

HHARNet: Taking inspiration from Inception and Dense Networks for Human Activity Recognition using Inertial Sensors

Hamza Ali Imran

Department of Computing
School of Electrical Engineering &
Computer Science,

National University of Sciences and Technology (NUST),
Islamabad, Pakistan
himran.mscs18seecs@seecs.edu.pk

Usama Latif

Operations Engineer
VAS, Apollo Telecom, Islamabad
usamalatif417@gmail.com

Abstract—Human Activity Recognition (HAR) is an important area of research in the light of enormous applications that it provides, such as health monitoring, sports, entertainment, efficient human-computer interface, child care, education, and many more. The use of Computer Vision for Human Activity Recognition has many limitations. The use of inertial sensors which include an accelerometer and gyroscopic sensors for HAR is becoming the norm these days considering their benefits over traditional Computer Vision techniques. In this paper, we have proposed a 1-dimensional Convolutions Neural Network which is inspired by two state-of-the-art architectures proposed for image classifications; namely Inception Net and Dense Net. We have evaluated its performance on two different publicly available datasets for HAR. Precision, Recall, F1-measure, and accuracies are reported.

Keywords—Human Activity Recognition (HAR), Activity Classification, Deep Neural Networks, Inertial Sensors based classification, Digital Signal Processing, Computer Vision Inspired 1D-CNN, Human Behavior Recognition, Convolutional Neural Network

I. INTRODUCTION

Human activity recognition has many applications which includes, but are not limited to, care for the elderly, child monitoring, education, entertainment, environmental support, sports, etc. Human Activity Recognition is part of the Body Area Network, which complements the Wireless Sensor Network [3]. Many drawbacks have been identified when traditional computer vision methods have been used for HAR [4] [5], which include: 1) confidentiality; 2) environmental impact; 3) higher operating costs; 4) occlusion; 5) reduced portability; and the list continues.

New advancements have recently been made in the field of recognition of human activity, which is the use of inertial sensors. This fast-paced growth of micro-electromechanical systems (MEMS) has paved the way for some major, applicable, and scientific breakthroughs in a wide range of research areas. Undoubtedly inertial sensors are one of the most significant MEMS sensors used collectively as Inertial Measurement Units (IMUs). Due to many positive features such as low cost, low power consumption, lightweight and portability, the use of inertial sensors has increased in all applications directly or indirectly connected to motion. By using advanced techniques, the data directly acquired for these sensors can be processed and used for complex motion analysis [6] [7].

The benefits of using sensor data exceed that of traditional computer-vision techniques. HAR uses many different types of sensors such as accelerometers, heart rate sensors, and gyroscopes, etc. The usage of data from these types of sensors covers all drawbacks of computer-vision techniques. The literature is composed of descriptive work of the applications inclusive of Machine Learning (ML) and Deep Neural Networks (DNN) Techniques used for HAR. An architecture working based on inertial sensor data has been proposed by us in this paper whose inspiration stems from computer-vision architectures similar to those proposed for image recognition.

SMARTPHONE Dataset [8], Wireless Sensor Data Mining (WISDM) Activity Prediction [9] are the two diverse datasets that were used for performance evaluation. SMARTPHONE dataset comprises of the readings from accelerometer sensor and gyroscopic sensor of a Smartphone while the rest of the datasets are only limited to that of accelerometers. We have used raw sensor values for our experiments regardless of the feature engineering which has been done by contributors of datasets. This makes our work completely a Deep Learning approach which removes the necessity of having domain knowledge. Complete details of these datasets can be found in Section III.

Our proposed architecture is composed of inception modules [10] which are a combination of one-dimensional convolutional layers. Furthermore, all the feature maps are also linked similarly to the DenseNet network [11]. The parameter of accuracy has widely been used as a measurement of performance for experiments similar to our work but the usage of unbalanced datasets nulls its characteristics. Therefore, we have reported F1 score, Precision, and Recall in conjunction with accuracy for these two datasets.

The presented paper is a part of the very initial work from our active research. Our contributions in this article are explained below

- We have presented a novel 1-dimensional convolutional neural network which is inspired by Inception and Dense Net Architectures.
- We have evaluated its performance on two different human activity recognition datasets.

