Decomposed “Spatial and Temporal” Convolution for Human Action Recognition in Videos

Khwaja Monib Sediqi[1], Hyo Jong Lee[1,2]
[1]Dept. of Computer Science & Engineering, Chonbuk National University
[2]Center for Advanced Image and Information Technology

Abstract

In this paper we study the effect of decomposed spatiotemporal convolutions for action recognition in videos. Our motivation emerges from the empirical observation that spatial convolution applied on solo frames of the video provide good performance in action recognition. In this research we empirically show the accuracy of factorized convolution on individual frames of video for action classification. We take 3D ResNet-18 as base line model for our experiment, factorize its 3D convolution to 2D (Spatial) and 1D (Temporal) convolution. We train the model from scratch using Kinetics video dataset. We then fine-tune the model on UCF-101 dataset and evaluate the performance. Our results show good accuracy similar to that of the state of the art algorithms on Kinetics and UCF-101 datasets.

1. Introduction

Recognition of human action in videos is a challenging task that has recently received a substantial amount of attention among researchers of computer vision [1]. Video based human action recognition has a vast and variant application area such as surveillance systems, robotics, health care and human-computer interaction. Unlike classification in still images which is concerned with spatial information only, video data contains temporal information, which makes the classification task more challenging.

Action recognition from a video stream can be defined as recognizing human actions automatically using a pattern recognition system with minimal human-computer interaction. Typically, an action recognition system analyzes certain video sequences or frames to learn the patterns of a particular human action in the training process and use the learn knowledge to classify novel actions during the test phase [2]

In this work we aim to decompose the 3D convolution of 3D ResNet-18 into 2D and 1D convolution, respectively. This task has recently been addressed in [3] by using 34-layers R(2+1)D net and utilizing Sports-1M video dataset.

2. Related Work

Video content analysis is one of the core problems in computer vision and has been studied for decades. Many research contributions in video processing have focused on developing spatiotemporal feature representation for video content analysis. A family of video content representation methods is based on shallow high-dimensional encodings of local spatiotemporal features. Some of these video representations include Spatiotemporal Interest Points (STIPS) [4], Histogram of Oriented Gradients (HOG) [5], Motion Boundary Histogram [6] Cuboids [7], and Action Bank [8]. These feature representations are hand-crafted and use different encoding technique such as those based on histogram or pyramids. Among these hand-designed representations, improved Dense Trajectories (iDT) [9] is widely considered the state of the art due to its bold results on video classifications.

With the breakthrough of deep learning in still-image recognition originated by AlexNet model [10] researchers devoted significant contribution to design similar model for video. 3D CNNs using temporal convolutions for recognizing human actions in video were arguably first proposed by Baccouche et al [11]. More recently 3D CNNs were shown to lead to strong action recognition results when trained on large datasets. 3D CNNs features also generalize well to other tasks, including action detection [12], video captioning [13] and gesture detection [14]

3. Decomposed Spatial and Temporal Convolutions

A possible theory is that a full 3D convolution maybe approximated by a 2D (Spatial) convolution followed by a 1D (Temporal) convolution, decoupling spatial and temporal modeling into two separate steps [3].

We decomposed the spatiotemporal convolution to two separate steps of spatial and temporal convolution by replacing *Ni* 3D convolution kernels of size $N_{i-1} \times t \times d \times d$ to $N_{i-1} \times 1 \times d \times d$ (spatial convolution) and $M_i \times t \times 1 \times 1$ (temporal convolution). While N_i denotes the number of convolutional filters applied in the *i*-th residual block ResNet, *t* and d represent temporal and spatial dimensions, respectively. The hyper-parameter *Mi* determines the dimensionality of the intermediate subspace to project signal between spatial and temporal convolution.

4. Experiment

In this section we describe the preliminary experiment on the Kinetics video dataset [15]. We constrain our experiment to shallow residual network. Owing to their good performance and simplicity, we use 18 layers of 3 dimensional residual network in our experiment. Table 1 provides the specifications of the 3D ResNet-18 that we considered in our experiment. In the table, convolutional residual blocks are shown in brackets, next to the number of times each block is repeated in the stack. Also, dimension given for filters and outputs are time, height, and width in this order. The purpose of this experiment is to explore the decomposed architecture of 3D ResNet-18 for action recognition. In this experiment we trained the 18-layers of the 3D ResNet network in a way that first we factorized the spatiotemporal convolution to spatial and temporal convolution. We use Kinetics video dataset as a primary benchmark for training since its large enough to train a network from scratch. To train the decomposed spatial and temporal ResNet-18 on the Kinetics dataset, we use SGD with a mini-batch size of 256 and train the model utilizing 4 GPUs (NVIDA Titan X). We set the weight decay to 0.001 and momentum of 0.9. We start from learning rate 0.1, and divide it by 10 for three times after the validation loss saturates. Batch normalization is applied to all convolutional layers. We then feed the network with clips consist of pre-extracted frames of size 112x112. We use stride 1x2x2 and got a different size of output. Regardless of the size of output produced by the last convolutional layer, the network applies global spatiotemporal average pooling to the final convolutional tensor, followed by a fully-connected (fc) layer performing the final classification. Since a model must also support transfer learning to other datasets, we implement transfer learning by fine-tuning the last layers of the model using UCF-101 video dataset [16].

<Table 1> R3D architectures considered in our experiments.

