KIPS Trans. Comp. and Comm. Sys.
Vol.11, No.3 pp.99~104
ISSN: 2287-5891 (Print), ISSN 2734-049X (Online)

https://doi.org/10.3745/KTCCS.2022.11.3.99

Teacher-Student Architecture Based CNN for Action Recognition

Yulan Zhao† · Hyo Jong Lee††

ABSTRACT

Convolutional neural network (CNN) generally uses two-stream architecture RGB and optical flow stream for its action recognition function. RGB frames stream display appearance and optical flow stream interprets its action. However, the standard method of using optical flow is costly in its computational time and latency associated with increased action recognition. The purpose of the study was to evaluate a novel way to create a two sub-networks in neural networks. The optical flow sub-network was assigned as a teacher and the RGB frames as a student. In the training stage, the optical flow sub-network extracts features through the teacher sub-network and transmits the information to student sub-network for baseline training. In the test stage, only student sub-network was operational with decreased in latency without computing optical flow. Experimental results shows that our network fed only by RGB stream gets a competitive accuracy of 54.5% on HMDB51, which is 1.5 times better than that on R3D-18.

Keywords : Two-Stream Network, Teacher-Student Architecture, CNN, Optical Flow, Action Recognition

동작 인식을 위한 교사-학생 구조 기반 CNN

Yulan Zhao† · 이 효 종††

요 약

대부분 첨단 동작 인식 컨볼루션 네트워크는 RGB 스트림과 광학 흐름 스트림, 양 스트림 아키텍처를 기반으로 하고 있다. RGB 프레임 스트림은 모양 특성을 나타내고 광학 흐름 스트림은 동작 특성을 해석한다. 그러나 광학 흐름은 계산 비용이 매우 높기 때문에 동작 인식 시간에 지연을 초래한다. 이에 양 스트림 네트워크와 교사-학생 아키텍처에서 영감을 받아 행동 인식을 위한 새로운 네트워크 디자인을 개발하였다. 제안 신경망은 두 개의 하위 네트워크로 구성되어있다. 즉, 교사 역할을 하는 광학 흐름 하위 네트워크와 학생 역할을 하는 RGB 프레임 하위 네트워크를 연결하였다. 훈련 단계에서 광학 흐름의 특징을 추출하고 교사 서브 네트워크를 훈련시킨 다음 그 특징을 학생 서브 네트워크를 훈련시키기 위한 기준선으로 지정하여 학생 서브 네트워크에 전송한다. 테스트 단계에서는 광학 흐름을 계산하지 않고 대기 시간이 줄어들도록 학생 네트워크만 사용한다. 제안 네트워크는 실험을 통하여 정확도 면에서 일반 이중 스트림 아키텍처에 비해 높은 정확도를 보여주는 것을 확인하였다.

키워드 : 양 스트림, 교사-학생 아키텍처, CNN, 광학 흐름, 동작 인식

1. Introduction

Video action recognition is an important function in the realm of computer vision with its applications including automated surveillance, self-driving vehicles and drone navigation. Convolutional neural networks (CNNs) have become standard of image classification[1,2]. Two-stream architecture CNN [3-5] has been extremely popular for action recognition exploiting RGB frames and optical flow as input stream then combining its feature to produce a final result. There are many models based on two-stream network such as optical flow guided feature(OFF) [13], hidden two-stream convolutional networks (H-TSCN) [14], and Actionflownet[15].

