HATRNet: Human Activity/Transition Recognition using Deep Neural Networks

Nicholas Gaudio, Akash Levy, Jonas Messner

March 18, 2019

1 Introduction

Human activity recognition (HAR) based on sensor data is a topic with great potential for customized healthcare. In developed countries today, most people own a smartphone with all the necessary sensor elements to perform HAR. We propose an end-to-end deep learning solution for categorizing accelerometer and gyroscope data into different activities/postural transitions to produce structured data suitable for health tracking.

Numerous implementations of human activity recognition have been demonstrated in prior studies. In [1], a support-vector machine (SVM) is used to classify activities and postural transitions. The approach taken does not employ an end-to-end strategy, and includes a lot of signal processing and feature extraction by hand. Frequency components of the data are obtained by performing a fast Fourier transform on the time-series data. The time and frequency data are then fed into a feature extractor which results in 561 features. These features are passed to the SVM model which predicts the activity with the feature dependent error rate of 3.22 % on the SBHAR (smartphone-based human activity recognition) dataset.

In [2], activities are classified by means of a convolutional neural network (CNN) by an architecture titled PerceptionNet. The PerceptionNet architecture automatically extracts the temporal dependencies of the time-series data and leverages the idea of late sensor fusion employing 1D convolutional (of filter size 1x15) and max pool layers in the early hidden layers followed by a 2D convolutional layer (of filter size 3x15) and global average pooling layer, where the sensor late fusion occurs. The PerpectionNet boasts the highest accuracy 97.25 % on the SBHAR test data compared to a CNN with early sensor fusion and LSTM model.

The goal of this work is to implement an end-to-end deep learning architecture without feature extraction. Our final end-to-end solution is similar to PerceptionNet [2], however we improve upon the existing work with: (1) more advanced preprocessing, (2) frequency data from fast Fourier transform over the inputs, (3) high-accuracy classification of transitions ([1] lumped all transitions into a single category, and [2] did not consider activity transitions at all), and (4) more thorough architectural optimization and hyperparameter search. Our code is made publicly available online under a permissive license[1]. We still incorporate the highly effective idea of late sensor fusion. While the reported error rates in previous works are already low, we distinguish the activities "sitting" and "standing" more accurately. Table 1 shows the confusion matrices reported in [1] and [2], respectively. These indicate the challenge of differentiating between the categories "sitting" and "standing". This misclassification can be a problem, especially when using activity recognition in health or nutrition applications. As pointed out in [3], around 114 kcal per day are additionally expended if performing work at a standing desk instead of the usual sitting desks in offices and schools. Our improvements to the architectures described in literature enable us to tackle this problem effectively.

[1] `https://github.com/akashlevy/HATRNet`

Table 1: Confusion matrices from [1] and [2].

	WA	WU	WD	SI	ST	LD	PT
WA	1834	64	5	3	2	0	1
WU	10	1743	51	5	5	0	16
WD	0	2	1671	1	7	0	1
SI	0	0	0	1875	**94**	6	3
ST	0	2	0	**109**	2049	0	1
LD	0	0	0	1	0	2148	2
PT	0	1	2	0	0	0	1036

