I3D-LSTM: A New Model for Human Action Recognition

To cite this article: Xianyuan Wang *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **569** 032035

I3D-LSTM: A New Model for Human Action Recognition

Xianyuan Wang, Zhenjiang Miao, Ruyi Zhang and Shanshan Hao

School of Computer and Information Technology, Beijing Jiaotong University, Haidian District, Beijing, 100044, China

Email: 17125164@bjtu.edu.cn

Abstract. Action recognition has already been a heated research topic recently, which attempts to classify different human actions in videos. The current main-stream methods generally utilize ImageNet-pretrained model as features extractor, however it's not the optimal choice to pretrain a model for classifying videos on a huge still image dataset. What's more, very few works notice that 3D convolution neural network(3D CNN) is better for low-level spatial-temporal features extraction while recurrent neural network(RNN) is better for modelling high-level temporal feature sequences. Consequently, a novel model is proposed in our work to address the two problems mentioned above. First, we pretrain 3D CNN model on huge video action recognition dataset Kinetics to improve generality of the model. And then long short term memory(LSTM) is introduced to model the high-level temporal features produced by the Kinetics-pretrained 3D CNN model. Our experiments results show that the Kinetics-pretrained model can generally outperform ImageNet-pretrained model. And our proposed network finally achieve leading performance on UCF-101 dataset.

1. Introduction

Great achievements have been made in image recognition tasks driven by the rapid development of neural networks and deep learning algorithms. The performance of image recognition achieved by CNN pretrained on ImageNet[1] is already much to our satisfaction. And existing main-stream methods of action recognition also adopt the same methods to learn the spatial features from RGB video frames and temporal features from optical flow. Though good performance can be achieved for video-based action recognition, our research works still find that it's not the optimal choice to pretrained on still image dataset ImageNet. That's mainly due to the fact that motional videos differ from still images to a large extent. As a result, the new Kinetics Human Action Recognition dataset[2] is taken birth especially for pretraining video recognition architectures. And the experiment results in the work[3] also show that almost every CNN architecture pretrained on Kinetics outperforms that pretrained on ImageNet.

Compared with image recognition task, sequence modeling is also an important factor in action recognition. When compared with the original Recurrent Neural Network(RNN), Long Short Term Memory(LSTM)[4] is more extensively applied in sequence modeling issue. Because LSTM can address the tough issue of gradient vanishing and alleviate gradient explosion problem to some extent.

In order to combine the advantage of the advanced feature extracor Kinectics-pretrained CNN and the powerful sequence modeling tools LSTM, we proposed a novel network. First, we use Inception 3D CNN[5] as the feature extractor, which is pretrained on Kinetics in the same way as I3D. Second, the output feature vectors of I3D are put into LSTM network to get high-level temporal features. At Last, a softmax classifier is introduced to make prediction of these high-level features. The network can be trained in a end-to-end way and the structure of the proposed network is shown as Figure 1.

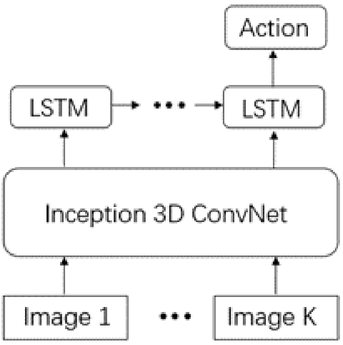

Figure 1. The proposed network mainly composes of two modules: Inception 3D network and LSTM network. I3D is for low-level spatial-temporal features extraction and LSTM is for high-level temporal features extraction.

