AcctionNet: A Dataset Of Human Activity Recognition Using On-phone Motion Sensors

James Bartlett [1 2 3] **Vinay Prabhu** [1] **John Whaley** [1]

Abstract

Smartphones have become ubiquitous in modern society. With almost everyone carrying a smartphone in their pocket, the availability of sensor data (accelerometer, gyroscope, etc.) has sky rocketed. How we can use all this sensor data to benefit smartphone users remains an open problem. We present a new human activity recognition dataset, AcctionNet, we hope provides one avenue to explore this wealth of data. AcctionNet comprises 10.3 million data-points and 13 different activity labels. AcctionNet is, as far as the authors are aware, the largest human activity recognition dataset that includes more than 5 labels. We benchmark the dataset using various state of the art imaging methods for time series, as well as standard deep learning models.

1. Introduction

Large scale datasets and architecture engineering are the two most crucial constitutive elements of the institution of deep learning. While sifting through the important milestones in deep learning (Wang et al., 2017), we learn of pivotal moments such as the release of the MNIST dataset (Bottou et al., 1994) along with the best performing *LeNet-4* Convolutional Neural Network (CNN) architecture that achieved $\sim 1\%$ error rate on the test set. The ImageNet Large Scale Visual Recognition Competition (ILSVRC) (Russakovsky et al., 2015) that entails object category classification across a thousand object categories on the *ImageNet dataset* containing millions of images has been instrumental in ushering in new CNN architectures such as AlexNet (Krizhevsky et al., 2012), ZFNet (Zeiler & Fergus, 2014), GoogLeNet (Szegedy et al., 2015) and ResNet (He et al., 2016). Recently, the Stanford Medical ImageNet project [1] was announced that contains 0.5 petabyte of clinical radiology data, comprising 4.5 million studies, and over 1 billion images to facilitate reproducible science in the domain of medical image machine learning research.

This paper captures our ongoing attempt to prepare a similar large-scale dataset which we term, `AcctionNet` in the domain of time-indexed activity recognition using the Inertial Measurement Unit - MicroElectroMechanical Systems (IMU-MEMS) sensors such as accelerometers and gyroscopes available on commercial smartphones. In doing so, we have sought to merge together datasets from previous isolated efforts and augment our own in-house generated dataset that results in a combined dataset containing around 11 million data-points over 17 different activity labels.

The goals of the paper are as follows:

1: To disseminate a large-scale labeled IMU-MEMS sensor time-series dataset of human activity recognition obtained by collating pre-existing datasets.

2: To provide baseline prediction accuracies obtained upon deploying some standard deep neural networks architectures.

3: To disseminate the results of our empirical investigations into the different time-series imaging techniques proposed in time-series analysis (TSA) literature for pre-processing the raw sensor data before feeding it into the chosen deep neural network.

The rest of the paper is organized as follows. In Section 2, we describe the methodology followed for the dataset preparation and describe some of the statistics of the dataset. In Section 3, we briefly describe the 2 standard deep neural network architectures we used for the supervised activity class prediction problem. In Section 4, we briefly survey the time series imaging techniques we used to pre-process the raw sensor data before being input to the deep neural network chosen and in Section-5 we present the results obtained. We conclude the paper in Section-6 along with a description of the ongoing efforts to further expand the dataset.

[1]Unify ID [2]University of California, Berkeley [3]work done as a summer 2017 intern at UnifyID. Correspondence to: Corresponding Author <james@unify.id, vinay@unify.id>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, 2017. JMLR: W&CP.

[1] `http://langlotzlab.stanford.edu/projects/medical-image-net/`

2. Dataset: Preparation methodology and statistics

We created the AcctionNet dataset by collating 6 existent datasets and post-processing them to achieve cross-dataset normalization. At the end of this phase, we had 51460 samples spanning 13 disparate activity labels. The datasets used were: (Shoaib et al., 2013), (Shoaib et al., 2014), (McCall et al.), the 2011 Opportunity Activity Recognition Challenge, (Kwapisz et al., 2011), and (Chen et al., 2015). These datasets were in varying formats, had multiple different sensors, and were sampled at different frequencies. Some of these datasets included 3 axis gyroscope measurements, in addition to the 3 axis accelerometer measurements common to all. To facilitate learning, the datasets were converted into a common format and resampled to 200 Hz. Additionally, activities that were the same semantically but labelled differently were relabelled to reflect their semantic similarity. The 13 activities in the dataset are: biking, walking downstairs, stationary (gym) biking, jumping, lunging, running, transitioning from sitting to standing, squatting, standing still, transitioning from standing to sitting, treadmill running, walking upstairs, and walking. An example image from each class is shown in 7. The data were resampled to 200 Hz using linear interpolation. The norms of the 3 axes for the accelerometer and the 3 axes for the gyroscope were taken leading to two additional columns of data: accelerometer magnitude and gyroscope magnitude. Each second of data which contained only one activity was extracted and formed into a 200×8 or 200×4 image depending on whether this data included gyroscope or not, where the 200 rows are the 200 samples at 200Hz and the columns are the acceleration in the x, y, z directions, the magnitude of the acceleration, and the same for the gyroscope if it was included. The resultant dataset was a cube of shape $51460 \times 200 \times 5$ including the activity label.

