Depthwise Separable Convolution for Human Activity Recognition

Nguyen Thi Hoai Thu, Dong Seog Han*
School of Electronics Engineering, Kyungpook National University

thunguyen@knu.ac.kr, *dshan@knu.ac.kr

Abstract

Human activity recognition (HAR) using wearable sensors has made great contributions in healthcare and surveillance domain. Recently, with the rapid growth of deep learning, deep learning methods are widely applied to the classification task in the HAR because of their ability to extract essential features automatically. However, this approach is restrained by the energy limitation of the wearable device as it requires high computational cost. In this paper, we propose a low energy expenditure HAR model that can be run on embedded devices. The proposed model uses a depthwise separable convolution instead of normal convolution to reduce the computational complexity. The experiment results show that the proposed model achieves a competitive performance compared to standard convolution method.

I. Introduction

Wearable sensor-based human activity recognition is the problem of understanding human behaviour by using data collected from sensors such as an accelerometer embedded in a smartphone [1]. Traditionally, to classify the human activities, conventional pattern recognition (PR) approaches have been used by adopting machine learning algorithms such as naive Bayes, support vector machine and hidden Markov models [2]. However, this method may heavily depend on heuristic handcraft feature extraction (e.g., statistical and structural features) which is usually limited by human domain knowledge [3]. Recently, with the fast development and advancement in feature representation of deep learning, which has achieved remarkable performance in many areas such as computer vision, natural language processing [4], several deep learning methods such as convolutional neural network (CNN), recurrent neural network (RNN) are applied for HAR [3]. However, because of the deep structure, most of deep learning methods are very expensive in their evaluation phase, which make them not convenient for mobile HAR because of the energy limitation. To overcome this challenge, we propose a CNN-based model using a depthwise separable convolution method for human activity recognition.

The depthwise separable convolution (DSC) was first introduced in [5] and is widely used for classification task in computer vision [6, 7]. The depthwise separable convolution is a factorized form of the standard convolution. A standard convolution is separated into a depthwise convolution and a 1x1 pointwise convolution. Instead of applying each filter to all the channels of the input like the standard convolution, the depthwise convolution layer applies one filter to one input channel, then a 1x1 pointwise convolution is employed to combine the outputs of the depthwise convolution. Depthwise separable convolution helps reduce not only the number of learnable parameters but also the computational cost in both training and testing process.

The standard convolution has a computational cost, C_{Standard}, of

$$C_{\text{Standard}} = H_K \cdot W_K \cdot M \cdot N \cdot H_O \cdot W_O \quad (1)$$

where H_K and W_K are the spatial dimension (height and width) of the kernels; M and N are the number of input and output channels; H_O and W_O are the spatial dimension of the output.

With the same input and output size, the computational cost of depthwise separable convolutional, C_{DSC}, can be defined as

$$C_{\text{DSC}} = H_K \cdot W_K \cdot M \cdot H_O \cdot W_O + M \cdot N \cdot H_O \cdot W_O \quad (2)$$

Making use of depthwise separable convolution gives us a reducing in computation of

$$\begin{aligned} C_{\text{Reduction}} &= \frac{C_{\text{Standard}}}{C_{\text{DSC}}} \\ &= \frac{H_K \cdot W_K \cdot M \cdot H_O \cdot W_O + M \cdot N \cdot H_O \cdot W_O}{H_K \cdot W_K \cdot M \cdot N \cdot H_O \cdot W_O} \\ &= \frac{1}{N} + \frac{1}{H_K \cdot W_K} \end{aligned} \quad (3)$$

Another advantage of this method is that it increases the non-linearity of the model by using more non-linear activation functions than the standard method without significantly increasing the number of convolutional layers. Table I contrasts the structure of the standard convolutional layer and the ReLU activation function to the factorized layer with depthwise convolution, a 1x1 pointwise convolution as well as ReLU after each convolutional layer.

TABLE I. STRUCTURE OF STANDARD CONVOLUTIONAL LAYER AND DEPTHWISE SEPARABLE CONVOLUTION

Standard Convolutional Layer	Depthwise Separable Convolution
3x3 Conv	3x3 Depthwise Conv
ReLU	ReLU
	1x1 Pointwise Conv
	ReLU

The architecture of our proposed convolutional neural network is shown in Table II. Every convolution-based layer is followed by a ReLU activation function as described in Table I. The last element of each input indicates the number of channels (the depth). The input in the first layer is the sensor data of 9 signals in 128 time steps. Max pooling layers with the pooling size of (2x1) are applied as a regularization method to avoid overfitting.