The optical flow computation for the action information in video is realized by calculating the displacement of the objects between each pair of adjacent frames. Each action in video frames lasts from 5 to 30 frames or longer in most video dataset. As the

※ This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education(GR 2019R1D1A3A03103736) and in part was supported by project for Cooperative R&D between Industry, Academy, and Research Institute funded Korea Ministry of SMEs and Startups in 20(Grant No S3114049).
※ 이 논문은 2021년 한국정보처리학회 ACK 2021의 우수논문으로 "FTSnet: 동작 인식을 위한 간단한 합성곱 신경망"의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 전북대학교 컴퓨터공학부 박사과정
†† 종신회원 : 전북대학교 컴퓨터공학부 교수
Manuscript Received : December 23, 2021
Accepted : January 24, 2022
* Corresponding Author : Hyo Jong Lee(hlee@jbnu.ac.kr)

frame increase in size the computational power of optical flow to process frame iterations increases. Subsequently, conventional optical flow process becomes a time consuming process with increase in latency and diminished function in real-time application.

The purpose of the study is to design and evaluate a novel CNN for action recognition based on two-stream network of teacher-student architecture [16]. 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 training stage, we trained the feature of the optical flow to teacher sub-network, and then transmitted the feature to student sub-network as a baseline to train the student sub-network. In the test stage, we only used the student sub-network to reduce latency eliminating computing optical flow.

2. Related Work

There are some significant progress with CNNs in computer vision tasks like object classification[2,17] and object detection [18]. When a large video datasets are published such as UCF101, HMDB51 [11,12], the CNNs are applied in action recognition [3,19,20].

2.1 Two-Stream Network

Traditional two-stream network [3] proposed by Simonyan *et al.* is a 2D CNN model with RGB frames and optical flow sub-networks. It uses video clips as the input stream, and decompose the clip into RGB frames. Then a stream of RGB frames functions as a spatial component while stream of optical flow calculates adjacent RGB frame as its temporal component. The spatial part carries appearance information about the objects, while the temporal part carries the movement information. Each stream is implemented by a sub-network and the result is combined by late fusion. In a research of two-stream network, Feichtenhofer *et al.* improved fusion of the two streams [4]. They initially focused on 2D CNNs, but transitioned to 3D CNNs for improved spatiotemporal features. Similarly, Diba *et al.* used C3D to learn motion from optical flow for end-to-end applications [5].

2.2 Teacher-Student Architecture

There are various designs of teacher-student networks [6-9]. The teacher-student architecture consists of two parallel CNNs, a large and a small models [6]. Traditionally, the large model has many nodes and parameters than the small one and often results in better outcome. The small model can process small dataset with fast result. In real-world applications, the large model spends costly resources and time, while the small network needs less resources, it only fits for little data for fast result. The distillation can transfer the knowledge learned by large model to small one, and the small model can use the distillation to access the problem and gain fast and accurate results. In process of transferring the distillation, the large model acts as a teacher and the small model as a student. In recent research, Kong et al. proposed the single learning student network to tackle the challenges of learning student networks with few data [7], and Bashivan et al. designed teacher guided architecture to gain more computational efficiency [8].

3. Proposed Method

Our model is based on two-stream network and teacher-student architecture [16] as shown in Fig. 1. We used video clips as input stream and extracted the RGB frames to feed the teacher-student network. There are two sub-networks in our architecture, teacher sub-network for optical flow branch and the student sub-