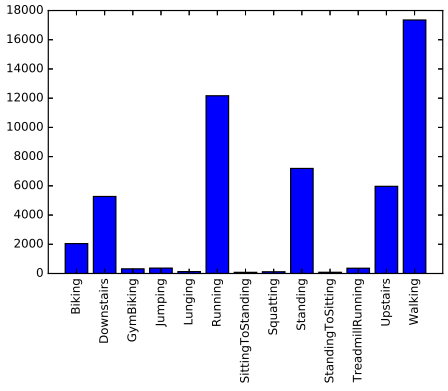


Figure 1. Label class counts in the AcctionNet dataset

The label class counts in the AcctionNet dataset are skewed and are as shown in fig 1.

3. Architectures of Deep neural networks used for supervised classification

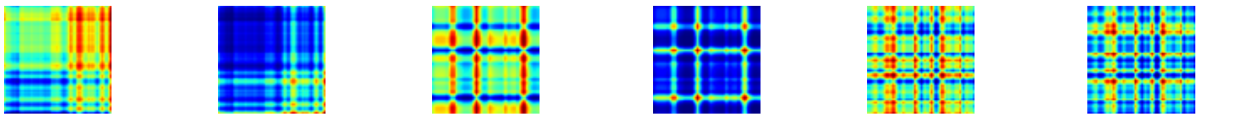
In this paper, we used 2 simple Convolutional Neural Network architectures, which we term as FCNN and CNN2 henceforth. As seen in fig 6(a), FCNN is a simple feed-forward architecture with 3 fully connected dense layers followed by a dropout (Srivastava et al., 2014) layer. The CNN2 architecture uses the standard `2Convolution-Nonlinear activation- Maxpooling - Dropout` paradigm and is as shown in Fig 6(b).

4. Pre-processing the sensor data: Imaging Methods

Inspired by the incredible accuracies achieved by convolution neural networks in the domain of computer vision, novel frameworks for encoding time series as different types of images have emerged recently such as the Gramian Angular Summation/Difference Fields (GASF/GADF) imaging technique and Markov Transition Fields (MTF). In this paper, we tried three different imaging methods to convert the sensor time series data into images before *feeding* them into the convolution neural networks described in Section 3.

4.0.1. Gramian Angular Summation/Difference Fields

Gramian Angular Fields (GAF), proposed in (Wang & Oates, 2015), encode angular information about the time series into an image. The image resulting from a GAF transformation is a quasi-Gramian matrix, computed on one of two inner-product spaces, $< x, y >= x \cdot y - \sqrt{1-x^2}\sqrt{1-y^2}$ and $< x, y >= y \cdot \sqrt{1-x^2} - x \cdot \sqrt{1-y^2}$, for Gramian Angular Summation Fields (GASF) and Gramian Angular Difference Fields (GADF) respectively. These Gramian matrices are equivalent to first converting a time series to polar coordinates such that $\phi_n = \arccos(x_n)$ and $r = \frac{t_n}{N}$ where $n \in \mathbb{N}$ and N is a normalization factor that scales all points to within the unit disc, then computing $\cos(\phi_i + \phi_j)$ (GASF) or $\sin(\phi_i + \phi_j)$ (GADF) for each $i, j \in \{0, ..., k\}$ where k is the number of samples in the given time series. Thus if the Gramian matrix is denoted as X then $X_{ij} = \cos(\phi_i + \phi_j)$ or $X_{ij} = \sin(\phi_i + \phi_j)$. Therefore, the image formed by this matrix represents angular information about the time series whilst still essentially encoding the original time series along the diagonal. The output images are downsampled to 64×64 using Piece-wise Aggregation Approximation (PAA) (Keogh & Pazzani, 2000). Six example images are shown in figure 2.

