A Proposed ConvXGBoost Model for Human Activity Recognition with Multi Optimizers

Rana Abdulrahman Lateef*
PhD Student at Computer Science Department, University of Technology, Baghdad.
Working at Computer Science Department, Baghdad College of Economic Sciences University, Baghdad, Iraq.
E-mail: Iraqranacsbaghdad@gmail.com; taught.rana.abdalrahman@baghdadcollege.edu.iq

Dr. Ayad Rodhan Abbas
Assistant Professor, Computer Science Department, University of Technology, Baghdad, Iraq.
E-mail: ayad.r.abbas@uotechnology.edu.iq

Received September 02, 2021; Accepted December 06, 2021
ISSN: 1735-188X
DOI: 10.14704/WEB/V19I1/WEB19114

Abstract

The wide use of smartphones and later smartwatches equipped with a set of sensors such as location, motion, and direction blaze the trail for researchers to better recognize human activity. However, researches on using inertial or motion sensors (i.e., accelerometer, gyroscope) for human activity recognition (HAR) has intensified and reside a great confrontation to be faced. Lately, many deep learning methods have been suggested to improve the human activity classification and discrimination performance to reach an optimal accuracy. Therefore, this paper applies a Convolutional eXtreme Gradient Boosting (ConvXGBoost), which combines Convolutional Neural Network (CNN) represented by AlexNet to learn the input features automatically, followed by XGBoost decision tree used to predict the class label and thereof recognize the performed activity. Human activities are collected from sensors as time series data. Therefore, we suggested using one-dimensional AlexNet (1D AlexNet) model instead of 2D. The AlexNet model is compiled with two optimizers Adam and Stochastic Gradient Descent (SGD) which are applied consecutively. The suggested architecture was trained and evaluated on the “WISDM Smartphone and Smartwatch Activity and Biometric Dataset” that consists of raw data for eighteen activities recorded from phone and watch. The experiments revealed that using multi optimizer with a convolutional neural network improved the accuracy of recognition by 5%. Moreover, a proposed ConvXGBoost model outperformed the performance of other models works with the dataset as mentioned above with an overall accuracy of 98-99% depends on the device used.
Keywords

HAR, deep learning, Convolutional Neural Network (CNN), Convolutional eXtreme Gradient Boosting, 1D AlexNet, Accelerometer, Gyroscope, Smartphone, Smartwatch, Adam, SGD, WISDM Smartphone and Smartwatch Activity and Biometric Dataset.

Introduction

HAR is the process of monitoring and analyzing a person’s behavior to understand the ongoing activities (L. Chen & Nugent, 2019). In recent years, HAR has become one of a very significant research field due to its wide applications in many areas, especially for healthcare (Wang, 2015), security (Ranasinghe et al., 2016), and military (Shakya et al., 2018). For example, recognizing activities such as running, resting, or walking of patients with heart disease or obesity is required to track an exercise routine as a portion of their treatment. The recognition of human activity can detect probable security threats through video surveillance (Kadim et al., 2020). Whereas providing exact information about soldier’s activities along with their locations and health states is highly useful for their performance and safety (Labrador & Yejas, 2013).

Essentially, HAR approach can be partitioned into two basic types: vision-based activity recognition and sensor-based activity recognition (Z. Chen et al., 2020). In vision-based activity recognition, a series of images or videos are taken using a camera-based surveillance system to analyze and monitor the activities done by people. Conversely, sensor-based activity recognition use an accelerometer, and gyroscope either placed on different position of the body or emerging in a smart device to collect the data generated by the movement of the human to monitor activity (Basly et al., 2020). Corresponding to an approach of feature extraction, the HAR system can be categorized as an imitative model based on hand-crafted features or a deep model. Recently, using deep learning techniques for feature learning has been gaining growing attention for their capability for extracting and designing features automatically. Deep learning can be considered as a branch or domain of machine learning (Abdulmunem et al., 2021; Fu, 2016). With the deep learning model, train a model becomes easier to recognize specific activities collected from raw sensor data in an efficient way. The scenario of HAR can be performed by a user with a device (e.g. smartphone, smartwatch) equipped with accelerometer and gyroscope sensors that continuously record sensor data and then fed to a learning model to perform activity recognition (Mutegeki, 2020). Nowadays, deep learning has become excessively used in fields as a classification problems (Young et al., 2018).
In this paper, we suggested using ConvXGBoost deep learning model as a classification approach with HAR problem, which is based on combining the performance of deep learning model represented by convolutional neural network represented by 1D Alex deep net and a scalable machine learning model represented by XGBoost which is the last layer in the model.

