ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse182942020.pdf
https://doi.org/10.30534/ijatcse/2020/182942020

End-to-End Human Activity Recognition using Convolutional Neural Network and Center Loss

Tam V. Nguyen[1], Manh Hoang[2], Binh A. Nguyen[2],Tung V. Nguyen[2], Phuong H. Lai[3], Nhan D. Nguyen[4]
[1]Dept. of Computer Science, Hanoi University of Science and Technology, Hanoi, Vietnam
tamnvhust@gmail.com
[2]ICT Department, FPT University, Hanoi, Vietnam. manhhhe130294@fpt.edu.vn, binhnase04865@fpt.edu.vn, tungnvhe130151@fpt.edu.vn
[3]Dept. of Computing Fundamentals, FPT University, Hanoi, Vietnam. phuonglh17@fe.edu.vn
[4]Dept. of Biomedical Engineering, Sungkyunkwan University, Suwon, South Korea. nhannd@skku.edu

ABSTRACT

Sensor-based human activity recognition (HAR) is an interesting research direction in the fields of healthcare, virtual reality, and other domains. In recent years, deep learning has been grown rapidly, and they have many successes in a wide range of domains such as computer vision, intelligent transportation, and human activity recognition. Previous studies that used deep learning-based methods to tackle the problem of human activity recognition did not consider the embedded features extracted from deep learning architectures, so it is hard to classify activities with similar patterns. In this paper, we would like to propose a deep learning-based method that takes the merits of Convolutional Neural Network (CNN) and Center Loss to recognize daily living activities. By stacking multiple CNN layers, we obtained an architecture robust for extracting features from sensor data, and applying Center Loss on embedded features makes our method more robust to discriminate between classes that have similar patterns. In experimental results, the proposed method achieves the accuracy rate of 94.2% F1-Score on the smartphone dataset.

Key words: Deep Learning; Human Activity Recognition; Convolutional Neural Network; LSTM; and Center Loss.

1. INTRODUCTION

In medical care, healthcare, sports monitoring, smart home, and many other applications, there is essential to capture activities that are performed by subjects. Many reasons make a recording of the problem of human activity recognition necessary daily such as a better understanding of patents' situations [1] or identifying training activities for athletes [2]. Due to its vast benefits, there have been many technologies and methods applied to on-body sensor-based human activity recognition [3, 4] in recent years.

Many studies used traditional machine learning to solve the problem of human activity recognition (HAR) [5]. Anguita et al. [6] used a Support Vector Machine classifier to classify seven motions, such as walking, running, and jumping. However, almost all traditional machine learning methods have a serious drawback, which is that input features have to be discriminative. Recently, with the power of deep learning technology [7], there are many studies [8] applied deep learning on the problem of human activity recognition. For instance, Wan et al. [9] presented a smartphone sensor-based Convolutional Neural Network (CNN) architecture to classify activities. Furthermore, the authors investigated the impact of hyper parameters over five models, i.e., CNN, Longest Shortest-Term Memory (LSTM), SVM, bidirectional LSTM (BLSTM), and multilayer perceptron by repeating more than 3,000 experiments. Francisco et al. [10] used the LSTM to extract the changes of each activity flow time to recognize human activities.

To obtain discriminative features, the well-known loss function used to tackle the problem of classification is Cross-Entropy loss [11]. If just using Cross-Entropy loss, embedded features extracted by a deep learning model are not strong enough for recognizing classes that have similar patterns. Many studies show that deep learning models' performance can be increased if adding constraints on hidden features. Wen et al. [12] combined Cross-Entropy with Center Loss to solve the problem of face recognition. Their results showed that the recognition performance is better than using only Cross-Entropy loss.

In this paper, we conduct a deep learning method to tackle the problem of human activity recognition. The biggest challenge in the problem is how we can classify classes that have similar patterns. The proposed method is described in detail in Sec. 2. Experimental results are shown in Sec.3, and conclusion will be explained in Sec. 4.

