RestHAR: Residual feature learning transformer for human activity recognition from multi-sensor data

Debaditya Shome
School of Electronics Engineering
KIIT University
Bhubaneshwar, Odisha, India 751024
Email: 1804372@kiit.ac.in

***Abstract*—In recent times, surge in the use of smartphones in our daily lives has created a huge opportunity for paving the road towards human-centric computing by utilizing the rich data which gets recorded by it's multiple sensors. Sensor-based human activity recognition has a tremendous amount of real-world applications such as health monitoring, surveillance, smart homes, and ambient assisted living. This paper presents a joint residual feature extractor and a transformer-based deep neural network for end-to-end human activity recognition using raw multi-sensor data captured from smartphones or wearable devices. Unlike conventional handcrafted feature extraction, this approach outperforms all present approaches showing state-of-the-art generalizable performance over multiple benchmark datasets. It achieves a test accuracy of 95.2% on the UCI HAR dataset and 96.4% test accuracy on the WISDM dataset.**

***Index Terms*—Deep learning, Activity Recognition, Transformers, Smartphones, Time-series**

I. Introduction

Smartphones and smart-wearables have significantly transformed human lives and taken over the world by a storm. With the advancements in their electronic designs, huge number of sensors are being embedded within them such as accelerometers, magnetometers, barometers, gyroscopes, proximity sensors, pressure, bluetooth, temperature and light sensors. Activity detection has emerged as a recent wave of context-aware customised applications in a variety of domains. All vertical industries can utilize the huge amount of data recorded in the sensors for applications such as monitoring patient's health, businesses tracking customers, tracking fitness progress, monitoring soldiers in military settings and many more. Due to this, human activity recognition (HAR) using multiple sensors has become an active area of research. Viewing from an algorithmic perspective, HAR involves mapping the raw windowed time-series data captured by these sensors as input to HAR algorithm which detects the smartphone user's activity and returns as output in real-time. Handcrafting such an algorithm is challenging and a time consuming procedure requiring a lot of domain expertise which makes it necessary to analyze every sensor in tiny details. Due to such challenges, Machine learning (ML) has been used widely in literature [1] for human activity recognition from sensors for it's capability to directly learn a mapping function between input and output. In [2], A combination of Support Vector Machine (SVM) and K-nearest neighbours (KNN) algorithm was used for

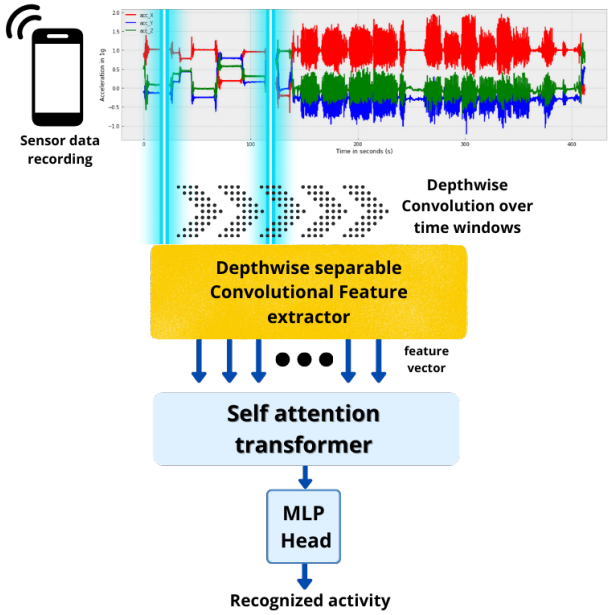

Fig. 1. Flow of the proposed system

sensor based HAR. In [3], the authors extracted statistical signal properties such as Mel-frequency cepstral coefficients as features and designed an SVM model combined with Particle swarm optimization method for HAR on motion sense dataset. In [4], the authors compared Random forest algorithm with other shallow learning models such as SVM, KNN, Naive Bayes, AdaBoost etc and showed that Random forest achieved higher accuracy on three sensor based HAR datasets compared to the other models. Although having higher number of sensors provides more data, but the above discussed ML based activity recognition algorithms may face challenges in terms of complexity to deal with such high dimensional data available. Neural networks are known for their ability to learn non-linear high dimensional mappings between raw inputs and outputs due to which a lot of work has been done on deep learning (DL) based HAR. In [5], multiple deep learning architectures were evaluated on the UCI HAR dataset which showed that a one-dimensional Convolutional neural network along with statistical features outperformed all approaches but upon evaluating the same model on a similar dataset called m-health,

