The Very Deep Multi-stage Two-stream Convolutional Neural Network for Action Recognition

Xiuju Gao [1, a], Hanling Zhang [2,b]

[1]College of Information Science and Engineering, Hunan University, Changsha 410082, China

[2]College of Information Science and Engineering, Hunan University, Changsha 410082, China

[a]gaoxiuju5@163.com, [b]jt_hlzhang@hnu.edu.cn

Keywords: action recognition; convolutional neural network; multi-stage training

Abstract. In this paper, we consider the very deep multi-stage two-stream convolutional neural network for action recognition in videos. The challenge of action recognition is to capture the appearance and motion information to describe various actions efficiently and to classify different levels of difficult videos correctly. The proposed new deep architecture we name the very deep two-stream convolutional neural network has preferable model capacity and it enables us to obtain appearance and motion information validly from image frames in videos. Besides, with the proposed multi-stage training strategy, multiple classifiers are jointly optimized to process samples at different difficulty levels. Finally, the Dynamic Random Forests classifier is employed to replace Softmax classifier or SVM, achieving a decent classification result. Our architecture is trained and evaluated on the standard video actions benchmarks of UCF-101, and it is competitive with the state of the arts.

Introduction

Recognizing human action is one of the important areas of computer vision research today. Its aim is to automatically analyze human action from the information acquired from a video. It applies in the situations including surveillance, video analysis, and robotics, and so on.

As a class of attractive deep learning models can automate the process of feature construction through learning features of various hierarchies by building high-level features from low-level ones [1,2], Convolutional Neural Networks (CNNs) have been primarily applied for object detection and action recognition. In this paper, we propose the very deep multi-stage two-stream convolutional neural network for action recognition. As Convolutional Networks (ConvNets) can capture the features such as edge, parts and combinations of low-level descriptions, some efforts [3,4] was made to go deeper with convolutions by increasing the depth and width of the network to expand model capacity. Besides, we notice that one classifier may give us an error result when the action is motion blur or partially occluded, and cascading few classifiers can improve the recognition performance demonstrated by [5]. Different classifiers are used in off-the-shelf networks including Softmax classifier, SVM, and so on. In this paper, we propose multi-stage training for the very deep two-stream ConvNets. The experimental result shows that the Dynamic Random Forests classifier achieves the best accuracy relative to Softmax classifier and SVM.

The contributions of this paper are as follows:

1) Instead of the inherent Softmax classifier or SVM, we verify that the Dynamic Random Forests classifier achieves impressive results in UCF-101 dataset.

2) We introduce multi-stage training for the very deep two-stream ConvNets and it can improve the recognition result by dealing with different qualities of action videos.

Related Work

A group of works [6,7] on human action and object recognition mainly focused on developing robust and descriptive features. With ConvNets enjoying great success in computer vision field, a number of attempts have been made to improve the original architecture of [8] to achieve better accuracy. Ouyang et al. [9] proposed multi-stage and deformable deep convolutional neural networks for object detection.

Multiple classifiers were jointly optimized to process samples at different difficulty levels and the deformation constrained pooling layer modeled the deformation of object parts with geometric constraint and penalty. Simonyan and Zisserman [10] captured the complementary information on appearance from still frames and motion between frames by using a two-stream ConvNet which incorporated spatial and temporal networks, demonstrating that a ConvNet trained on multi-frame dense optical flow was able to achieve good performance in spite of limited training data.

In order to consider the motion information, Ji et al. [11] developed a 3D CNN model for action recognition. This model extracted features from both spatial and temporal dimensions by performing 3D convolutions, capturing the motion information encoded in multiple adjacent frames. Wang et al. [12] built the network consisting of 3D convolutions and max-pooling operators over the video segments for automatic activity recognition from RGB-D videos.

To address relatively shallow models capacity is constrained by their depth. Simonyan and Zisserman [13] showed that a significant improvement on the prior-art configurations can be achieved by setting the depth to 16 - 19 weight layers. Szegedy et al. [14] proposed a deep convolutional neural network architecture codenamed Inception which allowed for increasing the depth and width of the network while keeping the computational budget constant for classification and detection in the ImageNet Large-Scale Visual Recognition Challenge 2014.

Method

In this section, we first introduce the architecture of our very deep two-stream ConvNet and then detail out the modification of the model. After these, we describe the training details. Finally, we present the classifier used in this paper.