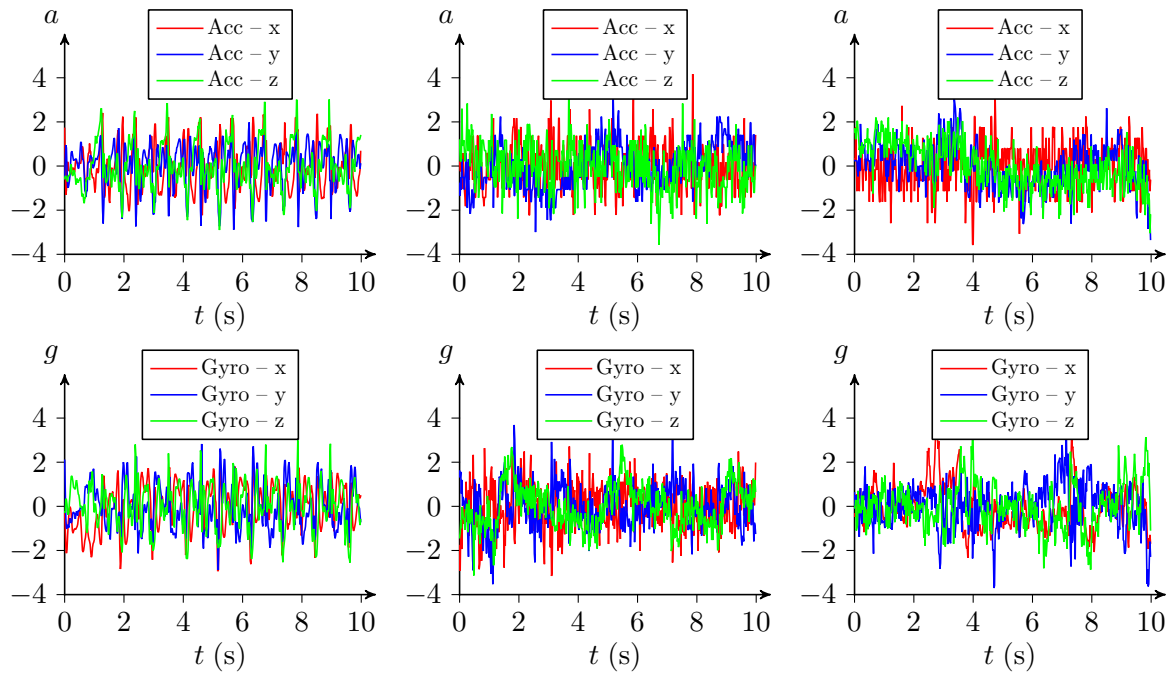
	WA	WU	WD	SI	ST	LD
WA	487	0	9	0	0	0
WU	2	468	0	0	0	1
WD	0	0	420	0	0	0
SI	0	2	0	443	**46**	0
ST	0	0	0	**16**	516	0
LD	0	0	0	0	0	537

WA: Walking, WU: Walking-Upstairs, WD: Walking-Downstairs, SI: Sitting, ST: Standing,
LD: Laying-Down, PT: Postural Transition

2 Dataset

Just as in [1], we use the SBHAR dataset with postural transitions. It contains 3-axial linear acceleration and 3-axial angular velocity, recorded at a rate of 50 Hz. The dataset is labelled with six activities (walking, walking upstairs, walking downstairs, sitting, standing, laying) and six postural transitions (stand-to-sit, sit-to-stand, sit-to-lie, lie-to-sit, stand-to-lie, and lie-to-stand). There are 1214 time-traces captured from a group of 30 volunteers aged 19-48 using the Samsung Galaxy S II smartphone. The dataset was captured at a rate of 50 Hz with the longest time-trace being of length 2032 data points and the shortest of length 73 data points.

Figure 1 shows time-series data of three samples from the dataset. All data traces are normalized to zero mean and unit variance. While the "walking" activity can be easily distinguished from the other two activities, differentiating "sitting" and "standing" activities from each other is much harder.