the performance degraded significantly which shows that hand crafted statistical features make the model less generalizable in real world sensor data which would have stochastic patterns and each device has different types of sensors for which the same feature extraction methods won't work. Moreover, these handcrafted feature extraction pipelines require a lot of domain expertise and only shallow features can be learnt from these approaches. Due to these reasons, most the present state of the art methods for human activity recognition are limited in terms of model generalization and real-world performance. Convolutional neural networks (CNNs) and Recurrent neural networks (RNNs) are the most used deep learning architectures for the HAR problem. It was mentioned in a study on HAR [6] that RNNs are better at recognizing short activities which have an order while CNNs are better for recognizing long term activity patterns. Zeng et al. [7] presented a standard CNN for classifying accelerometer data by separating data of each axis of the sensor into separate channels. In [8], the authors developed a weight sharing CNN approach for HAR for dealing with multi-modal data. In [9], recurrent architectures such as RNNs and LSTMs were explored for HAR. These recurrent models learn to map each time window of the data to an activity class where each timestep of that window is read sequentially one at a time and all the output scores of each timestep are aggregated together to get the activity label per window. A combination of CNN and LSTM was developed in [10] where the authors claim that the CNN layers are responsible for feature extraction from raw data and the LSTMs capture the temporal patterns from the extracted features. After reviewing the literature above, it can be seen that there is a need for end-to-end models with low complexity which can capture patterns from the raw sensor readings and can handle the unpredictable changes in the data in real-time deployed settings.

Motivated from the above reasons, this paper presents RestHAR, a robust self-attention based transformer neural network with only 0.4 million parameters for human activity recognition which shows state of the art performance on the UCI-HAR and WISDM dataset benchmark. To the best of the author's knowledge, this is the first work on transformer based time-series HAR from raw data of smartphone sensors.

II. Data preparation

The input data collected from the m sensors each with n axes record a 2D time-series sensor data. The first dimension of the data represents the t timesteps per window and the second dimension represents all the $m \times k$ sensor readings as channels. This data is converted into batches of size b. Thus the final training data is a 3D array of shape $(b, t, m \times k)$. For evaluation of our model, two benchmark datasets have been used which are discussed in this section.

A. UCI HAR dataset

The raw version of the UCI HAR dataset [11] is used in this work which consists of triaixal inertial signal data from accelerometer and gyroscope sensors. For the data acquisition phase of this dataset, 30 subjects with a waist-mounted smartphone performed 6 types of activities i.e, walking, walking downstairs, walking upstairs, laying down and standing. These experiments were recorded by a video camera for efficient labelling by multiple annotators. The dataset was further split into training split with 70% of the total data and testing split with 30% of the data for robust evaluation. The accelerometer and gyroscope readings were sampled at 50 Hz for obtaining time windows of 2.56 seconds each with 128 sensor readings per window (50% overlap).

B. WISDM dataset

The WISDM Smartphone and Smartwatch Activity and Biometrics Dataset [12] contains sensor recordings from 51 people who completed 18 tasks in three minutes each. Each participant wore a smartwatch on his or her dominant hand and carried a smartphone in their pocket. A custom Android app that operated on the smartphone and wristwatch was in charge of the data collection. The data was sampled at 20 Hz and 10 second windows were extracted, each window having 200 readings from the triaxial accelerometer sensor. The dataset was split in a ratio of 80% : 20% training and test data respectively.

III. Proposed architecture

A. Feature extraction network

It has been observed in literature than CNNs are powerful feature extractors in case of images. Recently, they have also shown great performance for extracting features from raw signals as well as time-series data and very deep CNN architectures have been experimented. Although increasing layers and making the network deeper is a easy way to increase accuracy, but it also has a limit and the later layers learn very less useful patterns thus resulting in decreased performance as shown in the ResNet paper [13]. Inspired from ResNet, RestHAR adapts a residual feature extraction network with depthwise separable convolutions as illustrated in figure 2. The main components of the feature extractor are discussed below.

1) **Depthwise separable 1D convolutions:** Unlike standard convolutions, depthwise separable convolutions are a two step process starting with a filtering stage where depthwise convolution operation is applied on each of the channels, which in our case are the $m \times k$ time series readings from the sensors. In the next stage, a pointwise convolutional operation takes place on the intermediate output from the first step. The advantages for using depthwise separable convolutions is the reduction in computational complexity as well as the fact that different sensor readings would yield different types of patterns, learning each of them separately boosts the quality of the features extracted by the network.