The Very Deep Two-Stream Convolutional Neural Network. Network architectures have great importance in feature extraction. We choose VGGNet [6] to design the network as VGGNet is a successful network for action recognition due to the fact that it uses relatively small convolutional kernel size (3 × 3), small pooling window (2 × 2) and deep structure (16 or 19 layers).

Two-stream convolutional neural network can capture some actions which are strongly associated with particular objects from still frames (the spatial recognition stream) and describe the motion between video frames (the temporal recognition stream). The input to a Spatial ConvNet is a fixed-size 224 × 224 RGB image and the input to a Temporal ConvNet is 10-frame stacking of optical flow fields and the same size as Spatial ConvNet. The image is passed through a stack of convolutional layers and two fully connected layers which have 4096 channels each are behind convolutional and pooling layers. All weight layers are equipped with the rectification non-linearity.

Optical flow is a crucial component of video classification approach because it encodes the pattern of apparent motion of objects in a visual scene. Our temporal stream ConvNet operates on multiple-frame dense optical flow, which is based on the minimization of a function containing a data term using the L1 norm and a regularization term using the total variation of the flow [15]. The feature allows discontinuities in the flow field, while being more robust to noise than the classical approach by Horn and Schunck [16].

Modification. Multi-stage classifiers have been widely used in object detection and achieved great success [17,18]. In this paper, we present a new deep architecture that can jointly train multiple classifiers through several stages of back-propagation for action recognition. The baseline very deep model and fully connected layers with multi-stage training are expressed in Fig. 1. Each stage handles samples at different difficulty levels.

As shown in Fig. 1, besides fc6, pool5 is connected to T extra fully connected layers of sizes 4096. The baseline deep model is first trained by excluding extra classifiers to reach a good initialization point. Then the extra classifiers are added stage-by-stage. At stage t, all the existing classifiers up to layer t are jointly optimized. Each round of optimization finds a better local minimum around the good initialization point reached in the previous training stages.

