GRU-based Attention Mechanism for Human Activity Recognition

Md. Nazmul Haque
Institute of Information Technology
University of Dhaka
Dhaka, Bangladesh
bsse0635@iit.du.ac.bd

M. Tanjid Hasan Tonmoy
Dept. of CSE
University of Dhaka
Dhaka, Bangladesh
2015-116-770@student.cse.du.ac.bd

Saif Mahmud
Dept. of CSE
University of Dhaka
Dhaka, Bangladesh
2015-116-815@student.cse.du.ac.bd

Amin Ahsan Ali
Dept. of CSE
Independent University of Bangladesh
Dhaka, Bangladesh
aminali@iub.edu.bd

Muhammad Asif Hossain Khan
Dept. of CSE
University of Dhaka
Dhaka, Bangladesh
asif@du.ac.bd

Mohammad Shoyaib
Institute of Information Technology
University of Dhaka
Dhaka, Bangladesh
shoyaib@du.ac.bd

***Abstract*—Sensor data based Human Activity Recognition (HAR) has gained interest due to its application in practical field. With increasing number of approaches incorporating feature learning of sequential time-series sensor data, in particular the deep learning based ones has performed reasonably in uniform labeled data distribution scenario. However, most of these methods do not capture properly the temporal context of time-steps in sequential time-series data. Moreover, the situation becomes worse for imbalanced class distribution which is a usual case for HAR using body-worn sensor devices. To solve this issues, we have integrated hierarchical attention mechanism with recurrent units of neural network in order to obtain temporal context within the time-steps of data sequence. The introduced model in this paper has achieved better performance with respect to the well-defined evaluation metrics in both uniform and imbalanced class distribution than the existing state-of-the-art deep learning based model.**

***Index Terms*—Human Activity Recognition, Attention Mechanism, Gated Recurrent Unit**

I. Introduction

Human Activity Recognition(HAR) is a domain of research aimed at recognizing human actions and movements from a series of observations. The increasing public adoption of smart devices with sensors such as accelerometer and gyroscope has created the opportunity to organize considerable amount of sensor data for classification of human activity. The research activities focused on HAR incorporates the compilation of sensor readings into sequential time-series data and develops models for recognition of activities by analyzing the acquired sequential sensor readings. HAR poses a variety of promising application domain which includes physical activity annotation in the field of medical data analysis [1], personal assistant system [2], augmented and virtual reality [3] and many others.

In the past years, the state-of-the-art solutions to HAR mainly based on the traditional machine learning techniques. This techniques mainly depend on heuristic based hand-crafted feature engineering that rely on low level representations. As the traditional machine learning models use low level representations, it lacks the characteristics of generalization [4], [5]. 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 [6] and recurrent [7] neural network models are become dominant over traditional methods in terms of performance. For example, to identify HAR, convolutional neural network(CNN) is used in [8] and [9] and recurrent neural network(RNN) is used in [7] and [10].

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 [11] which is required for better classification of activities. For sequence data, RNN performs better than CNN in most cases as it captures sequence information [12]. However, RNN faces long term dependency problems [13] 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 [14]. 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

978-1-7281-3445-1/19/$31.00

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.

The paper describes related work in section II where different approaches for recognition of human activity are discussed. Section III has been used for describing the proposed methodology. Section IV contains the results as well as interpretations of the outcome of the proposed method. Section V concludes the paper.

II. Related Work

The most referred work proposed in [15] use fast Fourier transform algorithm for feature extraction, useful in recognizing different activities, that produce satisfactory results with numerous sensors set on distinctive parts of the body in conjunction with various data mining algorithms. Different approaches like K-nearest neighbors [16], decision trees [17], multi-class support vector machine [18] are used to classify human activities. All of these approaches require the use of hand-crafted features and show poor results for classifying in similar type of activities like walking down and walking up.

In the modern age, deep learning has become prominent in the area of learning models that represent features from low-level to high-level abstraction used in [4], [5] which allow to extract features automatically without hand-crafted feature engineering. A common form of neural network called fully connected neural network (FCNN) with Principal Component Analysis based feature technique is used in [8] and [9] for HAR and sensor data. But FCNN is very expensive in terms of memory (weights) and computation (connections). It also has a great chance of overfitting problem as every node is connected with every node in every layer. To extract additional features a new technique called Shift-invariant sparse coding [9] was proposed and used in combination with FCNN and handcrafted features. Convolutional neural network (CNN or convnet) [19] with dropout [20] for reducing overfitting is a recent breakthrough for feature extraction. It is used by [21] in gesture recognition that give state-of-the-art result. A hierarchical model using convnets is proposed in [22]. To recognize human activity for unlabeled as well as labeled data, [23] used semi-supervised convnet model to learn discriminative hidden features. Where convnet learns to recognize features of an object and combine these features to recognize larger object. Recurrent Neural Networks (RNN) based appoach proposed in [7] to recognize human activities and abnormal behaviour, shows some promise but leaves room for improvements. When the sequence is long, RNN faces log term dependency problems. To solve this problem, a combination of convnet and long short term memory [10] is used in [24] that outperforms other models on the KTH dataset. Gated Recurrent Units [25] are a variant of RNN that also addresses long term dependency issues. Adam optimizer [26] is a popular choice for training such neural networks.

