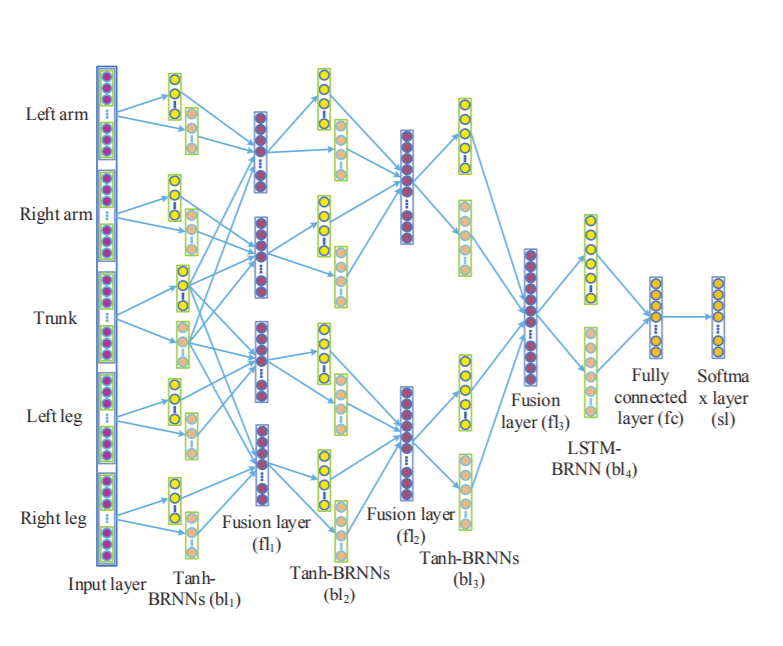
**Paper Review: “Hierarchical RNN for Skeleton Based AR”**

* Human actions can be represented by trajectories of skeleton joints
  + Traditional methods generally model spatial structure and temporal dynamics of human skeleton with hand-crafted features and recognize actions with well-designed classifiers
* Paper proposes end-to-end hierarchical RNN for skeleton based action recognition
  + Due to RNN modelling the long-term contextual information of temporal action sequence well
* Don’t take whole skeleton as input
  + Instead, divide it into 5 parts according to physical structure
  + Separately feed them to 5 subnets
* As number of layers increases, representations extracted by subnets are hierarchically fused to be the inputs of higher layers
* Final representations of skeleton sequences are fed into single-layer percepton and temporally accumulated output of the perceptron is the final decision
* Paper compares this with 5 other deep RNN architectures derived from the model to verify effectiveness of proposed network
  + Also compared with several other methods on 3 public datasets
  + Shows model achieves SotA performance with high efficiency
* HAR is important branch of computer vision, with applications in intelligent video surveillance, robot vision, HCI, etc.
* Human actions composed of motions of limbs and trunk which are represented by movements of skeleton joints in 3D
* Currently, reliable joint coordinates can be obtained from cost effective depth sensor using real-time skeleton estimation algorithms
* Skeleton-based HAR generally considered to be a time-series problem
  + Characteristics of body postures and dynamics over time are extracted to represent a human action
* Most existing skeleton-based AR methods explicitly model temporal dynamics of skeleton joints by using temporal pyramids and HMMs
  + TP methods generally restricted by width of time windows and can only utilize limited contextual information
  + For HMMs it is difficult to obtain temporal aligned sequences and corresponding emission distribs
* RNN and LSTM units recently used for HAR
  + Recent work uses just single layer RNN as a sequence classifier without part-based feature extraction and hierarchical fusion
* This paper, however, takes full advantage of deep RNN in modelling long-term context info of temporal sequences by proposing hierarchical RNN for skeleton-based AR
  + Temporal representations of low-level body parts are modelled by bi-directional RNNs and combined into representations of high-level parts
* Human body can be roughly decomposed into five parts: two arms, two legs, and one trunk
  + Human actions are composed of the movements of these body parts
  + Hence, paper divides human skeleton into the 5 corresponding parts and feeds them into 5 bidirectional RNN subnets (BRNNs) in first layer
* To model the movements from neighbouring skeleton parts, concatenate representation of trunk subnet w/ each of the other 4 subnets, then input those new 4 subnets’ concatenated results to 4 BRNNs in 3rd layer
* With similar procedure, representations of upper, lower, and whole body are then obtained in 5th and 7th layer, respectively (note that each concatenation is 1 layer and each BRNN is another)
* Finally, FC and softmax layer is performed on obtained representation to classify actions
  + LSTM neurons used in last BRNN to overcome vanishing gradient problem
* Also compares 5 other, deep RNN architectures derived from proposed model to verify effectiveness of proposed network and compare with several methods on 3 publicly available datasets
* Main contributions of paper summarized as follows
  + Arguably first to provide an end-to-end solution for skeleton-based AR by using hierarchical RNN
  + Verify effectiveness of necessary parts of proposed network by comparing w/ other 5 derived deep RNN architectures, including LSTM neurons in last BRNN layer and hierarchical skeleton-part fusion
  + Demonstrate proposed model can handle skeleton-based AR well w/o sophisticated preprocessing
* Use of RNN and perceptron can directly classify sequences without segmentation, unlike HMMs which need input sequences to be segmented and aligned
* Much previous work with RNNs just use them as a sequence classifier, while the paper proposes an end-to-end solution including both feature learning and sequence classifier
* With recurrent structure, RNN can model the contextual information of a temporal sequence
* LSTM contains one self-connected memory cell and three multiplicative units, which can store and access the long-range contextual information of a temporal sequence
* Bidirectional RNN (BRNN) proposed to utilize past and future context for every point in the sequence
  + Presents the sequence forwards and backwards to two separate recurrent hidden layers
  + These two recurrent hidden layers share the same output layer and can easily obtain the LSTM-BRNN just by replacing the non-linear units in BRNN with LSTM blocks
  + The outputs going to the output from each forward and backward layer are concatenated with each other
* Simple actions performed by only 1 of the 5 parts of the human physical structure, e.g. punching forward uses just one arm
  + Some actions come from moving the whole part of the upper or lower body, e.g. bending down