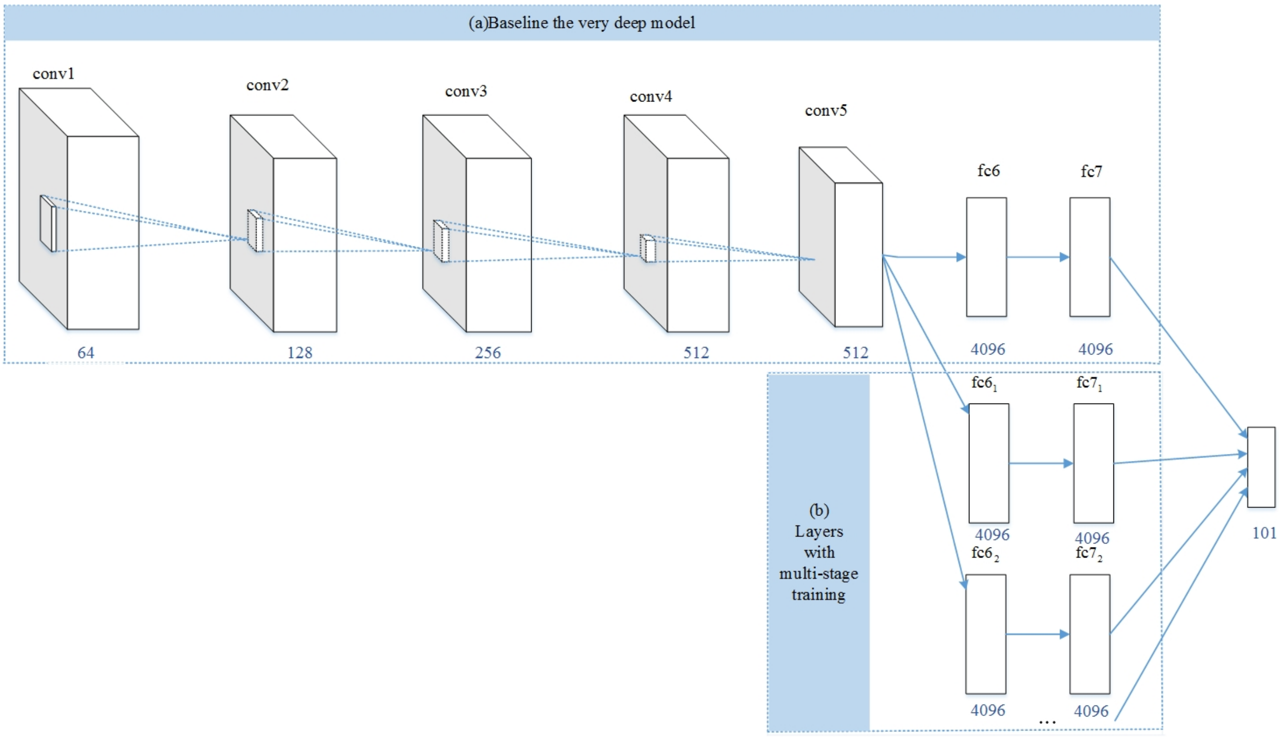

Fig. 1 The baseline very deep model and fully connected layers with multi-stage training.

Network Training. The initialisation of the network weights is important. Pre-training is an effective way to train deep models when training data are not enough. We use ImageNet models to pre-train the novel two-stream ConvNets.

We use mini-batch gradient descent based on back-propagation with momentum to train the new two-stream ConvNets. The batch size was set to 256, momentum to 0.9. The learning rate is set relatively small since it may miss the optimal solution when the value is large. For spatial net, the learning rate starts with 0.001, decreases to its 1/10 every 4,000 iterations, stops at 10,000 iterations. For temporal net, the learning rate starts with 0.005, decreases to its 1/10 every 10,000 iterations, stops at 30,000 iterations.

Classifier Training and Testing. Given a trained ConvNet and input images, we firstly extract the output of fc7 layer as features to train the Dynamic Random Forests (DRFs) classifiers and then test the recognition accuracy in the UCF-101 dataset [19].

Dynamic Random Forest is based on a sequential procedure that builds an ensemble of random trees by making each of them dependent on the previous ones. The DRF algorithm exploits the same idea with the adaptative resampling process of boosting and combines it with the randomization processes used in “classical” RF induction algorithms. The algorithm of Dynamic Random Forests (DRFs) is detailed in [20].

Classification Experiments

Dataset and Implementation Details. The UCF101 dataset contains 101 action classes and there are 13, 320 video clips. The videos are separated into 5 broad groups: Human-Object Interaction, Body-Motion Only, Human-Human Interaction, Playing Instruments and Sports. We adapt the three training/testing splits for evaluation as description in [19] and fuse the spatial and temporal networks with the weights of 1/3 and 2/3. We use the Caffe framework implementing our approach which is a kind of deep learning framework that focuses on cleanliness, readability, and speed.

Results. We demonstrate the recognition accuracy of three different classifiers on the UCF101 in Table 1. The classifiers include Softmax, SVM and the Dynamic Random Forests. We use the single classifier without multi-stage training here when we conduct this experiment. From Table 1, we see that the DRFs classifier outperforms the other two classifiers at least 0.3%. We use the DRFs as our classifier to do the next experiments.

Table 1: The recognition accuracy of three different classifiers without multi-stage training.

Classifier	Softmax	SVM	DRFs
Accuracy [%]	91.4	90.8	91.7

Table 2: Performance comparison with the state of the art on UCF101 dataset.

Method	Accuracy [%]
C3D (3 nets) + iDT + linear SVM [7]	90.4
LSTM with 30 Frame Unroll (Optical Flow + Image Frames) [21]	88.6
Very deep two-stream [4]	91.4
Very deep two-stream + DRFs	91.7
Very deep two-stream + multi-stage training	92.1

We further compare the result of our model with state of the arts methods in Table 2. We first compare with 3D Convolutional Networks learning spatiotemporal feature [7]. From Table 2, our result is better than C3D and F_{ST}CN model evidently. Besides that, we also list the result of Long Short Term Memory (LSTM) networks in [21] on the UCF-101 dataset. We see that our proposed models get 3.5% higher accuracy than LSTM model. Finally, we perform comparisons between the very deep two-stream in [4] with our model. Our model outperforms the very deep two-stream in [4] and is better than it by 0.7%. From the last two rows, we can know that our model is efficient for action recognition, and the multi-stage training can classify more difficult actions than single classifier because the multi-stage training gets 0.5% higher accuracy than only one classifier.

