IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 15, 2021, accepted March 27, 2021, date of publication April 2, 2021, date of current version April 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3070646

Boosting Inertial-Based Human Activity Recognition With Transformers

YOLI SHAVIT[1] AND ITZIK KLEIN[2], (Senior Member, IEEE)
[1]Faculty of Engineering, Bar-Ilan University, Ramat Gan 5290002, Israel
[2]Department of Marine Technologies, University of Haifa, Haifa 3498838, Israel

Corresponding author: Yoli Shavit (yolisha@gmail.com)

ABSTRACT Activity recognition problems such as human activity recognition and smartphone location recognition can improve the accuracy of different navigation or healthcare tasks, which rely solely on inertial sensors. Current learning-based approaches for activity recognition from inertial data employ convolutional neural networks or long short term memory architectures. Recently, Transformers were shown to outperform these architectures for sequence analysis tasks. This work presents an activity recognition model based on Transformers which offers an improved and general framework for learning activity recognition tasks. For evaluation purposes, several datasets, with more than 27 hours of inertial data recordings collected by 91 users, are employed. Those datasets represent different user activity scenarios with varying difficulty. The proposed approach consistently achieves better accuracy and generalizes better across all examined datasets and scenarios. A codebase implementing the described framework is available at: https://github.com/yolish/har-with-imu-transformer.

INDEX TERMS Human activity recognition, smartphone location recognition, inertial sensors, pedestrian dead reckoning, convolutional neural networks, Transformers, sequence analysis.

I. INTRODUCTION

Human activity recognition (HAR) and smartphone location recognition (SLR) aim to identify the user activity from sensory data. HAR measurements can be collected using video [1], utilizing channel state information of WiFi signals [2], [3], radar [4] or by sensors installed in wearable devices such as inertial sensors (accelerometers and gyroscopes) or ambient environment sensors (temperature and humidity). HAR has numerous applications, relying on one or more of these sensors, including surveillance [5], gesture recognition [6], [7], gait analysis [8], healthcare [9], [10], and indoor navigation [11], [12]. Due to its wide applicability it has been addressed and surveyed extensively in the literature [13]–[20]. In SLR, the user's actions are reflected through changes in the location of the smartphone. For example, consider a walking pedestrian where the smartphone is placed in their trouser's front right pocket (pocket mode). The pedestrian can remove the phone to send a message (texting mode) and then continue holding the phone while walking (swing mode).

The associate editor coordinating the review of this manuscript and approving it for publication was Larbi Boubchir.

HAR and SLR play a particularly important role in navigation solutions which rely solely on the smartphone inertial sensors. Specifically, HAR and SLR were shown to improve the accuracy of traditional pedestrian dead reckoning (PDR) by using it as a prior [21]–[25]. SLR was also shown to improve the performance of other navigation-related problems such as step length estimation [26]–[28] and adaptive attitude and heading reference system (AHRS) [29].

Given the emerging importance of HAR and SLR for navigation performance, different learning-based approaches were proposed to reason about inertial sensory data. Earlier methods for performing HAR for PDR relied on classical machine learning techniques [22]. Hand-crafted features were typically extracted from the raw signal and fitted with a classical machine learning classification methods such as Support Vector Machine and Decision Trees [21], [23]. More recently, feed forward networks (FFN) and long short term memory (LSTM) architectures were proposed for this task [24], [25], removing the burden of feature engineering while achieving improved accuracy. Recent detailed and extensive surveys describing traditional and deep learning techniques for HAR are available for the interested reader [13], [18]–[20]. There, various types of deep learning

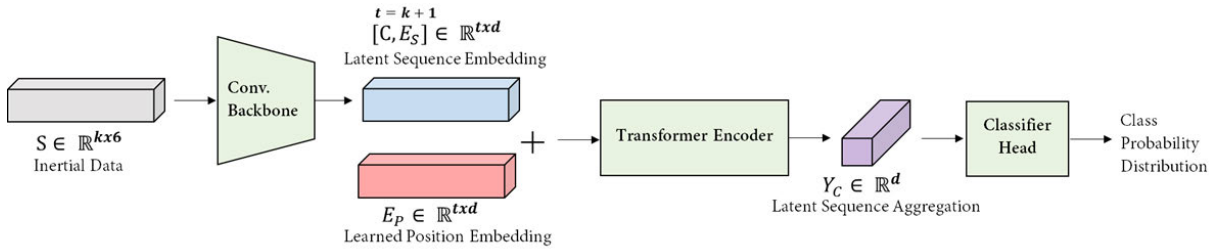

FIGURE 1. The proposed framework for inertial data classification with Transformers.

approaches such as convolutional neural networks (CNNs), recurrent neural networks, stacked autoencoders, temporal convolutional networks, and variational autoencoders are reviewed.

Recently, CNN and LSTM architectures were shown to improve SLR performance, compared to other learning-based approaches [30]. Methods coupling SLR with step length estimation proposed to use CNNs with or without LSTM [28], [30] or employed LSTM for SLR, similarly to previous works learning SLR for PDR [26]. Interestingly, CNN architectures with/without LSTMs yielded on-par performance, suggesting that LSTMs do not necessarily add an informative temporal aggregation, which is missing from CNNs, for this task [30].

