Video Human Action Recognition with Channel Attention on ST-GCN

Deyuan Zhang1*, Haoguang Wang2, Chao Weng3 and Xiangbin Shi4

1Shenyang Aerospace University, Shenyang, Liaoning, China
2Shenyang Aerospace University, Shenyang, Liaoning, China
3Shenyang Aerospace University, Shenyang, Liaoning, China
4Shenyang Aerospace University, Shenyang, Liaoning, China

*Corresponding author’s e-mail: dyzhang@sau.edu.cn

Abstract. Action recognition based on human skeleton information is a hot research topic in the field of computer vision, and ST-GCN graph convolutional network is widely used to extract spatial and temporal features of human skeleton to represent the human skeleton structure. However, in the process of extracting features, the weights on each channel of the feature are the same, so it is difficult to effectively discriminate the useful features from the useless ones. In this paper, we propose Channel Attention module, which learns the importance of each feature channel to perform human action recognition more effectively. Experimental results on Kinetics and NTU-RGB+D datasets show that Channel Attention module can achieve better accuracy.

1. Introduction

Human action recognition has been applied to many fields such as intelligent video surveillance[1] and intelligent rehabilitation medical treatment [2]. Because of the complexity and diversity of human action, human action recognition still faces great challenge. There are mainly two types of methods for solving human action recognition tasks, which are based on image information and human skeleton information. Image information research have good results in simple background conditions. However, in realistic situations, it may be affected by noise such as illumination. The action recognition method based on human skeleton information uses pose information to represent human features. This method can effectively reduce the impact of illumination.

Based on human skeleton information, graph convolutional network is widely used to extract spatial features. ST-GCN [3] is one of the widely used to extracts skeletal features and temporal sequence features by graph convolutional network in spatial dimension and temporal dimension respectively. However, in the process of feature extraction, the weight of each channel of the feature is the same, so it is difficult to effectively distinguish useful features and useless features. The attention mechanism can effectively solve this problem. GAT [5] proposes to use masked self-attention layer to assign different weights to neighboring nodes. This approach starts from proposing an attention mechanism in the spatial dimension. We decided to solve the problem of feature channel weights from the channel dimension and we introduce the attention mechanism in channel dimension to solve the problem of feature channel weight allocation more effectively.
In this paper we propose Channel Attention Module (CAM). We improve the feature maps by transforming the feature vectors in the channel dimension and in the temporal dimension through two operations, squeeze and excitation, which aggregate the information in terms of feature dimensions, and then use the transformed feature vectors as the weight coefficients of each feature map. In this way, each channel's feature map has its own weight, which enables more effective feature extraction and action recognition. In this paper, experiments are conducted on two datasets obtained by the depth camera and the pose estimation algorithm. The experimental results show that the model can achieve better results compared to ST-GCN.

2. Materials and Method

2.1. Framework of CAM-ST-GCN for human action recognition

In this paper, we focus on skeleton-based human action recognition algorithms, and the framework is shown in Figure 1. Firstly, given a sequence of images, openpose[4] algorithm is used to extract the human pose information. Secondly, CAM module based on channel attention mechanisms is introduced to extract spatial and temporal features. Finally, it will be classified to the corresponding action category by the standard SoftMax classifier.

![Figure 1 Framework overview of the network.](image)

2.2. Channel Attention Module (CAM)

CAM is proposed to give higher attention to more discriminative features by learning attention masks. It consists of two sub-modules: Channel Attention Module (CAM). CAM is formulated as follows.

\[
M_c = \text{sigmoid}(\text{elu}(\text{conv}(\text{avg}(f_{in}))))
\]  

(1)

The feature map is first transformed into a \(1\times1\times C\) vector by a global average pooling. This is followed by a \(1\times1\) convolution with dimension \(C/r\times C\). Then an ELU layer, which has a negative value compared to ReLU, pushes the mean value of the activation unit to 0, achieving the effect of batchnormlization and reducing the computational effort. We perform another \(1\times1\) convolution, and the dimension is \(C\times C/r\). Therefore, the dimension of the output \(M_c\) is \(1\times1\times C\). Finally, the sigmoid function is used to normalize the features to the interval \([0,1]\), and the normalized features are considered as attention weights. These weights will become effective during training by a data-driven approach, and the \(1\times1\) convolution parameters will be learned. Figure 2 shows the Structure of the CAM Attention Module.

![Figure 2 Structure of the CAM Attention Module](image)
The construction of each basic block is shown in Figure 3. Both spatial GCN and temporal GCN are followed by a batch normalization (BN) layer and a ReLU layer. A basic CAM-ST-GCN block is a sequence of a spatial GCN (Convs), a CAM module, and a temporal GCN (Convt). To stabilize the training and moderate the gradient propagation, a residual connection is added to each basic block.

![Residual connection](image)

Figure 3  Construction of each Basic Block

As shown in Figure 4, the overall architecture of the CAM-ST-GCN network is a stack of these basic blocks (Figure 3), each with 64, 64, 64, 128, 128, 128, 256, 256, 256 output channels. A data BN layer is added at the beginning for normalizing the input data. Finally, a global average pooling layer is executed to pool the feature mappings of different samples to the same size.

