recognition (HAR) is defined as identifying the physical activity of a person at a desired instant. It obtains its significance in several applications such as healthcare, sports and fitness, and assisted living systems.

Action recognition has been used in assessing the physical well-being of an individual as well as monitoring the rehabilitation progress of patients, such as paraplegics. In previous studies, gait analysis was used to detect step frequency and assess individuals for diagnosis, prognosis and progress of their rehabilitation [1]. Activity recognition also finds its way in the domain of sports and fitness since some people, such as athletes, are required to perform a set of activities that should be strictly

followed, to maintain a healthy body [2]. Similarly, obese people have to execute certain exercises and movements that would help in their calorie consumption [2]. Application of HAR in fitness is particularly relevant today since two-thirds of the world population is obese [3]. In all of these applications, further conclusions can be drawn for critical decision-making.

2.1 Ambient-Assisted Living (AAL)

It has been projected that 20% of the world population will belong to the senior citizen age group by the year 2050 [4], which opens up several challenges to the society. Since the elderly are more prone to diseases, this will cause a rapid increase in the diseases that our current healthcare systems can support. Due to shortage of caregivers and nursing homes, a huge majority of the elderly would still prefer to live independently in the comfort of their own homes. Hence, it is necessary to build systems and create services that assist this population while they age in place.

The general term that refers to concepts, products and services that have the goal of improving the quality-of-life (QoL) of individuals is ambient-assisted living (AAL). AAL systems usually employ intelligent technologies to assist individuals and ensure a better and a safer living environment. In AAL systems, monitoring the habitual physical activity is important for several reasons. By recording the daily activities of individuals, patterns and abnormalities in their behavior can be detected and aid into making inferences about their physical, mental and physiological well-being. Furthermore, by tracking their activities, a probabilistic model can be created to guide the intelligent system serving them.

2.2 State-of-the-Art in HAR

In this section, we discuss existing technologies used in HAR based on the platform used to gather data and infer human activity.

Vision-Based HAR
Vision-based systems have long been used in recognizing human activities and analyzing motion in general since they provide accurate characterization of the entire body [5, 6]. However, ideal results can only be obtained in controlled environments since video-based systems tend to suffer from problems such as data-association for multiple subjects. In addition, these systems are generally not immune to varying ambient conditions [6] and are computationally expensive due to the large amount of data being processed. Cameras also have limited fields of view, thus requiring installation of multiple cameras within an area. In most cases, the use of cameras is not practical in many environments because of their intrusiveness.

Sensor-Based HAR
In human activity recognition, human action can be inferred from a single sensor or from a set of sensors. Many HAR studies utilize sensor measurements from several locations on the body. However, most of these systems require the sensors to be attached firmly. Thus, wearing a network of sensors can be obtrusive and limits their real-world practicability.

There are also several attempts in using a single sensor for recognition and most of them make use of one tri-axial accelerometer. Accelerometers are widely used in motion sensing because of their low-power requirement and non-intrusiveness [7]. However, the classification performance is lower compared to when using multiple sensors. The most commonly used mobile sensors in HAR literature are accelerometers, gyroscopes, and magnetometers [7].

Smartphone-Based HAR

The smartphone is the latest technology that is being utilized for activity recognition due to its widespread use across various groups of people. Since smartphones are more integrated than other existing technologies in HAR, they can gain more acceptance due to their pervasiveness and non-intrusiveness. Hence, smartphones can be used as a cost-effective tool in pervasive healthcare to cut down healthcare costs due to the increasing population of the elderly [8].

