**Paper Review: “Deep, Convolutional, and Recurrent Models for Human Activity Recognition using Wearables”**

* Deep learning replacing hand-crafted feature extraction and classification techniques
* Paper explores deep, recurrent, and convolutional approaches to HAR over 3 datasets that contain movement data from wearable sensors
* Investigates the suitability of models w/ different model configurations for different tasks in HAR, explore impact of hyperparameters, and provide guidelines to implement deep learn in their problem setting
* Up until DL approaches, the dominant technical approach in HAR in ubiquitous computing includes sliding window segmentation of time-series data captured w/ body-worn sensors, manually designed feature extraction procedures, and a wide variety of supervised classification approaches
* More elaborate behaviours (of interest in medical applications) pose a significant challenge to this manually tuned approach
* DL substituting for manually designed feature extraction procedures is improvement, as these lack the robust physiological basis that benefits other fields such as speech recognition
* Paper provides unbiased and systematic exploration of the performance of state-of-the-art DL approaches on 3 recognition problems typical for HAR in ubicomp (ubiquitous computing)
* In >4k experiments, investigates suitability of each model (being one of deep, convolution, or recurrent NNs) for different tasks in HAR and how suitable hyperparameters are for each
* Shows that RNNs outperform state-of-the-art and allow for novel types of real-time application of HAR through sample-by-sample prediction of physical activities
* Movement data from body-worn sensors are multivariate time-series data w/ relatively high spatial and temporal resolution (e.g. 20Hz – 200Hz)
  + Analysis of this in ubicomp is typically following a pipeline-based approach
  + First step: segment the time-series data into contiguous segments (‘**frames**’), either by signal characteristics like signal energy or by sliding-window segmentation
  + For each frame, set of features is extracted, including statistical features or stem from frequency domain
* Restricted Boltzman Machines initially used in previous experiments but were outperformed by combination of PCA analysis and statistical feature extraction process
* Most popular approach so far in ubicomp relies on CNNs
  + Have been applied to specific problem domains such as detection of stereotypical movements in autism, where they significant improved upon the ‘SotA’ (state of the art)
* However, individual frames are usually treated as statistically independent
  + However, approaches that are able to exploit the temporal dependencies in time-series data appear as natural choice for modelling human movement captured w/ sensor data
* RNNs w/ LSTM cells have been explored thus far in 2 settings
  + Investigated recurrent approaches to identify individuals based on their movement data from mobile phone
  + Also, compared performance of RNN to CNNs in 2 HAR datasets to find using both CNNs then RNNs more effective
* Thus far RNNs not been applied to model movement data at the lowest possible level (sequence of samples from sensors)
* Sole reporting from other studies on peak performance of a model (with hyperparameters tuned) doesn’t reflect the overall suitability of a method for HAR in ubicomp
  + Remains unclear how much effort went into tuning other approaches it was compared to
* Datasets used in this study include manipulative gestures, repetitive physical activities, and medical application of HAR in Parkinson’s
* To explore suitability of each model for HAR, reasonable ranges for each of their hyperparameters are chosen and random sample model configurations
  + Impact of hyperparameters are explored for each approach
* DFNNs implemented w/ up to 5 hidden layers w/ softmax-group
  + Each hidden layer contains same # of units and corresponds to linear transform w/ ReLU active funct
  + Dropout and max-in norm used (after each mini-batch, incoming weights of each unit in the network are scaled to have max Euclidean length ‘din’)
  + Input data fed into network corresponds to frames of movement w/ each consisting varying # of ‘s’ samples from Rd, which are concatenated into vector Ff ϵ Rs\*d
  + Mini-batch approach used w/ 64 frames and stratified w.r.t. class distribution in training set
* CNNs aim to enable translational invariance w.r.t. precise location in time of each pattern within a frame
  + Each CNN contains at least 1 convolution layer, 1 pool layer, and 1 FC layer prior to top-level softmax group
  + Max-pool width fixed to 2 throughout all experiments
  + Dropout also used here w/ fixed ‘p’ probs for each layer through all experiments and used max-in normalisation
  + Instead of concatenating the different input dimensions, the matrix structure is retained (Ft ϵ Rs x Rd)
* Implemented RNNs based on LSTM cells w/ no peephole connections in order to exploit the temporal dependencies within movement data
* LSTM cells are designed to counter the effect of diminishing gradients if error derivatives are backpropagated through many layers “through time”
* Each cell keeps track of an internal state, i.e. its “memory”
  + Over time, they learn to output, overwrite, or null their internal memory based on current input and history of past internal states
  + Leads to system capabale of retaining info across hundres of time-steps
  + Both forward LSTMs and bi-directional LSTMs are implemented (the latter contains parallel recurrent layers that stretch both forwards and backwards of current time-step, followed by a layer that concatenates their internal states for timestep ‘t’)
* Forward LSTM contextualises current time-step based on those it has previously seen (inherently suitable for real-time applications where, at inference time, the ‘future’ is not yet known)
* Bi-directional LSTMs, on the other hand, use both future and past context to interpret the input at timestep ‘t’ (makes them suitable for offline analysis scenarios)
* Three different settings used for RNNs
  + First: input data fed into network at time ‘t’ corresponds to current frame of movement data (certain length in time + dimensions concatenated as in DFNN)
  + Second: real-time application, where each sample of movement data is presented to the network in the sequence in which it was recorded
  + Third: sees the application of bi-directional LSTMs to the same sample-by-sample prediction prob
* Common applications for RNNs are speech recog and NLP
  + There, the context for an input (e.g. a word) is limited to it’s surrounding entities (e.g. sentence or paragraph)
  + Usually trains on complete sentences
* In HAR, context of an individual sample of movement data is not well defined (not beyond immediate correlations w/ neighbouring samples)
  + Likely depends on type of movement and wider behaviour content
  + Affects the choice of window length for sliding window segmentation
* To extract a mini-batch, extract L samples at a position and increase the position by L steps, possibly wrapping around at the end of the sequence
* Problem comes w/ training on long sequences
  + Training a sufficiently large RNN may memorise the entire input-output sequence implicitly, leads to poor generalisation
* To avoid this, introduces ‘breaks’ where internal states of RNNs are reset to 0
  + After each mini-batch we decide to retain the internal state of RNN w/ carry over prob ‘pcarry’, otherwise reset to 0
  + Novel form of regularisation of RNNs
* After each epoch of training, performance evaluated on the validation set
  + Each model trained for >= 30 epochs, max of 300 epochs
  + After 30, training stops if no increase in valid performance for 10 subsequent epochs
  + Select epoch that showed the best valid-set performance and apply it to the test set
* First dataset: ‘Opportunity’
  + Contains manipulative gestures like opening/closing doors (short duration and non-repetitive)
* Second: ‘PAMAP2’
  + Contains prolonged and repetitive physical activities typical for systems aiming to characterise energy expenditure
* Third: ‘Daphnet Gait’
  + Corresponds to a medical application where participants exhibit a typical motor complication in Parkinson’s that is known to have large inter-subject variability
* fANOVA analysis framework used to estimate impact of each hyperparameter on the performance observed across all experiments
  + Determines extent to which each hyperparameter contributes to network’s performance
* fANOVA builds predictive model (random forest) of the model performance as a function of model’s hyperparameters
  + Then decomposed into marginal and joint interaction functions of hyperparameters
  + Percentage contributed to overall variability of network performance is obtained
* Opted to estimate the mean f1-scores
* Large spread of peak performances between models on OPP and DG datasets
  + >15% mean f1-scores between best performing (b-LSTMs) and worst (FCNN) on OPP dataset
  + Smaller but still considerable at 7% on PAMAP2
* Best performing on OPP (b-LSTMs) outperforms current state of the art by a considerable margin of 4% mean f1-score
* Meanwhile, setup of CNN discovered here outperforms previous best by 5% mean f1-score for this type of model
* Good performance of RNN approaches (which model movement at a sample level) holds potential for novel real-time applications in HAR, as they alleviate the need for segmentation of time-series data
* For RNNs, carry-over probability of ‘pcarry’ = 0.5 works well for most settings
* Bi-directional LSTMS outperform current SotA on OPP dataset
* RNNs outperform CNNs significantly on activities that are short in direction but that have a natural ordering, where recurrent approach benefits from the ability to contextualise observations across long periods of time

