Poster: Attention-based Spatio-Temporal Model for HAR Using Multivariate Time Series

Rui Xi, Ming Li, Daibo Liu, Mengshu Hou
Department of Computer Science and Engineering
University of Electronic Science and Technology of China
{ruix.ryan,uestcliming,dbliu.sky}@gmail.com, mshou@uestc.edu.cn

Abstract

Recognizing human activities in a sequence is a challenging area of research in ubiquitous computing. Modeling spatiotemporal dynamics from multivariate time series plays a key role in recognizing human activities, unfortunately, the existing approaches could not extract spatial and temporal dynamics simultaneously.

This poster presents ABST, a deep learning framework that extracts spatiotemporal dynamics for discriminating human activities. We demonstrate how to utilize two GRUs along two dimensions of inputs to automatically learn features. In addition, considering that each time step pays different levels of attention to the final prediction, on the topmost of framework, we apply an attention-gated recurrent layer. Experimental results show a promising recognition performance, and outperforms the state-of-the-art methods.

1 Introduction

With the explosive growth of wearables, and mobile sensing devices in general, wearables-based human activity recognition (HAR) has attracted extensive attention of researchers in edge computing[1]. For example, accurately inferring an individual's current activity is critical to understand the context and situation of a user in a given environment, and as a consequence, personalized services can be delivered. However, the key of HAR is to model the discriminative spatiotemporal dynamics of different activities.

In this poster, we propose a novel attention-based spatiotemporal model for recognizing complex human activities – ABST. First, through add an interaction between two gated recurrent unit(GRU)[2], along the dimensions of inputs, we could extract spatiotemporal features at each time. Furthermore, we use an attention-gated recurrent units to control how much information is incorporated from the extract features of each time step. However, higher attention value will make the model to focus more on the current hidden state and its inputs. At last, we choose the outputs of the final time step to generates a probability distribution by a softmax function. Then, the predicted activity is with the maximum value. Experimental results show that our proposed model performs a very competitive efficiency of HAR. Taking dataset OPPORTUNITY for instance, it achieves the best recognition performance at a F_m score of 85.22%, while the best score ever is only 71%. Furthermore, it also achieves the best performance that the weighted F_1 score is 94.41% with a very little variance of 0.21%, which is also lower than others.

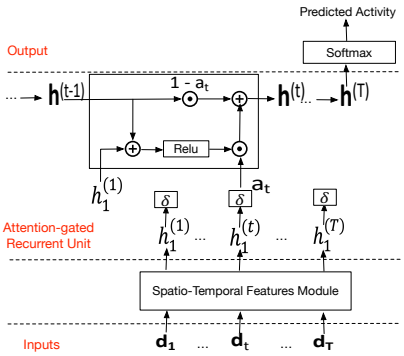

Figure 1. The graphical illustration of ABST.

2 ABST

As Fig. 1 depicts, ABST mainly consists of two module: spatiotemporal features extraction – for automatically extracting spatiotemporal features, and an attention-gated recurrent unit – for preserving the information of previous time steps with relevance modeled by attention scores. Next, we will give a detailed introduction to these two modules.

Spatio-Temporal Features Extraction: Given a sequence of time-series $D \in \mathbb{R}^{T \times n}$, herein, T is the time length and n means that D has n multivariate time-series each with a single component. We deploy recurrent cells along with all the dimensions: the temporal one of the sequence itself and the vertical one along the number of time-series. For efficiency and easy convergence, we choose GRU as the recurrent cell. According to GRU mechanism, we simplify its computations as follows, $(\mathbf{h}^{(t+1)}, \mathbf{m}^{(t+1)}) = GRU(\mathbf{H}, \mathbf{m}^{(t)}, \mathbf{W})$ wherein, the outputs $\mathbf{h}^{(t+1)}$, $\mathbf{m}^{(t+1)}$ are hidden vector and memory vector at time $(t+1)$, and $\mathbf{W}$

International Conference on Embedded Wireless Systems and Networks (EWSN) 2019
25–27 February, Beijing, China

ISBN: 978-0-9949886-3-8

represents a recurrent weight metric of the network. As the inputs, $\mathbf{H} \in \mathbb{R}^{2d}$ can be represented as $\mathbf{H} = \begin{bmatrix} I\mathbf{x}_i \\ \mathbf{h}^t \end{bmatrix}$, it is the concatenation of the new input $\mathbf{x}_i$ and the previous hidden vector $\mathbf{h}^{(t)}$, and I is a projection matrix.