Although HAR has been an active field of research over the past decade, very few works have successfully been deployed in mobile phones. There are still several challenges in designing a smartphone-based HAR system [9]. Two common challenges encountered in mobile-HAR are the variations in which smartphones can be positioned on the body as well as inherent diversity among humans. In previous studies, HAR models are usually only valid for a specific smartphone orientation and location. Since smartphones can be placed on different locations on the body at random placement orientations, these models are not valid in real-life scenarios. Thus, this warrants a learning algorithm that is independent of these variations. Furthermore, the manner in which different activities are performed varies from human-to-human. In most studies, the model is trained by the subject's own data, resulting to a subject-dependent prediction. However, it is sometimes inconvenient to retrain the system for each new user since collection of data can become difficult or nearly impossible in some scenarios. For example, a large volume of activities may need to be recognized, activities may be difficult to be simulated by the user, or subjects could be suffering from different medical conditions.

3 Deep Learning for Smartphone-Based HAR

In this study, the deep learning models used for classifying human activities based on sensor data from smartphones are Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) Networks which is a class of Recurrent Neural Networks (RNN). CNN is popular for time-series data or any data that has an underlying local dependency among its samples [10]. Likewise, LSTM is also capable of modelling the inherent dependencies in a time series data. Both models have the ability to automatically extract features from the raw data [10] which are more representative of the true nature of the input data. Unlike hand-engineered features, these features are better at discriminating between classes, since they are based on the data itself. For CNN, higher level features and hierarchical representations of the input are formed as you go deeper into the network.

3.1 Convolutional Neural Network

For convolutional neural networks, weights are shared across the input of each layer. A filter of weights is applied on a portion of the input and is replicated across the entire input space. This process is called convolution. The convolution process searches the occurrence of a certain feature associated with one filter in the input and outputs it into a feature map. One layer can output multiple feature maps that represent the presence of different features. The output of a convolutional layer is

$$x_i^{l,j} = \sigma\left(b_j + \sum\nolimits_{a=1}^{m} w_a^j x_{i+a-1}^{l-1,j}\right) \tag{1}$$

where $x_i^{l,j}$ is the output at the lth and jth feature map. The non-linear mapping σ is usually a ReLu function which is element-wise rectification.

After the convolutional and ReLu layers, a statistical vote is carried out over local regions in the input feature maps. We used max pooling as the statistical tool in this study. To perform max-pooling on top of a convolutional layer, the maximum value in a certain partition is obtained for all partitions of the convolved input. This operation gives rise to scale-invariant features of the input which is useful in recognizing activities that can be performed with varying intensities. The output of one max-pooling layer is given by

$$x_i^{l,j} = \max_{k=1}^{r}\left(x_{(i-1)\times s+k}^{l-1,j}\right) \tag{2}$$

Several stacks of these convolutional, ReLu and max-pooling layers in different permutations can be constructed depending on the application. Next, the output of the last layer will be flattened and fed into dense layers or fully-connected layers, similar to regular deep neural networks. Finally, the output layer is a softmax layer that will perform the final classification.

3.2 Long-Short Term Memory

Long-Short Term Memory (LSTM) is a recurrent neural network that allows us to model the temporal dynamics of the input signal more effectively since it addresses the problem of vanishing gradients. The problem of vanishing gradients arises when the output error is back propagated through several time steps. Updating the training parameters through each time step involves multiplying all the gradients. Hence, if the gradients are very small, the total product will be almost zero, and this will correspond to a zero improvement in the weights. Therefore, no learning takes place.

To regulate the problem of vanishing gradients, extra interactions are added. An LSTM cell has four main components namely, an input gate, a forget gate, an output gate and an intermediate cell state. The equations for these four components are

$$f_t = \sigma(W_f S_{t-1} + W_f X_t) \tag{3}$$

$$i_t = \sigma(W_i S_{t-1} + W_i X_t) \tag{4}$$

$$o_t = \sigma(W_o S_{t-1} + W_o X_t) \tag{5}$$

$$C_t = tanh(W_c S_{t-1} + W_c X_t) \tag{6}$$

Each of these gates is sum of old the state and the current input, each multiplied with their respective weights, and are passed to a sigmoid activation function. This allows us to control how far back in the past we want to recall. The intermediate cell state is obtained in a similar manner, but using *tanh* for the activation function.