Attention mechanism, introduced for sequence to sequence tasks tasks such as neural machine translation [27], speech recognition [28] has also been used for classification tasks in the domain of natural language processing [29]. The context vector computed by attention helps the network to learn where to focus on the representation generated by the encoder part for generating output sequence at each time step instead of compressing entire sequence to a fixed vector at once. Simplified form of attention mechanism [30] has been proposed for feed forward network which captures some long term dependencies.

The approaches described so far for activity recognition fail to capture the temporal context of sensor reading at different time steps of activity data which is required for better accuracy and generalization. Another approach proposed by [31] uses attention mechanism on top of a complex DeepConvLSTM architecture for finding relevant temporal context for activity recognition. In this work attention score is generated by applying attention after convolutional and pooling layers in DeepConvLSTM. This score does not reflect the hierarchy of simple features detected from raw sensor data and complex features detected from hidden state outputs in case of RNN (or from deeper layers in CNN). For finding relevant features for activity recognition, feature selection approaches used in [32], [33] can be applied. However, In this work, we propose an attention mechanism with GRU which distills more complex features that would be helpful for better classification.

III. 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 in the following sections.

A. Gated Recurrent Unit

GRUs are a variant of Recurrent Networks that have been shown to be 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 $h_{<t>}$ calculation is based on (3) with the input vector $X_{<t>}$ and previous hidden state $h_{<t-1>}$ going through update and reset gates in (1) and (2).

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

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

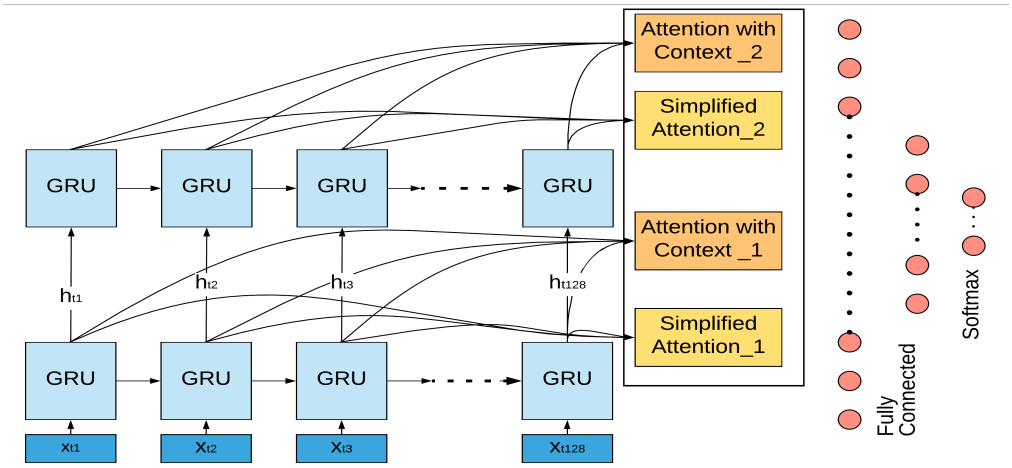

Fig. 1: Stacked 2 layer GRU model with simplified and context sensitive attention mechanism

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

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

B. Attention Mechanism for HAR

Context vector generated by attention allows a classification model to give varying levels of importance to the different temporal features generated by a recurrent network. Different variations of attention mechanisms such as simplified attention and hierarchical context based attention could be used for this type of tasks. These two types of attention could be used separately or a combination of their learned attention scores may be used. We take the latter approach.

1) Attention With Context: Because of susceptibility of sensor data to noise, context of a sensor value in relation to data on other time steps are helpful for creating a representation of data for classification. By using attention on both layers of stacked GRU, the model is able to exploit hierarchy of temporal features in the sensor data.