The proposed model is evaluated with a publicly available “WISDM Smartphone and Smartwatch and Activity Biometric Dataset.”

Related Work

In the last decade, researchers from various application domains have explored activity recognition by evolving different techniques and algorithms of either machine or deep learning with a different dataset. This section inspects some of the previous works with the same dataset used in this paper, and also works relates to the proposed approach but in other applications.

Weiss et al. (Weiss et al., 2016) worked on the dataset built by members of the WISDM (Wireless Sensor Data Mining) Lab. Five learning models are used in their work:, Random Forest, J48, IB3, Naïve Bayes, and Multi-Layer Perceptron. Each model uses a single sensor models using watch- gyroscope, watch -accelerometer, and phone - accelerometer. Their experiments revealed that smartwatch can recognize a specialized hand-based activities like “drinking” which cannot successfully be recognized by smartphone with an accuracy rate of 93.3% with smartwatch and 77.3% for a smartphone.

Later, Weiss et al. (Weiss et al., 2019a) Investigated the effectiveness of combining the accelerometer and gyroscope sensors for smartwatch and smartphone. Moreover, they evaluated eighteen different activities of daily living for their biometric identification in addition to biometric authentication with the Random Forest algorithm. Their study concluded that when a person has a smartphone and smartwatch the average accuracy is 94.4% using accelerometer sensor and 91% for each activity.

Benavidez et al. (Benavidez & McCreight, 2019) implemented both CNN and stacked Long Short Term Memory (stacked-LSTM) on the Weiss dataset to classify 18 activities. The results show that LSTM achieved an accuracy rate of 79% for watch vs. 74% for phone which is better than CNN with an accuracy rate of 74% for watch vs. 50% for phone.

Recently, the work set towards a combination of deep and machine learning to achieve good resultant accuracy. The CNN deep model is widely used in research either as a standalone
model or combined with other machine or deep models especially in the image domain (Agrawal & Mittal, 2020).

Thongsuwan et al. (Thongsuwan et al., 2021) produced a novel deep learning model named Convolutional eXtreme Gradient Boosting (ConvXGBoost) for classification issues. The proposed model consists of many stacked convolutional layers with neither fully connected nor pooling-layer. The model performance is evaluated with a different dataset from UCI of machine learning repository contain video or images. The results show that ConvXGBoost produced better results than applying each model alone; and also better than other models (e.g. MLP, SVC, and DTC), depending on the dataset.

Jiao et al. (Jiao et al., 2021) Suggested a CNN with XGBoost based on Adaptive Particle Swarm Optimization (APSO) that divide the particle into a subgroup. APSO optimization is applied to CNN and XGBoost simultaneously as a bidirectional optimization structure to optimize the hyperparameter of the overall architecture. The proposed is evaluated with three datasets. The results produced are better than other models.

In this work, using simple 1D Alex net reduce the number of calculation parameter. Compiling Alex-net with two optimizers named Adam and SGD then combining it with XGBoost achieved the best results.

**Dataset**

Weiss et al. (Weiss et al., 2019b) built a dataset in WISDM (Wireless Sensor Data Mining) Lab gathered from accelerometer and gyroscope sensors of smartwatch and smartphone to record raw time series data of 18 different activities of daily living include hand and non-hand controlled activities. The dataset is available publically on the Repository of UCI Machine Learning as the “WISDM Smartphone and Smartwatch Activity and Biometrics Dataset” since 2016 and, its updated version is available in 2019. The dataset consists of sensor data collected from fifty-one persons, each of them was required to implement eighteen activities for three minutes. Each participant holds a smartphone in his/her pocket and wears a smartwatch in their hand. The raw time-series sensor data was gathered from gyroscope and accelerometer equipped with the two devices at a rate of 20 HZ. The activities were subdivided into three groups to simplify analysis; First, No-hand activities: kicking, Jogging, stairs, walking, and standing. Second, Hand-controlled activities (General): playing catch, writing, dribbling, clapping, folding Clothes, and typing, brushing teeth. Finally, Hand-oriented activities (eating): eating a sandwich, eating pasta, eating soup, drinking, eating chips (Weiss et al., 2019c).
Methodology

1. AlexNet Model

The AlexNet model is a neural network similar to traditional neural networks such as 1D CNNs and 2D CNNs, which consist of 3 layers, input, hidden, and output layer. The CNN works as feature extractor that combines an artificial neural network with back propagation, which make the model complexity simpler and reduce the parameter. The premise of AlexNet made CNNs more deeply and the training result more precise. It consists of 5 convolution layers and 3 fully connected layers. The convolutional layers 1, 2, and 5 are connected with max-pooling layers (Zhang et al., 2021). AlexNet has some aspects compared with common CNN: since the drop out layer is applied after a fully connected layer, therefore ignoring some neurons randomly during the training process can mitigate the problem of overfitting, the nonlinear activation function (ReLU) is used to accelerate the process of forward propagation and solve the gradient explosion problem (Sun et al., 2016). Max-pooling layers are employed to increment the richness of features (Wu et al., 2018). Figure 1 shows the structure of AlexNet model used in this paper.