In this work, Transformers [31] are proposed for learning inertial-based HAR and SLR problems. Transformers implement an attention-based encoder-decoder architecture for sequence analysis. Attention mechanisms [32] learn to aggregate information from the entire sequence. By stacking attentional layers which scan the sequence, Transformers generate a position and context aware representations. This method was shown to outperform recurrent neural networks (RNNs) and LSTMs for various sequence-based problems in Natural Language Understanding and Computer Vision, achieving state-of-the-art performance [31], [33]–[36]. Here a Transformer-based architecture is presented for performing both HAR (classifying common user dynamics, such as walking, standing, running, stairs and so on) and SLR (classifying smartphone locations, such as talking, pocket or swing). The proposed approach is the first Transformer-based architecture to serve as a general framework for inertial based activity recognition tasks.

In order to evaluate the proposed approach, multiple HAR and SLR datasets collected by a total of 91 users with more than 27 hours of inertial data recordings are employed. Across all datasets, and considering various scenarios with changing difficulty, the proposed approach demonstrates a consistent boost to accuracy and robustness.

In summary, the main contributions of the paper are as follows:

1) Derivation of a framework for inertial data classification with Transformers. The proposed approach is the first to present a Transformer-based architecture to serve as a general framework for general activity recognition in both HAR and SLR tasks.
2) A Transformer network architecture design for handling inertial measurements, along with publicly available implementation.
3) Evaluation of the proposed framework for three commonly used classifications tasks: HAR, SLR and a combination of the two smartphone and human activity recognition (SHAR), demonstrating a consistent improvement in accuracy and a better generalization across datasets.

The rest of the paper is organized as follows: Section II describes the proposed Transformer-based architecture for classification with inertial data. Section III reviews the datasets used in this work while Section IV presents the results. Finally, Section V gives the conclusions of this research.

II. INERTIAL DATA CLASSIFICATION WITH TRANSFORMERS

An Inertial Measurement Unit (IMU) measures the specific force $\mathbf{f} \in \mathbb{R}^3$ and angular velocity $\mathbf{w} \in \mathbb{R}^3$ vectors over time. These two outputs are typically concatenated and aggregated depending on the sensor's recording frequency, such that a sample $\mathbf{S} \in \mathbb{R}^{k\times 6}$ represents a sequence of k measurements (i.e., recorded in a window of size k). In this work, the problem of activity recognition from inertial data is modelled as a sequence-to-one problem, where the input is a learned sequential embedding of the raw sensory measurements and their temporal positions. Following the success of Transformers in text classification [33], [37] and image recognition [36], [38], a Transformer Encoder is proposed for summarizing a sequence of (embedded) inertial measurements into a latent vector. A Multi-Layer Perceptron (MLP) with SoftMax can then be applied to output the class probability distribution, similarly to standard classifier heads used in sequence-to-one architectures [33], [36].

A. NETWORK ARCHITECTURE

The scheme proposed in this paper is depicted in Figure 1. Given a sample of inertial measurements $\mathbf{S} \in \mathbb{R}^{k\times 6}$, a series of four 1D convolutions is applied with GELU non-linearity. This step embeds the raw data in a higher dimension d, generating a latent embedding $\mathbf{E_S} \in \mathbb{R}^{k\times d}$ (latent features).

Similarly to state-of-the-art Transformer-based architectures for sequence classification [33], [36], *class token*

$\mathbf{C} \in \mathbb{R}^d$ is prepended to the embedded sequence. In addition, an embedding $\mathbf{E_{P_i}} \in \mathbb{R}^d$ for each position P_i in the sequence (including the class token) is learned and further added to the latent sequence representation. The initial input Z_0 to the Transformer Encoder is thus given by:

$$\mathbf{Z_0} = [\mathbf{C}, \mathbf{E_s}] + \mathbf{E_P} \in \mathbb{R}^{t \times d} \tag{1}$$

with $t = k + 1$.

A standard Encoder architecture [31] is employed, stacking L layers, each consisting of a self multi-head attention (MHA) layer and an MLP. In the proposed implementation, the MLP block includes two fully connected (FC) layers with a hidden dimension of $2 \cdot d$ and GELU non-linearity. The MHA operation is the core of the Transformer architecture. Given three sequences of length t and dimension d, namely, a query $\mathbf{Q} \in \mathbb{R}^{t \times d}$, a key $\mathbf{K} \in \mathbb{R}^{t \times d}$ and a value $\mathbf{V} \in \mathbb{R}^{t \times d}$, each head h computes a weighted aggregation of $\mathbf{V}$ with respect to Q:

$$\mathbf{h}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q_h}(\mathbf{K_h^T})}{\sqrt{d}})\mathbf{V_h} \in \mathbb{R}^{t \times d'} \tag{2}$$

with:

$$\mathbf{Q_h} = \mathbf{Q}\mathbf{W_h^Q} \in \mathbb{R}^{t \times d'} \tag{3}$$

$$\mathbf{K_h} = \mathbf{K}\mathbf{W_h^K} \in \mathbb{R}^{t \times d'} \tag{4}$$

$$\mathbf{V_h} = \mathbf{V}\mathbf{W_h^V} \in \mathbb{R}^{t \times d'} \tag{5}$$

where $d' = \frac{d}{n_h}$ and n_h is the number of heads. The matrices $\mathbf{W_h^Q}, \mathbf{W_h^K}, \mathbf{W_h^V} \in \mathbb{R}^{d \times d'}$ are linear projections from d to d'. The outputs of (2) from all the heads are concatenated across the channel dimension. The resulting updated representation is a weighted aggregation of the sequence in each position, based on the relative importance of all other positions. In self MHA (sMHA), $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ are taken to be the same sequence.