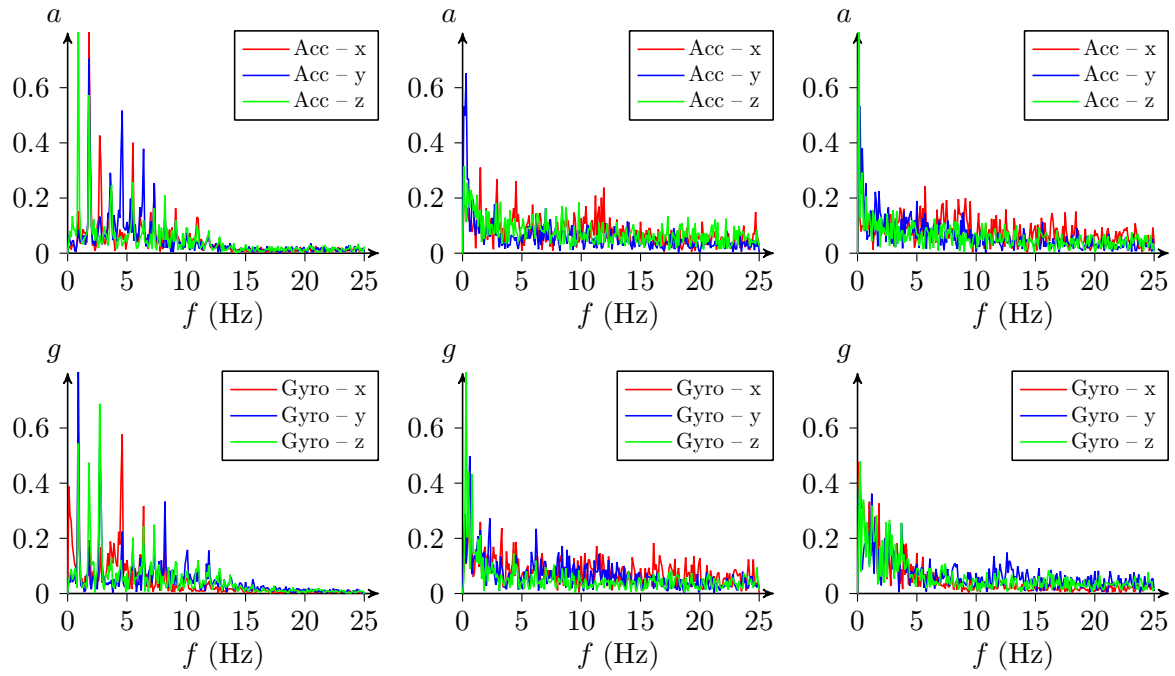


(a) Activity "walking". (b) Activity "sitting". (c) Activity "standing".

Figure 1: Time-series data of three samples.

Figure 2 shows the respective frequency data obtained by performing a fast Fourier transform on the normalized time-series data. It can be seen that the power spectral density plot of the

"walking" activity signal contains more energy at higher frequencies compared to the plots for 'sitting" and "standing". The power spectral densities of "sitting" and "standing" appear very similar to one another.



(a) Activity "walking". (b) Activity "sitting". (c) Activity "standing".

Figure 2: Frequency data of three samples.

In contrast to [1], HATRNet will involve no feature extraction beyond incorporation of frequency representations of the signals. As a result, our network will determine the features automatically, even from a relatively small dataset size of 1214.

Since our time-traces are all of varying length, a solution must be made to rectify these differences in length for training and testing for CNNs. Some ideas we have for this are:

- Zero-pad the time-series data to equal length
- Divide data into smaller, equal lengths e.g. 2s each (might be problematic for non-periodic transitions, throwing out data)
- FFT the time-series data (interpolation may be required), including both magnitude and phase
- Combine time and frequency data (requires zero-padding and interpolation)
- Use a sequence model, such as LSTM, which allows variable length time-series data

3 HATRNet

This chapter gives a short description of the proposed base architectures.

3.1 Fully-Connected Neural Network (FCNN)

The fully connected single layer, 20 neuron architecture was created to show the power of hand selected features. The authors of the UCI paper selected 561 features using techniques such as Butterworth low-pass filtering and variable calculation (i.e. mean, max). Sending in these 561 features into the just 20 neuron hidden layer network, the network was able to classify with an error rate of only 8.72 %.

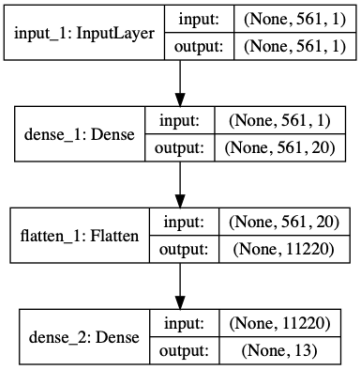


Figure 3: A densely connected network using the dataset's feature selection dataset.

