Human Activity Classification from Time-Series Data

Sai Surya Nattuva
University of Massachusetts Dartmouth

Author Note

Data Science Practicum Project
Advisor: Prof. Gokhan Kul
May 2022

Table of Contents

Abstract

Human activity can be observed and measured by employing different sensors on different parts of the body. An increase in the usage of wearable devices by people has led to increased research in Human Activity monitoring and analysis. This analysis has helped develop many technologies like fall detection, irregular heart rhythms, walking steadiness, posture, etc. This project focuses on classifying the activity performed by a human from time-series data generated from the user's phone's accelerometer. The time-series data is first converted into 2-D images and then fed to a CNN to learn five different activities from 4 features of the data. The project's primary focus is to leverage the predicting power of CNNs on time-series data with a small number of training samples and input channels.

Keywords: time-series, classification, cnn, deep neural networks, human activity

Human Activity Classification from Time-Series Data

The goal of this project is to tailor and train a SensorNet (Jafari, et al., 2019) Neural Network model to classify human activity from a time-series data collected from a smartphone's accelerometer. The user performs a series of moves while holding their smartphone in their dominant hand.

Related Work

(Li, et al., 2017) proposed a system to recognize concurrent activities from data generated by multiple sensors of different types. The system was designed to be a single model capable of working across different sensors deployed in different domains. The process is divided into two phases – feature extraction and decision making. The spatial features are extracted by passing the data to a CNN and then the temporal features by using an LSTM. These features are fed into a binary classifier whose output indicates if an activity is currently in progress.

(Radu, et al., 2017) studied the benefits of utilizing deep learning techniques to capture a user's activity and context. The authors have developed four neural network architectures that are either based on CNNs or DNNs. Two of these models follow a more traditional approach by training from a continuous stream of all sensors. While the other two models are deep model variants that are designed to learn specific modality regions. This novel approach enables the models to maximize the learning of intra-modal correlations.

(Yao, et al., 2017) have designed a low-power unified deep learning framework that can extract features from noisy signals and build a model that can be employed in a variety of fields. This framework is designed to work on time-series mobile sensing data. The approach is to extract the local interactions of similar sensors and global interactions between different sensors by employing a blend of Convolutional and Recurrent Neural networks. They tested the framework in different domains - car tracking, human activity recognition, user identification – all from motion sensors. This framework was designed to be installed on mobile devices and embedded systems.

Data Collection

The project focuses on the classification of the different punches thrown by the user during a practice session of boxing. In the experimental setup, the user held an android smartphone in his dominant hand (right hand in this case). The movements were recorded using an app called “Physics Toolbox Sensor Suite Pro". The app keeps recording the movements of the smartphone while the user is performing his routine. The app collects data from the accelerometer of the device which includes the X, Y, Z axes, and the G-force. The data is polled at 100Hz. We have recorded a practice boxing session for 77 seconds for this project. The data is generated with the time stamp and the values of the above measurements in a CSV format. The columns in the CSV are (in order) – Time, X, Y, Z, G-force. Since the app saves the time in “ss.mmmmmm” (seconds.microseconds) format, I have formatted the timestamps to GMT standard format using a custom Python script, to help with the next step.

Table 1
The features available data are described as below.

Feature	Measurement
TS	Timestamp of the data in the format (YYYY-MM-DD hh:mm:ss.mmmmmm)
gFx	Linear acceleration in the X-axis
gFy	Linear acceleration in the Y-axis
gFz	Linear acceleration in the Z-axis
TgF	G-force of the device