  + More complex ones are composed of motions of all 5 parts, e.g. running or swimming
* To effectively recognize various human actions, modelling the movements of these individual parts and their combinations is very necessary
* Proposed model differs in that it doesn’t need to model the spatial structure and temporal dynamics with hand-crafted features and well-designed classifiers
* Whole proposed hierarchical BRNN composed of 9 layers
* Layer 1:
  + 5 skeleton parts fed into 5 corresponding BRNN subnets
* Layer 2
  + Combine representations of trunk BRNN subnet with that of the other 4 subnets to obtain 4 new representations in order to model the neighbouring skeleton parts (since each arm and leg is a ‘neighbour’ with the trunk)
* Layer 3
  + These 4 representations are fed into 4 BRNN subnets, a la the first layer
* Layer 4
  + Representations of left-arm/trunk and right-arm/trunk BRNN are further combined to obtain upper body representations, with the same being done for the lower body
* Layer 5
  + Newly obtained 2 representations are fed into 2 BRNN subnets
* Layer 6
  + Outputs of these 2 subnets are fused again to represent the whole body
* Layer 7
  + Temporal dynamics of whole body representation further modelled by final BRNN
* These stacked BRNNs can be considered to extract spatial and temporal features of skeleton sequences
* Layer 8 and 9
  + FC layer and softmax layer are performed to classify the action
  + 
* Note that LSTM neurons are only adopted in the last layer of the architecture for one BRNN
  + First 3 BRNN layers all use ‘tanh’ activation function, which is a trade-off between improving representation ability and avoiding overfitting (in the case of using only LSTM units)
* Also, # of weights in LSTM is several times more than that in a ‘tanh’ neuron
  + Very easy to overfit network with limited training sequences
* BPTT algorithm used to obtain derivatives of objective function with respect to all the weights and minimize the maximum likelihood loss function by stochastic gradient descent
* Compare proposed network with 5 other architectures to very effectiveness of proposed model (which is called HBRNN-L)
  + 1st model: to prove importance of bidirectional connection, similar network w/ unidirectional connection proposed; hierarchically unidirectional RNN (HURNN-L)
  + 2nd model: to verify the role of part-based feature extraction and hierarchical fusion, compares a deep bidirectional RNN (DBRNN-L) which is directly stacked with several RNNS with the whole human skeleton as the input
  + 3rd model: compare deep unidirectional RNN (DURNN-L) which does not adopt both bidirectional connection and the hierarchical fusion
  + 4th and 5th models: to further investigate whether LSTM neurons in last recurrent layer are useful to overcome the vanishing/exploding gradient problem, investigate both DURNN-T and DBRNN-T, where they are both similar to their ‘-L’ counterparts but with ‘tanh’ activation function in all layers
* Note that all ‘B’ architectures have 5 learnable layers (4 recurrent/hidden and 1 fully-connected) and the number of neurons in FC layer is equal to that of action categories
* Three benchmark datasets are used to compare the model with other 5 architectures and recent work
* MSR action 3D dataset
  + Generated by Kinect-like depth sensor (widely used in HAR)
  + 20 actions performed by 10 subjects for 2 or 3 times
  + 557 valid samples with 22077 frames
  + Sequences captured at 15 FPS, each frame contains 20 skeleton joints
* Berkeley MHAD
  + Captured by multimodal acquisition system, optical motion capture system used to capture 3D position of active LED markers with frequency of 480Hz
  + Contains 659 sequences for 11 actions performed by 11 subjects with 5 repetitions of each action
* Motion capture dataset HDM05
  + Captured by optical marker-based technology with frequency of 120Hz containing 2337 sequences for 130 actions by 5 actors and 31 joints in each frame
* Varying joint numbers in each dataset for each of the 5 parts of the body (e.g. 4 joints for each arm in MSR, while 7 for each arm in MHAD)
* Human actions independent of absolute spatial position, hence normalise skeleton joints to unified coordinate system
* Adopt a Savitzky-Golay to preprocess data to improve signal-to-noise ratio of raw data (filters over 5 timesteps: current, previous 2, and next 2)
* Sample frames from sequences in fixed interval to reduce computation cost, as trajectories of skeleton joints vary smoothly
* When tested on MSR dataset, we can see that the proposed HBRNN-L achieves the best average accuracy and outperforms the other previously-created 4 methods in other papers with hand-crafted features
  + Performances of HURNN-L and DBRNN-L are promising
* Although 2 of the previous models achieve best performance in 2 of the 3 parts of the MSR dataset, the HBRNN-L model outperforms with average accuracy and is consistently better performing, indicating that it’s more robust to various data
* Fact that HBRNN-L obtains higher average accuracies than all other self-proposed 5 architecture variants prove the importance of bidirectional connection and hierarchical feature extraction
* Misclassifications mainly occur among several very similar actions
  + E.g. “Pick up and throw” often misclassified as “Bend….”