Each layer $l, l = 1..L$ in the Transformer Encoder performs the following computation by passing the input through a LayerNorm (LN) [39] before each module and adding it back with residual connections:

$$\mathbf{Z_l'} = sMHA(LN(\mathbf{Z_{l-1}})) + \mathbf{Z_{l-1}} \in \mathbb{R}^{t \times d} \tag{6}$$

$$\mathbf{Z_l} = MLP(LN(\mathbf{Z_l'})) + \mathbf{Z_l'} \in \mathbb{R}^{t \times d} \tag{7}$$

The output of the Transformer Encoder at the position of the class token represents a temporally aware aggregation of input sequence:

$$\mathbf{Y_C} = \mathbf{Z_L}[\mathbf{0}] \in \mathbb{R}^d \tag{8}$$

Y_C is provided as an input for a classifier head, consisting of LN and FC layers with GELU non-linearity and Dropout, reducing the dimension to $\frac{d}{4}$. A second FC layer maps $\frac{d}{4}$ to the number of classes. A Log SoftMax is applied on the output vector in conjunction with Negative Log Likelihood (NLL) loss to learn a multi-label classification task.

B. IMPLEMENTATION DETAILS

The proposed architecture is implemented in PyTorch [40]. The latent dimension d to 64 for the convolutional backbone and positional embedding. The Transformer Encoder consists of six layers with an eight-heads-sMHA block and an MLP. Finally, a Dropout $p = 0.1$ is used for both the Transformer Encoder and classifier head. The implementation of this framework (proposed architecture and its training and testing) is publicly available at: https://github.com/yolish/har-with-imu-transformer.

III. DATASETS

Three datasets are employed in order to evaluate the proposed approach. The first dataset represents the SLR problem (SLR dataset), containing five different common smartphone locations. This dataset was created by combining six different SLR datasets. The second dataset considers the HAR problem (HAR dataset). It consists of six different human dynamics, including the division of the stairs class into two separate classes: walking upstairs (Stairs up) and walking downstairs (Stairs down). The third dataset contains data with a combined SLR and HAR class labeling, which is referred to as SHAR (SHAR dataset). For example, the class ''Walking Pocket'' refers to a scenario of a human walking (HAR) with the smartphone placed in their pocket (SLR). This dataset includes 21 classes.

In total, the SLR, HAR and SHAR datasets contain 27.76 hours of recordings made by 91 people. Each dataset contains many different files. Each file has the name of the user, which made the recording and a description of its type (e.g. user1 walking texting), and can have a different time duration. When creating the unified dataset, all files from all users were marge into a single file. The specific class labels and data properties for each dataset are summarized in Table 1.

A. SLR

The SLR dataset consists of six different datasets. In all six datasets, the smartphone location was at least in one of five locations: Texting, Pocket, Swing, Talking and, Body, while the user was walking. In most of the datasets, there was no limitation on how the smartphone was held or on the walking characteristics. From this dataset, only the normalized accelerometer readings are used.

The first dataset [30], contains 164min of recorded data in four smartphone locations: Texting, Pocket, Swing and, Talking. It has three different sampling rates: 25, 50, and 100Hz, all recorded by a single user with a single smartphone. The second datatset [41], was created for PDR applications, not related to SLR, using eight people. Since the recordings were made while the users were walking with a smartphone, this dataset can also be used for SLR. It contains three smartphone locations: Pocket, Texting and, Body, with a total of approximately 70min of data, recorded at 200Hz. Similarly, the third dataset [26] was recorded to examine deep-learning PDR, but can also be used for the SLR problem. There,

TABLE 1. Description of the three different datasets employed in this research.

Dataset	User Labels	Classes	Sensors	Time [hour]	# Users
SLR	Texting, Pocket, Swing, Talking, Body	5	Accelerometers	10.8	57
HAR	Jogging, Sitting, Stairs down, Stairs up , Stationary, Walking	6	Accelerometers and Gyroscopes	7.25	24
SHAR	Biking Belt, Biking Pocket, Biking Uparm, Downstairs Belt Downstairs Pocket, Downstairs Uparm, Jogging Belt, Jogging Pocket Jogging Uparm, Sitting belt, Sitting Pocket, Sitting Uparm Standing Belt, Standing Pocket, Standing Uparm, Upstairs Belt Upstairs Pocket, Upstairs Uparm, Walking belt, Walking Pocket Walking Uparm	21	Accelerometers and Gyroscopes	9.71	10

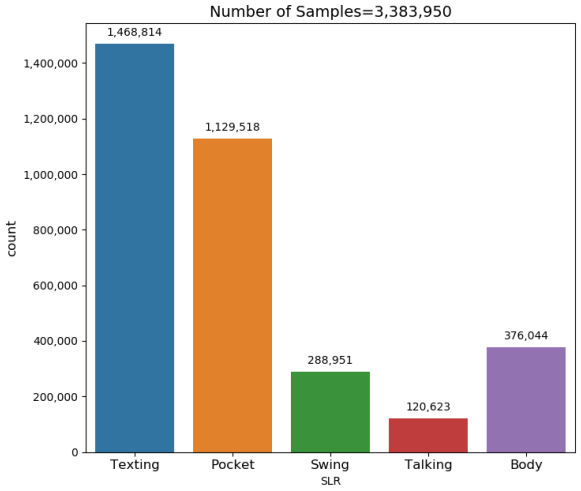

FIGURE 2. Samples class distribution in the SLR dataset.

