Commentary: Efficient dual attention SlowFast networks for video action recognition

Wanqing Yang
Z5325987

1. Introduction

In recent years, with the development of deep convolutional neural networks, action recognition has attracted the attention of many researchers. Action recognition data can be divided into image data and video data. The video data has one more time dimension than the image data. Video action recognition is widely used in various fields such as video surveillance and automatic driving. This paper mainly discusses video action recognition and its methods.

Although the current video motion recognition technology can achieve a relatively ideal effect, there are still the following problems. The first is that the specific temporal and spatial characteristics of the adaptive learning path are challenging to the neural network. The second is learning efficiency. For most 3D CNN architectures, a large number of floating point operations are required, which requires very high hardware performance.

Based on the above problems, this paper proposes a cross-modality dual attention fusion module (CMDA), and combines the two-stream efficient 3D action recognition model based on GhostNet, ShuffleNet, MobileNetV2 and ShuffleNetV2 to improve the existing motion recognition technology. Solving these problems will promote the development of video surveillance and autonomous driving, in which individuals are the biggest beneficiaries. This is because video surveillance can better ensure the safety of individuals and their property through efficient action recognition. In addition, automakers and traffic managers will benefit from the development of autonomous driving.

2. Methods

In order to improve the video action recognition method, this paper mainly proposed the fusion module CMDA and the Efficient 3D SlowFast networks model.

Because the Slow pathway in SlowFast cannot focus on the key parts of action recognition, the Fast pathway considers all parts indiscriminately. Therefore, this paper proposes a fusion module CMDA to make both Slow and Fast pathways have the attention to identify the key parts. CMDA is divided into two parts: attention mechanism and fusion method.

The first is the mechanism of attention, which is divided into space-time attention and efficient channel attention. Since video data has a temporal dimension, a temporal dimension is added to the self-attention proposed by Vaswani et al. (2017). Moreover, adjustable reduction parameters and residual connections are used to reduce the feature mapping between paths and improve the learning ability of the model. In order to reduce the complexity of the model and improve the efficiency of the model, a 1D Conv with adaptive kernel size is used to replace the FC layer in the original channel module and expand it to three-dimensional space.

The second is Fusion methods. In order to realize the information exchange between Slow pathway and Fast pathway, Fusion from fast to slow and Fusion from slow to fast are proposed in this paper. Because the number of frames on the Fast pathway is larger than that on the Slow pathway, down-sampling is performed along the time dimension of the Fast pathway. For Fusion from slow to fast, it does not have the interpretability of down-sampling. Therefore, this paper uses the single 1*1*1 Conv to reduce the channel size. Then, the nearest sampling method is used to reduce the amount of computation while maintaining the interpretability.

In order to improve the learning efficiency and reduce the requirements on hardware performance, this paper improved the original 3D CNN and designed several Efficient 3D SlowFast networks models. Taking GhostNet as an example, a CMDA fusion information is added after each of its first four phases, ensuring that the network is lightweight. And turn it into a 3D version for video recognition.

These two methods improve the learning efficiency of the model, reduce the requirement of hardware, and extend the traditional two-dimensional network to three-dimensional. However, the method still has some problems, such as down-sampling in fusion from fast to slow, often missing some key frames useful for identifying the action. One possible improvement is to use 3D temporal max-pooling, which does not break the interpretability of down-sampling, so only the time dimension needs to be considered.

3. Results

In this paper, the widely used Kinetics-400 and 20BN-Jester-v1 data sets are used for experiments. Kinetics 400 is a YouTube human action video with 400 human movements. The size of the training set is 224,919 and the size of the test set is 18,525. The 20BN-Jester-v1 dataset contains 27 predefined human gestures. The size of the training set is 118,562 and the size of the test set is 14,787.

In order to make a fair comparison between the proposed model and the traditional SlowFast, the data were preprocessed in accordance with SlowFast. This includes using 224*224 space sizes and random sampling. Since gestures such as Swiping Left and Swiping Right are difficult to recognize after flipping in data set 20BN-Jester-v1, these images are processed in this paper, including dithering enhancement of contrast and saturation, and indoor lighting is simulated.

Three experiments were carried out in this paper. In experiment 1, a comparison strategy was used on Kinetics-400 data set to compare the traditional SlowFast model with the improved model in four aspects: number of parameters, FLOPs, Top-1 Acc and Top-5 Acc. By comparison, we found that compared with the traditional SlowFast model, the improved Slowfast model significantly reduced the number of parameters and significantly improved the accuracy, which was in line with the anticipated advantages of the model. The comparison strategy was also used in experiment 2, which compared the 3D improved model of different efficient neural networks with the SlowFast model equipped with CMDA in four aspects: number of parameters, FLOPs, Top-1 Acc and Top-5 Acc on Kinetics-400 data set. By comparison, we found that the CMDA-equipped SlowFast model was significantly better than the recent lightweight SOTA approach. In Experiment 3, the above two models are also compared on the data set 20BN-Jester-v1, and the results show that the improved model is still better.

Based on the above analysis, the SlowFast model equipped with CMDA is more accurate and efficient than the recent lightweight SOTA approach, regardless of the data set. Users should therefore adopt the improved model.

Finally, in order to make the experimental results more convincing, we can also improve the experiment from the following aspects. First, we can increase the number of data sets and compare several models on different data sets. Second, we can compare the improved model with a CMDA-only model and a SlowFast only model.

4. Conclusions

This paper proposes a fusion strategy CMDA and Efficient 3D SlowFast networks to improve efficiency and expand dimensions by exchanging information on Slow and Fast pathways. The fusion strategy greatly improves the attention ability of Slow pathway and Fast pathway. Through the improvement of the existing two-dimensional efficient models such as GhostNet, MobileNetV2, ShuffleNet and ShuffleNetV2, they can be trained on the equipment with lower configuration, and the learning efficiency is increased and the training parameters are reduced.

However, this method still has some shortcomings. In the training process, we take random sampling for training, which may cause some frames with key actions to be ignored, thus resulting in a decline in the accuracy of training. In order to solve this problem, we can first conduct multiple sampling and train the data from multiple sampling, but this is time-consuming. Another possible improvement method is 3D temporal max-pooling, which does not break the interpretability of downsampling, so only the time dimension needs to be considered.

Secondly, although the efficiency of the model is improved, it is not improved enough compared with the traditional model to meet the requirements of reducing the equipment configuration.