Section II includes a literature review; Section III outlines the architecture proposed; Section IV discusses the architecture evaluation datasets used; Section V discusses the findings obtained; Section VI explains the training

process and also specifics software stack details; Section VII concludes the paper.

II. LITERATURE REVIEW

C. Xu, D et al. have created a Deep Neural Network they named InnoHAR [1]. It is based on the combination of Inception Neural Network and (Gated Recurrent Unit)) GRU layers. Experiments were conducted on 3 different datasets including OPPORTUNITY Dataset, PAMAP2 Dataset, SMARTPHONE Dataset. In the case of OPPORTUNITY Dataset, values from only on-body sensors which include inertial measurement units and 3-axis accelerometers were used. In the case of the PAMPA2 dataset accelerometer data was down-sampled to 33.3 Hz. The F1-measure achieved was 0.946, 0.935, and 0.945 for OPPORTUNITY, PAMAP2, and SMARTPHONE Datasets.

A One-dimensional convolution neural network for human activity recognition has been proposed [12]. Tri-axial accelerometer data (x,y,z) from a Smartphone was used and a magnitude vector was computed. The comparison was done with Random Forest Classifier. 92.71% accuracy was achieved. The dataset used had only 3 classes of Walk, Run, and Still.

In [13] feature engineering was done on inertial sensors data of Smartphone. Linear Discriminant Analysis (LDA) and Kernel Principal Component Analysis (KPCA) were used to get robust features. Deep Belief Neural Network (DBN) was then trained on those features. Results were compared with (Support Vector Machine) SVM and (Artificial Neural Network) ANN. A publicly available dataset was used. 95.85 % overall accuracy was achieved for proposed DBN.

Event detection through transformed accelerometer information was achieved by the researchers in [14]. They convert input signals straight into a reference coordinate system based on the rotation matrix (Euler Angle Conversion) derived from gyroscope orientation angles and orientation detectors. The benefit of the conversion is that it enables the identification of behavior and the detection; remaining impartial to orientation sensors. The number of classes was 5. The experiment was conducted with a phone in the hand of the person, pants pocket, or handbag. The results show that 84.77 percent of normal orientation impartial accuracy is achieved, which is 17.26 percent better than those without transformation of inputs

III. PROPOSED ARCHITECTURE

DenseNet [11] made use of features from all levels, resulting in features of all kinds ranging from small features to large features. The classification based on all of these features contributes to better model performance. InceptionNet Modules [10] uses various sizes of kernels. The advantage of this is that certain features can be best detected with a certain kernel size and making use of all possible options enhances the chances of detection. The combination of the structure of both modules should lead to an architecture that has both qualities and that should do better than existing architectures. We have named this module "InceptionDense". We have used three different sized kernels after 1x1 convolutions. The previous feature map outputs are also concatenated. For concatenation to take place, the spatial dimensions of all the feature maps (outputs after convolutions) are kept the same. Concatenation process

means to combine the feature maps together depth-wise. InceptionDense module is depicted in the figure (figure 1) given below.

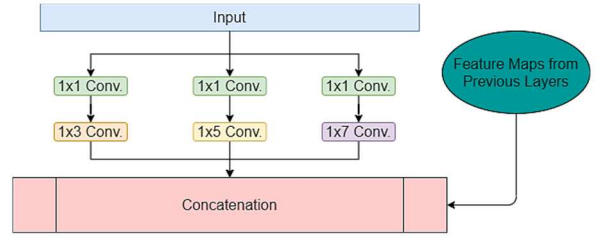


Fig. 1. InceptionDense - The Proposed Module

Figure 2 below shows the proposed architecture. We have used three InceptionDense modules of the type shown in figure 1. The number of modules is a hyperparameter which was selected empirically. The features channel from all the modules is being concatenated with the next module. There are no fully connected layers which reduce the number of parameters. We have made efficient use of 1x1 convolutions and reduced the number of filters to the number of classes which is 6 in the case of all the datasets we have used for the evaluation of our architecture. In the end, we have used Global Average Pooling before Softmax/Prediction. Global Average Pooling is means to calculate the average output of each feature map in the previous layer. Dropout of 0.50 was used after 1x1 layers before the InceptionDense Modules and before the last 1x1 layer which has 6 kernels.