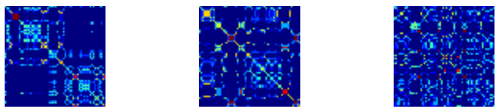


(a) GADF Biking (b) GASF Biking (c) GADF Walking (d) GASF Walking (e) GADF Squatting (f) GASF Squatting

Figure 2. Three examples of GADF (2(a), 2(c), 2(e)) and three examples of GASF (2(b), 2(d), 2(f)), all six images were created from the norm of the acceleration.

4.0.2. Markov Transition Fields

Markov Transition Fields (MTF), also proposed in (Wang & Oates, 2015), extend Markov Transition Matrices (MTM) to maintain information about temporal correlations. A Markov Transition Field is computed by dividing the time series into Q quantile bins, then computing a MTM for these bins, and finally stretching this $Q \times Q$ MTM into a $k \times k$ temporally correlated matrix, where k is the number of samples in a time series. The procedure for stretching a MTM into a MTF is as follows. Let q_i denote the quantile the time series is in at time i. Then for each $i, j \in \{0, ..., k\}$ compute the value of the MTF matrix at i, j as the value of the MTM at q_i, q_j. This forms a $k \times k$ MTF image. The output images are downsampled to 64×64 using Piecewise Aggregation Approximation (PAA) (Keogh & Pazzani, 2000). Three example images are shown in figure 3.

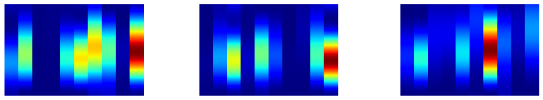


(a) Biking (b) Walking (c) Squatting

Figure 3. Three examples of MTF, all three images were created from the norm of the acceleration.

4.0.3. Spectrograms

Spectrograms have historically been used with much success in speech recognition, audio classification, and other time series analysis tasks, (Carbonneau et al., 2016). However, recently deep architectures trained on raw data have been outperforming training on "handcrafted" features like spectrograms (Thickstun et al., 2016). Nonetheless, we converted our human activity time series into spectrograms by computing spectrograms with 4096 bins and then cropping the images to frequencies below 20Hz, the typical range of frequencies for human motion (Antonsson & Mann, 1985). Three example images are shown in figure 4.



(a) Biking (b) Walking (c) Squatting

Figure 4. Three examples of spectrograms, all three images were created from the norm of the acceleration.

5. Baseline results on Dataset

A model was trained for each of the imaging techniques, and validated using 30-fold cross-validation. The mean accuracy for each model is shown in figure 5, with error bars corresponding to a 95% confidence interval. The model labeled "Raw" in the figure, corresponds to running the same convolutional architecture on a 200×4 *image* of the raw time series for each sensor axis. Each model in the figure labeled with the name of an imaging method corresponds to running a convolutional neural network on that imaging method. Additionally, standard deep learning methods were tested on the dataset. We tested both a fully connected network with 4 layers, and a deep recurrent model with Gated Recurrent Units (GRU) (Cho et al., 2014). For the fully connected neural network, we convert our 200×4 image into a 800 length vector. A fully connected network with ReLU activations performed the best on this dataset with a mean accuracy of 89.9%, followed closely by convolutional neural networks run on GADF imaging of the time series, which had a mean accuracy of 88.2%. A fully connected network with tanh activations, a convolutional model on raw data with ReLU activations, the convolutional model run on GASF images, a convolutional model on raw data with tanh activations, a convolutional model run on MTF images, a recurrent model (GRU), and a convolutional model run on spectrogram images, had mean accuracies of 86.7%, 85.3%, 84.5%, 84.1%, 78.1%, 65.1%, and 54%, respectively. Although a fully connected architecture with ReLU activations performed the best in terms of accuracy, we will show that this architecture is more susceptible to adversarial attacks than other architectures due to its relative linearity because of the ReLU activations.

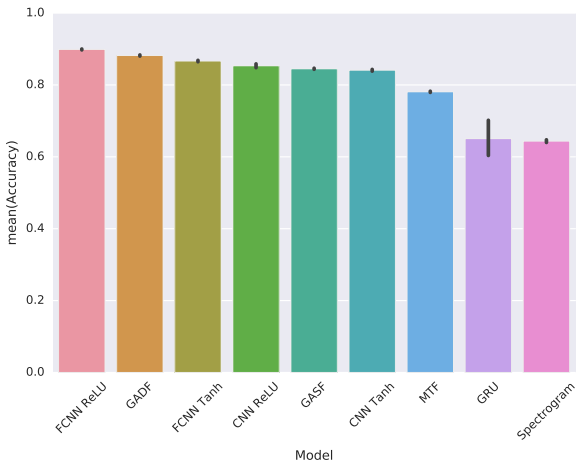


Figure 5. Barplot of mean accuracies for from left to right, fully connected network with ReLU activations, GADF, FCNN with tanh activations, CNN on raw data with ReLU activations, GASF, CNN on raw data with tanh activations, MTF, deep GRU and spectrogram.