The current cell state is computed as the sum of intermediate cell state times the input gate and the previous cell state times the forget gate. The new state will be the *tanh* of the cell state multiplied by the output gate.

$$c_t = (i_t * C_t) + (f_t * c_{t-1}) \tag{7}$$

$$h_t = o_t * tanh(c_t) \tag{8}$$

3.3 Bayesian Optimization Using GP

Bayesian optimization is a type of a sequential model-based optimization (SMBO) algorithm that uses previous observations of the loss function f in determining the next point in the hyperparameter space to sample f for. It relies on sequentially building a model for f for varying hyperparameter sets, by using smooth functions called Gaussian processes. This allows us to predict the expected performance of the network for a certain set, as well as the uncertainty of the prediction.

The posterior distribution of f is updated for every observed value of $f(\mathrm{x})$ corresponding to a hyperparameter set x that maximizes an acquisition function until desired convergence is reached. The most common acquisition function found in literature is the expected improvement (EI) which is defined as

$$EI(\mathrm{x}) = \mathrm{E}[\max\{0, f(\mathrm{x}) - f(\hat{\mathrm{x}})\}] \tag{9}$$

where $\hat{x}$ is the current optimal hyperparameter set. Maximizing this function gives us the set that improves f the most.

We can compute the expected improvement for the GP model by using integration by parts

$$EI(\mathrm{x}) = \begin{cases} (\mu(\mathrm{x}) - f(\hat{\mathrm{x}}))\Phi(z) + \sigma(\mathrm{x})\Phi(z), & \sigma(\mathrm{x}) > 0 \\ 0, & \sigma(\mathrm{x}) = 0 \end{cases} \tag{10}$$

$$z = \frac{\mu(\mathrm{x}) - f(\hat{\mathrm{x}})}{\sigma(\mathrm{x})}$$

where $\mu(\mathrm{x})$ is the expected value of f, while $\Phi(z)$ and $\Phi(z)$ are the cumulative distribution and probability density function of the standard normal distribution, respectively.

From this closed form solution, we see that the EI is high when the expected value of the loss, $\mu(\mathrm{x})$, is greater than the current best value $f(\hat{\mathrm{x}})$. Likewise, EI is high when the uncertainty $\sigma(\mathrm{x})$ is high around x. Hence, by maximizing EI, we get the points that gives a higher value of f as well as points in the region of the hyperparameter space that were not explored yet. This allows us to build the model for hyperpameter performance more efficiently.

4 Simulation Results

4.1 Dataset

Sensor Activity Dataset [11] is a publicly available dataset which consisted of accelerometer, gyroscope, and magnetometer readings by five Samsung Galaxy SII (i9100) from ten participants, while performing seven ambulatory activities. The participants performed walking, sitting, standing, jogging, biking, walking upstairs and walking downstairs for 3–4 min, while the smartphones are placed in five on-body locations namely, belt, left trousers pocket, right trousers pocket, upper arm and wrist.

Data was collected at a sampling rate of 50 Hz, which was observed to be sufficient in recognizing physical activities in the past [12]. The sensor stream was segmented by a sliding time window of 2 s with 50% overlap. The choice of both the time window length and the amount of overlap has been shown to be effective in physical activity recognition [12].

4.2 Location-Independent Prediction

Table 1 shows the summary of the recognition performance of CNN for different smartphone locations, as well as the overall performance. It can be seen that the best-performing smartphone location is the left pocket location with an accuracy of 98.78%. The overall performance for all locations is 97.37%. Similarly, Table 2 shows the summary of the recognition rates of the LSTM model. It can be observed that both left and right pockets gave the most accurate predictions having recognition rates of 98.38% and 97.49%, respectively. In this case, the recorded per-location results of both

Table 1. Classification rates for each location using CNN

	Belt	Left pocket	Right pocket	Upper arm	Wrist	Overall
Accuracy	97.01%	99.07%	97.89%	96.88%	96.27%	97.37%

models listed in Tables 1, 2 tell us that ambulatory motion is best captured when sensor data is collected from the trousers pocket of the subject.