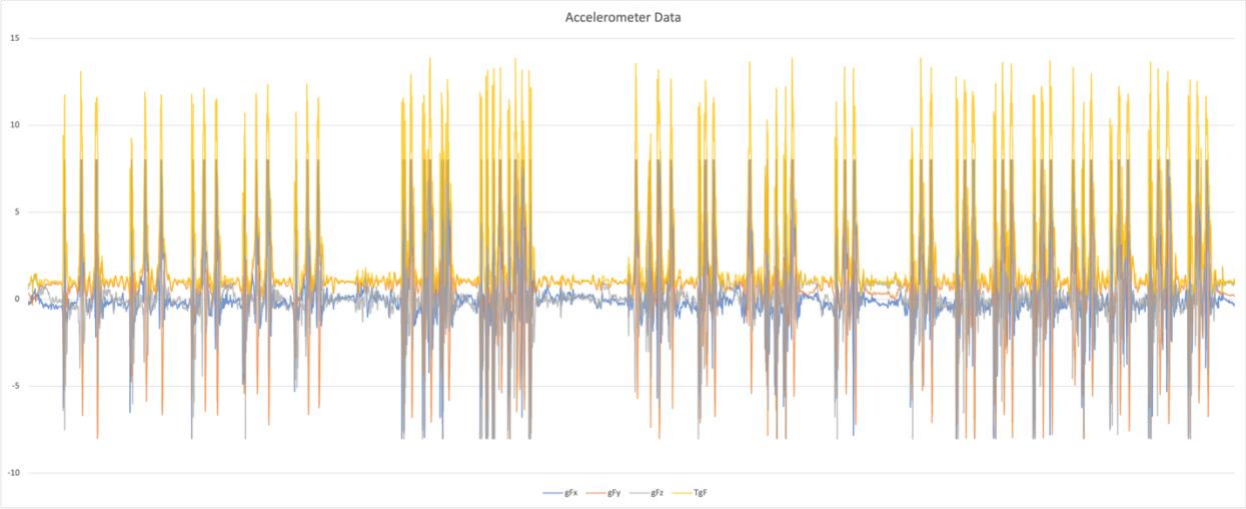

Figure 1 Time series data with 4 modalities – gFx, gFy, gFz, TgF on the Y-axis and the elapsed time on the X-axis

Data Labeling

Since the data generated by the smartphone is unlabeled, I have used an open-source data labeling tool “Label-Studio” to label the data. Label-Studio is a graphical user interface where we can drag along the timeline of the data and label those regions appropriately.

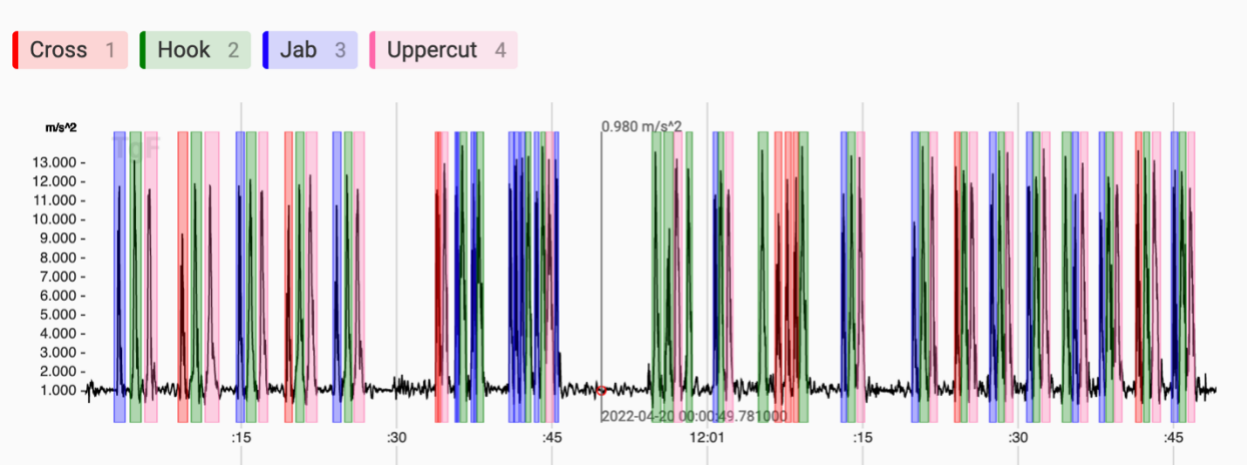

Figure 2 Label Studio interface snippet

Label-Studio can take a CSV as input time-series data that can be added as a labeling project in the application. Once we finish the labeling tasks, we can export the annotations in the form of a JSON file. The JSON file contains the annotations as an array of labels assigned to regions of the time-series data denoted by the start and stop times. To add the labels to all the data points in the data, I have written a Python script to read the regions from the JSON file and assign the labels to the appropriate data points in those regions. I have assigned a constant label for the rest of the unlabeled regions of the time-series data. In this way, every data point in the time series has a label associated with it.