Let us denote $\mathbf{h}_1 = \{\mathbf{h}_1^{(t)}\}$, $\mathbf{h}_2 = \{\mathbf{h}_2^{(t)}\}$ the input hidden vectors along with temporal dimension and vertical dimension, respectively. Meanwhile, in order to make the two GRU process to influence each other, we denote the input hidden vector $\mathbf{H} = \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{bmatrix}$ without $\mathbf{x}$. Therefore, the outputs of each dimension can be rewritten as

$$\begin{aligned} (\mathbf{h}_1^{'}, \mathbf{m}_1^{'}) &= GRU(\mathbf{H}, \mathbf{m}_1, \mathbf{W}_1) \\ (\mathbf{h}_2^{'}, \mathbf{m}_2^{'}) &= GRU(\mathbf{H}, \mathbf{m}_2, \mathbf{W}_2) \end{aligned} \tag{1}$$

Furthermore, due to that HAR is a sequential problem, we set the temporal dimension as the priority dimension, moreover, we also use its corresponding outgoing vectors to estimate the target. Therefore, the model first computes the transform for the vertical dimension obtaining the output hidden vectors $\mathbf{h}_2^{'}$. Then, it concatenates the output $\mathbf{h}_2^{'}$ and the input hidden vector $\mathbf{h}_1$ to form a new $\mathbf{H}$ as $\begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2^{'} \end{bmatrix}$. Subsequently, the output hidden vector $\mathbf{h}_1^{'}$ can be obtained.

Attention-gated Recurrent Units: On the consideration that each frame of a sequence of an activity pays different levels of attention on the final prediction, and only part of frames contain the most discriminative information. As the topmost of Fig. 1 shows, reference to [5], we integrate an attention weight to control how much information at the current time is incorporated for the final prediction. For the output $\mathbf{h}_1^{(t)}$ at time t, its attention score $a_t = \delta(\mathbf{w}\mathbf{h}_1^{(t)} + b)$, herein, $\mathbf{w}$ is the weight vector and b is the bias term. δ is a function to normalize the score in a value range of 0 to 1.

For classification, we choose the outputs at the final time step as inputs of softmax function to compute probability distributions. Then, the activity corresponding to the maximal probability is the predicted result.

3 Proof-of-concept Results

We implement ABST using Tensorflow library, and train it in a fully-supervised way by applying Adam to optimize all trainable parameters. Furthermore, Dropout and $L2$ regularization are both used to mitigate overfitting.

Recognition Performance: As TABLE 1 lists, we summarize the quantitative comparisons against other state-of-the-arts. From it, we could intuitively observe that our proposed model achieves the best performance on OPPORTUNITY and PAMAP2 datasets. As the last two columns shows, ABST achieves the best performance of 94.41% weighted F_1 score with a little bit of variance (0.21%). Moreover, it also achieves a great improvement by more than 14% F_m score. In addition, for PAMAP2, the proposed model achieve 94.1% F_m score, also outperforms other methods whose best F_m score is 93.7%.

Furthermore, Fig. 2 illustrates the percentages of each activity (red line), the recognition performance of D^2CL (blue line) and this proposed model (green line). We can observe that the training data is severely imbalanced that the activity 0 is almost 70% while others are very few. Even though, ABST (the green line) performs much more stable that the difference between the best and the worst is only 20%, while the difference for D^2CL is more than 50%.

Table 1. Comparisons on OPP. and PAMAP2 datasets.

	PAMAP2	OPPORTUNITY	
Performance	F_m	F_m	F_w
CNN [3]	0.937	0.591	0.894
DeepConvLSTM[4]	-	0.704	0.915
D^2CL[6]	0.932	0.7107	0.9197
ABST	**0.9410**	**0.8522**	**0.9441**

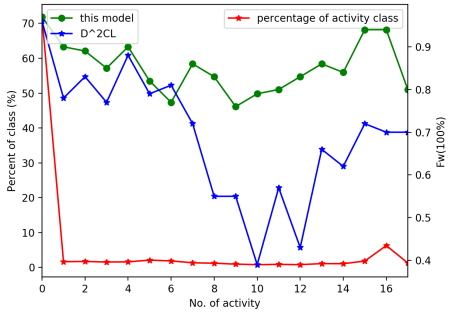

Figure 2. Performance on different gesture classes of OPPORTUNITY.

4 Conclusion

We present ABST, an efficient deep learning framework for activity recognition from multivariate time series. It is capable of extracting spatiotemporal dynamics by applying GRUs in two dimensions of inputs in parallel. Meanwhile, it combines an attention weight to control how much information at a time step is incorporated for the final prediction. Experimental results show that ABST can achieve remarkable performance in comparison with other state-of-the-arts.

5 References

[1] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *ESANN*, 2013.

[2] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[3] N. Y. Hammerla, S. Halloran, and T. Ploetz. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv preprint arXiv:1604.08880*, 2016.

[4] F. J. Ordóñez and D. Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.

[5] W. Pei, T. Baltrušaitis, D. M. Tax, and L.-P. Morency. Temporal attention-gated model for robust sequence classification. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 820–829. IEEE, 2017.

[6] R. Xi, M. Hou, M. Fu, H. Qu, and D. Liu. Deep dilated convolution on multimodality time series for human activity recognition. In *Proceedings of International Joint Conference on Neural Networks*. IEEE, 2018.