![Figure 1 1D AlexNet architecture](image)

2. eXtreme Gradient Boosting (XGBoost) Model

XGBoost, suggested by Chen and Guestrin (T. Chen & Guestrin, 2016), is a powerful tree boosting machine learning method for regression and classification based on the structure of gradient boosting. However, it is different from gradient boosting in utilize of Taylor extension to sacrificical the Loss function. The model owns a leading tradeoff variance and bias with a slighter decision tree commonly used to gain more elevate in accuracy. Given a dataset of $s$ samples and $m$ features, which depicted as $D = \{(x_i, y_i)\} (|D| = s, x_i \in$
\( P^m, y_i \in P \), where \( y \) represent the true value and \( x \) represent the eigenvalue. The approach works by aggregate the outcomes of Z trees as the last value been predicted, which can be depict as:

\[
\hat{y}_i = \sum_{z=1}^{Z} f_z(x_i), \; f_z \in A
\]  

(1)

Where \( A \) represents a group of decision trees stated as follow:

\[
A = \{ f(x) = \omega_{q(x)} \} \quad (q: P^m \rightarrow T, \omega \in P^T )
\]

Where \( f(x) \) represents one of the trees, \( \omega_{q(x)} \) represents weight for the leaf nodes, \( T \) represents the count of leaf nodes, and \( q \) represents the architecture of every tree, which directs the sample to the conformable leaf node. Thus, the portend value of XGBoost is the aggregate of the leaf nodes values for every tree. The aim is to learn a set of \( z \) trees. Therefore, the goal function was minimized as follows: (T. Chen & Guestrin, 2016).

\[
L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{z=1}^{Z} \Omega(f_z)
\]  

(2)

Where \( l \) is define a loss that represent the difference between the true value \( y_i \) and the estimated values \( \hat{y}_i \). The regularization \( \Omega \) is utilize to determine the retribution of the decision tree in order to avoid overfitting and it depict as follows: (Pang et a., 2019)

\[
\Omega(f) = \gamma T + \frac{1}{2} \lambda \| \omega \|^2
\]  

(3)

Where, \( \gamma \) represent a hyperparameter that monitors a model complexity, \( \lambda \) is the retribution coefficient for weight of leaf node \( \omega \), which is ordinary fixed. The two symbols \( \gamma \) and \( \lambda \) define a complication of the model and are commonly set experimentally. Through training, a new tree is combined to fit the lagging of the prior iteration. Thus, if the model has \( t \) trees, it can be represent as follows:

\[
\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_z(x_i)
\]  

(4)

Replacing (4) into the goal function in equation (2) produces the function:

\[
L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_z(x_i)) + \Omega(f_z)
\]  

(5)
Then, XGBoost implements the Taylor extension of the objective function, picks the initial three expression, eliminates small short expression with high order, and finally mutate the objective function into: (T. Chen & Guestrin, 2016)

\[
L^{(t)} \approx \sum_{i=1}^{n} \left[ l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i \; f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_k)
\] (6)

Where \(g_i\) and \(h_i\) represent the first and second derivative of loss function respectively. The lagging between the forecast outcome \(\hat{y}^{t-1}\) and \(y_i\) does not impact on the optimization of the goal function, so it is eliminated.

\[
\bar{L}^{(t)} = \sum_{i=1}^{n} \left[ g_i \; f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_k)
\] (7)

The repetition of the tree model is converted into the repetition of the leaf nodes, and the outcome value of the best leaf node is \(w_j^* = -\frac{G_j}{H_j + \lambda}\) where \(G_j\) is \(\sum_{i=1}^{l_j} g_i\) and \(H_j\) is \(\sum_{i=1}^{l_j} h_i\), so substituting the best value into the goal function, the final goal function gained is:

\[
Goj = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T
\] (8)

This approach upholds column sampling in both minimizing computational complexity and overfitting.