2. THE PROPOSED METHOD

The process of human activity prediction follows the general patterns of the classification system and can be divided into three phases. This process related to a set of stages ranging from data collection of raw data which are selected from sensors to generate efficient classification models of human activities. In the data collection stage, participants who wear a

smartphone (Samsung Galaxy S II) on the waist are required to perform six activities as "walking, walking up stair, walking down stair, sitting, standing, and laying". Tri-axis accelerometer and gyroscope sensor data are collected during performing the experiments. From raw data, we apply some methods to reduce noise and smooth origin data. Then, we slide a fixed-length window over the pre-processed data to get samples for training, evaluating and testing sets. Finally, we use deep learning algorithms to predict what kind of activities. Specially, deep learning model utilized in this paper is the Convolutional Neural Network with joint loss functions, i.e., Cross-Entropy Loss and Center-Loss. Figure 1 illustrates the process of human activity classification.

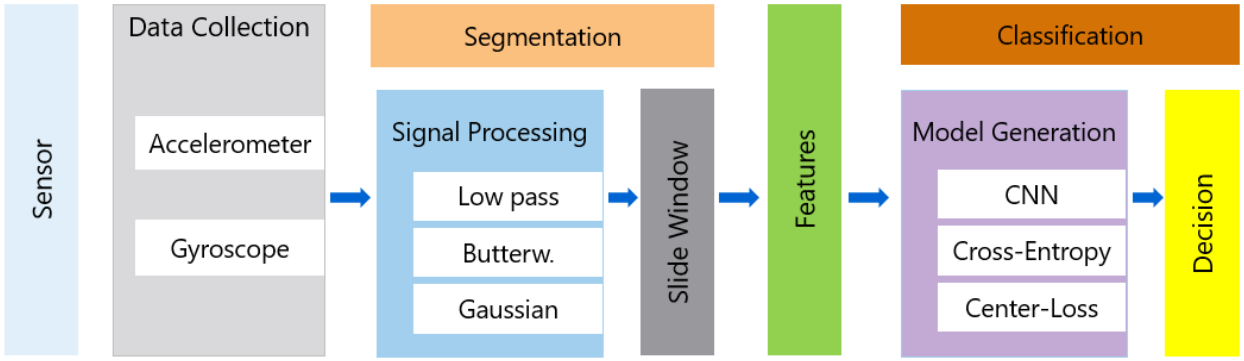


Figure 1: Set of steps for human activity recognition based on deep learning algorithm.

Table 1: The detailed our network architecture.

Input: 128× 6
Conv1d: 3×64, stride 1, padding same, LeakReLu
Conv1d: 3×64, stride 1, padding same, LeakReLu
MaxPooling: 2, stride: 2
Conv1d: 3×128, stride 1, padding same, LeakReLu
Conv1d: 3×128, stride 1, padding same, LeakReLu
Conv1d: 3×128, stride 1, padding same, LeakReLu
MaxPooling: 2, stride: 2
FC: 512, LeakReLu
FC: 256, LeakReLu
FC: 64, LeakReLu
FC: 6

To predict human actions, the selected classifier is based on deep learning methods. Our network architecture is inspired by VGG 19 [13, 14]. By stacking multiple CNN layers, we have a novel deep learning architecture that has a good performance on Human Activity Recognition. One disadvantage of deep learning is overfitting, to prevent this phenomenon, we dropped some CNN layers rapidly. By doing this, our architecture is simple and still enough capability to perform the task of classifying. The proposed network architecture is showed in Table 1.

