A Review of Human Activity Recognition via Wearable Sensors Using Neural Networks

Gregory Walsh, gw2g17@soton.ac.uk, MSc. Data Science

***Abstract*—Automated human activity recognition (HAR) via wearable sensors, the process of determining the activity of a person from wearable sensor data, has been a subject of research for many years. Recently, a subset of neural network (NN) models have emerged as state-of-the-art methods for performing HAR on wearable sensor data. In this paper, details of these approaches, as described in the literature on HAR, are reviewed. Specifically, recent papers which tackle HAR problems using multilayer perceptrons (MLP), convolutional neural networks (CNN) and recurrent neural networks (RNN). Original contributions and potential gaps in the recent literature are identified, and areas of further research are recommended.**

I. Introduction

Human activity can be described as a sequence of various body movements, and so it follows that the data generated by sensors measuring such activity is inherently chronological in nature. Since the goal of HAR via wearable sensors is to find a mapping from a set of chronological sensor measurements to a sequence of actions, one can consider HAR via wearable sensors as a sequence to sequence problem.

In performing sequence to sequence HAR, researchers face two key challenges. Many of the powerful classification methods, for example, support vector machines (SVM), are unsuited to time series inputs, since they work on the assumption that samples are statistically independent of one another [1]. To circumvent this restriction, the HAR problem is often relaxed to a straightforward classification problem by discretising the input time series data into fixed length windows of data, which are then treated as statistically independent (with an unquantified loss of information as a consequence) [2]. Furthermore, prior to the emergence of NN learning techniques, feature extraction often involved the manual specification and evaluation of a mix of statistical features which can be a labour intensive process requiring domain expertise [2].

In recent research, RNNs have been shown to address both of these issues, and CNNs have been shown to address the latter, whilst delivering competitive or state-of-the-art results [3], [4]. As explained by Hammerla et al., when using RNNs, specifically Long Short-Term Memory (LSTM) networks, the assumption that observations are independent need not be made, since these models are capable of maintaining, and updating as required, an internal state from one input sample to the next [3]. The need to perform manual feature engineering when using CNNs or LSTM networks is also eliminated since NNs are capable of extracting relevant features automatically, as discussed by Yang et al. [4].

In this paper, three NN models are briefly described, and their performance on a variety of wearable sensor HAR datasets, as reported in the literature, is evaluated. Furthermore, important original contributions and potential shortcomings which have been identified in the literature are discussed. Areas of potential future research, for the purpose of a summer Master's project, are also given.

II. Models for Classification

A. Multilayer Perceptrons

MLP classifiers are the simplest form of "deep" NN classifiers. Since they have no internal state, as with SVM and other similar models, any time series data must first be discretised into fixed length input windows and labelled with an associated target class.

MLPs appear infrequently in modern HAR research, and where they do, they have been used as a baseline model for comparison against more complex CNNs and LSTM networks [3]. This is somewhat surprising since they are capable of achieving results which are competitive with other high-performance models. For example, Hammerla et al. report for an MLP a weighted F_1 score [5] of 0.888 on the Opportunity gesture dataset [6] which exceeds the score of 0.851 reported by Yang et al. with their CNN approach (which until recently was the top score) [3], [4]. With such strong results, it seems reasonable to consider further HAR research to understand if their full potential has yet been reached.

B. Convolutional Neural Networks

CNN classifiers can be used in tasks where the input consists of data in which physical proximity between values in the input space, for example, adjacent pixel values in an image, corresponds to some latent relationship between those values [7]. Their success is in part due to their ability to automatically extract useful discriminating features, without requiring domain expertise, and this has made them attractive choices for HAR problems [4], [8]–[10]. Secondly, they show translational invariance meaning they are somewhat insensitive to spatial translations of extracted features, for example when recognising an object in an image, the size of the object will not significantly affect the network's ability to correctly classify it [7]. Drawing a parallel to HAR, sub-actions within an activity, such as individual steps when walking, may be closer or further apart in time depending on the subject, and so classification techniques which can deal with scaling in time/space make for attractive research candidates.

In the recent literature on HAR with CNNs, researchers consistently discretised the times series sensor data into equal length windows as a preprocessing step. However, there was less consistency in how these windows were subsequently prepared for the CNN [3], [4], [8]–[10]. For example, Jiang et al. re-imagine HAR as an image recognition problem by generating a 2D "activity image" for each discrete window of data, details of which can be found in their paper [8]. Whilst the method (which they call "DCNN") does produce reasonable results, it fails to outperform an SVM. Only when the model is stacked with an SVM (which they call DCNN+) do they achieve higher results, which naturally leads to speculation on whether the combination of different models is what results in higher performance, rather than some special property of their activity image, or of CNNs.

