FTSnet: 동작 인식을 위한 간단한 합성곱 신경망

조옥란, 이효종
전북대학교 컴퓨터공학부
e-mail : zhaoyulan27@naver.com, hlee@jbnu.ac.kr

FTSnet: A Simple Convolutional Neural Networks for Action Recognition

Yulan Zhao, Hyo Jong Lee*
Dept. of Computer Science and Engineering, Jeonbuk National University
*Corresponding author

Abstract

Most state-of-the-art CNNs for action recognition are based on a two-stream architecture: RGB frames stream represents the appearance and the optical flow stream interprets the motion of action. However, the cost of optical flow computation is very high and then it increases action recognition latency. We introduce a design strategy for action recognition inspired by a two-stream network and teacher-student architecture. There are two sub-networks in our neural networks, the optical flow sub-network as a teacher and the RGB frames sub-network as a student. In the training stage, we distill the feature from the teacher as a baseline to train student sub-network. In the test stage, we only use the student so that the latency reduces without computing optical flow. Our experiments show that its advantages over two-stream architecture in both speed and performance.

1. Introduction

Video action recognition is an important field in the community of computer vision with many societal applications such as surveillance, pilotless automobile, and more. Since convolutional neural networks (CNNs) have achieved great attentions in image classification [1, 2], they become the main method for action recognition gradually. Two-stream architecture [3] has been extremely popular which takes RGB frames and optical flow as input stream then fuses their scores as the final result.

However, the computation of optical flow requires hundreds of iterations on each pair of adjacent frames, it is a complex and complicated process which need too much resource so that it increases action recognition latency and limits in real-time applications.

In this paper, we propose a CNN for action recognition which base on both two-stream network and teacher-student architecture. There are two sub-networks in our neural networks, the optical flow sub-network as teacher and the RGB frames sub-network as student. In training stage, we distill the feature from teacher as baseline to train student. In test stage, we only use the student so that the latency reduces without computing optical flow.

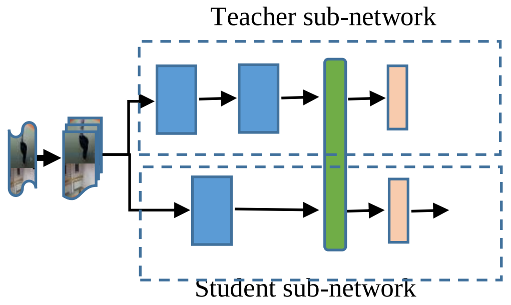
2. Related Work

Simonyan and Zisserman proposed the traditional two-stream network is 2D CNN model which has RGB frames and optical flow two sub-networks [3]. It takes video clips as the input stream, and clip is decomposed into spatial and temporal parts. The spatial part carries information about objects which is in the form of RGB frames stream. The temporal part conveys the movement of the objects which is in the form of optical flow stream. Each stream is implemented by a sub-network and the result is combined by late fusion. Feichtenhofer improved the two-stream network to fuse the two streams. They focused on 2D CNNs in the initial work, but they made transition to 3D CNNs because of the better spatiotemporal features.

The teacher-student architecture [4] consists of two parallel CNNs, a cumbersome or large model and a small one. The cumbersome network has much more parameters then the small network. In actual application, the large one will occupy more resources but its accuracy is better, the small network needs less resources and it fits for little data or fast result. The distillation is the knowledge acquired by large model can be transferred to small one. The small model can only be used to access the problem for fast and accurate results after distillation of knowledge. In process of distillation, the large model is as a teacher and the small is as the student. In recent research, Shumin K. *et al*. proposed single learning student network [5], and Pouya B. designed teacher guided architecture [6].

3. Methods

Our strategy is based on two-stream and teacher-student architecture as Figure 1. In training stage, we calculate the optical flow stream in teacher sub-network which possesses important motion information firstly, and train the network for action recognition and then freeze its weights. Secondly, we leverage the knowledge of optical flow to train student sub-network. In test stage, we only use student network with RGB stream to avoid optical flow computation to save the resources and time.



