Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition

Yong Du, Wei Wang, Liang Wang
Center for Research on Intelligent Perception and Computing, CRIPAC
Nat'l Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

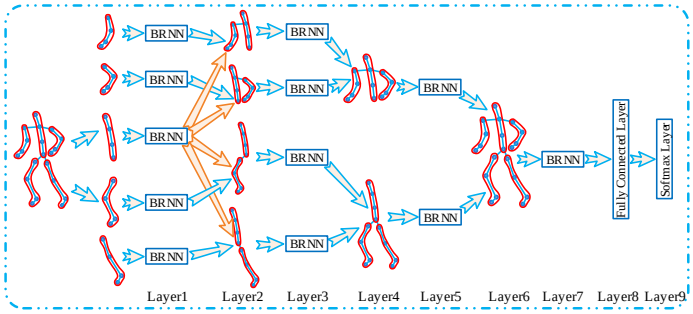
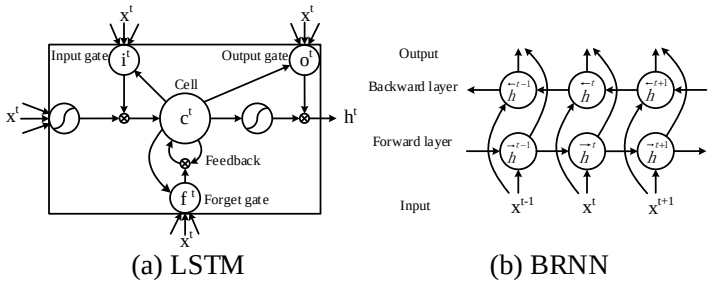


Figure 1: An illustrative sketch of the proposed hierarchical recurrent neural network. The whole skeleton is divided into five parts, which are fed into five bidirectional recurrent neural networks (BRNNs). As the number of layers increases, the representations extracted by the subnets are hierarchically fused to be the inputs of higher layers. A fully connected layer and a softmax layer are performed on the final representation to classify the actions.



(a) LSTM (b) BRNN

Figure 2: The architectures of long-short term memory (LSTM) and bidirectional recurrent neural network [1].

Human actions can be represented by the trajectories of skeleton joints. Traditional methods generally model the spatial structure and temporal dynamics of human skeleton with hand-crafted features and recognize human actions by well-designed classifiers. Most of the existing skeleton based action recognition methods explicitly model the temporal dynamics of skeleton joints by using Temporal Pyramids (TPs) [4] and Hidden Markov Models (HMMs) [5]. The TPs methods are generally restricted by the width of the time windows and can only utilize limited contextual information. As for HMMs, it is very difficult to obtain the temporal aligned sequences and the corresponding emission distributions. Recently, recurrent neural networks (RNNs) with Long-Short Term Memory (LSTM, Fig. 2) [1] neurons have been used for action recognition [2, 3]. All this work just uses single layer RNN as a sequence classifier without part-based feature extraction and hierarchical fusion.

In this paper, considering that recurrent neural network can model the long-term contextual information of temporal sequences well, we propose an end-to-end hierarchical RNN for skeleton based action recognition. Instead of taking the whole skeleton as the input, we divide the human skeleton into five parts according to human physical structure, i.e., two arms, two legs and one trunk, and feed them into five bidirectionally recurrently connected subnets (BRNNs, Fig. 2) in the first layer. To model the actions from the neighboring skeleton parts, we concatenate the representation of the trunk subnet with those of the other four subnets, respectively, and then input these concatenated results to four BRNNs in the third layer as shown in Fig. 1. With the similar procedure, the representations of the upper body, the lower body and the whole body are obtained in the fifth and seventh layers, respectively. Up to now, we have finished the representation learning of skeleton sequences. Finally, a fully connected layer and a softmax layer are performed on the obtained representation to classify the actions. It should be noted that, to overcome the vanishing gradient problem when training RNN [1], we adopt LSTM neurons in the last BRNN layer.

The objective function of our proposed model is to minimize the maximum likelihood loss function:

$$\mathcal{L}(\Omega) = -\sum_{m=0}^{M-1} \ln \sum_{k=0}^{C-1} \delta(k-r) p(C_k|\Omega_m) \quad (1)$$

where $\delta(\cdot)$ is the Kronecker function, r denotes the groundtruth label of the sequence Ω_m, and M denotes the number of the training samples. The back-propagation through time (BPTT) algorithm [1] is used to train our model.

In order to verify the effectiveness of the proposed network, we compare with other five different architectures derived from our proposed model. As illustrated before, we propose a hierarchically bidirectional RNN (**HBRNN-L**) for skeleton based action recognition (the suffix "-L" means that only the last recurrent layer consists of LSTM neurons, and the rest likewise). To prove the importance of the bidirectional connection, a similar network with unidirectional connection is proposed, which is called hierarchically unidirectional RNN (**HURNN-L**). To verify the role of part-based feature extraction and hierarchical fusion, we compare a deep bidirectional RNN (**DBRNN-L**), which is directly stacked with several RNNs with the whole human skeleton as the input. Furthermore, we compare a deep unidirectional RNN (**DURNN-L**) which does not adopt both the bidirectional connection and the hierarchical fusion. To further investigate whether LSTM neurons in the last recurrent layer are useful to overcome the vanishing/exploding problem in RNN, we examine another two architectures **DURNN-T** and **DBRNN-T**. Here DURNN-T and DBRNN-T are the similar networks to DURNN-L and DBRNN-L, but with the tanh activation function in all layers. All the six architectures have the same number of learnable layers.

In the experiments, we compare with the five derived deep RNN architectures mentioned above to verify the effectiveness of the proposed network, and also compare with several methods on three publicly available datasets. Experimental results demonstrate that our proposed method achieves the state-of-the-art performance. And final discussion about the computational efficiency demonstrates that our proposed model can perform well with high computational efficiency (for testing, about 52.46 ms per sequence, 234 frames per sequence).

[1] Alex Graves. *Supervised sequence labelling with recurrent neural networks*, volume 385. Springer, 2012.

[2] Alexander Grushin, Derek D Monner, James A Reggia, and Ajay Mishra. Robust human action recognition via long short-term memory. In *IJCNN*, pages 1–8. IEEE, 2013.

[3] Grégoire Lefebvre, Samuel Berlemont, Franck Mamalet, and Christophe Garcia. Blstm-rnn based 3d gesture classification. In *ANNML*, pages 381–388. Springer, 2013.

[4] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *CVPR*, pages 588–595. IEEE, 2014.

[5] Di Wu and Ling Shao. Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition. In *ICCV*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.