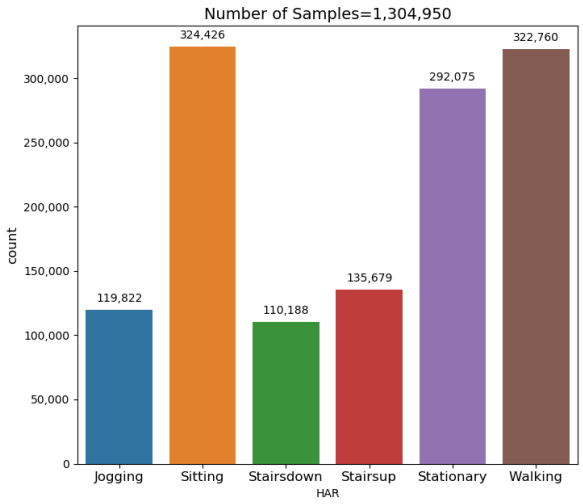

FIGURE 3. Samples class distribution in the HAR dataset.

eight people were recorded in a sampling rate of 100Hz about 240min, while the smartphone was in Pocket or Texting classes. The fourth dataset [30], has recordings of four locations: Texting, Pocket, Swing and, Talking, made by six people, with three different sampling rates (25, 50, and 100Hz), and six different smartphones, yielding a total of 15min of recorded data. The fifth dataset [42], was recorded for HAR applications and is included also in the HAR dataset (Section III-B). For the SLR dataset, only the walking part is used. The dataset was recorded by 24 people using a smartphone in their pocket with a sampling rate of 50Hz. The last (sixth) dataset [43], was also created for an HAR research. There, the goal was to evaluate HAR performance with smartphones' and smartwatches' recordings. For the SLR dataset, walking from the smartwatches were employed since they share the same dynamics as smartphones in a swing motion. The Body class was also used for this purpose. The recordings where made using ten people, with a sampling rate of 50Hz, for a total of 48min.

To summarize, the combined SLR dataset has 3, 383, 950 samples (in each accelerometer axis) recorded by 57 people using different sampling rates of $25-200$Hz. The distribution of the samples in each class is shown in Figure 2. Note that no pre-processing was performed on the raw data. All of the datasets were stacked together to a single one regardless of the sampling rate they were recorded by. The motivation for doing so was to make the network robust to the sampling rate since in real time the user is free to choose the sampling rate between a predefined range.

B. HAR

The HAR dataset was collected with an iPhone 6s kept in the participant's front pocket [42], [44]. 24 participants (10 women and 14 men) with varying age, weight, and height performed six activities in 15 trials in the same environment and conditions: Walking, Jogging, Sitting, Standing, Stairs down and, Stairs up. This dataset has 1, 304, 950 samples in each accelerometer and gyroscope axis. The sampling rate was 50Hz leading to a total of 435min. The distribution of the samples in each class is shown in Figure 3.

C. SHAR

The SHAR dataset is derived from a dataset created by [43] in order to evaluate how and when various motion sensors, which are available on a smartphone, can best be exploited (individually or combined) for improving activity recognition. To that end, [43] have collected recordings with a sampling rate of 50Hz of ten participates during seven physical activities: Walking, Running, Sitting, Standing, Jogging, Biking, Upstairs (walking upstairs) and Downstairs (walking downstairs). Each of those participants was equipped with five smartphones in five different locations: right jeans pocket, left jeans pocket, on the belt position towards the right leg using a belt clip, on the right upper arm and, on the right

wrist. The SHAR dataset is constructed with combined SLR and HAR labels (classes) capturing both the physical activity and smartphone location. For this purpose, the right and left pocket recordings from [43] were first united under the Pocket label. Recordings from [43] were then labelled based on three smartphone locations and seven human activities: [Pocket, Belt, Uparm (upper arm)] x [Walking, Running, Sitting, Standing, Jogging, Biking, Upstairs, Downstairs], yielding a total of 21 classes. Each class has 87, 00 samples except the three upstairs classes (Upstairs Belt, Upstairs Pocket, Upstairs Uparm) with have 60, 900 samples. Thus, this dataset has 1, 748, 700 samples in each accelerometer and gyroscope axis resulting in 583min of recorded data.

IV. RESULTS

A. EXPERIMENTAL SETUP

For evaluation purposes, the method proposed in this paper is further compared to a CNN model shown to achieve the best performance on different SLR tasks [30]. This model consists of a convolutional encoder and a classifier head. The encoder includes two 1-dimensional convolutional layers with RELU non-linearity followed by Dropout and max pooling. The classifier head consists of two FC layers with RELU non-linearity. Log SoftMax is applied on the output of the final FC layer.

Each dataset is arbitrarily split into a train set and test set (where 85% of the samples, on average, are selected for the train set), while ensuring all classes are represented in both sets.

Both models (CNN model and the proposed approach) are optimized using Adam, with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-10}$. A batch size of 128 and a weight decay of 10^{-4} are employed. An initial learning rate of $\lambda = 10^{-4}$ is used and further reduced by half every m epochs depending on the experiment (m is set to the same value for both models). Note that in order to support a fair comparison, all hyper-parameters, except for the number of epochs, are not fine-tuned and are kept fixed for both models. Each model is trained for up to 30 epochs for small datasets and for up to 80 epochs for larger datasets. The full configuration used for training is available in the shared codebase.

For ease of description, from here on the CNN model and the proposed approach are refereed to as IMU-CNN and IMU-Transformer, respectively. The following sections describe different experiments aimed at evaluating performance, robustness and generalization, across different activity recognition scenarios.