Creating 2-D Images

We are using 4 features namely - gFx, gFy, gFz, and TgF, of the data to train the Neural Network model. The raw stream of data is first normalized column-wise using the l2 norm to normalize the scale of the data. The normalized data is converted into 2-D images by sliding a 64-size window through a transposed representation of the data. The window is slid across the data with a span size of 32. This step effectively generates 64-size 2-D representations of the data which are of the shape (4, 64) where 4 is the number of features and 64 is the size of the window. Since each image has only 1 label associated with it, the label at the end time stamp of the image window is considered the label for the image. This process created 330 images from the available time-series data.

Neural Network Architecture

The Deep Neural Network architecture that I have used to build the model is adapted from the research paper SensorNet: A Scalable and Low-Power Deep Convolutional Neural Network for Multimodal Data Classification (Jafari, et al., 2019). This architecture is optimized to run smoothly on low-power devices like SoCs, or embedded systems. The NN architecture consists of 5 convolutional layers, 3 pooling layers, and 1 fully connected dense layer. The output layer is a Softmax layer that predicts the class of the passed input images.

The filters in the convolutional layers generate the shared features between the different dimensions or modalities of the data. Since it is assumed that there is no spatial correlation among the different modalities of the data, a filter of size (4, 5) is used in the first convolutional layer, and filters of size (1, 5) are used in the subsequent convolutional layers. To pass these learned features to the dense layer, there are 3 max-pooling layers that flatten the image with a pool size of (1, 2) after the second, fourth, and fifth convolutional layers. This flattened image is then passed to the final 2 fully connected layers of the network. First is a 64-node dense layer, which is fully connected to a 5-node (as we have 5 output classes) Softmax output layer that gives the probability scores for the different output classes.

I have chosen the “ReLU” activation function, learning rate of 0.001, the “Adam” optimizer and the “Sparse Categorical Crossentropy” as the loss function for the NN layers based on the results of hyperparameter tuning, which is explained in a further section.

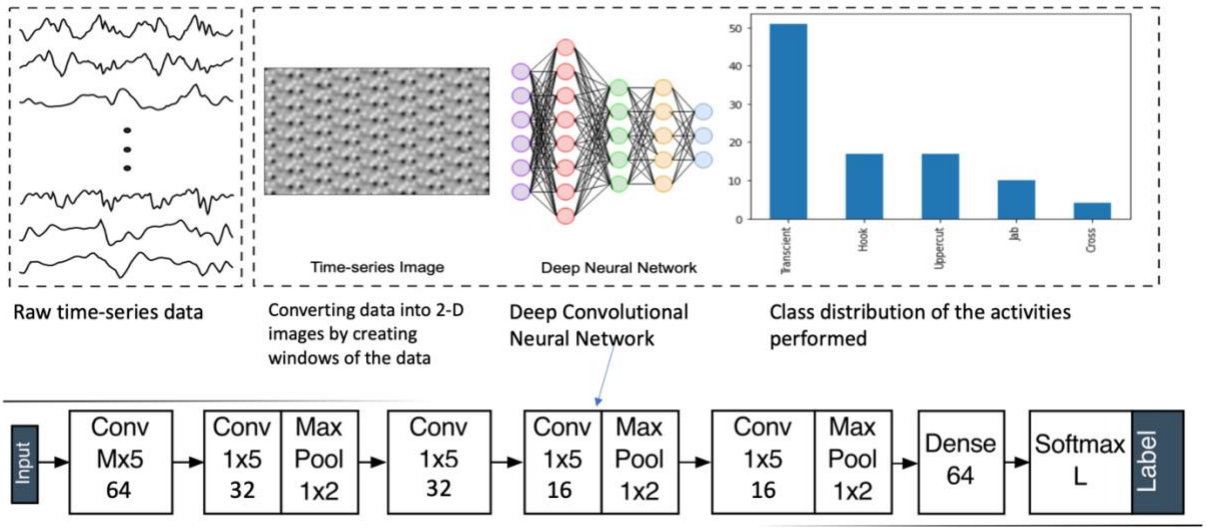

Figure 3 SensorNet Architechture

Table 2
This table contains the details of the architecture of the Neural Network.