Layer name	Output size	R3D - 18
Conv1	L x 56 x 56	3 x 7 x7, 64, stride 1 x 2 x 2
Conv2_x	L x 56 x 56	$\begin{bmatrix} 3x3x3, 64 \\ 3x3x3, 64 \end{bmatrix} x\ 2$
Conv3_x	$\frac{L}{2}$x 28 x 28	$\begin{bmatrix} 3x3x3, 128 \\ 3x3x3,128 \end{bmatrix} x\ 2$
Conv4_x	$\frac{L}{4}$x 14 x 14	$\begin{bmatrix} 3x3x3,256 \\ 3x3x3, 256 \end{bmatrix} x\ 2$
Conv5_x	$\frac{L}{8}$x 7 x 7	$\begin{bmatrix} 3x3x3, 512 \\ 3x3x3,512 \end{bmatrix} x\ 2$
	1x 1 x 1	Spatiotemporal pooling, fc layer with softmax

5. Result

In this section we present our experiment result. We train the network from scratch using Kinetics video dataset. The dataset is provided in three splits as of the training, validation and test sets. We train the network using the train split of the dataset, validate it on the validation split and provide result on the test split of the dataset. Figure 1 shows the training and validation accuracy of the model on the training and validation split of the Kinetics dataset. We then fine-tuned the network by training the last fully connected (fc) layer of the model by using benchmark dataset UCF-101. We test the network on the test split of the UCF-101 dataset and report the result. Table 2 illustrates the top-1 clip and top-5 clip accuracy for both Kinetics and UCF-101 dataset. Our preliminary result indicates that a shallow decomposed spatial and temporal convolutional ResNet achieves similar result to that of the state of the art results as of the other networks for action recognition. It is notable that our model shows good performance without overfitting because of using large-scale Kinetics dataset.

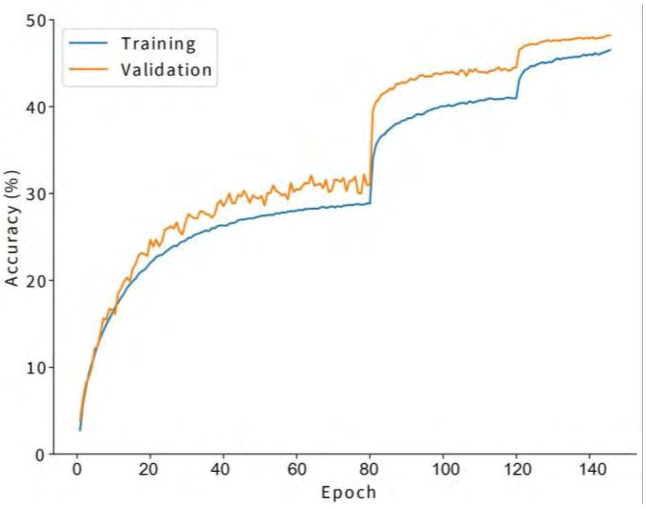

Figure 1: Training of the model on the Kinetics dataset. Kinetics dataset is large enough to train networks from scratch. Thus, helped to avoid overfitting.

<Table 2>Decomposed S&T Conv. ResNet-18 Accuracy on the Kinetics and UCF-101 datasets.

Method	Kinetics		UCF101	
	Top-1	Top-5	Top-1	Top-5
PoTion [17]	-	-	-	65.2
I3D-RGB [15]	67.5	87.2	-	95.6
I3D-Two-Stream [15]	**75.7**	**92.0**	-	**98.0**
De. S&T ResNet-18 (ours)	66.5	87.1	66.7	88.2

6. Conclusion

In this research we explored decomposed “spatial and temporal” convolution utilizing 3D ResNet-18. We trained our model from scratch using the Kinetics dataset. Our shallow model achieved comparable result to that of the state of the art on Kinetics and UCF-101. We provided two model, one pre-trained (on Kinetics) and the other fine-tuned (on UCF-101). We hope that our analysis will inspire new ideas and help improve the efficacy and modeling for spatiotemporal convolution. Our future work will be devoted to search and improve further state of the art architectures for human action recognition.

Acknowledgment. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (GR 2016R1D1A3B03931911)

7. References

[1] A. Karpathy, G.Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, Proc, 2014.

[2] A. Gilbert, J. Illingworth, and R. Bowden, "Action recognition using mined heirarachical compound

features," *IEEE Transactions on Pattern Analysis and Machine Inteligence*, vol. 33, no. 5, pp. 883 - 897, 2011.

[3] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun and Manohar Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *CVPR*, 2017.

[4] I. Laptev and T. Lindeberg, "Space-time interest points," in *ICCV*, 2003.

[5] N. Dalal and B Triggs, "Histogram of Oriented Gradients for Human Detection," *Proc. CVPR*, vol. 2, pp. 886 - 893, 2005.

[6] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European Conference on Computer Vision*, 2006.

[7] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features.," in *Proc. ICCV VS-PETS*, 2005.

[8] S. Sadanand and J. Corso, "Action bank: A high level representation of activity in video," in *CVPR*, 2012.

[9] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *ICCV*, 2013.

[10] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[11] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential Deep Learning for Human Action Recognition," in *Springer Berlin Heidelberg*, Berlin, Heidelberg, 2011.

[12] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," in *CVPR*, 2016.

[13] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in *CVPR*, 2016.

[14] Z. Qiu, T. Yao, and T.Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *CVPR*, 2017.

[15] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017.

[16] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human action classes from videos in the wild," in *CRCV-TR-12-01*, 2012.

[17] Vasileios Choutas, Philip Weinzaepfel, Jerrom Revaud and Cordelia Schmid, "Potion: Pose motion representation for action recognition," in *CVPR*, 2018.