B. THE EFFECT OF WINDOW SIZE

An IMU sample contains measurements aggregated over a window of time. The size of the window (k), determines the length of the sequence passed to the model. In general, in HAR, SLR or SHAR tasks, it is desirable to work with the smallest window size which still achieves a target accuracy. In the datasets employed, the sampling rate varies between 25-200Hz. Considering the slowest sampling rate, a window size of 50 corresponds to a time duration of two seconds. Larger window sizes increases the probability of a mode change during a single window, which is an undesired behavior. With this motivation in mind, the IMU-CNN and IMU-Transfomer models were trained using decreasing windows sizes starting from 50: 50, 38, 26 and 14. Since the SHAR dataset represents both HAR and SLR tasks, it was chosen for this analysis (training and testing).

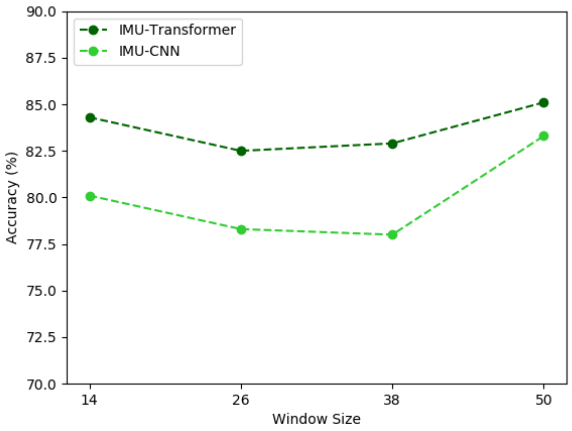

FIGURE 4. Classification accuracy of the SHAR dataset with different windows sizes.

Figure 4 shows the results obtained with the two models. The best accuracy is achieved when using a window of 50 samples, with a gap of 1.8% in favor of the IMU-Transformer model. For smaller windows the improvement gap grows to 4.2% ($k = 26, 38$) and 4.9% ($k = 14$). For both models, a similar trend emerges where a more notable degradation occurs when using $k = 26, 38$. This can be explained by events of class mixture per sequence, which are less significant and frequent when using the smallest window size (14). When considering the variance in accuracy, the intra-difference (within smaller windows) is smaller compared to the inter-distance, with respect to the original window size. In addition, a more significant degradation in performance is observed with the IMU-CNN model. Specifically, when comparing the performance between $k = 50$ and $k = 14, 26, 38$, the accuracy decreases in 4.5% on average with IMU-CNN, compared to only 1.9% with IMU-Transformer. Hence, the IMU-Transformer model not only consistently improves the performance, regardless of the window size, but is also more robust to smaller window sizes, making it a favorable option for real-time applications. Based on the results above, a window size of 50 was selected for further experiments and analysis.

C. ACTIVITY RECOGNITION PERFORMANCE ACROSS DIFFERENT DATASETS

In order to evaluate the performance per task, the IMU-CNN and IMU-Transformer models were trained and tested on the SLR, HAR and SHAR datasets. Table 2 gives the accuracy for each dataset and the mean accuracy across datasets.

TABLE 2. Results obtained for the SLR, HAR and SHAR datasets. The accuracy (%) for a CNN model (IMU-CNN) and the proposed architecture (IMU-Transformer) is reported per dataset and overall (mean performance).

Experiment	Window Size	IMU-CNN Accuracy	IMU-Transformer Accuracy
SLR	50	96.5%	97.4%
HAR	50	86.2%	89.6%
SHAR	50	83.3%	85.1%
Overall	50	88.7%	90.7%

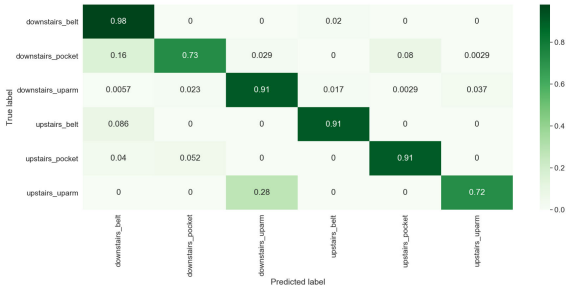

FIGURE 5. Confusion matrix for the SHAR stairs dataset using a CNN model (IMU-CNN).

IMU-Transformer consistently achieves better accuracy, with a 2% improvement on average. The performance of both models depends on the dataset, with a decreasing accuracy for the SLR, HAR and SHAR datasets, respectively. In addition both models achieve a significantly higher accuracy ($> 8\%$) on the SLR dataset, compared to the two other tasks. These results are consistent with previous observations on the SLR dataset (distinct patterns between classes that are relatively easy to learn [30]) and suggest that the IMU-Transformer model learns IMU data better than the IMU-CNN model, regardless of how challenging the specific dataset is.