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

$$\alpha_{<t_i>} = \frac{\exp(e_{<t_i>}^T \cdot e_s)}{\sum_t \exp(e_{<t_i>} e_s)} \quad (5)$$

$$c_{<i>} = \sum_t \alpha_{<t_i>} h_{<t_i>} \quad (6)$$

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.

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

$$\alpha_{<t_i>} = \frac{\exp(e_{<t_i>})}{\sum_t \exp(e_{<t_i>})} \quad (8)$$

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

C. Proposed Model Architecture

We propose a two layer stacked GRU architecture with attention applied to the hidden state outputs of each recurrent layer. 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 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 d_1 and d_2. 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.

IV. Experiment Result and Description

A. Dataset Description

In this experiment, we use the benchmark HAR dataset [34] 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 (β_1=0.9 and β_2=0.999) for backpropagation. Moreover, the initial α is set to 0.001 which is presented in [35] with λ = 0.2 (decay based on validation loss). The dropout probability d_1 and d_2 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.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

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 this experiment, we have compared our proposed model architectures with a state-of-the-art baseline CNN based method [6]. The column two of TABLE I presents the comparison according to the experimental setup described in the HAR dataset. It is observed from this table that 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.

We have demonstrated the impact of class imbalance in TABLE I (Column 3-8) and TABLE II for different methods. In both cases (1/2 and 2/3 class drop), our stacked GRU with attention wins for all classes in terms of accuracy. Besides these, to test the generalization ability of the proposed method, we perform five fold training and test as mentioned in the implementation detail subsection. These results is demonstrated in TABLE III. In this case, we consider AUC as a performance metric as it performs better for imbalance dataset [36]. From this table, it is also observed that when imbalanced is injected, the proposed method wins for every classes individually which demonstrated the superiority of the proposed method.

Note that, the improvement of the proposed method with the baseline CNN model observed in the aforementioned tables is not that significant. The reason for such slight improvements can be explained by the simplicity of the dataset. In the dataset, the activities are separable by a large margin which is shown in Fig. 4. The figure is generated with dimension reduction using t-SNE [37]. So that when we create class imbalance,there is no significant drop in accuracies for both of the methods and the difference between these two methods in terms of accuracy also remains small. We believe, the improvement will be much larger for more challenging dataset.

We demonstrated the confusion matrix of the proposed method and baseline CNN in Fig. 2 and 3 respectively. The confusion matrix indicates that misclassification occurs rarely and also within the vicinity of ground truth class even in imbalanced class scenario. However, from these two figures, it is observed that the proposed method has fewer misclasssification compare to the baseline CNN.

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.

TABLE I: 1/2 Drop of Specific Class Data in Training Set and Measurement using given Test Data in terms of accuracy (%)

Model Architecture	Performance (Accuracy %)	Class - 1	Class - 2	Class - 3	Class - 4	Class - 5	Class - 6
CNN	94 .022	89.37902952	88.49677638	90.02375297	90.6345436	92.33118426	92.09365456
GRU + Attention	93.79	90.90600611	92.50084832	90.77027486	92.195453	92.36511707	92.05972175
Stacked GRU (2 Layers) + Attention	94.16355	92.02578894	91.78825925	92.53478113	91.9239905	93.07770614	94.46895148

TABLE II: 2/3 Drop of Specific Class Data in Training Set and Measurement using given Test Data in terms of accuracy (%)

Model	Class - 1	Class - 2	Class - 3	Class - 4	Class - 5	Class - 6
Baseline (CNN)	90.56667798	89.68442484	89.41296233	89.88802172	91.44893112	91.58466237
2-Stacked GRU + Attention	90.73634204	91.38106549	92.26331863	92.02578894	92.56871395	93.31523583

TABLE III: Subject-wise Class Drop in K-Fold and Measurement using given Test Data in terms of AUC

Model Architecture	k - fold	Class - 1	Class - 2	Class - 3	Class - 4	Class - 5	Class - 6
Baseline (CNN)	0	0.989897104	0.988005576	0.989058376	0.990233114	0.986666755	0.992600584
	1	0.982893556	0.986482948	0.984618942	0.981161394	0.99192285	0.984878402
	2	0.980885725	0.987111175	0.98570861	0.987078789	0.99061635	0.983950708
	3	0.989421716	0.988371167	0.984282503	0.983464878	0.99116295	0.982648438
	4	0.990690272	0.990301677	0.99302656	0.986080276	0.987868021	0.989988681
	Average	**0.9867576746**	**0.9880545086**	**0.9873389982**	**0.9856036902**	**0.9896473852**	**0.9868133626**
Stacked GRU (2 Layers) + Attention	0	0.9891089821	0.9917239879	0.9937438938	0.9885901347	0.9892608509	0.9915665231
	1	0.9898042512	0.9845693577	0.9922561365	0.9879438003	0.9901914546	0.9906566782
	2	0.992904892	0.990711794	0.992136949	0.988169278	0.991100119	0.992741053
	3	0.988042117	0.992132516	0.990958787	0.987103942	0.989934521	0.993510194
	4	0.991783724	0.993914724	0.989005982	0.992895329	0.992081285	0.992756454
	Average	**0.9903287933**	**0.9906104759**	**0.9916203497**	**0.9889404968**	**0.9905136461**	**0.9922461805**

