Human Activity Recognition with DeepConvLSTM

Arafat Ibne Yousuf
170104140@aust.edu

Nurnaby siddiqui
170104151@aust.edu

Foisal Reza
170104152@aust.edu

***Abstract*—Smartphones are current scientific innovations that allow us to track our activities as well as interact with one another. We require sensor data from a smartphone to identify motions like walking, going upstairs, walking downstairs, sitting, standing, and laying in order to track activities (multi-class classification). Engineered characteristics generated through heuristic methods have historically been used to address human activity recognition (HAR) challenges. Deep convolutional neural networks, according to current research, are well adapted to automate feature extraction from raw sensor inputs. Human actions, on the other hand, are built up of complex sequences of motor motions, and capturing these temporal dynamics is critical for HAR success. We present a general deep framework for activity detection based on convolutional and LSTM recurrent units, which: (i) is suitable for multimodal wearable sensors; and (ii) can perform sensor fusion naturally. Our findings show that our framework outperforms competing frameworks for deep non-recurrent networks by an average of 4% on the challenge dataset, outperforming some previously reported results by up to 9%.**

***Index Terms*—human activity recognition; wearable sensors; deep learning; machine learning; sensor fusion; LSTM; CNN; neural network**

I. Introduction

Recognizing human activities (e.g., from simple hand gestures to complex activities, such as "cooking a meal") and the context in which they occur from sensor data is at the core of smart assistive technologies, such as in smart homes [1], in health support [2]. Some simple activity-aware systems are now commercial in the form of fitness trackers or fall detection devices. However, many scenarios of high societal value are still elusive, such as providing "memory prosthesis" to people with dementia, inserting subtle cues in everyday life in the right context to support voluntary behaviour change (e.g., to fight obesity), or enabling natural human-robot interaction in everyday settings. These scenarios require a minute understanding of the activities of the person at home and out and about.

This work is motivated by two requirements of activity recognition: enhancing recognition accuracy and decreasing reliance on engineered features to address increasingly complex recognition problems. Human activity recognition is challenging due to the large variability in motor movements employed for a given action. For example, the OPPORTUNITYchallenge that was run in 2011 aiming at recognising activities in a home environment showed that contenders did not reach an accuracy higher than 88% [3].

Human activity recognition (HAR) is based on the assumption that specific body movements translate into characteristic sensor signal patterns, which can be sensed and classified using machine learning techniques. In this article, we are interested in wearable (on-body) sensing, as this allows activity and context recognition regardless of the location of the user. Deep learning refers broadly to neural networks that exploit many layers of non-linear information processing for feature extraction and classification, organised hierarchically, with each layer processing the outputs of the previous layer. Deep learning techniques have outperformed many conventional methods in computer vision [4].

Convolutional neural networks (CNNs) [5] are a type of DNN (deep neural network) with the ability to act as feature extractors, stacking several convolutional operators to create a hierarchy of progressively more abstract features. Such models are able to learn multiple layers of feature hierarchies automatically (also called "representation learning"). Long-short-term memory recurrent (LSTMs) neural networks are recurrent networks that include a memory to model temporal dependencies in time series problems. The combination of CNNs and LSTMs in a unified framework has already offered state-of-the-art results in the speech recognition domain, where modelling temporal information is required [6]. This kind of architecture is able to capture time dependencies on features extracted by convolutional operations.

Deep learning techniques are promising to address the requirements of wearable activity recognition. First, performance may chiefly be improved over existing recognition techniques. Second, deep learning approaches may have the potential to uncover features that are tied to the dynamics of human motion production, from simple motion encoding in lower layers to more complex motion dynamics in upper layers. This may be useful to scaling up activity recognition to more complex activities.

II. Related Works

A large number of researchers have conducted considerable work in exploring different sensing technologies and proposed a number of methods to model and recognize human activities Several scholars used the machine learning (ML) method to predict Human Activities using Human Activity Recognition with Smartphones Dataset.

L Bao and Intille [7] collected experimental data from twenty volunteers and compared the recognition rate of several different classifiers. Experimental results showed that decision tree classifier can obtain the best performance with an accuracy of 84.0%.

Ravi et al. [8]carried out a study to explore the possibility of activity recognition with a single triaxial accelerometer.

In the phase of data collection, they used an accelerometer mounted onto the pelvic region of an individual to collect raw accelerometer data about eight activities, including standing,walking, running, upstairs, and downstairs, sitting, vacuuming, and brushing teeth.