* For Berkeley MHAD dataset, can see that HBRNN-L achieves 100% accuracy with simple preprocessing and performs better than the 5 derived RNN architectures
  + Again, proves advantages of proposed model
  + Meanwhile, these 6 architectures achieve higher accuracy than 6 architectures from previous studies, which means that proposed model provides effective end-to-end solution for modelling temporal dynamics in action sequences
* For HDM05 dataset, proposed HBRNN model obtains SotA accuracy with standard deviation of 0.5 and 96.92% accuracy, with HURNN-L, DBRNN-L, and DURNN-L also obtaining excellent results
* Typical misclassifications here come between “Grab” and “Deposit” related skeletal sequences
  + Find that these categories of actions share similar spatial and temporal variations and map to the same 3 sub-actions in chronological order
  + Minor differences between the two make it difficult to distinguish these two kinds of actions
* Experiment showed that ‘-L’ models are easy to overfit while ‘-T- models always underfit during training, possibly due to the vanishing gradient problem
  + To overcome overfitting in ‘-L’ models, adopt strategies like adding input noise, weight noise, and early stopping (dropout doesn’t work well here), while for ‘\_T’ models use retraining strategy by tuning learning rate and adding various levels of input and weight noise
* HURNN-L achieves comparable perform to HBRNN-L but runs faster, so more suitable to online applications
* In future, can consider combining more features than skeletal joints into network, e.g. object appearance

**Significant Points and Takeaways from Paper**

* Skeleton-based HAR generally considered to be a time-series problem
  + Characteristics of body postures and dynamics over time are extracted to represent a human action
* Paper takes full advantage of deep RNN in modelling long-term context info of temporal sequences by proposing hierarchical RNN for skeleton-based AR
  + Divides human skeleton into the 5 corresponding parts and feeds them into 5 bidirectional RNN subnets (BRNNs) in first layer
  + Temporal representations of low-level body parts are modelled by BRNNs and combined into representations of high-level parts (e.g. output of arm BRNN fused with body trunk)
  + As number of layers increases, representations extracted by subnets are hierarchically fused to be the inputs of higher layers (representing upper body, lower body, and eventually full-body)
  + Final representations of skeleton sequences are fed into single-layer FC and softmax layer
* LSTM neurons used in last BRNN to overcome vanishing gradient problem
  + Note that LSTM neurons are only adopted in the last layer of the architecture for one BRNN
* Use of RNN and perceptron can directly classify sequences without segmentation, unlike HMMs which need input sequences to be segmented and aligned
* Much previous work with RNNs just use them as a sequence classifier, while the paper proposes an end-to-end solution including both feature learning and sequence classifier
* Bidirectional RNN (BRNN) proposed to utilize past and future context for every point in the sequence
* Compare proposed network with 5 other architectures to verify effectiveness of proposed model, with changing bidirectional to unidirectional connections, hierarchical fusion vs stacked RNN model, and using all ‘tanh’ activ functions or LSTM on last layer
* Three benchmark datasets are used to compare the model with other 5 architectures and recent work
* Adopt Savitzky-Golay to preprocess data to improve signal-to-noise ratio of raw data (filters over 5 timesteps: current, previous 2, and next 2)
* HBRNN-L model outperforms other previously-created 4 methods in other papers with hand-crafted features with average accuracy and is consistently better performing, indicating that it’s more robust to various data
* HBRNN-L obtains higher average accuracies than all other self-proposed 5 architecture variants proves the importance of bidirectional connection and hierarchical feature extraction
* Experiment showed that ‘-L’ models are easy to overfit while ‘-T- models always underfit during training, possibly due to the vanishing gradient problem