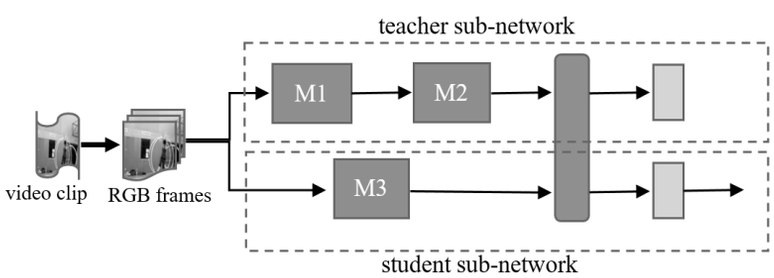

Fig. 1. Architecture of Action Recognition

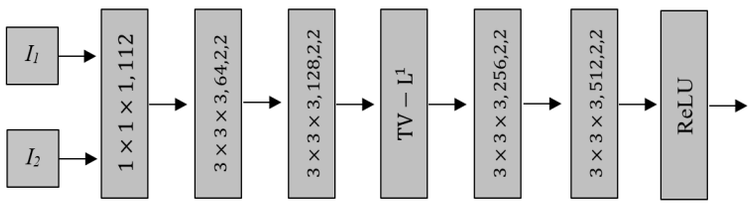

Fig. 2. M1: The Module of Optical Flow Extraction

network for RGB frames branch. We divided the processing of network into two stages. In a training stage, we calculated the optical flow stream in teacher sub-network to possess important motion information. Then information is used to train the network for action recognition and then freeze on its finalized weights. We then used the knowledge of optical flow to train the student sub-network. In the test stage, we only used student sub-network with RGB frames stream to avoid optical flow computation to save resources and time.

3.1 Teacher Sub-network

There are two modules in the teacher sub-network, the optical flow extraction module is named as M1 and action recognition module as M2. The module M1 calculates optical flow from the sequential RGB frames. The module M2 extracts the features of optical flow.

In the module M1, we don't employ complex networks [21,22] to compute optical flow, its extraction is based on the brightness consistency assumption between the sequential images I_1 and I_2, with small change in moment of the object. The approximating optical flow is formulated as shown in Equation (1).

$$I_2(x,y) = I_1(x+\triangle x, y+\triangle y) \tag{1}$$

$I_1(x,y)$ denotes the object at the location (x,y) of the image at time t. $I_2(x,y)$ is the object at the location in the image after time $\triangle t$, and $(\triangle x, \triangle y)$ is the object spatial pixel displacements in x and y axes, respectively. It can be approximated with a Taylor series as shown Equation (2).

$$I_2 = I_1 + \frac{\delta I}{\delta x}\triangle x + \frac{\delta I}{\delta y}\triangle y \tag{2}$$

Zach *et al.* proposed TV-L^1 method [10,11] to calculate the optical flow approximately. The total variational method estimates the optical flow by an iterative optimization method. The tensor $u \in R^{2\times W\times H}$ is the x and y directional optical flow for each location of the object in the image. The method first computes the gradient in both x and y directions: ∇I_2. The initial optical flow is set to $u=0$. The ρ denoting the image residual between I_1 and I_2, the iterative optimization is performed with updating u,v,p as shown in Equation (3)-(6).

$$u = v + \theta \bullet dvg(p) \tag{3}$$

$$v = u + \begin{cases} \lambda\theta\nabla I_2 & \rho < \lambda\theta|\nabla I_2|^2 \\ -\lambda\theta\nabla I_2 & \rho > \lambda\theta|\nabla I_2|^2 \\ -\rho\dfrac{\nabla I_2}{|I_2|^2} & |\rho| \le \lambda\theta|\nabla I_2|^2 \end{cases} \tag{4}$$

$$p = \frac{p + \frac{\tau}{\theta}\nabla u}{1 + \frac{\tau}{\theta}|\nabla u|} \tag{5}$$

$$(dvg(p)) = \begin{cases} p^1_{i,j} - p^1_{i-1,j} & \text{if } 1<i<N \\ p^1_{i,j} & \text{if } i=1 \\ -p^1_{i-1,j} & \text{if } i=N \end{cases} + \begin{cases} p^2_{i,j} - p^2_{i,j-1} & \text{if } 1<j<M \\ p^2_{i,j} & \text{if } j=1 \\ -p^2_{i,j-1} & \text{if } j=M \end{cases} \tag{6}$$

where hyper-parameter θ is the weight of the TV-L^1 regularization term, λ is weight of the data term, τ is the time-step and $\tau \le \frac{1}{8}$. The dual vector field p is used to minimize the energy.

We computed the optical flow using formula mentioned above via residual networks as shown in Fig. 2.

The module M2 extracts the features of optical flow denoted as F_{of} in Fig. 3. Subsequently, we trained the 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 its network by back propagation.