Lee and Masc [9]built a system that consisted of a biaxial accelerometer, a gyroscope and a digital compass to infer a user's location and classify siting, walking and standing activities.

Yun et al. [10] proposed to use a triaxial accelerometer, a triaxial angular rate sensor and a magnetometer to collect foot motion related data for estimating foot kinematics. Their system was able to extract information about foot orientation, velocity, acceleration, position and gait phase, and obtain relatively low estimation error

III. Project Objectives

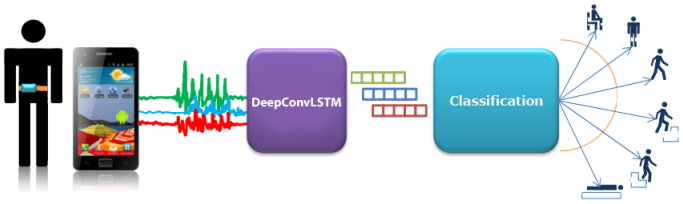

Fig. 1: The Proposed System Pipeline

We introduce a new DNN framework for wearable activity recognition, which we refer to as DeepConvLSTM. This architecture combines convolutional and recurrent layers. The convolutional layers act as feature extractors and provide abstract representations of the input sensor data in feature maps. The recurrent layers model the temporal dynamics of the activation of the feature maps. In this framework, convolutional layers do not include a pooling operation . In order to characterise the benefits brought about by DeepConvLSTM, we compare it to a "baseline" non-recurrent deep CNN. Both approaches are defined according to the network structure depicted in. For comparison purposes, they share the same architecture, with four convolutional layers and three dense layers. The input is processed in a layer-wise format, where each layer provides the representation of the input that will be used as data for the next layer. The number of kernels in the convolutional layers and the processing units in the dense layers is the same for both cases. The main difference between DeepConvLSTM and the baseline CNN is the topology of the dense layers. In the case of DeepConvLSTM, the units of these layers are LSTM recurrent cells, and in the case of the baseline model, the units are non-recurrent and fully connected. Therefore, performance differences between the models are a product of the architectural differences and not due to better optimisation, preprocessing or ad hoc customisation.

IV. Methodologies / Model

DeepConvLSTM is a DNN, which comprises convolutional, recurrent and softmax layers. Firstly, sensor data are transformed through four convolutional operations. Convolutional layers process the input only along the axis representing time. The number of sensor channels is the same for every feature map in all layers. These convolutional layers employ rectified linear units (ReLUs) to compute the feature maps. Layers 6, 7 and 8 are recurrent dense layers. Recurrent dense layers adapt their internal state after each time step. The activation of the recurrent units is computed using the hyperbolic tangent function. The output of the model is obtained from a softmax layer (a dense layer with a softmax activation function), yielding a class probability distribution for every single time step. The shorthand description of this model is: C(64) - C(64) C(64) - C(64) - R(128) - R(128) - R(128) - Sm, where C(F, L) denotes a convolutional layer L with F feature maps, R(n) a recurrent LSTM layer with n cells and Sm a softmax classifier. Models are trained in a fully supervised way, backpropagating the gradients from the softmax layer through to the convolutional layers. The number of parameters to optimize in a DNN varies according to the type of layers it comprises and has a great impact on the time and computing power required to train the networks. The number and size of the parameters in the networks are presented below. .

Model: "model"

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 128, 6, 1)]	0
conv2d (Conv2D)	(None, 124, 6, 64)	384
conv2d_1 (Conv2D)	(None, 120, 6, 64)	20544
conv2d_2 (Conv2D)	(None, 116, 6, 64)	20544
conv2d_3 (Conv2D)	(None, 112, 6, 64)	20544
reshape (Reshape)	(None, 112, 384)	0
lstm_1 (LSTM)	(None, 112, 128)	262656
dot_1 (Dropout)	(None, 112, 128)	0
lstm_2 (LSTM)	(None, 112, 128)	131584
dot_2 (Dropout)	(None, 112, 128)	0
lstm_3 (LSTM)	(None, 128)	131584
dot_3 (Dropout)	(None, 128)	0
act_smx (Dense)	(None, 6)	774

Total params: 588,614
Trainable params: 588,614
Non-trainable params: 0

Fig. 2: Model of Implementation Human Activity Recognition using Deep Learning

For the sake of efficiency, when training and testing, data are segmented on mini-batches of a size of 100 data segments. Using this configuration, an accumulated gradient for the parameters is computed after every mini-batch. Both models are trained with a learning rate of 10e3 and a decay factor of 0.9. Weights are randomly orthogonally initialized. We introduce a drop-out operator on the inputs of every dense layer, as a form of regularization. This operator sets the activation of randomly selected units during training to zero with probability p = 0.5.