2. Related Work

2.1. 3D Convolution for Human Action Recognition

3D CNN[6] has always been a typical research method since it's first introduced for action recognition task. In this work, 3D CNN model is adopted to directly extract spatial features and temporal features from raw video data. Another typical research work named C3D[7]. Based on spatial-temporal features learned by 3D CNN, C3D can achieve leading performance on 4 main-stream benchmarks by just using a simple linear classifier. Recently, with the appearance of large size video datasets like ActivityNet[8] and Kinetics, some other works[3,5,7] involving 3D CNN outperform state-of-the-art models after pretraining on huge video dataset. As we all know, 2D CNN model almost dominate the domain of image recognition task because it can automatically extract spatial features by performing 2D convolution, which is more accurate and efficient than traditional handcrafted features. However, experiments in research work[3] has already validated that even very shallow 3D CNN models pretrained on Kinetics can outperform very deep 2D CNN models to a large extent for action recognition task. Recently a novel Two-Stream Inflated 3D ConvNet (I3D) model[5] ,which expand convolution and pooling kernels of Inception module in GoogLeNet[9] into 3D, achieves best accuracy in UCF-101 dataset[10].

2.2. Recurrent Neural Networks for Action Recognition

RNN is so powerful that they are widely used in sequence modelling task like video human action recognition. LSTM is an improved network based on RNN, which can prevent gradient vanishing problem and gradient explosion problem during the training process. To date, LSTM is extensively used for learning motion features in video-based action recognition. LRCN[11] model is both deep in spatial dimension and temporal dimension in that CNN is used for learning spatial features and LSTM is used for learning temporal features. What's more, the length of input and output of LRCN model can be variant. What's more, another work[12] concerning LSTM regards a video as an ordered sequence, which feeds the output feature of underlying CNN network into LSTM network, achieves a leading performance in UCF-101 dataset.

Despite of the fact that LSTM network and 3D CNN are getting more attention in the area of action recognition, very few current research works make proper use of them. In our work we find that

3D CNN is better for learning temporal features between adjacent video frames while LSTM is better for modelling high-level sequence features. As a result, a novel model named I3D-LSTM is proposed in our work.

3. The Proposed Model

In this work, our proposed model can be mainly decomposed into two modules: Three Dimension Inception(I3D) network and Long Short Term Memory(LSTM) work. Here I3D is applied for extracting spatial features and capturing low-level motion features within adjacent frames. And then the output feature learned by underlying I3D model will serve as the input of LSTM network, which is mainly responsible for modelling high-level spatial features. As a result, the feature of input videos can be well learned and represented in that our proposed method can learn both low-level and high-level features very well.

3.1. The Inflated Inception Module

The fact that previous works about 3D CNN usually suffer from overfitting is mainly caused by two factors bellow: (1) It's not until recent several years that huge video dataset which has enough data for pretraining 3D CNN network come out. (2) Due to the complicated structure of 3D convolution kernels, the previous 3D CNN architectures can not build deeper just like image classification models ResNet [13]and the like.

GoogLeNet is well known as a kind of advanced image classification architectures, which is constructed by 9 inception modules and 5 pooling layers. Inception v1 is a basic module which is served as component of GoogLeNet. In our work, we apply the RGB-I3D network proposed in [5]. Inception v1 is chosen as backbone to extract spatial-temporal feature within adjacent frames. Then we inflate all these 2D convolution kernels and pooling kernels into 3D, namely we transform all square filters to cubic filters. The operation of inflating the filters is showed in Figure 2. At last I3D will be pretrained on Kinetics video dataset to improve the generality of model and avoid overfitting.