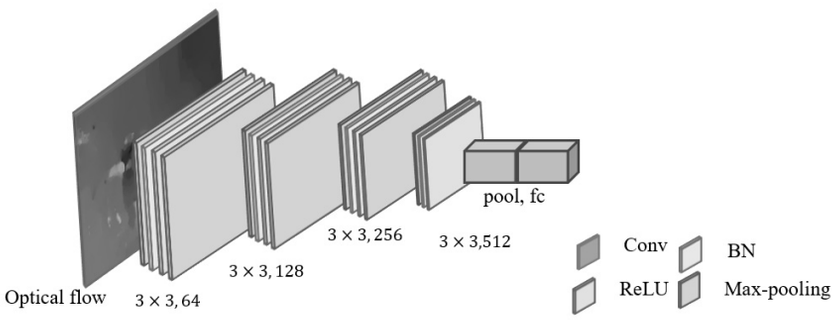

Fig. 3. M2: The Module of Optical Flow Feature Extraction

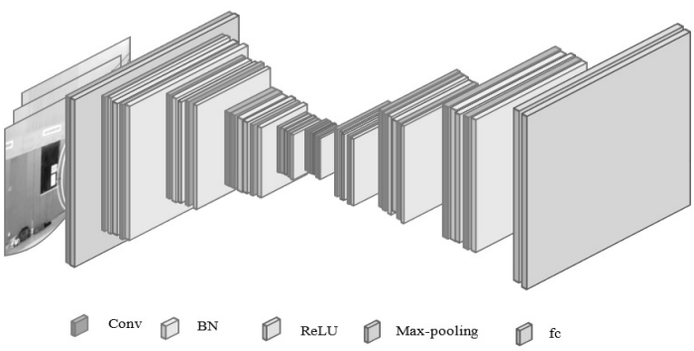

Fig. 4. M3: The Module of RGB Frames Action Recognition

3.2 Student Sub-network

The module M3 in student sub-network extracts the feature of RGB frames as F_{rgb} , and receives the feature of optical flow from the teacher sub-network as F_{of} shown in Fig. 4.

We used the loss function of Mean Squared Error (MSE) on both F_{rgb} and F_{of} to back propagate its network. Thereof, the RGB stream feature can simulate the features of optical flow stream to train the early part of student sub-network. We then used a loss function of cross-entropy between the predictive class denoted as P_{rgb} and the true class T *with* MSE to train the student sub-network entirely as shown in Equation (7).

$$L_{RGB} = CrossEntropy(P_{rgb}, T) + \lambda \| F_{rgb} - F_{of} \|^2 \qquad (7)$$

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

4. Experiment

We focused on the popular dataset for action recognition: HMDB51 [11]. It 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 an input stream and a sample of 224×224 cropped image.

In the action recognition modules, we used the SGD optimization method with a weight decay of 0.0005, momentum of 0.9, and an initial learning rate of 0.1. The accuracy of our experiment on HMDB51 was 54.5%. Table 1 includes R3D-18 with single RGB, TV-L[1] and two-stream network modules and its accuracy.

The optical flow frames are shown in Fig. 5. The three columns on the left are RGB frames extracted from video clips, and the two images on the right are two optical flow frames u extracted from the adjacent two RGB frames.

In the first row, the action class is ‘pour’. There is a significant movement in frames with obvious change