Our network architecture consists of 5 convolutional layers with (size, number of filters) is (3, 64), (3, 64), (3, 128), (3, 128) and (3, 128) respectively, the stride is one, padding is same. Two pooling layers in the network architecture are max pooling layer with (size, stride) is the same (2, 2). The following are three fully connected layers with (the number of hidden units): 512, 256, and 64 respectively. Passing the fully connected layer with 64 neurons, from each sample with the shape of 128×6, we receive a 64-d vector, a high-level feature descriptor representing the input signal. Finally, this vector is pushed through the last fully connected layer with 6 neurons, which in turn, outputs a 6-d vector $\boldsymbol{s} = (s_1, s_2, \dots, s_6)$ representing the score of each class. In this case, we have 6 classes of the problem of human activity recognition. When the cross-entropy loss is used in the training phase, the softmax activation function is the applied to the score vector s, and it in turn outputs a new 6-d vector $\boldsymbol{u} = (u_1, u_2, \dots, u_6)$, that is defined as follows:

$$u_i = \frac{e^{s_i}}{\sum_{k=1}^{6} e^{s_k}} \quad (1)$$

Thus, we finally obtain the vector of 6 elements representing the probability distribution over six activity classes. The class label, which corresponds to the highest probability, will be selected as the final answer. Cross-Entropy loss is applied to $\boldsymbol{u}$ and the label to calculate the distance between the prediction and the label. To get hidden features at the third fully connected layer are more discriminative, we applied center loss on the output of this layer. So center loss consists 6 center points. Each center is a 64-d vector that represents the center hidden features of one class. The mathematical formulation of the objective function is illustrated as follows:

$$L = L_s + \lambda_1 \times L_c + \lambda_2 \times \|w\|_2^2 \quad (2)$$

where, L_s is Cross-Entropy Loss, L_c is Center Loss and $\|w\|_2^2$ is a regulation loss which is used for preventing overfitting phenomenon. λ_1 and λ_2 scalars are used for balancing three-loss components. The conventional cross-entropy loss can be considered as a special case of this join supervision, if λ_1 is set to 0.