2) **Residual blocks:** The core of the feature extraction network is the stack of R Residual blocks connected

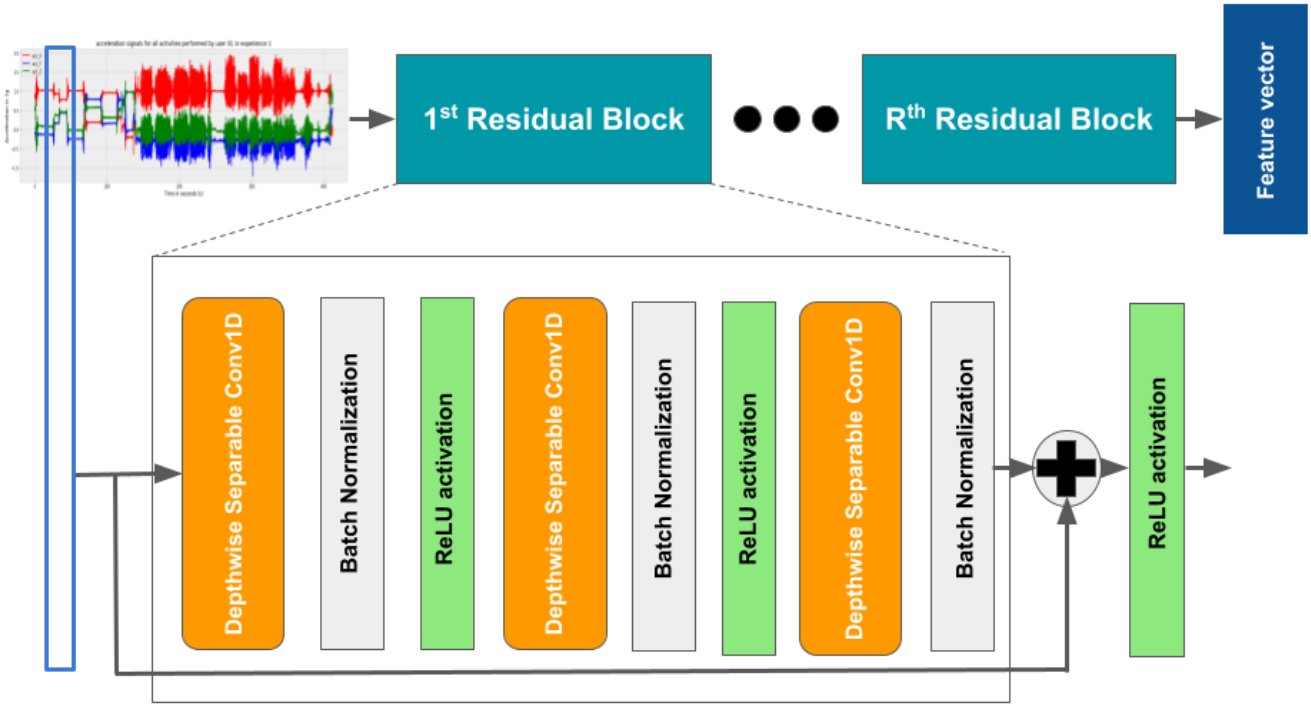

Fig. 2. Depthwise separable residual feature extractor network

together with skip connections over each of these blocks. As we don't know the ideal number of layers (or residual blocks) for a neural network, which may vary depending on the dataset's complexity, generalizability of the architecture gets limited in traditional CNNs. Adding skip connections to our network enables the network to learn negligible weights for the layers that are not relevant and do not contribute value to overall accuracy, rather than using the number of layers as an essential hyperparameter to tune. Each block consists of three depthwise separable 1D convolutional layers with Batch Normalization layers in between. The first two convolutional layers are ReLU activated and the last layer gets concatenated with the skip connection from the previous residual block and the resulting output gets ReLU activated and passed further into the next residual block.

B. Transformer based classifier

The transformer model proposed in this paper has an architecture as shown in figure 3. The dynamic features extracted at each window from the sensors are passed into a block of stacked n transformer encoders. The underlying design of each transformer encoder is shown in figure 4. The feature vector first passes through a layer normalization block followed by a Multi-Head attention block with h parallel attention heads. An MLP head is added to each of these transformer encoder which consists of f Feed-forward layers and f can be changed as an hyperparameter. Two skip connections are made from the input stage and after multi-head attention module to ensure that richer patterns are captured also at the later encoders rather than only the first few encoders. Output of each transformer encoder serves as the feature vector for the next encoder. Final output of the n_{th} encoder goes into a classfier made up of feed forward layers which classifies that particular time window into the recognized activity.