In contrast, other researchers opt to adapt CNNs to the 1D case, corresponding to the signals' temporal dimension [3] [10] [4] [9]. Interestingly, Ronao et al. show that that pooling size ($l = 2, \ldots, 15$) had very little effect on accuracy [9]. Similarly, the CNN proposed by Ordóñez and Roggen lacks pooling and yet it is still capable of almost state-of-the-art performance on several public datasets [10]. This perhaps hints that the translation invariance characteristic of CNNs with pooling layers, which is so useful in image recognition, may be less applicable for certain HAR problems. However, the HAR datasets which were analysed contain either simple short duration activities, such as opening a door, repetitive behaviours such as walking, or static activities such as sitting, rather than compound activities, such as making a cup of tea, which require a longer sequence of unique actions. As such, it may be that translation invariance is not significant when classifying primitive activities (with short durations or short cycles). To establish if this is the case, it may be necessary to evaluate CNNs on HAR datasets containing more complex activities.

With respect to convolution parameters, experiments on CNNs (all with pooling) of Ronao et al. found that short kernel lengths ($< 0.18s$ equivalent) are detrimental to network performance [9]. From the sensor sampling frequency and kernel lengths given in other papers, we calculate corresponding durations of 0.16s [4], 0.1-0.3s [3] and 0.16s [10], which fall close to or within the range recommended by Ronao et al. of 0.18-0.28s. Because of the consistency seen in kernel lengths across the literature, it is the author's opinion that further research into this hyperparameter should be given low priority.

Finally, an artefact of HAR with CNNs (resulting from the use of fixed sized windows) was observed wherein sequential predictions sometimes rapidly switch between classes at rates significantly greater than the rate at which different actions are performed, as shown by Ordóñez and Roggen in Fig. 6 of [10]. Only Yang et al. discuss a method to deal with the issue, borrowed from an earlier HAR paper [11], which smooths the predictions by taking an unweighted vote using a fixed number of windows either side of the current window. Since the method is fairly simple, there may be potential for improvement, for example by applying a weighted scheme based on distance to the window being classified.

C. Long-Short-Term-Memory Networks

LSTM networks provide a method for mapping sequential input data, such as a time series sensor data, to a variety of output formats, including another sequence, a vector, or simply a binary output [12]. Most noteworthy is their capability, once trained, to take into account data which has already passed through the network from previous inputs when making later predictions. This is made possible by the inclusion of a persistent internal mutable memory [13]. Given they are capable of directly modelling the sequential nature of time series sensor readings they make a natural candidate for research into HAR models.

A thorough review of LSTM networks is given by Hammerla et al. in which they present results obtained from three different network designs. Two unidirectional LSTM networks (networks which sequentially take in samples and output a response at each step) are presented first. In the first variation (named "LSTM-F"), predictions for some time t_i are made by feeding in, one sample at a time, a set of size L multi-sensor samples $\{\boldsymbol{x}_{i+1-L}, \ldots, \boldsymbol{x}_i\}$ stretching from time t_{i-L} to t_i. They do not describe whether the internal state of the LSTM units is persisted from one window to the next. Consequently, this method may suffer from the prediction switching artefact seen with CNNs, however, Hammerla et al. make no mention of this issue. They do however also introduce a further unidirectional LSTM network (named "LSTM-S") in which the only input to the network at time t_i is $\boldsymbol{x}_t$, thereby making the model suitable real-time classification.

Lastly, they present a bi-directional LSTM network, which like the LSTM-S model, does not use windows and so it makes no assumption of independence between samples, but in addition has a set of LSTM units into which the signals are fed in reverse order. Bi-directional LSTM layers are frequently used in situations where post hoc information can modify the significance or meaning of some earlier component of the signal, for example, our understanding of the word "fired" in the sentence, "she fired the clay", is different to that in, "she fired the employee" due to the change of noun at the end. This bi-directional model, as Hammerla et al. state, is only suitable for retroactive classification when one has all the data [3].

The reported relative performance of these models across several datasets is inconsistent, with each model taking the top spot in one of three tests, often by a significant margin (and with state-of-the-art performance in the Opportunity dataset). Partially on this basis, Hammerla et al. hypothesise that differences between these datasets (for example containing scripted versus unscripted activities) may cause variations in performance with different models, and on this basis make some recommendations about when to apply particular models. However, since a different performance metric is used for each dataset (total F1 score, mean F1 score by activity, and weighted F1 score by activity prevalence) and since these metrics are not directly comparable (since they each use different formulations of the harmonic mean [5]) it is not certain that the difference in rankings is due to performance of the models rather than the formulations of the statistics. On this basis, it is the opinion of the author that further, more carefully controlled comparisons of performance across multiple datasets, with standardised performance measures, are required to make reliable recommendations as to which methods are best suited to which situations.

In Ordóñez and Roggen's LSTM model, as with the LSTM-F model proposed by Hammerla et al, at each time t_i a trailing window of length L is used to select a set of past samples which will be fed into the model [10]. However, the window is first processed by a set of convolutional layers, which outputs a set of filter maps, collectively of length L and width d equal to the number of sensors multiplied by the number of kernels. Predictions for time t_i are then made by inputting samples (size $1 \times d$) one at a time into the LSTM network. Using this model, Ordóñez and Roggen reported, what was at the time, state-of-the-art results on the Opportunity and Skoda datasets, comfortably outperforming both their own baseline CNN and the previous top result published by Yang et al. by several percentage points [4]. What is not clear though is the degree to which this performance gap is due to the combination of the CNN and LSTM working together versus the behaviour of the LSTM layers alone since they make no comparison with an LSTM network without CNN preprocessing.