V. Dataset

Accelerometer and gyroscope tri-axial sensor data were collected from 30 volunteer subjects who performed six different activities while the smartphone was in their pockets. These sensor data were sampled at a rate of 50 Hz, and were separated into windows of 128 values, with 50% overlap; the 128-real value vector stands for one example for one activity (for each acc and gyro axis). There are a total of 7352 examples for the training data (from 21 randomly selected subjects), and 2947 examples for the test data (from the remaining 9 subjects).

- WALKING - 1
- CLIMBING UP THE STAIRS - 2
- CLIMBING DOWN THE STAIRS - 3
- SITTING - 4
- STANDING - 5
- LAYING - 6

VI. Data visualization

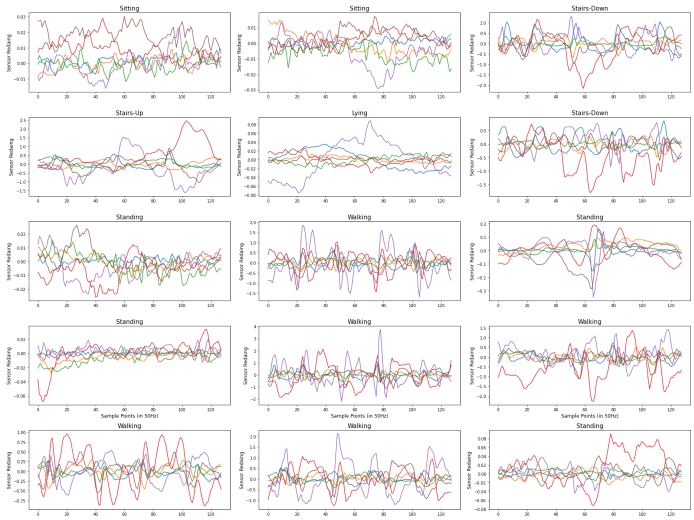

Fig. 3: Sensor reading for different activities

VII. Result Analysis

In this section, we present the results and discuss the outcome. We show the performance of the approaches and also evaluate some of their key parameters to obtain some insights about the suitability of this approach for the domain.

TABLE I: Model Comparison

REF	MODEL	ACCURACY
[11]	SVM	96.40
[12]	AdaBoost	94.33
[13]	CNN	85.10
[14]	Deep-Res-Bidir-LSTM	90.20
	DeepConvLSTM	**98.45**