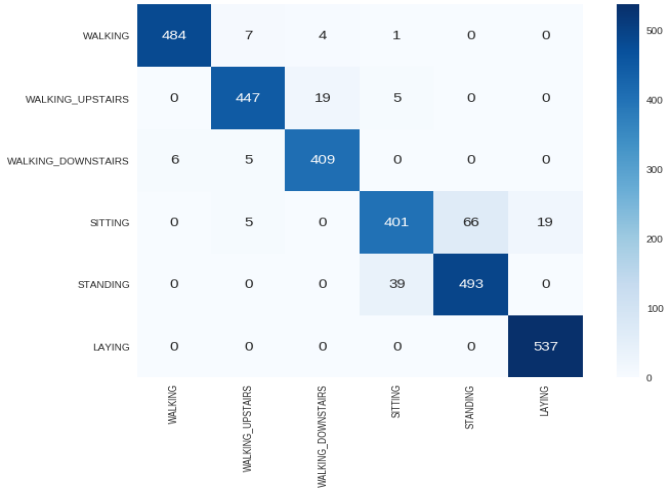

Fig. 2: Confusion Matrix when half of the training data for class 'Walking Downstairs' is dropped for stacked GRU model

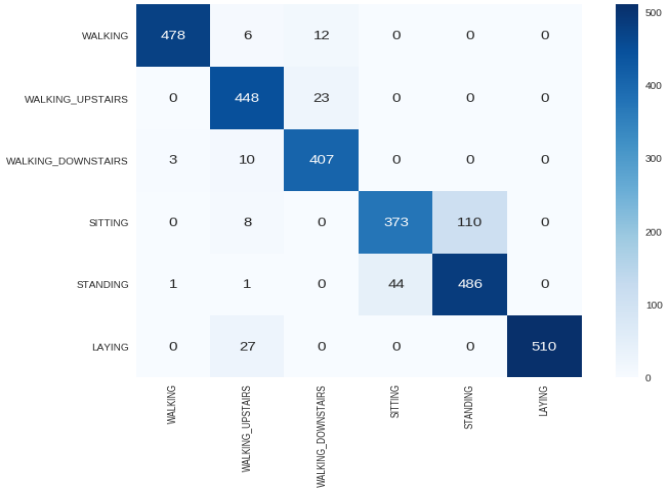

Fig. 3: Confusion Matrix when half of the training data for class 'Walking Downstairs' is dropped for CNN model

V. Conclusion

The proposed method performs reasonably well for the recognition of human activity while not being sensitive to class imbalance in the training data as it learns temporal features using attention mechanism. More parrallelizable models could be developed in the future by constructing embedding from the temporal data. In this paper, we intend to show the effectiveness of the proposed mechanism using a benchmark HAR dataset. We believe the proposed method will also perform better for more complex datasets, collected from heterogeneous devices that have more varied class distribution in terms of activity duration and feature complexity which we will address in future.

Acknowledgment

This research is supported by the University Grants Commission, Bangladesh under the Dhaka University Teachers Research Grant No Reg/Admin-3/54292-94.

References

[1] H. Song, D. Rajan, J. J. Thiagarajan, and A. Spanias, "Attend and diagnose: Clinical time series analysis using attention models," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

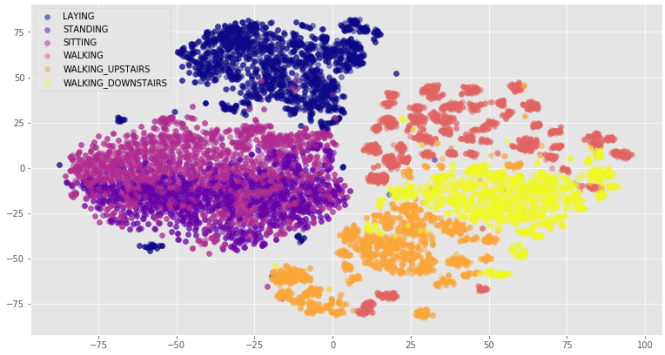

Fig. 4: Visualization of the separation of activities

[2] F. Montalto, C. Guerra, V. Bianchi, I. De Munari, and P. Ciampolini, "Musa: Wearable multi sensor assistant for human activity recognition and indoor localization," in *Ambient Assisted Living*, pp. 81–92, Springer, 2015.