Ordóñez and Roggen's reasoning behind combining convolutional layers with LSTM layers is based on the success of this approach in voice to text problems. However, this reasoning is fairly weak, after all, the rate at which audio data is captured (typically 44.1KHz or 48KHz) is several orders of magnitude greater than the speaking rate of the average person (about two words per second) [14], and so extracting features via convolution and pooling is reasonable. On the contrary, with HAR via wearable sensors, the sampling rate is much lower, typically 30Hz [3], [4], [9], [10], which is over a factor of one thousand times smaller than the audio sampling rate. It is therefore questionable that preprocessing with a CNN would bring the same level of benefit to HAR with LSTMs as it does to the voice to text domain. Indeed, Hammerla et al. present results of an LSTM network without CNN preprocessing which outperforms Ordóñez and Roggen's hybrid network on the Opportunity dataset [3]. On this basis, it is the opinion of the author that additional experiments ought to be carried out comparing LSTM networks with and without CNN preprocessing to ascertain if adding CNN preprocessing does in fact improve the performance of LSTM networks on HAR problems.

III. Conclusion

In this paper, important contributions to the field of HAR via wearable sensors, and potential shortcomings in the research have been identified. Whilst it is clear from the published results that CNNs and LSTM networks can achieve state of the art performance, there is a lack of substantive evidence on the types of HAR problems to which CNNs versus LSTM networks are best suited. In particular, it has been noted that there exists a lack of consistency between researchers in the performance evaluation of new approaches, particularly with respect to which datasets to use, to how preprocessing is performed, and to which performance statistics are selected. There has also been minimal research published on the performance of standard MLPs on HAR problems. In light of these facts, additional carefully controlled experimentation comparing the performance of CNN, MLP, LSTM, and other hybrid models on a variety of HAR datasets (from scripted simple activities to unscripted complex compound activities) is recommended.

An exploration of smoothing methods, with the goal of reducing the negative effects of the switching artefacts which occur when using discretised time series data, is another potential area for research. Such research may also be applicable to other classification problems where, like HAR with CNNs/MLPs, sequential input data is first discretised into windows. Potential avenues for research include weighted voting schemes, and model stacking wherein a secondary model, given a neighbourhood of windows around the current window, is trained to predict the true value.

Finally, the author was unable to find published results on HAR using gated recurrent unit (GRU) networks. Like LSTM networks, GRU networks are naturally suited to performing classification on sequences, and in some instances they outperform LSTM networks [15]. Therefore evaluating GRU models on HAR datasets, alongside other models, could constitute an original and useful area of research.

References

[1] C. J. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Min. Knowl. Discov.*, no. 2, pp. 121–167.

[2] O. D. Lara and M. A. Labrador, "A Survey on Human Activity Recognition using Wearable Sensors," *IEEE Commun. Surv. Tutorials*, no. 3, pp. 1192–1209.

[3] N. Y. Hammerla, S. Halloran, and T. Ploetz, "Deep, Convolutional, and Recurrent Models for Human Activity Recognition using Wearables," in *Proc. 25th Int. Jt. Conf. Artif. Intell.*, pp. 1533–1540.

[4] J. B. Yang, M. N. Nguyen, P. P. San, X. L. Li, and K. Shonali, "Deep Convolutional Neural Networks On Multichannel Time Series For Human Activity Recognition," in *Proc. 24th Int. Conf. Artif. Intell.*, pp. 3995–4001.

[5] V. Van Asch. (2013) Macro and Micro-Averaged Evaluation Measures - UNPUBLISHED. [Online]. Available: https://pdfs.semanticscholar.org/1d10/6a2730801b6210a67f7622e4d192bb309303.pdf

[6] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Förster, G. Tröster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, H. Sagha, H. Bayati, M. Creatura, and J. Del R. Millàn, "Collecting complex activity datasets in highly rich networked sensor environments," in *INSS 2010 - 7th Int. Conf. Networked Sens. Syst.* IEEE, jun, pp. 233–240.

[7] Y. LeCun and Y. Bengio, "The handbook of brain theory and neural networks," M. A. Arbib, Ed., 1998, ch. Convolutional Networks for Images, Speech, and Time Series, pp. 255–258.

[8] W. Jiang and Z. Yin, "Human activity recognition using wearable sensors by deep convolutional neural networks," in *MM 2015 - Proc. 2015 ACM Multimed. Conf.*, pp. 1307–1310.

[9] C. A. Ronao and S.-B. Cho, "Human activity recognition with smartphone sensors using deep learning neural networks," *Expert Syst. Appl.*, pp. 235–244.

[10] F. J. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors (Switzerland)*, no. 1, pp. 115–149, jan.

[11] H. Cao, M. N. Nguyen, C. Phua, S. Krishnaswamy, and X.-L. Li, "An integrated framework for human activity classification," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, ser. UbiComp '12, 2012, pp. 331–340.

[12] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3104–3112.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[14] R. Allen and S. Anderson, *Speech in American Society*.

[15] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Gated feedback recurrent neural networks," in *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15, 2015, pp. 2067–2075.