D. CHALLENGING MODES AND GENERALIZATION

Due to the underlying dynamics some modes are more challenging to learn than others. For example, the stairs related dynamics expressed in stairs-up and stairs-down scenarios. In order to evaluate the proposed approach in this scenario, the SHAR dataset is subset by taking samples only from the following six classes: Downstairs Belt, Downstairs Pocket, Downstairs Uparm, Upstairs Belt, Upstairs Pocket, Upstairs Uparm. Figure 5 presents the confusion matrix for the IMU-CNN model, showing a total accuracy of 86.6%. Four out of six modes received more than 91% yet, both Downstairs Pocket and Upstairs Uparm achieved only 73% and 72%, respectively. In particular, about 28% of the Upstairs Uparm samples were misclassified as Downstairs Uparm. In a similar manner, Figure 6 shows the confusion matrix for the IMU-Transformer model, achieving a total accuracy of 92.3% corresponding to a 5.7% improvement. Focusing on the Upstairs Uparm mode, the 28% misclassification of the IMU-CNN model is reduced to 7.5%. Another important aspect of model performance is its ability to generalize across datasets. For this purpose, the stairs experiment is further extended by training on the SHAR dataset but evaluating on

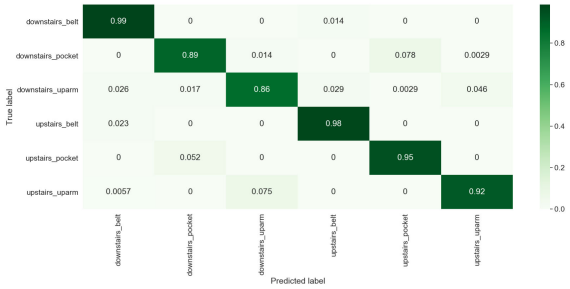

FIGURE 6. Confusion matrix for the SHAR stairs dataset using the proposed approach (IMU-Transformer).

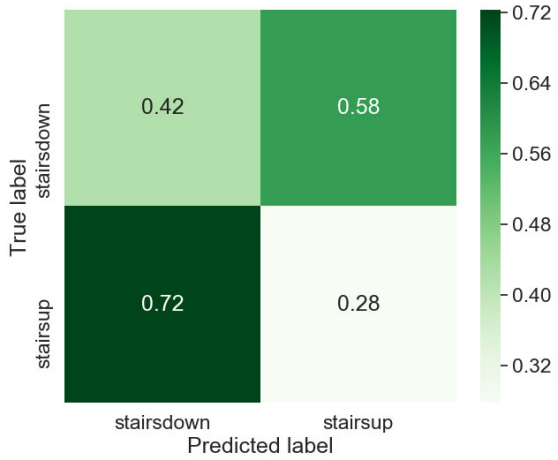

FIGURE 7. Confusion matrix for the SHAR dataset with HAR as the test dataset using a CNN model (IMU-CNN).

the HAR test set. In order to obtain class compatibility, the six stairs-related SHAR classes are collapsed into two classes: Stairs up (for Upstairs Belt, Upstairs Pocket and, Upstairs Uparm) and, Stairs down (for Downstairs Belt, Downstairs Pocket and, Downstairs Uparm). The HAR dataset is then subset with these two classes. The results of this experiment are depicted in Figures 7-8 for the IMU-CNN and IMU-Transformer models, respectively. The IMU-CNN reached a total accuracy of 33.8%, where most of the error (72% of the samples) is attributed to the misclassification of Stairs up as Stair down. IMU-Transformer significantly improved the total accuracy (80.2%) and obtained a symmetrical behaviour between the two classes.

E. SUMMARY OF RESULTS

The results of the experiments described in this paper are summarized in Table 3. The proposed approach and a CNN model were extensively evaluated on three datasets with more than 27 hours of recordings, collected by 91 users, considering (1) the target tasks, namely SLR, HAR and SHAR, (2) the effect of the window size (SHAR-50/38/26/14 in Table 3), (3) challenging dynamics (SHAR-Stairs in Table 3) and (4) generalization (SHAR/HAR-Stairs in Table 3).

For each experiment conducted, the difference between the accuracy of the proposed approach (IMU-Transformer)

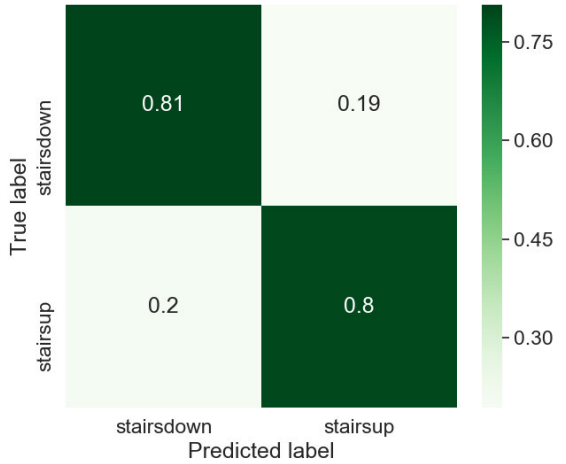

FIGURE 8. Confusion matrix for the SHAR dataset with HAR as the test dataset using the proposed approach (IMU-Transformer).

TABLE 3. A summary of the experiments described in this paper. The proposed Transformer-based architecture achieves a consistent improvement compared to the CNN-based model.

Experiment	Window Size	Reported Accuracy Improvement
Activity Recognition Datasets		
SLR	50	0.9%
HAR	50	3.4%
SHAR	50	1.8%
SHAR-Window	38	4.9%
SHAR-Window	26	4.2%
SHAR-Window	14	4.2%
Challenging Activity Recognition Scenarios		
SHAR-Stairs	50	5.7%
SHAR/HAR-Stairs	50	46.4%

and the accuracy of the CNN architecture (IMU-CNN) is reported, where positive values indicate improvement. The IMU-Transformer model consistently improves the classification accuracy compared to the IMU-CNN model, with a notable improvement in challenging scenarios.