![CAM-ST-GCN Network Architecture](image)

Figure 4  CAM-ST-GCN Network Architecture

3. Results & Discussion

3.1. Experimental setup

The experiments were conducted on two large-scale action recognition datasets: Kinetics [7] and NTU-RGB+D [8]. The Kinetics dataset contains approximately 300,000 video clips retrieved from YouTube. We evaluate the recognition performance based on the top-1 and top-5 classification accuracies recommended by the dataset authors. The NTU-RGB+D dataset contains 56,000 action clips from 60 action classes. The clips were all completed by 40 volunteers captured in a restricted laboratory environment while recording three camera views. The dataset has two benchmarks 1) X-Sub of the
same camera view for different volunteers. 2) X-View of the same volunteers for different camera views. We follow this convention and report the top-1 recognition accuracy on both benchmarks.

We use OpenPose to obtain pose information and evaluate recognition performance based on top-1 and top-5 classification accuracies recommended by the dataset authors. All experiments were conducted on a PyTorch deep learning framework using 1 GPU.

3.2. Analysis of experimental results
As shown in Table 1, the experimental results of five behavioral recognition methods based on pose information are compared on the Kinetics dataset. Feature Enc represents a feature encoding method for manual extraction of pose features; Deep LSTM and Baseline TCN represent deep learning methods that extract pose features in the time domain only; ST-GCN represents the fusion of temporal-spatial features; CAM-ST-GCN represents the addition of CAM attention module to extract features before fusing temporal and spatial features. It can be seen that the top1 and top5 action recognition accuracies of CAM-ST-GCN are optimal. It proves that the CAM attention module is more effective for extracting features.

| Method          | TOP-1 | TOP-5 |
|-----------------|-------|-------|
| Feature Enc[9]  | 14.9% | 25.8% |
| Deep LSTM[8]    | 16.4% | 35.3% |
| Baseline TCN[10]| 20.3% | 40.0% |
| ST-GCN[3]       | 30.7% | 52.8% |
| CAM-ST-GCN      | 31.6% | 53.5% |

As shown in Table 2, the experimental results of the CAM-ST-GCN model and the ST-GCN model are compared on the NTU-RGB+D dataset, and the CAM-ST-GCN model outperforms the ST-GCN model on both evaluation benchmarks, proving once again that the introduction of the CAM attention module is more effective for human action recognition.

| Method     | X-sub | X-view |
|------------|-------|--------|
| ST-GCN[3]  | 81.5% | 88.3%  |
| CAM-ST-GCN | 81.9% | 88.8%  |

4. Conclusion
In the problem of skeleton-based human action recognition, the most central thing is how to accurately obtain the skeletal information of the human body and extract the features effectively. In this paper, we use a Channel Attention Module (CAM) to extract features of human skeletal maps more efficiently. This module solves the problem that graph convolution does not effectively discriminate useful features from useless features in the process of feature extraction. It obtains the importance of each feature channel by means of learning, and then boosts the useful features and suppresses the useless features according to this importance. We conducted extensive experiments on two large datasets on Kinetics and NTU-RGB+D, and the experiments compared the effects of introducing the attention mechanism and not using it. The experimental results show that the human action recognition results are the best when the CAM attention module is introduced. The experimental results show that the correct rate of
behavior recognition for the latter is better than the former due to the former, where the former is better than the latter in terms of the improvement of the correct rate of behavior recognition.

Acknowledgments
This work was supported in part by Liaoning Provincial Department of Education Science Foundation (JYT2020050).

References
[1] Foroughi, H., Aski, B. S., & Pourreza, H. (2008) Intelligent video surveillance for monitoring fall detection of elderly in home environments. In: 11th international conference on computer and information technology, Vancouver. pp. 219-224.
[2] Ghesu, F. C., Georgescu, B., Neumann, D., et al. (2017) Intelligent medical image landmark detection: U.S. Patent 9,569,736.
[3] Yan, S., Xiong, Y., Lin, D. (2018) Spatial temporal graph convolutional networks for skeleton-based action recognition. In: 32th AAAI conference on artificial intelligence. New Orleans. pp. 7455-7459.
[4] Veličković, P., Cucurull, G., Casanova, A., et al. (2018) Graph attention networks. In: 6th International Conference on Learning Representations. pp. 10903-1091.
[5] Cao, Z., Hidalgo, G., Simon, T., et al. (2018) OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. IEEE T Pattern Anal, 43:172-186.
[6] Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2019) Skeleton-Based Action Recognition with Multi-Stream Adaptive Graph Convolutional Networks. IEEE T Circ Syst Vid, 29:3247-3257.
[7] Kay, W., Carreira, J., Simonyan, K., et al. (2017) The kinetics human action video dataset. In: 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Hawaii. pp. 6950-7015.
[8] Shahroudy, A., Liu, J., Ng, T. T., et al. (2016) Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: 29th Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas. pp. 1010-1019.
[9] Fernando, B., Gavves, E., Oramas, J., et al. (2015) Modeling video evolution for action recognition. In: 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston. pp. 5378-5387.
[10] Kim, T. S., Reiter, A. (2017) Interpretable 3D Human Action Analysis with Temporal Convolutional Networks. In: 16th IEEE conference on computer vision and pattern recognition workshops (CVPRW), Honolulu. pp. 1623-1631.