II. Experimental Results

To analyze the performance of the proposed HAR model, we conducted our experiment on the UCI HAR dataset [8]. The dataset contains raw data collected from 3-axial linear acceleration and 3-axial angular velocity embedded in a smartphone at a constant sampling frequency of 50 Hz. There are six basic activities: three static postures (standing, sitting, lying) and three

TABLE II. STRUCTURE OF DEPTHWISE SEPARABLE CONVOLUTION BASED NETWORK

Input	Operator	Filter Shape
128x9x1	Conv (stride = 1, pad = 2)	(3x3x1)x32
128x9x32	Depthwise Separable Conv (stride = 1)	(3x3x1)x32 (1x1x32)x32
126x7x32	MaxPooling	Pool 2x1
63x7x32	Conv (stride = 1)	(3x3x32)x64
61x5x64	Depthwise Separable Conv (stride = 1)	(3x3x1)x64 (1x1x64)x64
59x3x64	MaxPooling	Pool 2x1
29x3x64	Conv (stride = 1)	(3x3x64)x128
27x1x128	MaxPooling	Pool 2x1
13x1x128	Flatten	
1664	Fully connected	1664x100
100	Softmax	Classifier

dynamic activities (walking, walking downstairs, walking upstairs). The data is into fixed-length windows of 2.56 seconds and 50% overlap (128 readings/window). The windows, then are fed into CNN models as virtual 2D images with a size of (128x9x1). In order to evaluate the performance of our proposed method, a comparison between standard CNN and depthwise separable convolutional network is provided in Table III. The average accuracy is obtained by running 40 experiments for each model. The computational cost shown in Table III is the total number of multiplications of one window data in convolution-based layers for a forward pass. Total number of trainable parameters of the networks are also calculated. The confusion matrix of depthwise separable convolution based HAR model in one experiment is shown in Fig. 1.

TABLE III. COMPARISON OF AVERAGE ACCURACY AND COMPUTATIONAL COST BETWEEN CNN AND DEPTHWISE CNN

Model	Average Accuracy (%)	Parameters	Computation Cost (millions)
Standard CNN	92.524 (±1.848)	305,354	21.123
Depthwise Separable Conv (proposed model)	91.857 (±1.398)	265,858	1.679

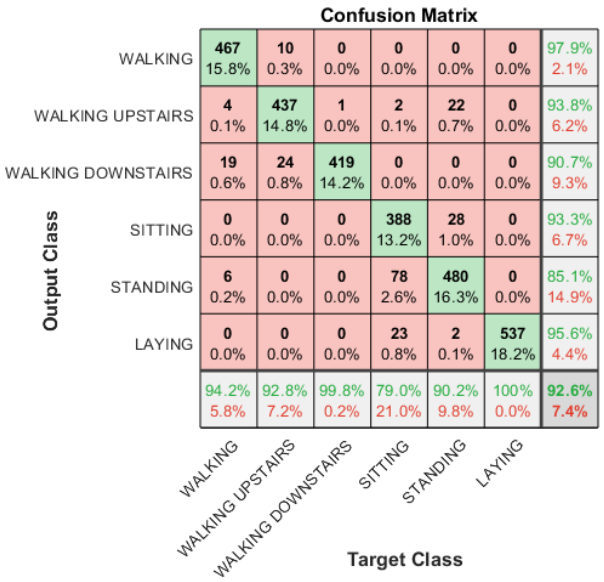

Fig. 1. Confusion matrix of depthwise separable convolution-based HAR model in one experiment.

From Table III, it can be seen that using depthwise separable convolution-based CNN compared to standard CNN only reduces accuracy by 0.7% while save tremendously on the number of multiplications (a reduction of 12 times) and number of learnable parameters. The experimental results and complexity analysis validate that the proposed HAR model can be used for mobile application with reasonable accuracy.

III. Conclusion

In this paper, we presented a CNN-based HAR model which uses a depthwise separable convolution method. The proposed model reduces the complexity significantly while still provides accurate classification result for human activity recognition. In future work, we intend to improve the accuracy of the recognition and try other CNN architectures to apply to the mobile HAR application.

Acknowledgment

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2016-0-00564, Development of Intelligent Interaction Technology Based on Context Awareness and Human Intention Understanding).

References

[1] O. S. Eyobu and D. S. Han, "Feature Representation and Data Augmentation for Human Activity Classification Based on Wearable IMU Sensor Data Using a Deep LSTM Neural Network," *Sensors* 18, no. 9 (2018): 2892.

[2] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications Surveys Tutorials*, vol. 15, no. 3, pp. 1192–1209, Third 2013.

[3] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*, vol. 119, pp. 3 – 11, 2019.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–44, 05 2015.

[5] L. Sifre, "Rigid-motion scattering for image classification," Ph. D. thesis, 2014.

[6] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv*:1502.03167, 2015.

[7] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, abs/1704.04861, 2017.

[8] Anguita D, Ghio A, Oneto L, Parra X, Reyes-Ortiz JL, "A public domain dataset for human activity recognition using smartphones," *Esann*, 2013.