Table 1. Experimental Results

CNNs	Accuracy (%)
R3D-18 RGB	34.5
R3B-18 TV-L[1]	36.5
Two-stream	46.6
Ours	54.5

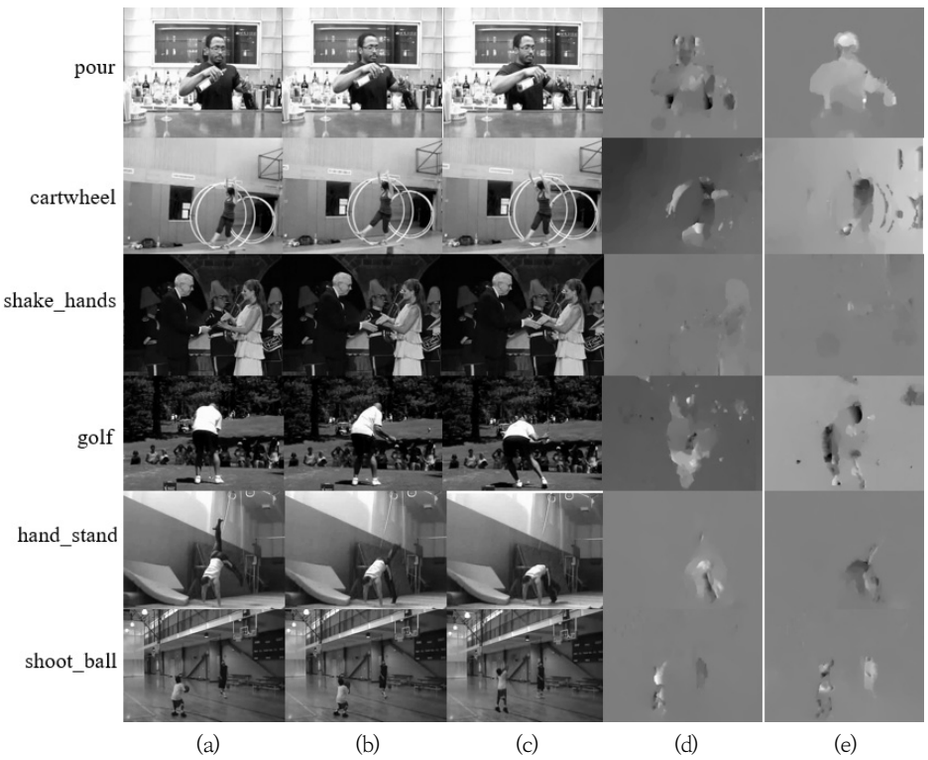

Fig. 5. Examples of Optical Flow: For Every Row, The Image (d) is The Optical Flow Between Frame (a) and (b), While Image (e) is The Optical Flow Between (b) and (c).

in optical flow trajectory in frame (d) and (e). In the second row, the action class is 'cartwheel', with the objects including a girl and a cartwheel. We can see the change of movement information from (d) and (e). In the third row, the class is 'shake_hands'. The main motion is of two people shaking hand. There is no obvious movement change of object between frames (b) and (c) from the optical flow image (e). In the forth row, the main motion is hitting a golf ball by the man. We can see obvious movement changes from (d) and (e). In the fifth row, the class is 'hand_ stand'. The movements and posture of the active man have changed significantly. In the sixth row, the class is 'shoot_ball'. The action changes of the boy's shooting and the coach's posture can be seen from (d) and (e).

5. Conclusion

We introduced an architecture for action recognition based on teacher-student neural network. The student model only took video clips as an input stream but was able to extract both the appearance and the motion information. The model worked by training a sub-network to minimize the loss between its features and the features of optical flow stream and combining cross entropy loss for action recognition. Our architecture resulted in higher accuracy in HMDB51 benchmark compared to other popular methods. The proposed method is sensitive and affected by lighting changes and camera motion between frames. In future studies, we will improve these defects and expand to other video dataset such as UCF101 [12], and improve our network in architecture to get better performance.

References

[1] C. Szegedy, W. Liu, Y. Jia, P. Sermane, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the Computer Vision and Pattern Recognition*, pp.1-9, 2015.

[2] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," in *Proceedings of the Computer Vision and Pattern Recognition*, pp.2285-2294, 2016.

[3] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proceedings of the Neural Information Processing*, pp.568-576, 2014.