Fig. 4: Accuracy vs Loss

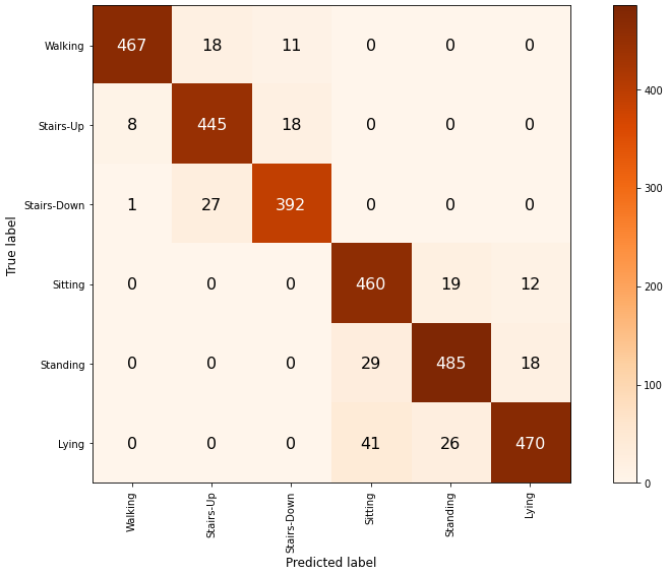

Fig. 5: Confusion Matrix

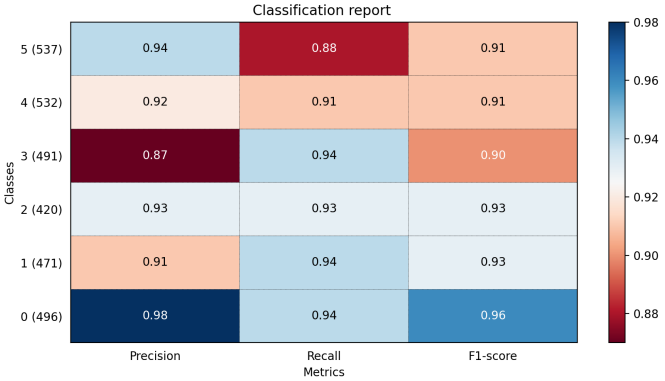

Fig. 6: Classification Report

VIII. Conclusion Future Direction

As we can see all the features except standing and sitting can be separated very easily. Stacking almost always boosts your accuracy as explained in case above, it does comes at the cost of extra training time. But overall within a short time (1-1.5 min) the smartphone has enough data to determine what its user is doing. In addition this insights have been extracted from only two smartphone sensors which probably could be accessed by most of our Apps. With enough data these models could be also used to classify the users and their activity patterns everyday. These could be used to asses a user's medical condition and contribute to maintain a healthy activity cycle.

References

[1] Rashidi, Parisa, and Diane J. Cook. "Keeping the Resident in the Loop: Adapting the Smart Home to the User." IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 39, no. 5, Sept. 2009, pp. 949–59. IEEE Xplore, https://doi.org/10.1109/TSMCA.2009.2025137.

[2] Avci, A.; Bosch, S.; Marin-Perianu, M.; Marin-Perianu, R.; Havinga, P. Activity Recognition Using Inertial Sensing for Healthcare, Wellbeing and Sports Applications: A Survey. In Proceedings of the 23rd International Conference on Architecture of Computing Systems (ARCS), Hannover, Germany, 22–23 Febuary 2010; pp. 1–10.

[3] Chavarriaga, R.; Sagha, H.; Calatroni, A.; Digumarti, S.; Millán, J.; Roggen, D.; Tröster, G. The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition. Pattern Recognit. Lett. 2013, 34, 2033–2042

[4] Lee, Honglak, et al. "Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations." Proceedings of the 26th Annual International Conference on Machine Learning, Association for Computing Machinery, 2009, pp. 609–16. ACM Digital Library, https://doi.org/10.1145/1553374.1553453.

[5] LeCun, Y.; Bengio, Y. Chapter Convolutional Networks for Images, Speech, and Time Series. In The Handbook of Brain Theory and Neural Networks; MIT Press: Cambridge, MA, USA, 1998; pp. 255–258.

[6] Sainath, T.; Vinyals, O.; Senior, A.; Sak, H. Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks. In Proceedings of the 40th International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, 19–24 April 2015; pp. 4580–4584.

[7] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in International conference on pervasive computing, pp. 1–17, Springer, 2004

[8] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman, "Activity recognition from accelerometer data," in Aaai, vol. 5, pp. 1541–1546, Pittsburgh, PA, 2005

[9] S.-W. Lee and K. Mase, "Activity and location recognition using wearable sensors," IEEE pervasive computing, vol. 1, no. 3, pp. 24–32, 2002.

[10] X. Yun, J. Calusdian, E. R. Bachmann, and R. B. McGhee, "Estimation of human foot motion during normal walking using inertial and magnetic sensor measurements," IEEE transactions on Instrumentation and Measurement, vol. 61, no. 7, pp. 2059–2072, 2012

[11] Bernardino Romera-Paredes, Hane Aung, and Nadia Bianchi-Berthouze. A one-vs-one classifier ensemble with majority voting for activity recognition. In European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), 2013

[12] Attila Reiss, Gustaf Hendeby, and Didier Stricker. A competitive approach for human activity recognition on smartphones. In European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), 2013.

[13] Yang, Jianbo, et al. "Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition." Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015. www.aaai.org, https://www.aaai.org/ocs/index.php/IJCAI/IJCAI15/paper/view/10710.

[14] Zhao, Yu, et al. "Deep Residual Bidir-LSTM for Human Activity Recognition Using Wearable Sensors." Mathematical Problems in Engineering, vol. 2018, Dec. 2018, p. e7316954. www.hindawi.com, https://doi.org/10.1155/2018/7316954.