In addition to accuracy, the mean runtime was also evaluated (using the SHAR dataset). When run on a GPU (Tesla V100 16Gb), the IMU-CNN model classifies a sample with a window size of 50 in 0.48ms on average versus 4.76ms with the IMU-Transformer model. When tested on a CPU, the average inference time increases, with 3.93ms and 14.3ms for the IMU-CNN and IMU-Transformer, respectively. In both cases the classification runtime is negligible compared to the runtime in which a classification is expected, even when considering the highest sampling rate (200Hz with a window size of 50 corresponding to 250ms). Integrating the proposed framework in a real-time application involves deploying the trained model (weights) and executing its forward pass with an inertial signal aggregated over a predefined time window.

V. CONCLUSION

This paper presents a deep learning framework for activity recognition. The proposed approach employs Transformers for performing sequence aggregation using attention, which have been successfully used for sequence analysis tasks in other domains. To-date, this is the first time a Transformer architecture is employed for inertial-based activity recognition. Three types of datasets, with more than 27 hours of recordings, collected by 91 users, were used for an extensive evaluation: 1) smartphone mode location recognition, created from six different datasets, 2) human activity recognition and, 3) combined smartphone location and human activity recognition. In addition more challenging scenarios were addressed: 1) classification of stairs up/down motion in three different smartphone locations and, 2) testing the second dataset with a network trained on the third dataset for stairs only data.

Throughout multiple experiments, representing different activity recognition scenarios and settings, the proposed approach demonstrates an improved prediction accuracy that can be transferred better between datasets compared to a CNN-based solution. While this approach is an order of magnitude slower (4.76ms vs 0.48ms), its runtime is negligible compared to the runtime in which a classification is expected, even when considering the highest sampling rate.

An immediate extension of the proposed framework is to evaluate it with more user and smartphone modes. In addition, further enhancements can be made to the proposed framework, leveraging on Transformers acceleration techniques. Finally, transfer learning can be investigated to evaluate whether a model trained on one dataset can serve as a better starting point for new models trained on incoming datasets.

Since the proposed approach performs classification of inertial data it can be directly applied for other inertial-based classification tasks. In addition, it can be further adapted to handle other sensory data collected in a sequential manner for activity recognition, by simple modifications to the CNN backbone. In order to support results reproduction and an easy transfer to other domains, the implementation of the proposed framework is available at: https://github.com/yolish/har-with-imu-transformer.

REFERENCES

[1] B. Ni, G. Wang, and P. Moulin, "RGBD-HuDaAct: A color-depth video database for human daily activity recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 1147–1153.
[2] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Device-free human activity recognition using commercial WiFi devices," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1118–1131, May 2017.
[3] H. Li, X. He, X. Chen, Y. Fang, and Q. Fang, "Wi-motion: A robust human activity recognition using WiFi signals," *IEEE Access*, vol. 7, pp. 153287–153299, 2019.
[4] X. Li, Y. He, and X. Jing, "A survey of deep learning-based human activity recognition in radar," *Remote Sens.*, vol. 11, no. 9, p. 1068, May 2019.
[5] W. Lin, M.-T. Sun, R. Poovandran, and Z. Zhang, "Human activity recognition for video surveillance," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2008, pp. 2737–2740.
[6] X. Zhang, X. Chen, Y. Li, V. Lantz, K. Wang, and J. Yang, "A framework for hand gesture recognition based on accelerometer and EMG sensors," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 41, no. 6, pp. 1064–1076, Nov. 2011.
[7] R. Xu, S. Zhou, and W. J. Li, "MEMS accelerometer based nonspecific-user hand gesture recognition," *IEEE Sensors J.*, vol. 12, no. 5, pp. 1166–1173, May 2012.
[8] N. Y. Hammerla, S. Halloran, and T. Ploetz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," 2016, *arXiv:1604.08880*. [Online]. Available: http://arxiv.org/abs/1604.08880