Layer	Configuration
Conv Mx5 64	64 node convolutional layer
	4x5 filter size
Conv 1x5 32	32 node convolutional layer
	1x5 filter size
MaxPool 1x2	Max pooling layer
	Pool size (1, 2)
Conv 1x5 32	32 node convolutional layer
	1x5 filter size
Conv 1x5 16	16 node convolutional layer
	1x5 filter size
MaxPool 1x2	Max pooling layer
	Pool size (1, 2)
Conv 1x5 16	16 node convolutional layer
	1x5 filter size
MaxPool 1x2	Max pooling layer
	Pool size (1, 2)
Dense 64	64-node fully connected layer
Softmax L	5-node Softmax output layer

Model Pipeline

Now that we have described the input processing and NN architecture, the model pipeline is described in this section. The pipeline consists of 4 steps namely – input processing, data segmentation, model training, and model performance evaluation. The steps are described briefly below.

Input Processing. This step is where the normal input time series data is converted into 2-D images along with the corresponding labels.

Data Segmentation. This step is where we segment the data into training, validation, and testing sets. The proportions of the training, validation and testing data sets are 55%, 15%, and 30% respectively. The segmentation is done in the chronological order of the input images.

Model Training. This step is where we train the Neural Network model on the training data and then validate the learnings of the model with the validation data. The model I have used for this architecture is a Keras Sequential model, stacked with the layers as mentioned in the "Neural Network Architecture" section. The model is compiled with the "Adam" optimizer and a learning rate of 0.001. The model is trained with a batch size of 32 for 150 epochs and the loss function is "Sparse Categorical Crossentropy".

Model Performance Evaluation. This step is where we test the model performance by predicting the labels for the testing data. I have considered the F-1 Score and the ROC-AUC plots as the basis for measuring the performance of the model.

Hyperparameter Tuning

After the model pipeline was setup, I have trained the hyper parameters for the NN by making a grid of parameters to optimize by utilizing the "RandomizedSearchCV" algorithm from "SKLearn" Python module with 3 cross-fold validation. The "accuracy score" is used as the measure for optimizing the hyperparameters. The parameters that I have selected for tuning are as below.

Table 3

This tables lists out the hyperparameters optimized for the Neural Network using RandomizedSearchCV.

Tuned Parameter	Range	Optimized Value
Number of filters	[16, 32, 64, 128]	**64**
Activation function	[ReLU, Tanh]	**ReLU**
Epochs	[50, 100, 150, 200]	**150**
Batch size	[16, 32, 64, 128]	**32**
Loss function	["CategoricalCrossEntropy", "SparseCategoricalCrossEntropy"]	"**SparseCategoricalCrossEntropy**"
Optimizer	["Adam", "RMSProp", "SGD"]	**"Adam"**
Learning Rate	[0.1, 0.01, 0.001, 0.0001]	**0.001**

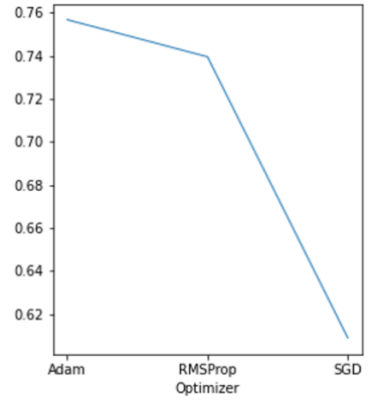

Figure 4(a)

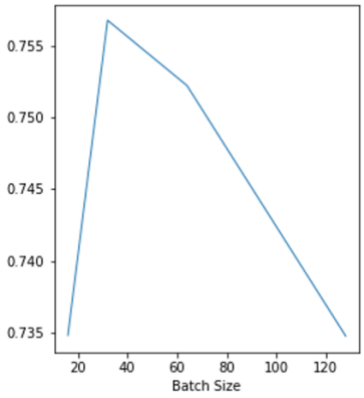

Figure 4(b)