3. EXPERIMENTAL RESULTS

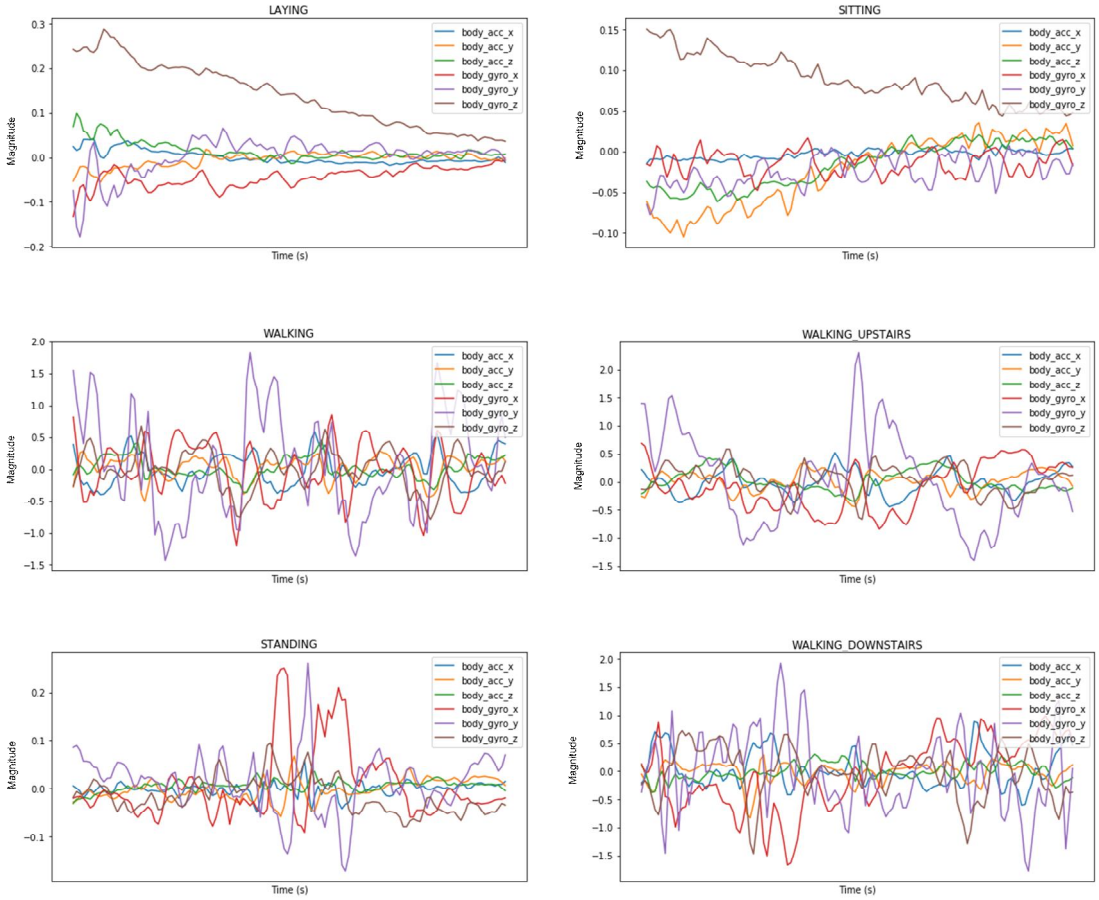


Figure 2: Some examples of the training set. Examples of LAYING class and SITTING class have similar trends, and some instances of the WALKING class and WALKING_UPSTAIRS class have similar patterns.

The dataset consists of accelerometer and gyroscope tri-axis sensor data were collected from 30 volunteers within an age bracket of 19-48 years. The sensor signals (accelerometer and gyroscope) were pre-processed by applying noise filters and then sampled in fixed-width sliding windows of 2.56 sec and 50% overlap (128 readings/window). The 128-real value vector stands for one sample for one activity. Overall, the dataset has 7352 samples for training data (from 21 randomly selected subjects) and 2947 examples for test data (from 9 remaining subjects). We sampled 5147 instances for the training set and 2205 instances for the validation set from training data, which is used for tuning the hyper-parameters and selecting the model. The formula sample is described by Eq. (3). Figure 2 presents some examples of the smartphone dataset.

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (3)$$

To evaluate the quality of the model's predictions, we used F1-Score (macro) to compare our experimental results with other previous studies' results. Besides F1-Score, other evaluation metrics, i.e., accuracy, recall, and precision, are considered to give a comprehensive view of our model's ability for the problem of human activity recognition. Their metrics' mathematics definitions are described as bellows:

$$Acc = \frac{1}{N}\sum_{i=1}^{N}(y_i == y'_i) \quad (4)$$

$$Precision = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FP_c} \quad (5)$$

$$Recall = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FN_c} \quad (6)$$

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (7)$$

Where, N is the of samples and C is the number of classes, y_i is the label of sample i and y'_i is a prediction of sample i. TP_c is the number samples of class c have right prediction. TheFP_c is the number of samples which don't belong class c but are predicted as class c. FN_c is the number of samples which belong class c but aren't predicted as class c.

We applied models SVM, LSTM, BLSTM, CNN and our methods for dataset in Wan's method [12] for the comparison.

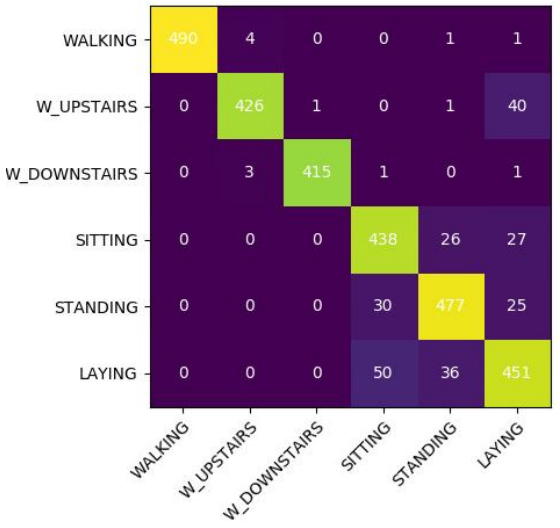
In Tab. 6, we compare the performance of the proposed structures with previous studies in terms of accuracy, precision, recall, and F1-score. The results illustrate that our method conducts the excellent recognition performance. In details, our best setting achieves the accuracy, precision, recall, and F1-score rate at 94.0%, 94.2%, 94.2%, and 94.2%, respectively. In addition, Tab. 2 shows the positive impacts of using Center Loss on the results of the prediction. It is obvious that using the combination of Center Loss and Cross-Entropy Loss can improve considerably the quality of prediction. To get a detail about the quality of classifying classes having similar patterns, Table 3 is used for presenting the recall metric of each class. Four classes that are STANDING, LAYING, WALKING, and WALKING_UPSTAIRS have great improvements, likely STANDING class can reduce 3.3% recall error, or WALIKING_UPSTAIRS class can increase by 4.7 % recall precise.