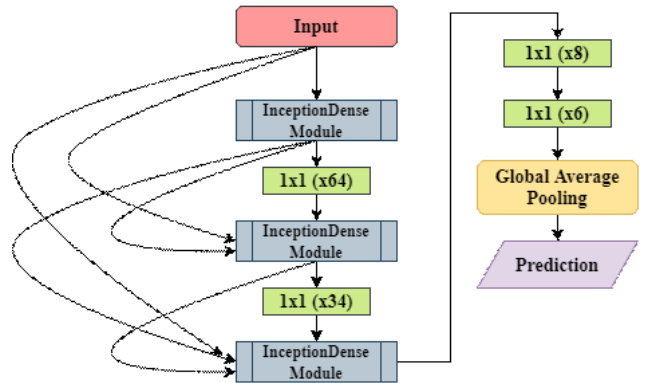


Fig. 2. Proposed Architecture

IV. DATASET USED

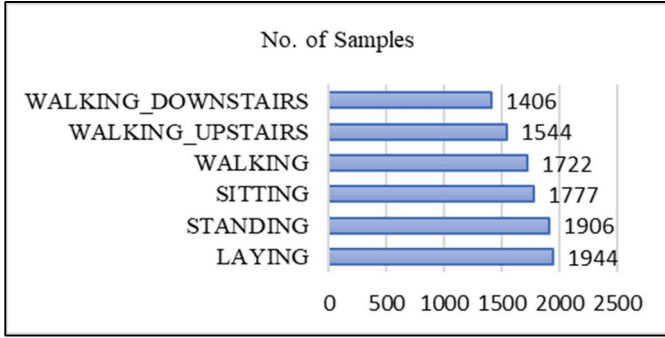
These three diverse datasets consisting of SMARTPHONE Dataset [8] and WISDM Activity Prediction [9] have been used to assess the performance of our architecture. Details regarding these datasets can be found below. Splitting of these datasets has been done as follows: 70% training and 15% validation and 15% testing.

A. SMARTPHONE Dataset

Dataset [8] consists of data taken from a gyroscopic sensor and an accelerometer of a Smartphone named Galaxy SII which was given to the volunteers to be worn on their waist. A total of 30 volunteers were involved in the data collection experiment; having ages between 19 to 48 and performing 6 activities whose class distribution can be seen in the following table (table 1). A feature vector having 561 features was already created by the team in both the time and frequency domain by the use of feature engineering. We have used raw values for sensors to train our Neural Network and have distributed the dataset into 70% training

and 15% validation and 15% testing resulting in a balanced dataset.

TABLE I. SMARTPHONES DATASET CLASS DISTRIBUTIONS

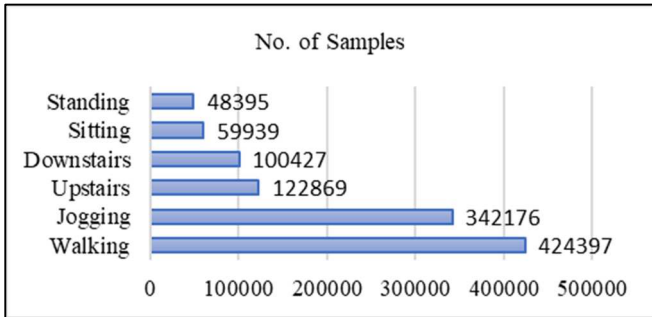


If compared to the other dataset, this dataset is fairly balanced.

B. WISDM Activity Prediction Dataset

This Dataset [9] consists of results obtained from controlled conditions in a laboratory and raw entries were 1,098,207. No missing values were found therefore requires no pre-processing. Class distribution can be seen in table 2.

TABLE II. WISDM ACTIVITY PREDICTION DATASET CLASS DISTRIBUTION



As seen from the above histogram, the dataset is highly unbalanced.

V. RESULTS

Accuracy has been largely used in the related work but the drawback with the majority of the inertial sensor collected dataset for HAR is its unbalance. By using these unbalanced datasets, Accuracy has become a less reliable parameter for measurement of performance. The same point has been stressed in [1] as well. Better parameters for measurement are F1 measure, Precision, and Recall. A confusion matrix for a test set of values for which ground truth is known is a table used commonly to visualize the performance of a classification model or classifier. A confusion matrix is a summary of the classification problem prediction results. The number of correct and wrong forecasts are summarized by count values and split by class. Figure 3 below shows the Confusion Matrix for SMARTPHONE Dataset.

