GRU- based Attention Mechanism for Human Activity Recognition

Introduction

Human Activity Recognition (HAR) is a domain of research aimed at recognizing human actions and movements from a series of observations. As the traditional machine learning models use low level representations, it lacks the characteristics of generalization. High level abstraction along with low level representations are necessary for a likely solution to this generalization problem.

Deep learning based methods deal with both low and high level representations of data. Therefore, recently two variants of deep learning methods namely convolutional and recurrent neural network models are become dominant over traditional methods in terms of performance.

Although deep learning based methods show promising performance, conventional sliding window based approach for CNN is unable to fully capture the temporal context of the sensor reading which is required for better classification of activities. For sequence data, RNN performs better than CNN in most cases as it captures sequence information. However, RNN faces long term dependency problems when the sequence is long enough. Note that, the sequence information found in HAR data is usually long. So, it is necessary to capture the long term dependency information for better classification.

Gated Recurrent Unit (GRU) is a variant of RNN which incorporate long term dependency information. It is expected that GRU will perform better in case of HAR data as it is able to capture temporal context of sensor data. It is noteworthy to mention here that, all temporal context are not equally important for classification, some are more important than others. Hence, it is necessary to give more attention to the important temporal context than others. Moreover, during the acquisition of HAR data, usually the training data found for different activities are not equal which causes class imbalance problem. The emphasis on important temporal context also helps to solve this problem. To capture the nature of different continuous movements and to extract the salient features, in this paper, we propose an attention mechanism based GRU model architecture. This architecture plays a crucial role in capturing context of the sensor reading and extracts the class imbalance tolerance characteristics. The main contributions of this paper are as follows:

- We propose to use a hierarchical temporal attention with GRU for capturing important temporal contexts.
- The hierarchical model propose to use here is parallelizable.
- The model is able to handle class imbalance problem.

Proposed method

The proposed method combines several building blocks for constructing the network. We use Gated Recurrent Units, two different types of attention mechanism which are described below:

A) Gated Recurrent Unit

GRUs are a variant of Recurrent Networks that is able to capture long term dependencies in temporal data while not suffering from similar vanishing gradient problem as regular RNN and requiring fewer parameters than LSTM. Hidden state h calculation is based on (3) with the input vector X and previous hidden state h going through update and reset gates in (1) and (2).

$$Z_{<t>} = \sigma(W_{zx} \cdot X_{<t>} + W_{zh} \cdot h_{<t-1>} + b_z) \quad (1)$$

$$\Gamma_t = \sigma(W_{\Gamma x} \cdot X_{<t>} + W_{\Gamma h} \cdot h_{<t-1>} + b_\Gamma) \quad (2)$$

$$h_{<t>} = (1 - Z_{<t>}) \Theta h_{<t-1>} + Z_{<t>} \Theta \tanh(W_{hx} \cdot X_{<t>} + W_{hh} \cdot (h_{<t-1>} \Theta \Gamma_t) + b_h) \quad (3)$$

Here, σ is used in (1) and (3) which refers to sigmoid function and Θ in (3) represents element-wise multiplication.

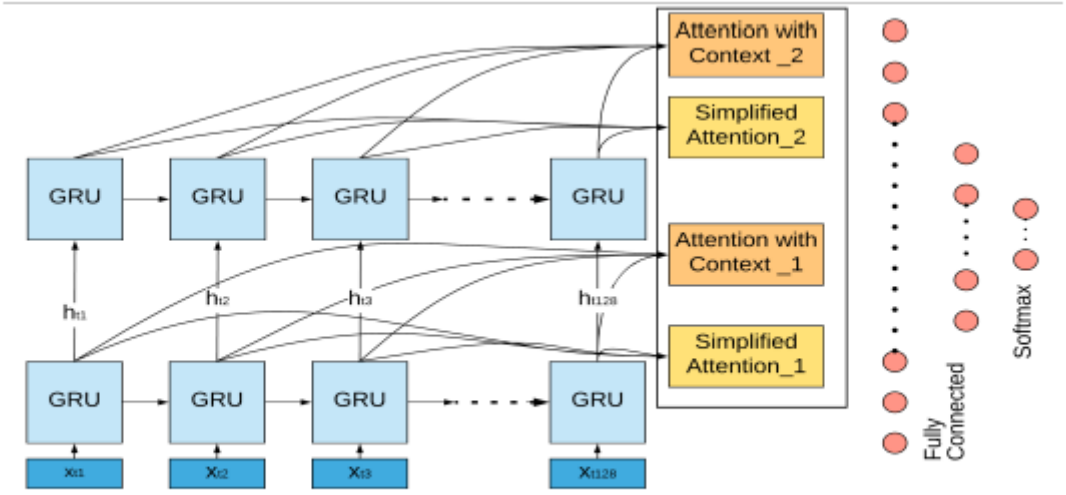

Fig. 1: Stacked 2 layer GRU model with simplified and context sensitive attention mechanism

able to exploit hierarchy of temporal features in the sensor data.

$$e_{<ti>} = \tanh(W_{ac} \cdot h_{<ti>} + b_{ac}) \quad (4)$$

$$\alpha_{<ti>} = \exp(e^{T}_{<ti>} \cdot e_s) / \Sigma_t \exp(e_{<ti>} e_s) \quad (5)$$

$$c_{<i>} = \Sigma_t \alpha_{<ti>} h_{<ti>} \quad (6)$$

W_{ac} and b_{ac} in (4) are parameters to be learned. e_s in (5) allows the preservation of context information and is learned jointly when training the network. A summation of the relative weights of the time steps is generated as the context vector in (6).