4.3 Subject-Independent Prediction

In this section, we test the generalization ability of the models using the leave-one-subject-out validation. Using CNN, Table 3 lists the test accuracy for each subject,

Table 2. Classification rates for each location using LSTM

	Belt	Left pocket	Right pocket	Upper arm	Wrist	Overall
Accuracy	96.51%	98.38%	97.49%	96.03%	95.17%	97.17%

Table 3. Leave-one-subject-out classification rates using CNN

Test subject	Classification rate
Subject 1	85.49%
Subject 2	93.33%
Subject 3	90.73%
Subject 4	91.04%
Subject 5	92.44%
Overall	90.61%

when the model is trained with the data, from while the remaining subjects. The overall accuracy of 90.61% is the average of the five classification rates for each of the five test subjects, verifying the subject-independent recognition ability of CNN.

On the other hand, it can be seen from Table 4 that LSTM has significantly lower classification rates for the leave-one-subject-out recognition compared to CNN. Hence, it is less capable of providing subject-independent recognition. This is due to the fact that CNN has convolutional and max-pooling layers that inherently extract scale- and

Table 4. Leave-one-subject-out classification rates using LSTM

Test subject	Classification rate
Subject 1	77.07%
Subject 2	80.88%
Subject 3	83.04%
Subject 4	81.10%
Subject 5	78.02%
Overall	80.02%

shift-invariant features, while LSTM is only concerned with the temporal dependencies in the data. Both models were only trained for eight epochs, and the performance can still improve by training the models further.

4.4 Hyperparameter Optimization

In this study, we use *gp_minimize* from the Scikit-Optimize, or skopt, library, which is an implementation of Bayesian Optimization using Gaussian Processes.

The network is first evaluated at an initial set of hyperparameters, with 1000 hidden neurons, 2 hidden layers, and 1×10^{-3} learning rate. The network is then updated and evaluated on each of the hyperparameter setting at each call. The valid range of the

Table 5. List of valid ranges of hyperparameters to search in during optimization

	Lower bound	Upper bound
No. of hidden neurons	10	1000
No. of hidden layers	1	5
Learning rate	1.00E-06	1.00E-02

Table 6. Summary of validation accuracy during hyperparameter optimization

No. of hidden neurons	No. of hidden layers	Learning rate	Validation accuracy
559	5	5.70E-04	0.9886
1000	5	7.60E-04	0.9857
1000	1	1.70E-04	0.9839
1000	5	5.13E-05	0.9839
1000	5	1.30E-04	0.9838
1000	1	3.73E-05	0.982
425	2	3.10E-04	0.981
1000	1	2.30E-05	0.981
1000	5	1.60E-04	0.981
24	4	8.20E-04	0.9801

three hyperparameters that were considered are listed in Table 5. The hyperparameter setting that gave the highest validation accuracy of 98.86% is a network of 559 hidden neurons, 5 hidden layers and learns at a rate of 5.7×10^{-4}. Table 6 shows the summary for different calls during optimization while exploring the hyperparameter space, listing the top ten optimal hyperparameter settings. By plotting the results in a table, trends in the data will be more apparent, for which Gaussian processes are satisfactorily used.

5 Conclusion and Future Work

Smartphones can be an unobtrusive means of gaining contextual information from the user. In this study, Convolutional Neural Networks and Long-Short Term Memory Networks were examined in classifying activities of daily living (ADL) from three smartphone sensor signals.

Using CNN, we were able to achieve location- and subject-independent recognition which can be attributed to the presence of the convolutional and max-pooling layers in the network. The CNN model achieved an overall accuracy of 98.38% and 90.61% for location- and subject-independent recognition, respectively. For the LSTM model, we were able to achieve a location-independent recognition accuracy of 97.17%, which is slightly lower than that obtained with CNN. However, the overall accuracy of LSTM for the subject-independent recognition using the leave-one-subject-out training is 80.02%, which proves that LSTM is generally less capable of generalizing to different

subjects compared to CNN. Finally, we investigated the application of Bayesian Optimization Using Gaussian Processes in finding the optimal or near-optimal hyperparameter values.