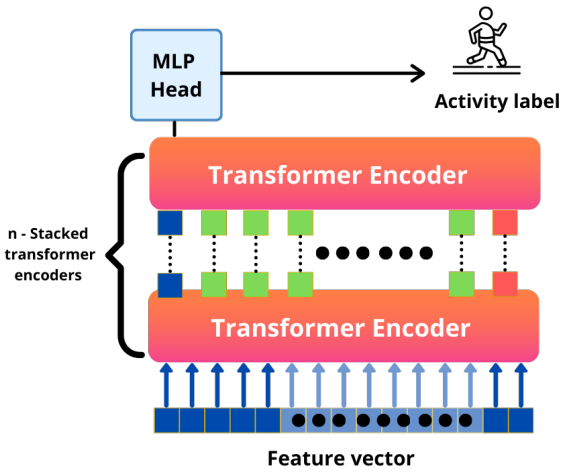

Fig. 3. RestHAR architecture

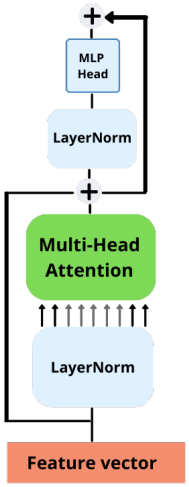

Fig. 4. Transformer encoder

IV. Training Mechanism

The feature extraction network and the transformer networks are connected together with shared weights. The models are trained together and jointly optimized using the following configuration.

Loss function: The loss function L used in this paper is the Binary cross entropy loss function. As seen in equation 1 We compute the loss for each class label per observation which indicates and add the results. Here, M represents the total number of activity classes which equals to six.

$$L = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{1}$$

Optimizer: For the optimization of the loss function discussed above, the Adam optimizer [14] is used, which is used widely for it's fast and efficient performance with very low memory requirements.

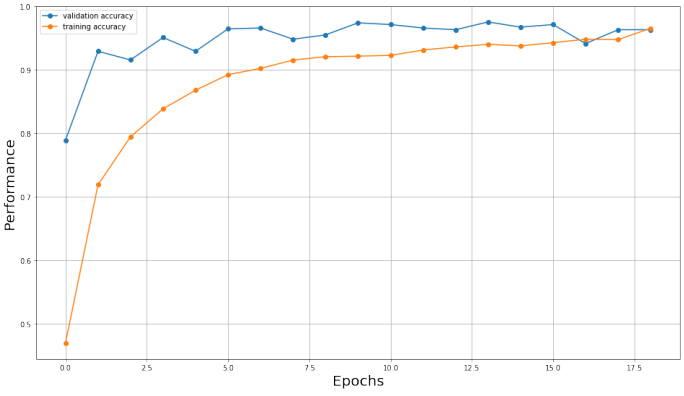

Fig. 5. Training and validation accuracy on UCI HAR dataset

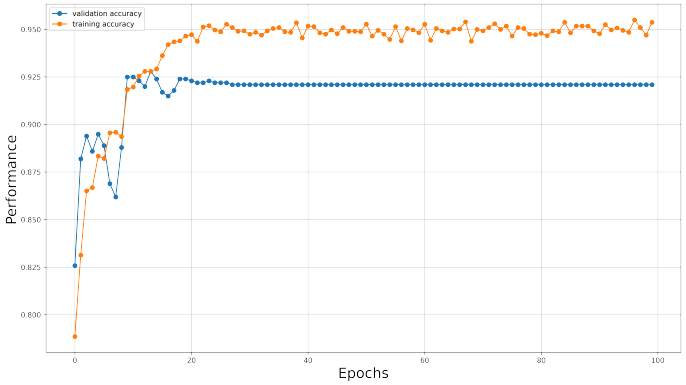

Fig. 6. Training and validation accuracy on WISDM dataset

V. Results and Evaluation

In this section, we use the two benchmark datasets discussed in section II to evaluate the performance of our model. As seen in figure 5 and figure 6, the model learns well with zero overfitting on both datasets almost at same performance. Although the time windows chosen were 2.56 seconds on UCI HAR and 10 seconds on WISDM dataset, the model's performance doesn't degrade, thus showing the capability of the model to learn global dependencies such as CNNs and local dependencies such as RNNs. The confusion matrices in figure 7 and 8 figure illustrate the class-wise performance of the model. It can be clearly seen that the model shows very less misclassification even between those types of activities which are hard to distinguish, such as 8% and 10.5% misclassification rate in recognizing walking upstairs and downstairs as seen in figure 7 and 8 respectively.