2) Simplified Attention: Sensor data of human activities differ from natural language data in that words are context dependent in most cases where sensor data may not depend on the context. Simplified attention does not have to learn the e_s parameter.

$$e_{<ti>} = \tanh (W_{as} \cdot h_{<ti>} + b_{as}) \quad (7)$$

$$\alpha_{<ti>} = \exp(e_{<ti>}) / \Sigma_t \exp(e_{<ti>}) \quad (8)$$

Context vector is obtained using (6) using the relative weight $\alpha_{<ti>}$ obtained from (8)

C) Proposed Model Architecture

A two layer stacked GRU architecture with attention applied to the hidden state outputs of each recurrent layer is proposed. The stacked layers help to learn more complex features from the sensor data. The attention scores from both layers are concatenated to create a hierarchy context vectors before feeding it to the densely connected layers. Both simplified and context-based attention scores are computed independently. Batch normalization is used before applying attention and after the concatenation of the attention scores. There are three fully connected layers after attention module. The first two layers are used to learn respective weights for the different types of features obtained from the attention modules. Rectified Linear Unit (ReLU) is used for the activation of these layers. Dropout is also performed for regularization with probability d1 and d2. Dropout prevents a neural network from becoming heavily dependent on a specific weight of a single neuron by turning off a fraction of the neurons randomly during training. For classifying human activities, softmax activation is applied to the final layer. The model architecture is illustrated in Fig. 1. We train the model with learning rate α and decay factor λ. Note that, a simpler model using a single GRU layer with same attention mechanism may also be used. Such a model requires fewer parameters to learn and resulting in less computational complexity. However, such simpler model may result in lower performance. In the proposed model architecture, sequential computation overhead may be reduced by computing the attention scores as well as the hidden states for the second GRU layer in parallel. In this

case we can achieve better performance with much less time. We use attention mechanism with GRU. Attention score may be calculated on derived features of CNN layers which may lead to the loss of context information. In comparison, using attention mechanism on the hidden states computed directly from sensor data facilitates better learning of the temporal contexts. In our proposed model, we use two types of hierarchy: firstly, hierarchy with two different types of attention (e.g. simplified and with context) and secondly, attentions from multiple layers of GRU. Using such a hierarchy of attention allows the incorporation of less complex features (low level information) with more complex features (high level representation of features) required for classification.

Experiment result and description

A) Dataset Description

In this experiment, we use the benchmark HAR dataset which provides time series information of six different activities(walking, walking upstairs, walking downstairs, sitting, standing and laying). Two types of sensor namely accelerometer and gyroscope are used to capture these information and randomly partitioned into 70% train and 30% test sets. Data has been preprocessed with noise-filters and sampled in fixed length sliding windows of 2.56 sec and 50% overlap which yields 128 readings per window. The training data is generated with a total of 7352 examples incorporating data of 21 randomly selected individuals. On the other hand, test set is composed of 2947 examples incorporating the remaining 9 individual subjects selected for this specific dataset.

B) Implementation Detail

During training, Adam optimizer is used with its default parameters ($\beta1=0.9$ and $\beta2=0.999$) for backpropagation. Moreover, the initial α is set to 0.001 which is presented in [35] with $\lambda = 0.2$ (decay based on validation loss). The dropout probability d1 and d2 are set to 0.25 and 0.1 respectively in the fully connected layers. The standard accuracy metric defined in 9 is used to evaluate the methods used in this experiment.

$$\text{Accuracy} = (T\,P + T\,N) / (T\,P + T\,N + F\,P + F\,N) \qquad (9)$$

where, TP, TN, FP and FN are defined as True Positive, True Negative, False Positive and False Negative respectively.

To evaluate the performances of the methods in the class imbalance scenario, we first drop half of the training data and second drop two-third of the training data. In both cases,

we keep unchanged test data (mentioned in the Dataset Description subsection). For measuring the performance of the methods in the class imbalance scenario, we use area under the curve (AUC) as an evaluating metric in this case. We also performed five fold training (in each case, we consider 17 individuals information of all classes except the imbalanced class) and test (in each case, test data is fixed with 9 individuals as mentioned in the dataset description subsection) with 50 percent dropout of the training data of a specific class that is considered as an imbalanced class.

C) Result and Discussion

In terms of accuracy, our proposed stacked GRU with attention performs better from both the proposed simplified model (with GRU + Attention) and the baseline CNN based method. Such an improvement is expected as we consider hierarchy in the attention model and two layer stacked GRU. Based on the results in this experiment, it is evident that the proposed method is better in terms of accuracy and AUC metric. This method is able to produce such result by giving the attention to both simple and complex features which are learned from the proposed combination of two attention mechanisms.