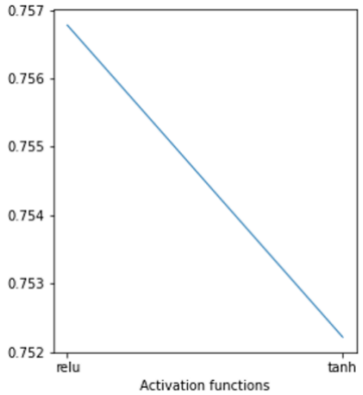

Figure 4(c)

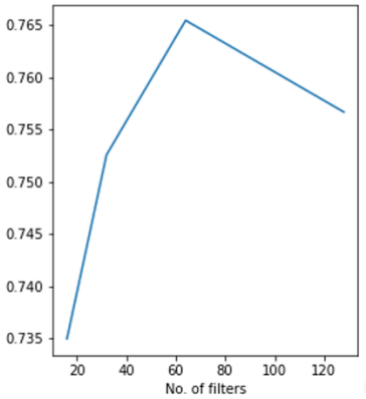

Figure 5(d)

Parameter tuning accuracy improvement from 4(a) Optimizer, 4(b) Batch Size, 4(c) Activation Function, 4(d) No. of filters,

Model Performance

The model performance is tested by evaluating the predictions of the model on the testing data. I have used “F-1 score” and “ROC-AUC” score and the “ROC” curves as the means to measure the model performance.

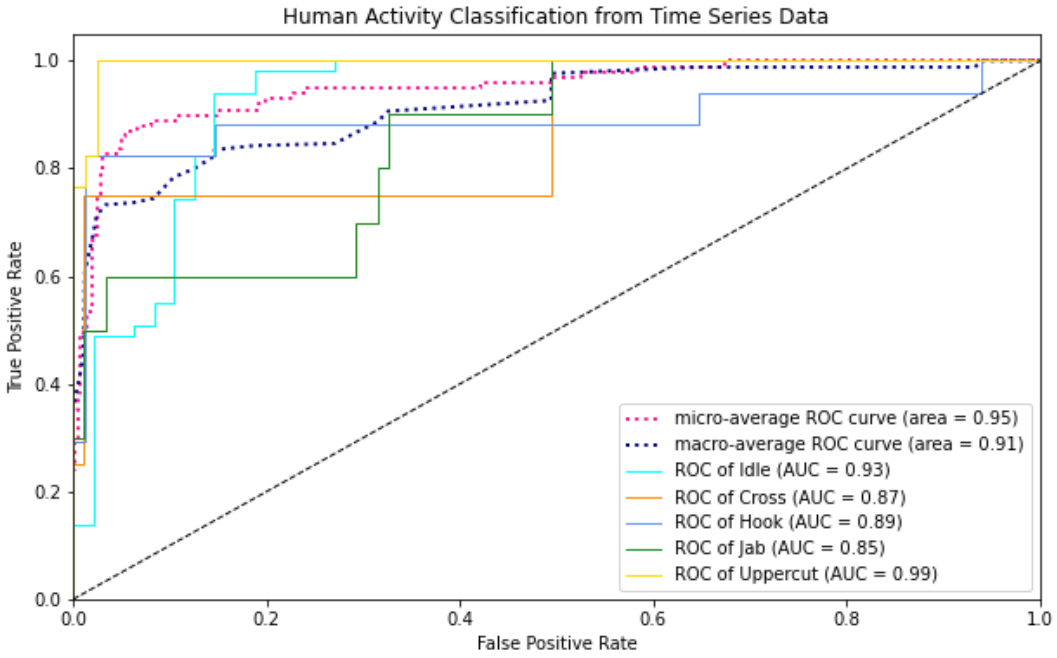

Figure 6 ROC-AUC curves for the output labels

Table 4

The following table lists the ROC-AUC accuracies and the F-1 scores of the model.