[4] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the Computer Vision and Pattern Recognition*, pp.1933-1941, 2016.
[5] A. Diba, A. Pazandeh, and L. V. Gool, "Efficient two-stream motion and appearance 3D CNNs for video classification," *arXiv:1608.08851*, 2016.
[6] G. Hinton, O. Vinyal, and J. Dean, "Distilling the knowledge in a neural network," in *Neural Information Processing Deep Learning Workshop*, 2014.
[7] S. Kong, T. Guo, S. You, and C. Xu, "Learning student networks with few data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol.34, No.4, pp.4469-4476, 2020.
[8] J. P. Bashivan, M. Tensen, and J. J. DiCarlo, "Teacher guided architecture search," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.5320-5329, 2019.
[9] D. Shah, V. Trivedi, V. Sheth, A. Shah, and U. Chauhan, "ResTS: Residual deep interpretable architecture for plant disease detection," *Information Processing in Agriculture*, https://doi.org/10.1016/j.inpa.2021.06.001.
[10] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime TV-L1 optical flow," in *DAGM 2007: Pattern Recognition*, Vol.4713, pp.214-223, 2007.
[11] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp.2556-2563, 2011.
[12] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.
[13] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, and W. Zhang, "Optical flow guided feature: A fast and robust motion representation for video action recognition," in *Proceedings of the Computer Vision and Pattern Recognition*, pp.1-9, 2018.
[14] Y. Zhu, Z. Lan, S. Newsam, and A. G. Hauptmann, "Hidden two-stream convolutional networks for action recognition," *arXiv preprint arXiv:1704.00389*, 2017.
[15] J. Y.-H. Ng, J. Choi, J. Neumann, and L. S. Davis, "Actionflownet: Learning motion representation for action recognition," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp.1616-1624, 2018.
[16] Y. Zhao and H. Lee,"FTSnet: A simple convolutional neural networks for action recognition," in *Proceedings of the Annual Conference of KIPS(ACK) 2021*, pp.878-879, 2021.
[17] K. He, X. Zhang, S. Ren, and J. Sun,"Deep residual learning for image recognition," in *Proceedings of the Computer Vision and Pattern Recognition*, pp.770-778, 2016.
[18] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim, "Robust facial landmark detection via recurrent attentive-refinement networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp.57-72, 2016.
[19] Z. Wang, Q. She, and A. Smolic, "ACTION-Net: Multipath excitation for action recognition," in *Proceedings of the Computer Vision and Pattern Recognition*, pp.13214-13223, 2021.
[20] L. Wang, Z. Tong, B. Ji, and G. Wu, "TDN: Temporal difference networks for efficient action recognition," in *Proceedings of the Computer Vision and Pattern Recognition*, pp.1895-1904, 2021.
[21] T. Hui, X. Tang, and C. C. Loy, "A lightweight optical flow CNN-Revisiting data fidelity and regularization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.43, No.8, pp.2555-2569, 2021.
[22] K. Luo, C. Wang, S. Liu, H. Fan, J. Wang, and J. Sun, "UPFlow: Upsampling pyramid for unsupervised optical flow learning," in *Proceedings of the Computer Vision and Pattern Recognition*, pp.1045-1054, 2021.

Yulan Zhao

https://orcid.org/0000-0002-9469-5119
e-mail : zhaoyulan27@naver.com

She received a M.S. degree in computer science from Northeast Electric Power Univ. in 2009. She is currently pursuing the Ph.D. degree in the Department of Computer Science and Engineering with Jeonbuk National Univ. Her research interests are computer vision, image processing, artificial intelligence and action recognition.

Hyo Jong Lee

https://orcid.org/0000-0003-2581-5268
e-mail : hlee@jbnu.ac.kr

He received a Ph.D. degree in computer science from the University of Utah in 1991. He has been a professor at Jeonbuk National Univ. since 1991. His research interests include computer graphics, image processing, and parallel processing, and artificial intelligence.