A. SMARTPHONE Dataset Results

The accuracy achieved for this dataset was 89.44 % on test data. The confusion matrix is given in the figure (figure 3) Detailed Performance Report is shown in figure 4. It can be observed from the Confusion matrix that the neural network struggles the most in differentiating between “Sitting” and “Standing” classes with a misclassification count of 182. The best F1-measure is for class “WALKING” which is 0.99 and worse is for “STANDING” which is 0.74.

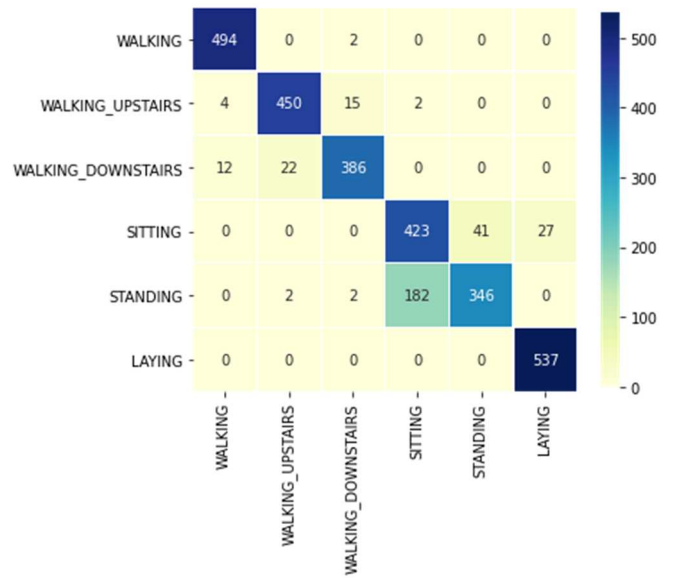


Fig. 3. Confusion Matrix SMARTPHONE Dataset

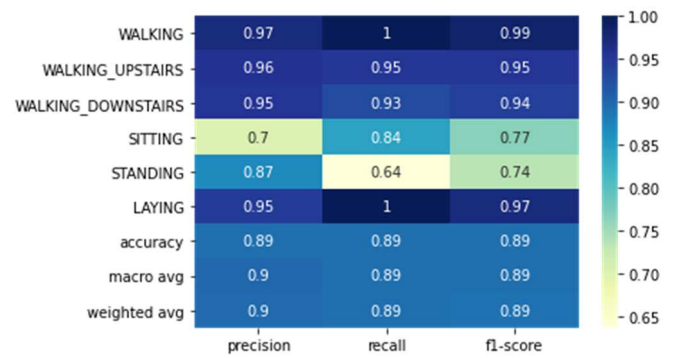


Fig. 4. Performance Report SMARTPHONE Dataset

B. WISDM Activity Prediction Dataset

The accuracy achieved for this dataset was 95.26 % on test data. The confusion matrix is given in the figure below (figure 5). Detailed Performance Report is shown in figure 6. The best F1-measure is for class “Walking” which is 0.99 and the worse is for class “Upstairs” which is 0.85.



Fig. 5. Confusion Matrix WISDM Activity Prediction Dataset

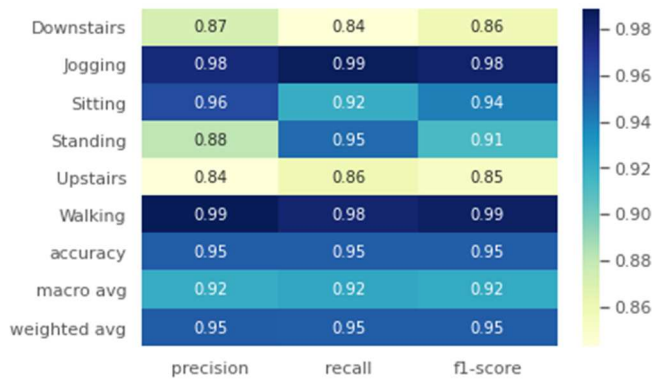


Fig. 6. Performance Report WISDM Activity Prediction Dataset

VI. IMPLEMENTATION DETAILS

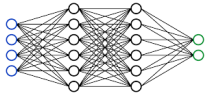
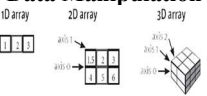
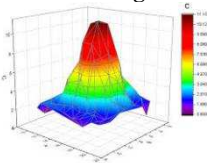
A. Training Process

"Adam" and "RMSprop" optimizers were used in our experiments which gave us the result that "RMSprop" is preferred for SMARTPHONE dataset but "Adam" performed better for other dataset.