**Significant Points and Takeaways from Paper**

* More elaborate behaviours in HAR (of interest in medical applications) pose a significant challenge to the manually tuned approach of feature extraction
  + DL substituting for manually designed feature extraction procedures is improvement, as these lack the robust physiological basis that benefits other fields such as speech recognition
* Individual frames in time-series data are usually treated as statistically independent
  + However, approaches that are able to exploit the temporal dependencies in time-series data appear as natural choice for modelling human movement captured w/ sensor data
* Sole reporting from other studies on peak performance of a model (with hyperparameters tuned) doesn’t reflect the overall suitability of a method for HAR in ubicomp
* In HAR, context of an individual sample of movement data is not well defined (not beyond immediate correlations w/ neighbouring samples)
  + Likely depends on type of movement and wider behaviour content
  + Affects the choice of window length for sliding window segmentation
* Problem comes w/ training on long sequences
  + Training a sufficiently large RNN may memorise the entire input-output sequence implicitly, leads to poor generalisation
  + To avoid this, introduces ‘breaks’ where internal states of RNNs are reset to 0
  + After each mini-batch we decide to retain the internal state of RNN w/ carry over prob ‘pcarry’, otherwise reset to 0 (novel form of regularisation of RNNs)
  + For RNNs, carry-over probability of ‘pcarry’ = 0.5 works well for most settings
* fANOVA analysis framework used to estimate impact of each hyperparameter on the performance observed across all experiments
  + Determines extent to which each hyperparameter contributes to network’s performance
* Good performance of RNN approaches (which model movement at a sample level) holds potential for novel real-time applications in HAR, as they alleviate the need for segmentation of time-series data
* RNNs outperform CNNs significantly on activities that are short in direction but that have a natural ordering, where recurrent approach benefits from the ability to contextualise observations across long periods of time