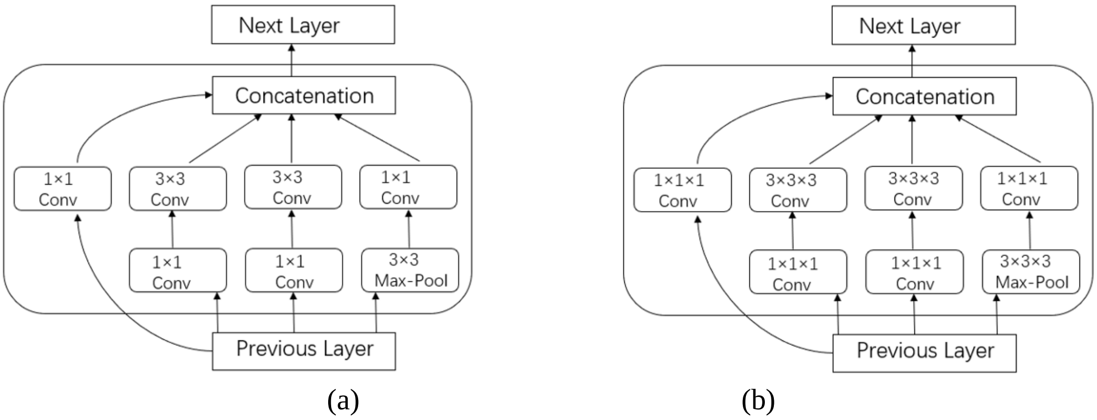

Figure 2. (a) is the inception module before inflation, the convolution kernels and pooling kernels are square. (b) is inception module after inflation, the convolution kernels and pooling kernels are cubic.

3.2. The Long Short Term Memory Network

In consideration of the fact that I3D is mainly powerful for learning low-level temporal features and spatial features, we utilize LSTM network to model high-level temporal features. In practice, we abandon the softmax classifier in original I3D model, and then we put the features after the last 1×1×1 convolution layer into the LSTM network. The structure of LSTM is showed in the Figure 3, where σ and tanh are activation function, c is the state of cell, h represents the hidden state, x is the input signal. In order to update the state of LSTM cells, 3 gates(forget gate, input gate, output gate) are used to determine what message should be remained or removed from the history information. The computation process of forget gate can be expressed by Equation(1).

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

where W_f and b_f stand for the weight matrix and bias respectively, and the hidden state of last timestep h_{t-1} is concated with the current timestep input x_t. The σ function will output a probability value vary from 0 to 1, which helps the forget gate to remember or forget the past state with a probability. Similarly, input gate can be formulated as Equation(2) and Equation(3) separately. And Equation(4) represents the process of updating the cell state.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\tilde{c}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{3}$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \tag{4}$$

The output gate can be formulated as Equation(5) and Equation(6)

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t * \tanh(c_t) \tag{6}$$

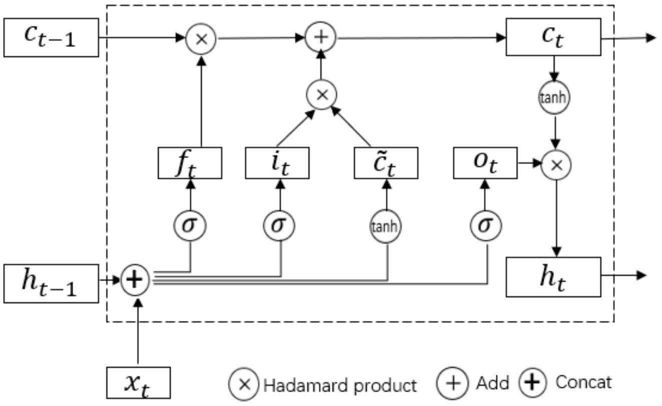

Figure 3. The structure of LSTM cell. Input gate, forget gate and output gate are used to forget or remember the history message.

4. Experiment and Results

4.1. Dataset and Implementation

The proposed network in our paper is evaluated on UCF-101 human action recognition dataset, which consists of 13320 trimmed videos with 101 action classes, and all these videos are collected from YouTube. In our experiments, 9537 videos of UCF-101 dataset are chosen for training and the others are for testing. Then we utilize I3D as our backbone to extract spatial-temporal features，and we pretrain I3D on Kinetics dataset. The Kinetics dataset contains 400 classes of human actions, with over 400 unique video for each class. Every video has duration of about 10 seconds and is trimmed precisely.

4.2. Experiments Details

In order to make I3D network a better feature extractor we first pretrain the model on the huge video dataset Kinetics. With the help of open-source I3D model, the complicated I3D network is easy to converge and advance its generality. First of all, with the reference of RGB-I3D model in the work[5], we try to reproduce the experiments result and achieve 94.3% accuracy on UCF-101 with RGB modality. Inspired by previous works, we try to combine LSTM network with I3D to model high-level temporal features. We also utilize UCF-101 dataset to validate our proposed model and achieve the accuracy of 95.1%, which exceeds the original I3D model and other main-stream methods. The experiment environment we rely on is Ubuntu 16.04 LTS, Tensorflow1.2.0 and NVidia P40 GPU.