Table 2: Comparison with recent studies in terms of accuracy, precision, recall and F1

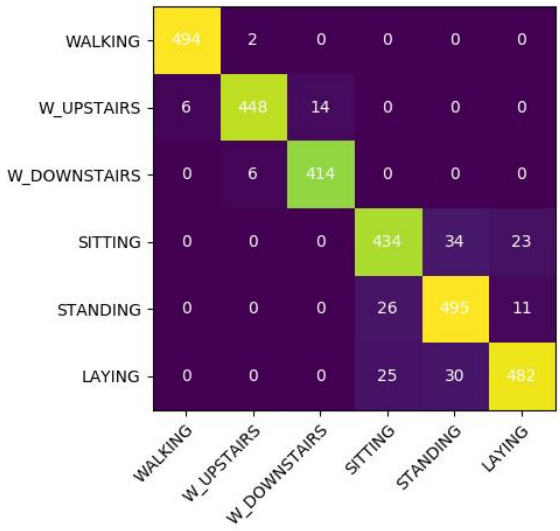
Model	**Accuracy**	**Precision**	**Recall**	**F1**
SVM [15]	0.905	0.905	0.899	0.899
LSTM [15]	0.890	0.891	0.889	0.889
BLSTM [15]	0.894	0.894	0.894	0.893
CNN [15]	0.927	0.932	0.928	0.929
Our Architecture + Cross-Entropy Loss	0.916	0.922	0.919	0.920
Our Architecture + Cross-entropy Loss + Center Loss	**0.940**	**0.942**	**0.942**	**0.942**

Table 3: Comparison between two proposed methods on the recall of each activity.

Activity	**Recall**	
	Our Architecture + Cross-Entropy Loss	Our Architecture + Cross-Entropy Loss + Center Loss
SITTING	**0.892**	0.884
STANDING	0.897	**0.930**
LAYING	0.840	**0.898**
WALKING	0.988	**0.996**
WALKING_UPSTAIRS	0.910	**0.957**
WALKING_DOWNSTAIRS	**0.988**	0.986



(a)



(b)

Figure 3: Confusion matrixes of our proposed models. (a) the confusion matric of our model which uses our architecture and cross-entropy loss. (b) the confusion matric of our model which uses our architecture, cross-entropy loss and center loss.

To demonstrated adding Center Loss can reduce the rate of missing prediction of similar classes, we plot confusion matrixes of our methods before and after using Center Loss in Figure 3. Moreover, the number of samples that belong, STANDING class, predicted to the SITTING class, and LAYING class are reduced dramatically. The phenomenon is the same in the case of the LAYING class.

4. CONCLUSION

In this paper, we presented a deep learning method with joint loss functions to classify human activities. The proposed method is robust to recognize classes having similar patterns. By evaluating on the real-world human activity recognition dataset, the proposed approach outperforms some recent studies in terms of F1-score. In details, it achieves an F1-score rate of 94.2% on Smartphone Dataset. Furthermore, embedded features extracted by CNN have their own cluster

characteristics due to using Center-Loss, which enables to add and predict new activity classes without retraining the entire model.