In the future, we wish to recognize higher level activities, as well as composite ones, to challenge the classification ability of deep learning models. Furthermore, computational and resource expenditures during training and testing can also be considered, and different regularization methods can be explored.

Acknowledgement. The authors acknowledge the financial support of the University of the Philippines and Department of Science and Technology through the Engineering for Research and Development for Technology (ERDT) Program.

References

1. Zhang, Y., Markovic, S., Sapir, I., Wagenaar, R.C., Little, T.D.: Continuous functional activity monitoring based on wearable tri-axial accelerometer and gyroscope. In: 2011 5th International Conference on Pervasive Computing Technologies for Healthcare and Workshops, Pervasive Health 2011, pp. 370–373 (2011). https://doi.org/10.4108/icst.pervasivehealth.2011.245966
2. Yamansavascilar, B., Amac Guvensan, M.: Activity recognition on smartphones: efficient sampling rates and window sizes, 1–6 (2016). https://doi.org/10.1109/percomw.2016.7457154
3. Altini, M., Penders, J., Amft, O.: Energy expenditure estimation using wearable sensors: a new methodology for activity-specific models. In: Proceedings—Wireless Health 2012, WH 2012 (2012). https://doi.org/10.1145/2448096.2448097
4. Rashidi, P., Mihailidis, A.: A survey for ambient-assisted living tools for older adults. IEEE J. Biomed. Health Inform. **17**(3) (2013)
5. Khan, A.M., Tufail, A., Khattak, A.M., Laine, T.H.: Activity recognition on smartphones via sensor-fusion and KDA-based SVMs. Int. J. Distrib. Sens. Netw. 1–14 (2014). https://doi.org/10.1155/2014/503291
6. Zhu, C., Sheng, W.: Multi-sensor fusion for human daily activity recognition in robot-assisted living. In: Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction—HRI 2009 (2009). https://doi.org/10.1145/1514095.1514187
7. San Buenaventura, C., Tiglao, N.: Basic human activity recognition based on sensor fusion in smartphones. In: IFIP/IEEE IM 2017 Workshop: 1st Workshop on Protocols, Applications and Platforms for Enhanced Living Environments (2017)
8. Vavoulas, G., Pediaditis, M., Chatzaki, C., Spanakis, E., Tsiknakis, M.: The mobifall dataset: fall detection and classification with a smartphone. Int. J. Monit. Surveill. Technol. Res. **2**, 44–56 (2016). https://doi.org/10.4018/ijmstr.2014010103
9. Pires, I., Garcia, N., Pombo, N., Flórez-Revuelta, F.: From data acquisition to data fusion: a comprehensive review and a roadmap for the identification of activities of daily living using mobile devices. Sensors **16**(2), 184 (2016). https://doi.org/10.3390/s16020184
10. Zebin, T., Scully, P.J., Ozanyan, K.B.: Human activity recognition with inertial sensors using a deep learning approach. In: 2016 IEEE Sensors (2016). https://doi.org/10.1109/icsens.2016.7808590
11. Shoaib, M., Bosch, S., Incel, O., Scholten, H., Havinga, P.: Fusion of smartphone motion sensors for physical activity recognition. Sensors **14**(6), 10146–10176 (2014). https://doi.org/10.3390/s140610146

12. Wen, J., Loke, S., Indulska, J., Zhong, M.: Sensor-based activity recognition with dynamically added context. In: Mihaela, U., Valeriy, V. (eds) 12th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services MOBIQUITOUS 2015. International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, pp. e4.1–e4.10, Coimbra, Portugal, 22–24 July 2015 (2015). https://doi.org/10.4108/eai.22-7-2015.2260164