4.3. Analysis of Results

So as to prove that the proposed method in our work can outperform current main-stream models in general, we study and summarize some main-stream methods and make a thoughtful comparison about their advantages and disadvantages. C3D is a classic method for using 3D convolution kernels, which is natural for process signal with spatial-temporal features like videos. Nevertheless, the complicated structure of C3D architecture hinders itself from being deeper. Two-Stream[14] and TSN[15] are typical method for using spatial RGB modality and optical field temporal modality, however the two-stream networks overlook the fact that LSTM is better for sequence modelling rather than CNN networks. What's more, the LRCN model utilizes the CNN for learning spatial features and LSTM for learning spatial features, but it ignores the fact that 2D CNN is only powerful for extracting spatial features but not for extracting spatial-temporal features between adjacent frames. Consequently, in our research work we think highly of both low-level temporal features and high-level temporal features. In practice, we use Kinetics-pretrained I3D model as low-level spatial-temporal features extractor and LSTM as high-level spatial features modelling tool. The performance of main-stream methods and our proposed method are clearly showed in Table 1, and all of these approaches mentioned in the table is evaluated on UCF-101 video dataset.

Table 1. Compare with the state-of-the-art models

Model	UCF101
iDT[16]	86.4%
C3D[7]	85.2%
LRCN[11]	82.3%
TSN[15]	94.2%
Two-Stream[14]	88.0%
I3D[5]	94.3%
Ours	95.1%

5. Conclusion

In this paper, a new architecture that can achieve leading performance on UCF-101 video dataset is introduced. First we adopt Kinetics-pretrained I3D model to learn low-level features between adjacent frames, and then we utilize LSTM network to model high-level spatial features. The performance on the UCF-101 dataset also validates that the novel model introduced in our paper is more efficient and advanced than some other main-stream methods. In the future research, we will keep making efforts to advance the performance of our model on some other benchmark dataset.

Acknowledgments

This work is supported by the NSFC 61672089, 61273274, 61572064, and National Key Technology R&D Program of China 2012BAH01F03.

References

[1] Deng J , Dong W , Socher R , et al. ImageNet: A large-scale hierarchical image database[C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009.

[2] Kay W , Carreira J , Simonyan K , et al. The Kinetics Human Action Video Dataset[J]. 2017.

[3] Hara K , Kataoka H , Satoh Y . Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?[J]. 2017.

[4] Hochreiter S , Schmidhuber, Jürgen. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[5] Carreira J , Zisserman A . Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset[J]. 2017.

[6] Ji S , Xu W , Yang M , et al. 3D Convolutional Neural Networks for Human Action Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1):221-231.

[7] Tran D , Bourdev L , Fergus R , et al. Learning Spatiotemporal Features with 3D Convolutional Networks[J]. 2014.

[8] Heilbron F C , Escorcia V , Ghanem B , et al. ActivityNet: A large-scale video benchmark for human activity understanding[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, 2015.

[9] Szegedy C , Liu W , Jia Y , et al. Going Deeper with Convolutions[J]. 2014.

[10] Soomro K , Zamir A R , Shah M . UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild[J]. Computer Science, 2012.

[11] Donahue J , Hendricks L A , Rohrbach M , et al. Long-term Recurrent Convolutional Networks for Visual Recognition and Description[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 39(4):677-691.

[12] Ng Y H , Hausknecht M , Vijayanarasimhan S , et al. Beyond Short Snippets: Deep Networks for Video Classification[J]. 2015.

[13] He K , Zhang X , Ren S , et al. Deep Residual Learning for Image Recognition[J]. 2015.

[14] Simonyan K , Zisserman A . Two-Stream Convolutional Networks for Action Recognition in Videos[J]. 2014.

[15] L. Wang, Y. Xiong, Z. Wang, et al., Temporal segment networks: Towards good practices for deep action recognition, in European Conference on Computer Vision, 20–36, Springer (2016).

[16] Wang H , Schmid C . Action Recognition with Improved Trajectories[C]// 2013 IEEE International Conference on Computer Vision. IEEE, 2014.