ACKNOWLEDGEMENT

This work is supported by FPT University, Hanoi, Vietnam; and Sungkyunkwan University, Suwon, South Korea. Moreover, we would like to say thanks to Prof. Giao N. Pham, Dept. of Computing Fundamentals, FPT University, Hanoi, Vietnam, who supported, revised and gave advises for this paper.

REFERENCES

1. G. Fortino, R. Giannantonio, R. Gravina, P. Kuryloski and R. Jafari. **Enabling Effective Programming and Flexible Management of Efficient Body Sensor Network Applications**, IEEE Transactions on Human-Machine Systems, Vol. 43, No. 1, pp. 115-133, Jan. 2013. https://doi.org/10.1109/TSMCC.2012.2215852
2. O. C. Ann and L. B. Theng. **Human activity recognition: A review**, 2014 IEEE International Conference on Control System, Computing and Engineering (ICCSCE 2014), BatuFerringhi, 2014.
3. M. Munoz-Organero and A. Lotfi. **Human Movement Recognition Based on the Stochastic Characterisation of Acceleration Data**, Sensors, Vol. 16, No. 9, Sep. 2016.
4. Y. Zhang, Z. Zhang, Y. Zhang, J. Bao, Y. Zhang and H. Deng. **Human Activity Recognition Based on Motion Sensor Using U-Net**, IEEE Access, Vol. 7, pp. 75213-75226, 2019.
5. Kaghyan, Sahak and Sarukhanyan, Hakob. **Activity recognition using k-nearest neighbor algorithm on smartphone with triaxial accelerometer**, International Journal of Informatics Models and Analysis (IJIMA), No. 1, pp. 146-156, 2012.
6. Anguita D., Ghio A., Oneto L., Parra X., Reyes-Ortiz J.L. **Human Activity Recognition on Smartphones Using a Multiclass Hardware-Friendly Support Vector Machine**, Ambient Assisted Living and Home Care, IWAAL, 2012.
7. Dargan, Shaveta, Munish Kumar, MaruthiAyyagari, and Gulshan Kumar. **A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning**, Archives of Computational Methods in Engineering Online First, pp. 1-22, 2019.
8. X. Li, Y. He, and X. Jing. **A Survey of Deep Learning-Based Human Activity Recognition in Radar**, Remote Sensing, Vol. 11, No. 9, May 2019.
9. Wan, S., Qi, L., and Xu, X. **Deep Learning Models for Real-time Human Activity Recognition with Smartphones**, Mobile Network Application,Vol. 25, pp. 743–755, 2020. https://doi.org/10.1007/s11036-019-01445-x
10. F. Ordóñez and D. Roggen. **Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition**, Sensors, Vol. 16, No. 1, Jan. 2016.
11. Zhilu Zhang and Mert R. Sabuncu. **Generalized cross entropy loss for training deep neural networks with noisy labels**, in Proc. of the 32nd International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, pp. 8792–8802, 2018.
12. Wen Y., Zhang K., Li Z., Qiao Y. **A Discriminative Feature Learning Approach for Deep Face Recognition**, ECCV, Lecture Notes in Computer Science, vol. 9911, 2016.
13. C. Mohamed, B. Nsiri, S. Abdelmajid, A. Mokhtari, and B. Benaji. **Deep Convolutional Networks based on encoder-decoder architecture for automatic Optic Disc segmentation in retina images,** International Journal of Advanced Trends in Computer Science and Engineering, Vol. 9, No. 2, pp. 2078-2084, 2020. https://doi.org/10.30534/ijatcse/2020/181922020
14. A. S. Alon, and J. L. Dioses. **A Machine Vision Detection of Unauthorized On-Street Roadside Parking in Restricted Zone: An Experimental Simulated Barangay-Environment**, International Journal of Emerging Trends in Engineering Research, Vol. 8, No. 4, pp. 1056-1061, April 2020. https://doi.org/10.30534/ijeter/2020/17842020
15. W. Wan, Y. Zhong, T. Li and J. Chen. **Rethinking Feature Distribution for Loss Functions in Image Classification**, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, pp. 9117-9126, 2018