B. Software Stack

The entire experiment was performed in Google's Collaboratory, GPU runtime environment. The list of major packages used along with their version numbers is shown in the table (table 3 below)

TABLE III. SOFTWARE STACK USED

Platform	Google Colaboratory	
Category	Name	Version
Deep and Machine Learning Libraries 	Tensorflow	2.3.0
	Keras	2.4.3
	Sklearn	0.22.2.post1
Data Manipulation 	Numpy	1.18.5
	Pandas	1.0.5
Plotting 	Matplotlib	3.2.2
	Seaborn	0.10.1

VII. CONCLUSION

Recognition of Human Behaviors is an active area of study with applications in different areas. We have proposed a novel 1D-CNN which is inspired by InceptionNet [10] and DenseNet [11] and named it HHARNet. The architecture performance has been evaluated on SMARTPHONE Dataset [8] and WISDM Activity Prediction Dataset [9]. The accuracy of 89.44%

and 95.26% has been achieved respectively. Moreover, F1-measure, Recall, and Precision are also reported. The best F1-measure is for class "WALKING" which is 0.99 and worse is for "STANDING" which is 0.74 in case of SMARTPHONE Dataset. The best F1-measure is for class "Walking" which is 0.99 and the worse is for class "Upstairs" which is 0.85 for WISDM Dataset.

REFERENCES

- [1] C. Xu, D. Chai, J. He, X. Zhang and S. Duan, "InnoHAR: A Deep Neural Network for Complex Human Activity Recognition", IEEE Access, vol. 7, pp. 9893-9902, 2019.
- [2] H. Nweke, Y. Teh, M. Al-garadi and U. Alo, "Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges," Expert Systems with Applications, vol. 105, pp. 233-261, 2018.
- [3] R. Gravina, P. Alinia, H. Ghasemzadeh and G. Fortino, "Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges", 2019.
- [4] H. Qazi, U. Jahangir, B. Yousuf and A. Noor, "Human action recognition using SIFT and HOG method", 2017 International Conference on Information and Communication Technologies (ICICT), 2017.
- [5] S. Ke, H. Thuc, Y. Lee, J. Hwang, J. Yoo and K. Choi, "A Review on Video-Based Human Activity Recognition", Computers, vol. 2, no. 2, pp. 88-131, 2013.
- [6] Song-Mi Lee, Sang Min Yoon and Heeryon Cho, "Human activity recognition from accelerometer data using Convolutional Neural Network", 2017 IEEE International Conference on Big Data and Smart Computing (BigComp), 2017.
- [7] M. Hassan, M. Uddin, A. Mohamed and A. Almogren, "A robust human activity recognition system using smartphone sensors and deep learning", 2018.
- [8] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. A Public Domain Dataset for Human Activity Recognition Using Smartphones. 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2013. Bruges, Belgium 24-26 April 2013.
- [9] J. Lockhart, G. Weiss, J. Xue, S. Gallagher, A. Grosner and T. Pulickal, "Design considerations for the WISDM smart phone-based sensor mining architecture", Proceedings of the Fifth International Workshop on Knowledge Discovery from Sensor Data - SensorKDD '11, 2011. Available: 10.1145/2003653.2003656
- [10] C. Szegedy et al., "Going deeper with convolutions," in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1-9.
- [11] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
- [12] Song-Mi Lee, Sang Min Yoon and Heeryon Cho, "Human activity recognition from accelerometer data using Convolutional Neural Network", 2017 IEEE International Conference on Big Data and Smart Computing (BigComp), 2017.
- [13] M. Hassan, M. Uddin, A. Mohamed and A. Almogren, "A robust human activity recognition system using smartphone sensors and deep learning", Future Generation Computer Systems, vol. 81, pp. 307-313, 2018.
- [14] Heng, X., Wang, Z., & Wang, J. (2014). Human activity recognition based on transformed accelerometer data from a mobile phone. International Journal of Communication Systems, 29(13), 1981-1991. doi:10.1002/dac.2888