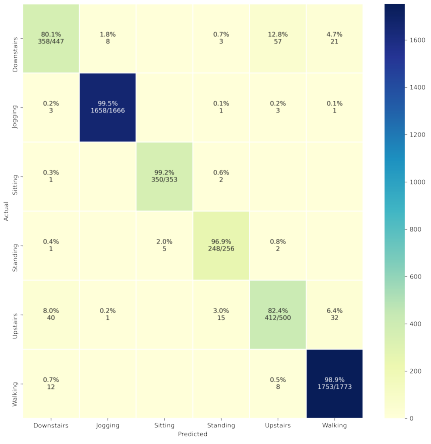

Fig. 7. Confusion matrix on WISDM dataset

VI. Conclusion

In this paper, a novel end to end deep learning model for human activity recognition is proposed which shows state of

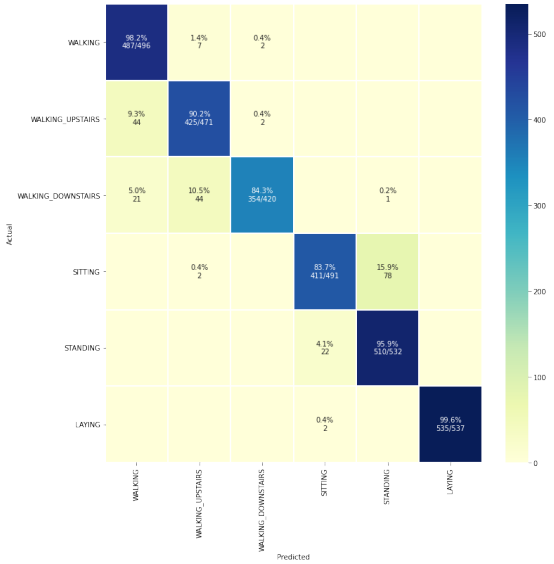

Fig. 8. Confusion matrix on UCI HAR dataset

the art performance in terms of generalizability and accuracy on two benchmark datasets (UCI HAR and WISDM). It is observed that upon both datasets that a few minimal configuration of our architecture with lesser number of blocks is capable to get such high accuracies, which shows that the architecture is deployable in low powered edge devices and wearables which can handle models with lesser parameters.

REFERENCES

[1] S. Ramasamy Ramamurthy and N. Roy, "Recent trends in machine learning for human activity recognition—a survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1254, 2018.

[2] I. Menhour, M. Abidine, and B. Fergani, "A new activity classification method k-svm using smartphone data," in *2019 International Conference on Advanced Electrical Engineering (ICAEE)*, pp. 1–5, IEEE, 2019.

[3] M. Batool, A. Jalal, and K. Kim, "Sensors technologies for human activity analysis based on svm optimized by pso algorithm," in *2019 International Conference on Applied and Engineering Mathematics (ICAEM)*, pp. 145–150, IEEE, 2019.

[4] A. Wang, H. Chen, C. Zheng, L. Zhao, J. Liu, and L. Wang, "Evaluation of random forest for complex human activity recognition using wearable sensors," in *2020 International Conference on Networking and Network Applications (NaNA)*, pp. 310–315, IEEE, 2020.

[5] H. Nematallah and S. Rajan, "Comparative study of time series-based human activity recognition using convolutional neural networks," in *2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pp. 1–6, IEEE, 2020.

[6] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*, vol. 119, pp. 3–11, 2019.

[7] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu, and J. Zhang, "Convolutional neural networks for human activity recognition using mobile sensors," in *6th international conference on mobile computing, applications and services*, pp. 197–205, IEEE, 2014.

[8] S. Ha and S. Choi, "Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors," in *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 381–388, 2016.

[9] A. Murad and J.-Y. Pyun, "Deep recurrent neural networks for human activity recognition," *Sensors*, vol. 17, no. 11, p. 2556, 2017.

[10] F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[11] D. Anguita, A. Ghio, L. Oneto, X. Parra, J. L. Reyes-Ortiz, *et al.*, "A public domain dataset for human activity recognition using smartphones.," in *Esann*, vol. 3, p. 3, 2013.

[12] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SigKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74–82, 2011.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.