**Proposed ConvXGBoost**

In this work, the proposed ConvXGBoost model was evaluated with a dataset described above, where 70% of data was used for the training stage and 30% for testing. The reason to choose this dataset is that it consists of sensor raw data collected using two devices, a smartphone and a smartwatch. The procedure implemented in this work was first accomplished by a preprocessing step that performs standardization on the data and then perform sliding window by splitting data into fixed windows of 10 seconds (i.e. 200 data points) since the observations were recorded at 20 Hz (i.e. 20 data points per second), after that an AlexNet model has been implemented as a convolutional model used to extract deep features. The model is compiled with two optimizers Adam and SGD which applied respectively. Finally, the obtained feature vector is fed to XGBoost model to classify activities. Figure 2 below represent the block diagram of proposed ConvXGBoost architecture for human activity recognition.
Experimental Results and Discussion

The results of activity recognition are depicted and analyzed in this part. Classification accuracy is used to present the results based on the percentage of classification that truly identifies the performed activity. Each model is evaluated with two sensors, an accelerometer and a gyroscope using watch, phone, and watch-phone datasets as shown in Table 1. The results demonstrate that a proposed ConvXGBoost model outperformed other previous models with accuracy rate over 98%. In addition, the results revealed that using phone and watch data separately can achieved better results than using them together. Applying two optimizers with AlexNet model improved the rate of accuracy with about 5%. The resulting learning curve presented in figure 3 show the accuracy and loss for each dataset. The confusion matrices depict in figure 4 show that the model can effectively discriminate between activities that request similar hand movement like eating soup versus eating sandwich, and distinguish between dribbling and kicking.

| Table 1 Overall accuracy of proposed ConvXGBoost model. |
|-----------------|-----------------|-----------------|-----------------|
| Model           | Phone (%)       | Watch (%)       | Phone and watch (%) |
| AlexNet with Adam | 91.5            | 92.5            | 92.6            |
| AlexNet with Adam and SGD | 97.2            | 96.4            | 97.7            |
| ConvXGBoost     | 99.3            | 99.2            | 98.5            |
Figure 3: learning curve for applying ConvXGBoost model

(a) Resultant Confusion Matrix for phone data
### Resultant Confusion Matrix for watch data

<table>
<thead>
<tr>
<th>Activity label</th>
<th>Activity code</th>
</tr>
</thead>
<tbody>
<tr>
<td>walking</td>
<td>0</td>
</tr>
<tr>
<td>jogging</td>
<td>1</td>
</tr>
<tr>
<td>stairs</td>
<td>2</td>
</tr>
<tr>
<td>Sitting</td>
<td>3</td>
</tr>
<tr>
<td>standing</td>
<td>4</td>
</tr>
<tr>
<td>typing</td>
<td>5</td>
</tr>
<tr>
<td>teeth</td>
<td>6</td>
</tr>
<tr>
<td>soup</td>
<td>7</td>
</tr>
<tr>
<td>chips</td>
<td>8</td>
</tr>
</tbody>
</table>

### Resultant Confusion Matrix for phone watch data

<table>
<thead>
<tr>
<th>Activity label</th>
<th>Activity code</th>
</tr>
</thead>
<tbody>
<tr>
<td>pasta</td>
<td>9</td>
</tr>
<tr>
<td>drinking</td>
<td>10</td>
</tr>
<tr>
<td>sandwich</td>
<td>11</td>
</tr>
<tr>
<td>kicking</td>
<td>12</td>
</tr>
<tr>
<td>catch</td>
<td>13</td>
</tr>
<tr>
<td>dribbling</td>
<td>14</td>
</tr>
<tr>
<td>writing</td>
<td>15</td>
</tr>
<tr>
<td>clapping</td>
<td>16</td>
</tr>
<tr>
<td>folding</td>
<td>17</td>
</tr>
</tbody>
</table>

Figure 4 Confusion Matrix for applying ConvXGBoost. With its activity code and labels
Conclusions and Future Works

To elevate the accuracy of HAR, a ConvXGBoost model depends on hyperization between 1D AlexNet –like CNN and XGBoost suggested. The model is basically composed of feature extraction using 1D AlexNet with multi optimizer and feature classification using XGBoost. The two-stage model guarantees that 1D AlexNet can fully extract features. In addition, uses XGBoost decision tree defeats the lack of a single classifier and efficiently discriminates features. The proposed model outperformed the performance of other models that implemented with “WISDM Smartphone and Smartwatch Activity and Biometrics Dataset” to recognize eighteen human activities. The experimental results show that applying two successive optimizers with AlexNet model improved accuracy by 5% compared to applying one optimizer. Moreover, it performed better with watch data than with phone and combined phone-watch data. On contrary, ConvXGBoost performed better with phone and watch data than with combined phone-watch data. The significant-good results achieved with the proposed model make it effective in biometric identification and biometric authentication.

In future work, the proposed model may implement another human activity datasets to study its effectiveness. In addition, may modify the suggested model to employ a hyperparameter optimization algorithm with the XGBoost model.

References