6. Conclusion and ongoing work

We have contributed a human activity recognition dataset of significant scale by collating together 6 independent datasets. We subsequently carried out an empirical investigation on the efficacy of state of the art time series imaging techniques along with 2 standard deep learning architectures and establish the baseline results for the dataset. We are currently in the process of collating more independent datasets used in the literature before and also appending our in-house developed dataset, all with the goal of trying to create an *ImageNet* like dataset for the field.

Acknowledgements

We would like to thank Divyansh Agarwal for adding to this work by collecting an original data and collating more publicly available datasets. The full dataset can be found here: UnifyID dataset. We would also like to thank UnifyID for funding this research.

References

Antonsson, Erik K and Mann, Robert W. The frequency content of gait. *Journal of biomechanics*, 18(1):39–47, 1985.

Bottou, Léon, Cortes, Corinna, Denker, John S, Drucker, Harris, Guyon, Isabelle, Jackel, Lawrence D, LeCun, Yann, Muller, Urs A, Sackinger, Edward, Simard, Patrice, et al. Comparison of classifier methods: a case study in handwritten digit recognition. In *Pattern Recognition, 1994. Vol. 2-Conference B: Computer Vision & Image Processing., Proceedings of the 12th IAPR International. Conference on*, volume 2, pp. 77–82. IEEE, 1994.

Carbonneau, Marc-André, Granger, Eric, Attabi, Yazid, and Gagnon, Ghyslain. Feature learning from spectrograms for assessment of personality traits. *CoRR*, abs/1610.01223, 2016. URL http://arxiv.org/abs/1610.01223.

Chen, Chen, Jafari, Roozbeh, and Kehtarnavaz, Nasser. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pp. 168–172. IEEE, 2015.

Cho, Kyunghyun, Van Merriënboer, Bart, Gulcehre, Caglar, Bahdanau, Dzmitry, Bougares, Fethi, Schwenk, Holger, and Bengio, Yoshua. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Keogh, Eamonn J. and Pazzani, Michael J. Scaling up dynamic time warping for datamining applications. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, pp. 285–289, New York, NY, USA, 2000. ACM. ISBN 1-58113-233-6. doi: 10.1145/347090.347153. URL http://doi.acm.org/10.1145/347090.347153.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Kwapisz, Jennifer R, Weiss, Gary M, and Moore, Samuel A. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, 12(2):74–82, 2011.

McCall, Corey, Reddy, Kishore K, and Shah, Mubarak. Macro-class selection for hierarchical k-nn classification of inertial sensor data.

Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, Huang, Zhiheng, Karpathy, Andrej, Khosla, Aditya, Bernstein, Michael, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Shoaib, M., Scholten, H., and Havinga, P. J. M. Towards physical activity recognition using smartphone sensors. In *2013 IEEE 10th International Conference on Ubiquitous Intelligence and Computing and 2013 IEEE 10th International Conference on Autonomic and Trusted Computing*, pp. 80–87, Dec 2013. doi: 10.1109/UIC-ATC.2013.43.

Shoaib, Muhammad, Bosch, Stephan, Incel, Ozlem Durmaz, Scholten, Hans, and Havinga, Paul J. M. Fusion of smartphone motion sensors for physical activity recognition. *Sensors*, 14(6):10146–10176, 2014. URL http://www.mdpi.com/1424-8220/14/6/10146.

Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1): 1929–1958, 2014.

Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott, Anguelov, Dragomir, Erhan, Dumitru, Vanhoucke, Vincent, and Rabinovich, Andrew. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

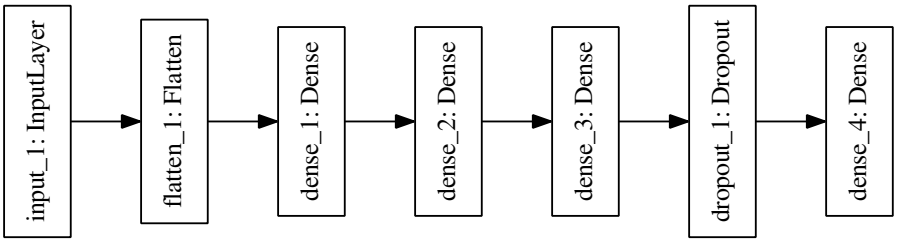
Thickstun, J., Harchaoui, Z., and Kakade, S. Learning Features of Music from Scratch. *ArXiv e-prints*, November 2016.