3.2 Convolutional Neural Network (CNN)

A sanity check convolutional neural network was was created to ensure the plausibility of utilizing a CNN on the raw time-series data. Four 2D convolutional layers of filter size 12x3 were employed with a dropout rate of 0.40 to reduce over-fitting the training set. Filter sizes of 12x3 were performed to begin to fuse three channels' (of six) time-series data while maintaining same padding. The purpose of this CNN was solely to show that a CNN has the ability to properly classify the raw data. In the upcoming weeks, we will continue our literature search and begin to tune our CNN network for optimal performance.

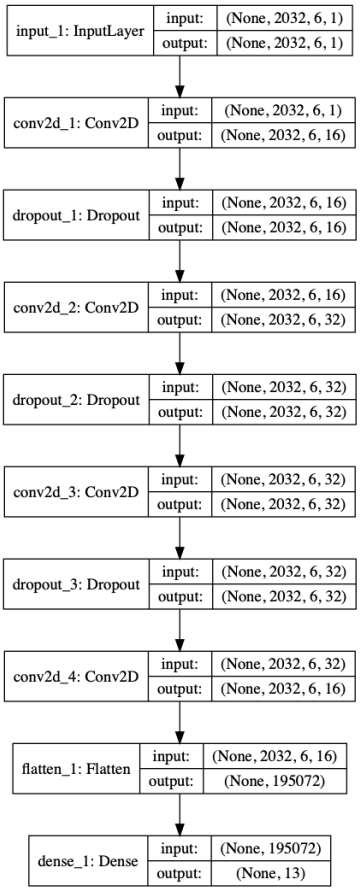


Figure 4: A sanity convolutional neural network using the raw time-series data.

3.3 Recurrent Neural Network (RNN)

We plan to test sequence models on our dataset, starting with simple recurrent neural networks (RNNs). In these models, we will pass activations across individual time samples in order to gain accuracy, so each layer will have multiply-sized inputs/outputs (see Figure 5). We have developed a simple model with a 50-unit recurrent layer with a ReLU activation function, and a 13-unit fully connected layer with a softmax activation function. Currently, this model produces low accuracy (44 %). However, we still need to tune the hyperparameters and make sure that we have configured the model correctly. We plan to continue to refine this model further for our final project report.

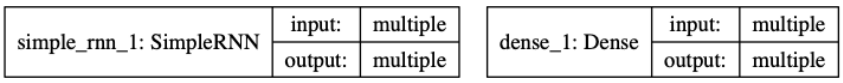


Figure 5: A sanity-check recurrent neural network using the raw time-series data.

4 Summary

All initially experimented architectures show classification promise. Our baseline, which we will attempt to surpass, even with end-to-end learning, is the 561 feature selected SVM which achieved an error rate of 3.22 %. With a dataset size of only 1214 time traces, end-to-end learning will be a challenge. To aid this, we will investigate also training on the frequency (with phase) information of our signals. We will also be furthering our literature understanding

to gain insight into more domain specific network architectures, as it has been found that hyperparameters such as non-square filter sizes may lead to increased classification accuracy.

Table 2: Comparison of neural network accuracy.

	FCNN	CNN	RNN	[1] (SVM)
Error Rate	8.72 %	24.05 %	56.04 %	3.22 %

References

[1] J.-L. Reyes-Ortiz, L. Oneto, A. Samà, X. Parra, and D. Anguita, "Transition-aware human activity recognition using smartphones," *Neurocomput.*, vol. 171, pp. 754–767, Jan. 2016.

[2] P. Kasnesis, C. Z. Patrikakis, and I. S. Venieris, "Perceptionnet: A deep convolutional neural network for late sensor fusion," *CoRR*, vol. abs/1811.00170, 2018.

[3] C. Reiff, K. Marlatt, and D. Dengel, "Difference in caloric expenditure in sitting versus standing desks," *Journal of physical activity & health*, vol. 9, pp. 1009–11, 09 2012.