(Figure1) the architecture of FTSnet

3.1 Teacher Sub-network

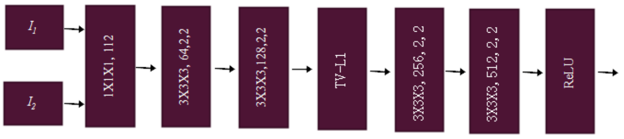
We use the video clips as input stream, then extract the RGB frame from clips. The optical flow extraction module calculates optical flow from RGB frame firstly in the teacher branch, and then feature extraction module computes the features of optical flow which denotes as F_{of}. Then we train teacher sub-network to classify actions using optical flow stream with a cross entropy loss function between the predicted class labels P_{of} and the true class labels T. Finally, the F_{of} is transmitted to the student sub-network to train it by back propagation.

$$L_{of} = CrossEntropy(P_{of}, T) \qquad (1)$$

The optical flow extraction bases on the brightness consistency assumption. That is approximating optical flow is formulated as followed:

$$I_2(x,y) = I_1(x+\Delta x, y+\Delta y) \qquad (2)$$

where $I_1(x,y)$ denotes the pixel at the location *(x, y)* of a frame at time *t*, $I_2(x,y)$ means the frames changed after time($t+\Delta t$), and is the spatial pixel displacements in *x* and *y* axes respectively. And we use the famous equation TV-L^1 [7] to extract optical flow via 3D residual networks as figure 2.



(Figure2) the architecture of FTSnet

3.2 Student Sub-network

The student sub-network extracts the feature of RGB frames that names as F_{rgb}, and receives the feature of optical flow as F_{of} from the teacher sub-network. We use the loss function Mean Squared Error (MSE) on both F_{rgb} and F_{of} to back propagating in network. Thereof, the RGB stream feature can simulate the features of optical flow stream to train the prior part of student sub-network.

And then we use a loss function about cross-entropy between the prediction of class that denotes P_{rgb} and the true class T combining of MSE to train the student sub-network entirely as Equation (2).

$$L_{RGB} = CrossEntropy\left(P_{rgb}, T\right) + \alpha \left\| F_{rgb} - F_{of} \right\|^2 \quad (3)$$

where α is the scalar weight as the influence of motive feature.

4. Experiment

We focused on the popular dataset for action recognition: HMDB51. HMDB51 consists of 51 action classes with more than 6800 videos and 3 splits for training and test. There are 3570 clips in training set and 1530 clips in test set. In our experiment, we extracted 25 RGB frames from each video clip randomly as input stream and sample a 112X112 crop in image. We choose the 3D ResNeXt-18 as the elementary CNNs modules, and the SGD optimization method with a weight decay of 0.0005, momentum of 0.9, and an initial learning rate 0.1.

We compared the performance of FTSnet with single RGB frames stream, single optical flow by TV-L1 and two-stream in table 1.

(Table 1) the Result of FTSnet

CNNs	HMDB51 Accuracy (%)
R3D-18 RGB	34.5
R3D-18 TV-L^1	36.5
Two-stream	46.6
ours(R3D-18)	54.5

5. Conclusion

In this paper, we introduced FTSnet, a simple model which only took RGB frames as input stream but extracted both appearance and motion information from stream. It implemented by training a sub-network to minimize the loss between its features and the features of optical flow stream, combined with the cross entropy loss for action recognition. Our result showed that our single-stream FTSnet architecture outperformed RGB and optical flow streams on popular benchmarks HMDB51. In future research, we will extend the other video dataset such as UCF101, and improve our network in architecture to get better performance.

REFERENCES

[1] Christian Szegedy, Wei Liu, Yangaing Jia, Pierre Sermane, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going Deeper with Convolutions". In CVPR, 2015.

[2] Ali Diba, Ali Mohammad Pazandeh, and Luc Van Gool. "Efficient two-stream motion and appearance 3D CNNs for video classification". In ECCV workshop, 2016.

[3] Karen Simonyan and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos". In NIPS, 2014.

[4] Geoffrey Hinton, Oriol Vinyal and Jeff Dean. "Distilling the Knowledge in a Neural Network". In NIPS, 2014.

[5] Shumin Kong, Tianyu Guo, Shan You and Chang Xu. "Learning Student Networks with Few Data". In AAAI, 2020.

[6] Pouya Bashivan and MarkTensen, "Teacher Guided Architecture Search". In ICCV, 2019.

[7] C. Zach, T. Pock, and H. Bischof . "A Duality Based Approach for Realtime TV-L1 Optical Flow". In DAGM, 2007.