[9] P. Vepakomma, D. De, S. K. Das, and S. Bhansali, "A-Wristocracy: Deep learning on wrist-worn sensing for recognition of user complex activities," in *Proc. IEEE 12th Int. Conf. Wearable Implant. Body Sensor Netw. (BSN)*, Jun. 2015, pp. 1–6.
[10] W. Taylor, S. A. Shah, K. Dashtipour, A. Zahid, Q. H. Abbasi, and M. A. Imran, "An intelligent non-invasive real-time human activity recognition system for next-generation healthcare," *Sensors*, vol. 20, no. 9, p. 2653, May 2020.
[11] Z. Sun, X. Mao, W. Tian, and X. Zhang, "Activity classification and dead reckoning for pedestrian navigation with wearable sensors," *Meas. Sci. Technol.*, vol. 20, no. 1, Jan. 2009, Art. no. 015203.
[12] S. Guo, H. Xiong, X. Zheng, and Y. Zhou, "Indoor pedestrian trajectory tracking based on activity recognition," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 6079–6082.
[13] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognit. Lett.*, vol. 119, pp. 3–11, Mar. 2019.
[14] A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, and P. Havinga, "Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey," in *Proc. 23th Int. Conf. Archit. Comput. Syst.*, Feb. 2010, pp. 1–10.
[15] L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu, "Sensor-based activity recognition," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 6, pp. 790–808, Nov. 2012.
[16] L. Minh Dang, K. Min, H. Wang, M. Jalil Piran, C. Hee Lee, and H. Moon, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognit.*, vol. 108, Dec. 2020, Art. no. 107561.
[17] M. Shoaib, S. Bosch, O. Incel, H. Scholten, and P. Havinga, "A survey of online activity recognition using mobile phones," *Sensors*, vol. 15, no. 1, pp. 2059–2085, Jan. 2015.
[18] S. Ramasamy Ramamurthy and N. Roy, "Recent trends in machine learning for human activity recognition—A survey," *Wiley Interdiscipl. Reviews: Data Mining Knowl. Discovery*, vol. 8, no. 4, Jul. 2018, Art. no. e1254.
[19] A. Sargano, P. Angelov, and Z. Habib, "A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition," *Appl. Sci.*, vol. 7, no. 1, p. 110, Jan. 2017.
[20] F. Demrozi, G. Pravadelli, A. Bihorac, and P. Rashidi, "Human activity recognition using inertial, physiological and environmental sensors: A comprehensive survey," *IEEE Access*, vol. 8, pp. 210816–210836, 2020.
[21] R. Yang and B. Wang, "PACP: A position-independent activity recognition method using smartphone sensors," *Information*, vol. 7, no. 4, p. 72, Dec. 2016.
[22] M. Elhoushi, J. Georgy, A. Noureldin, and M. Korenberg, "Online motion mode recognition for portable navigation using low-cost sensors," *Navigation*, vol. 62, no. 4, pp. 273–290, Dec. 2015.
[23] I. Klein, Y. Solaz, and G. Ohayon, "Pedestrian dead reckoning with smartphone mode recognition," *IEEE Sensors J.*, vol. 18, no. 18, pp. 7577–7584, Sep. 2018.
[24] S.-H. Fang, Y.-X. Fei, Z. Xu, and Y. Tsao, "Learning transportation modes from smartphone sensors based on deep neural network," *IEEE Sensors J.*, vol. 17, no. 18, pp. 6111–6118, Sep. 2017.
[25] X. Zhou, W. Liang, K. I.-K. Wang, H. Wang, L. T. Yang, and Q. Jin, "Deep-learning-enhanced human activity recognition for Internet of healthcare things," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6429–6438, Jul. 2020.
[26] C. Chen, X. Lu, A. Markham, and N. Trigoni, "IONet: Learning to cure the curse of drift in inertial odometry," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 1–9.
[27] Q. Wang, L. Ye, H. Luo, A. Men, F. Zhao, and Y. Huang, "Pedestrian stride-length estimation based on LSTM and denoising autoencoders," *Sensors*, vol. 19, no. 4, p. 840, Feb. 2019.
[28] I. Klein and O. Asraf, "Stepnet—-Deep learning approaches for step length estimation," *IEEE Access*, vol. 8, pp. 85706–85713, 2020.
[29] E. Vertzberger and I. Klein, "Attitude adaptive estimation with smartphone classification for pedestrian navigation," *IEEE Sensors J.*, vol. 21, no. 7, pp. 9341–9348, Apr. 2021, doi: 10.1109/JSEN.2021.3053843.
[30] I. Klein, "Smartphone location recognition: A deep learning-based approach," *Sensors*, vol. 20, no. 1, p. 214, Dec. 2019.
[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.
[32] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent., (ICLR)*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, 2015, pp. 2–16.
[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.
[34] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
[35] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision—ECCV*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 213–229.
[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*. [Online]. Available: http://arxiv.org/abs/2010.11929
[37] W.-C. Chang, H.-F. Yu, K. Zhong, Y. Yang, and I. Dhillon, "X-BERT: Extreme multi-label text classification with using bidirectional encoder representations from transformers," 2019, *arXiv:1905.02331*. [Online]. Available: http://arxiv.org/abs/1905.02331
[38] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," 2020, *arXiv:2012.12877*. [Online]. Available: http://arxiv.org/abs/2012.12877
[39] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*. [Online]. Available: http://arxiv.org/abs/1607.06450
[40] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. Alche-Buc, E. Fox, and R. Garnett, Eds. Vancouver, BS, Canada: Curran Associates, 2019, pp. 8026–8037.
[41] H. Yan, Q. Shan, and Y. Furukawa, "RIDI: Robust IMU double integration," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 621–636.
[42] M. Malekzadeh, R. G. Clegg, A. Cavallaro, and H. Haddadi, "Mobile sensor data anonymization," in *Proc. Int. Conf. Internet Things Design Implement.*, Apr. 2019, pp. 49–58.
[43] M. Shoaib, S. Bosch, O. Incel, H. Scholten, and P. Havinga, "Fusion of smartphone motion sensors for physical activity recognition," *Sensors*, vol. 14, no. 6, pp. 10146–10176, Jun. 2014.
[44] M. Malekzadeh, R. G. Clegg, A. Cavallaro, and H. Haddadi, "Protecting sensory data against sensitive inferences," in *Proc. 1st Workshop Privacy Design Distrib. Syst.*, Apr. 2018, pp. 1–6.

YOLI SHAVIT received the B.Sc. degree in computer science and life science from Tel Aviv University, the M.Sc. degree in bioinformatics from Imperial College London, and the Ph.D. degree in computer science from the University of Cambridge. She is currently a Postdoctoral Researcher with Bar-Ilan University and a Principal Research Scientist with the Huawei Tel Aviv Research Center, Toga Networks, a Huawei Company. Her research interests include algorithms in deep learning and their applications to real-life domains and visual and multi-sensor localization problems.

ITZIK KLEIN (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in aerospace engineering and the Ph.D. degree in geo-information engineering from the Technion–Israel Institute of Technology, in 2004, 2007, and 2011, respectively. He is currently an Assistant Professor with the University of Haifa, heading the Department of Marine Technologies, Autonomous Navigation and Sensor Fusion Laboratory. His research interests include navigation, novel inertial navigation architectures, autonomous underwater vehicles, sensor fusion, smartphone based navigation, and estimation theory.