[3] D. Van Krevelen and R. Poelman, "Augmented reality: Technologies, applications, and limitations," *Vrije Univ. Amsterdam, Dep. Comput. Sci*, 2007.

[4] Y. Bengio *et al.*, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[5] L. Deng, "A tutorial survey of architectures, algorithms, and applications for deep learning," *APSIPA Transactions on Signal and Information Processing*, vol. 3, 2014.

[6] C. A. Ronao and S.-B. Cho, "Human activity recognition with smartphone sensors using deep learning neural networks," *Expert Syst. Appl.*, vol. 59, pp. 235–244, Oct. 2016.

[7] D. Arifoglu and A. Bouchachia, "Activity recognition and abnormal behaviour detection with recurrent neural networks," *Procedia Computer Science*, vol. 110, pp. 86–93, 2017.

[8] T. Plötz, N. Y. Hammerla, and P. L. Olivier, "Feature learning for activity recognition in ubiquitous computing," in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[9] C. Vollmer, H.-M. Gross, and J. P. Eggert, "Learning features for activity recognition with shift-invariant sparse coding," in *International conference on artificial neural networks*, pp. 367–374, Springer, 2013.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, Nov. 1997.

[11] Y. Meng and A. Rumshisky, "Context-aware neural model for temporal information extraction," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018.

[12] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of cnn and rnn for natural language processing," *arXiv preprint arXiv:1702.01923*, 2017.

[13] Y. Bengio, P. Frasconi, and P. Simard, "The problem of learning long-term dependencies in recurrent networks," in *IEEE international conference on neural networks*, pp. 1183–1188, IEEE, 1993.

[14] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[15] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *International conference on pervasive computing*, pp. 1–17, Springer, 2004.

[16] W. Wu, S. Dasgupta, E. E. Ramirez, C. Peterson, and G. J. Norman, "Classification accuracies of physical activities using smartphone motion sensors," *Journal of medical Internet research*, vol. 14, no. 5, 2012.

[17] A. M. Khan, "Recognizing physical activities using wii remote," *International Journal of Information and Education Technology*, vol. 3, no. 1, p. 60, 2013.

[18] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *International workshop on ambient assisted living*, pp. 216–223, Springer, 2012.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 06 2014.

[21] S. Duffner, S. Berlemont, G. Lefebvre, and C. Garcia, "3d gesture classification with convolutional neural networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5432–5436, IEEE, 2014.

[22] J. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[23] M. Zeng, T. Yu, X. Wang, L. T. Nguyen, O. J. Mengshoel, and I. Lane, "Semi-supervised convolutional neural networks for human activity recognition," in *2017 IEEE International Conference on Big Data (Big Data)*, pp. 522–529, IEEE, 2017.

[24] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *International workshop on human behavior understanding*, pp. 29–39, Springer, 2011.

[25] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *CoRR*, vol. abs/1406.1078, 2014.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[27] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2015.

[28] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, pp. 577–585, 2015.

[29] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy, "Hierarchical attention networks for document classification," in *HLT-NAACL*, 2016.

[30] C. A. Raffel and D. P. W. Ellis, "Feed-forward networks with attention can solve some long-term memory problems," *CoRR*, vol. abs/1512.08756, 2015.

[31] V. S. Murahari and T. Plötz, "On attention models for human activity recognition," in *Proceedings of the 2018 ACM International Symposium on Wearable Computers*, pp. 100–103, ACM, 2018.

[32] N. X. Vinh, S. Zhou, J. Chan, and J. Bailey, "Can high-order dependencies improve mutual information based feature selection?," *Pattern Recognition*, vol. 53, pp. 46–58, 2016.

[33] S. Sharmin, M. Shoyaib, A. A. Ali, M. A. H. Khan, and O. Chae, "Simultaneous feature selection and discretization based on mutual information," *Pattern Recognition*, vol. 91, pp. 162–174, 2019.

[34] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," 01 2013.

[35] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[36] C. X. Ling, J. Huang, and H. Zhang, "Auc: a better measure than accuracy in comparing learning algorithms," in *Conference of the canadian society for computational studies of intelligence*, pp. 329–341, Springer, 2003.

[37] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.