Conclusions

This paper proposes the very deep multi-stage two-stream convolutional neural network that learns feature extraction and classification for action recognition in videos. The model incorporates separate spatial and temporal streams based on ConvNets, using RGB and optical flow as inputs to capture both appearance and motion information. The whole model considers the depth of convolutional networks achieving excellent model capacity. We choose VGG16 as our ConvNets, and jointly train multiple classifiers through multi-stage of back-propagation dealing with different difficulty levels action classification. Adding Dynamic Random Forests classifier to the model, experiments results on UCF-101 show that the whole model has competitive performance.

References

[1] R. Girshick: Fast R-CNN[C]. 2015 IEEE International Conference on Computer Vision (ICCV). IEEE Computer Society, (2015):1440-1448.

[2] R. Girshick, J. Donahue, T. Darrell, et al: Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]. Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE, (2014):580-587.

[3] X. Glorot, A. Bordes, Y. Bengio: Deep Sparse Rectifier Neural Networks[J]. Journal of Machine Learning Research, (2010), 15.

[4] L. Wang, Y. Xiong, Z. Wang, et al: Towards good practices for very deep two-stream ConvNets. ArXiv 1507.02159, (2015).

[5] S. Chauhan, and Prof. Prema K.V: Performance evaluation of multistage classifier. International Journal of Emerging Technology and Advanced Engineering. Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 3, March 2013).

[6] L. Sun, K. Jia, D.Y. Yeung: Human Action Recognition Using Factorized Spatio-Temporal Convolutional Networks[C]. International Conference on Computer Vision (ICCV), (2015).

[7] D. Tran, L. Bourdev, R. Fergus, et al: Learning Spatiotemporal Features with 3D Convolutional Networks[C]. IEEE International Conference on Computer Vision. IEEE, (2015).

[8] A. Krizhevsky, I. Sutskever, and G.E. Hinton: ImageNet Classification with Deep Convolutional Neural Networks[J]. Advances in Neural Information Processing Systems, (2012), 25(2):2012.

[9] W. Ouyang, X. Wang, X. Zeng, et al: DeepID-Net: Deformable deep convolutional neural networks for object detection[J]. Developmental Medicine & Child Neurology, (2015), 46(5):358–360.

[10] K. Simonyan, and A. Zisserman: Two-Stream Convolutional Networks for Action Recognition in Videos. ArXiv:1406.2199v2 [cs.CV] 12 Nov (2014).

[11] S. Ji, M. Yang, and K. Yu: 3D convolutional neural networks for human action recognition.[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, (2013), 35(1):221-31.

[12] K. Wang, X. Wang, L. Lin, et al: 3D Human Activity Recognition with Reconfigurable Convolutional Neural Networks[C]. Proceedings of the ACM International Conference on Multimedia. ACM, (2015):97-106.

[13] K. Simonyan, and A. Zisserman: Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Eprint Arxiv, (2014).

[14] C. Szegedy, W. Liu, Y. Jia, et al: Going deeper with convolutions[C]. Computer Vision and Pattern Recognition. IEEE, (2015):1-9.

[15] Javier Sanchez, Enric Meinhardt-Llopis, and Gabriele Facciolo: TV-L1 optical flow estimation. Image Processing On Line, 3 (2013), p. 137–150. http://dx.doi.org/10.5201/ipol.2013.26.

[16] B.K.P. Horn, and B.G. Schunck: "Determining optical flow": a retrospective[J]. Artificial Intelligence, (1993), 59(93):81–87.

[17] C. Kaynak, and E. Alpaydan: Multistage classification by cascaded classifiers. Proceedings of the 12th IEEE International Symposium on Intelligent Control 16-18 July 1997, Istanbul, Turkey.

[18] C. Premebida, O. Ludwig, M. Silva, et al: A cascade classifier applied in pedestrian detection using laser and image-based features[C]. Conference Record - IEEE Conference on Intelligent Transportation Systems. (2010):1153-1159.

[19] K. Soomro, A.R. Zamir, and M. Shah: UCF101: A dataset of 101 human actions classes from videos in the wild. http://crcv.ucf.edu/data/UCF101.php

[20] A. Ozcift: Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis [J]. Computers in Biology & Medicine, (2011), 41(41):265-71.

[21] J.Y.H. Ng, M. Hausknecht, S. Vijayanarasimhan, et al: Beyond short snippets: Deep networks for video classification[C]. Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on. IEEE, (2015):1613-1631.