Wang, Haohan, Raj, Bhiksha, and Xing, Eric P. On the origin of deep learning. *arXiv preprint arXiv:1702.07800*, 2017.

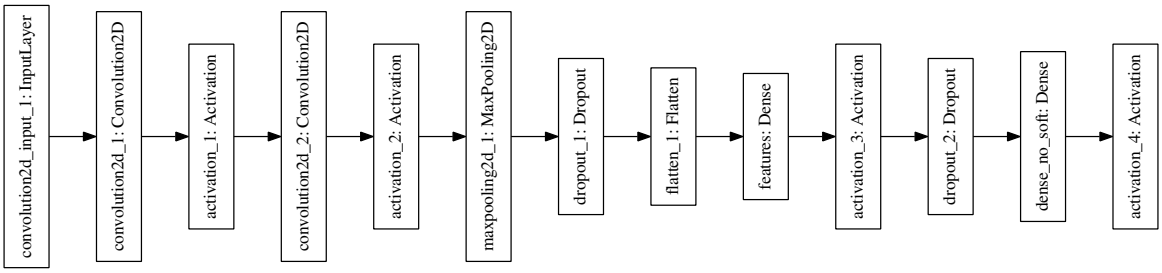
Wang, Zhiguang and Oates, Tim. Imaging time-series to improve classification and imputation. *CoRR*, abs/1506.00327, 2015. URL http://arxiv.org/abs/1506.00327.

Zeiler, Matthew D and Fergus, Rob. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Appendix: Visualizations

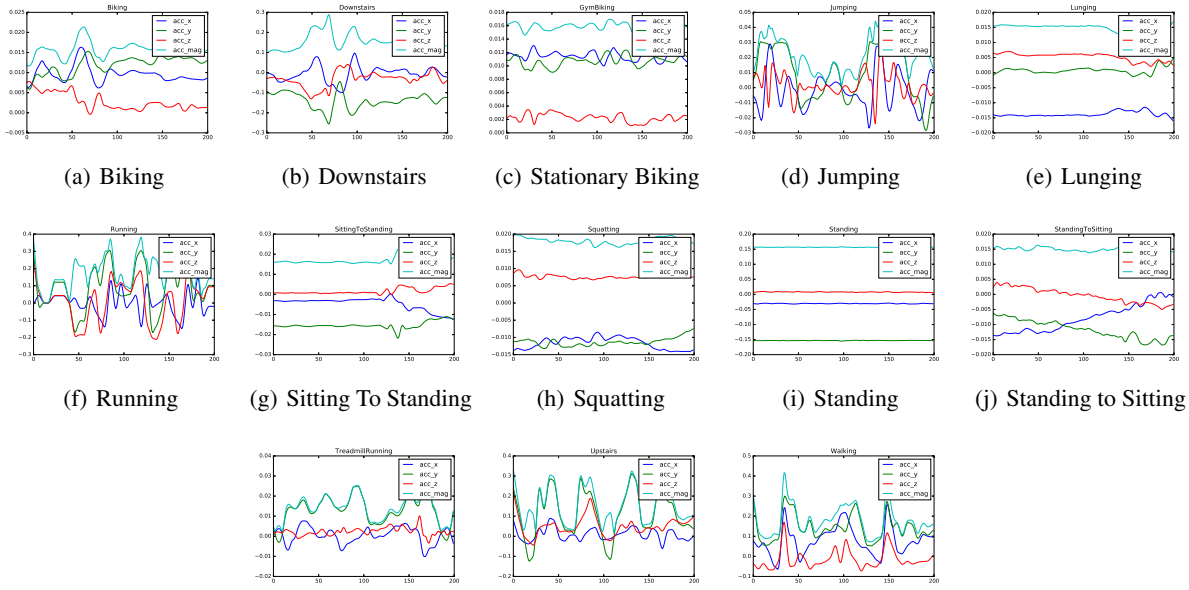


(a) The FCNN architecture



(b) The CNN2 architecture

Figure 6. This figure presentas the CNN architectures used. We term 6(a) as FCNN in the paper and 6(b) and CNN2



(a) Biking (b) Downstairs (c) Stationary Biking (d) Jumping (e) Lunging

(f) Running (g) Sitting To Standing (h) Squatting (i) Standing (j) Standing to Sitting

(k) Treadmill Running (l) Upstairs (m) Walking

Figure 7. An example of each class in the dataset, shown with all three accelerometer axes plotted, in addition to accelerometer magnitude plotted.