Output Label	ROC-AUC (Higher is better)	F-1 Score (Higher is better)
Idle	0.93	0.88
Cross	0.87	0.33
Hook	0.89	0.87
Jab	0.85	0.56
Uppercut	0.99	0.89
Overall	**0.95 (micro-average)**	**0.84**

Table 5
The classification report for the testing data is as below.

	Precision	Recall	F1-score	Support
Idle	0.87	0.90	0.88	51
Cross	0.50	0.25	0.33	4
Hook	0.93	0.82	0.87	17
Jab	0.62	0.50	0.56	10
Uppercut	0.81	1.00	0.89	17
accuracy			**0.84**	99
macro avg	0.75	0.70	0.71	99
weighted avg	0.83	0.84	0.83	99

With these metrics, we can conclude that the model's prediction accuracy is 95% (based on the micro-averaged ROC-AUC score) in predicting the labels.

Discussion

The SensorNet architecture (Jafari, et al., 2019) is a state-of-the-art Neural Network architecture designed for the task of classifying time series data. It also does not depend on the order of the features and is also fine-tuned to work well on multimodal time-series data. Additionally, SensorNet feeds on 2-D representations of the normalized raw input sensor data, which eliminates the need for preprocessing the signals before they are fed to the network. The window size can also be conveniently resized to suit the length and frequency of the activities performed. We can also decide the step size of the windowing function to control the overlap of the windows.

With the prior mentioned reasons in mind, I will explain my motive for selecting a Neural Network for this classification task. A statistical model would be trained on every single data point as opposed to windows of the data used to train the neural network. I think that this kind of sampling would train the model on learning patterns in the data at every timestamp rather than learning an action performed over a duration. SensorNet architecture is optimized to be computationally inexpensive due to its simplistic and straightforward architecture without sacrificing accuracy.

Future Work

In its current state, there is a lot of scope for improvement in this project. First, we can significantly improve the training and validation loss and accuracies of the model by generating more data and by collecting it from multiple users. This would enable us to generalize the model further and would enable the model to capture a broader pattern from data generated from punches thrown at different magnitudes and different angles by different people.

Second, the neural network is already optimized to be computationally inexpensive, but by utilizing TensorFlow GPU as the backend, we could possibly try a wider array of neural network architectures and also optimize the models with a larger hyperparameter grid to finely tune the model learning.

Third, since the SensorNet model is trained to capture human activities from time-series data, we can extend this architecture by tweaking the output layer to detect anomalies in the data stream. This process could include adding a bias in the output layer of SensorNet or training an anomaly detection model which can identify anomalies based on the outputs of SensorNet.

Lastly, this could be packaged to act as a backend to a real-time activity detection system. This kind of system could possibly be used for sparring, identifying illegal moves, etc., or it can be extended to other areas of human interaction classification or even adapt to other time-series classification tasks. This approach would be especially helpful as it requires neither signal processing nor an extensive knowledge of the variables.

Works Cited

Jafari, A., Ganesan, A., Thalisetty, C. S., Sivasubramanian, V., Oates, T., & Mohsenin, T. (2019). SensorNet: A Scalable and Low-Power Deep Convolutional Neural Network for Multimodal Data Classification. In *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS–I: REGULAR PAPERS,* (Vol. 66, pp. 274-287). IEEE.

Li, X., Zhang, Y., Zhang, J., Chen, S., Marsic, I., Farneth, R. A., & Burd, R. S. (2017). *Concurrent activity recognition with multi- modal CNN-LSTM structure.* Retrieved from Arxiv: https://arxiv.org/ abs/1702.01638

Radu, V., Tong, C., Bhattacharya, S., Lane, N. D., Mascolo, C., Marina, M. K., & Kawsar, a. F. (2017). Multimodal Deep Learning for Activity and Context Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1-27.

Yao, Hu, S. a., Zhao, S. a., Zhang, Y. a., Abdelzaher, A. a., & Tarek. (2017). DeepSense: A Unified Deep Learning Framework for Time-Series Mobile Sensing Data Processing. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 351–360). International World Wide Web Conferences Steering Committee.