Deep Alternative Neural Network: Exploring Contexts as Early as Possible for Action Recognition

Jinzhuo Wang, Wenmin Wang, Xiongtao Chen, Ronggang Wang, Wen Gao
School of Electronics and Computer Engineering, Peking University

Task and Motivations

We aim at automatically assigning high-level concepts to each untrimmed video, known as action recognition task. Our motivation comes from several limitations of existing approaches.

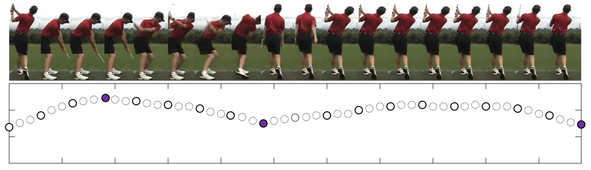
1. Existing deep architectures often mine contexts in the latter layers, lacking contexts of low level features. In many computer vision tasks, especially action recognition, contexts of low-level features and its evolutions are crucial for distinguishing similar instances.
2. For action recognition in videos, popular methods tend to extract features of entire frames which contain many irrelevant information. Such setting may burden deep network training.

Contributions

Based on the left discussions about our motivation and existing limitations, we make several contributions as follows.

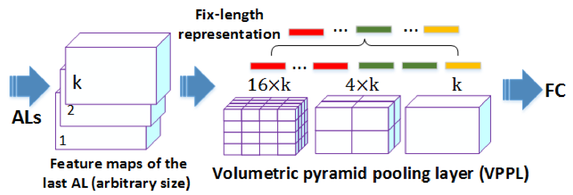
1. A novel deep alternative neural network (DANN) for action recognition in videos, which mines contexts and its evolutions at the beginning of networks.
2. An adaptive method to determine the temporal size of input video clip for deep network based on the density of optical flow energy, instead of manual choices used in relevant methods.
3. A volumetric pyramid pooling layer to resize the output of input video clips of arbitrary sizes to fixed-size before fc layers.

Adaptive Input



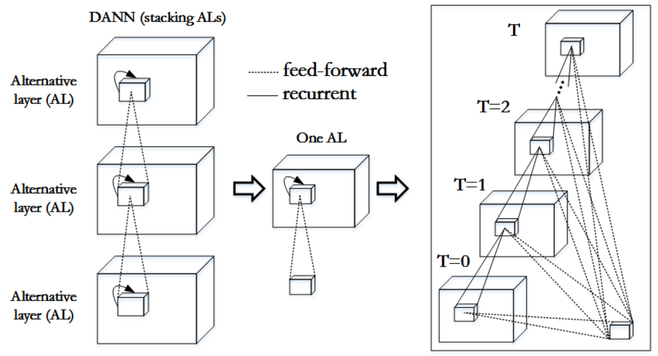
The input size of deep networks in temporal domain is often determined empirically since it is hard to evaluate all the choices. We instead introduce an adaptive method to automatically select the most discriminative video fragments using the density of optical flow energy. We attempt to preserve as much as motion information and appropriate range dependencies while not breaking their semantic structures in temporal domain. Many evidences show that motion energy intensity induced by human activity exhibits regular periodicity. We leverage this point for segmentation to pursue adaptive input.

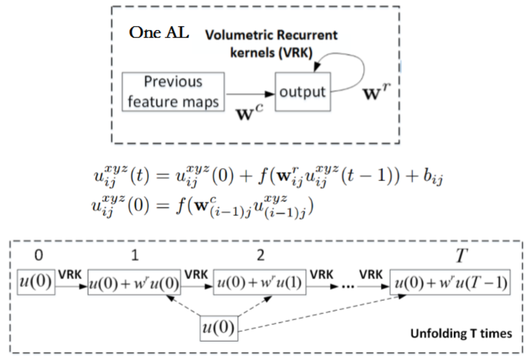
VPP Layer



The AL accepts input videos of arbitrary sizes and produces outputs of variable sizes. However, the fully connected layers require fixed-length vectors. To adopt DANN for input video clips of arbitrary sizes, we replace the last pooling layer with a volumetric pyramid pooling layer (VPPL) inspired by the success of spatial pyramid pooling layer (SPPL). The outputs of the VPPL are kM-dimensional vectors, where M is the number of bins and k is the number of kernels in the last alternative layer, which are then sent to the fc layers.

Architecture





The key component of DANN is the alternative layer (AL), which consists of a standard volumetric convolutional layer followed by a designed recurrent layer. Volumetric convolution is performed to extract features from local spatiotemporal neighborhoods on feature maps in the previous layers. Then a recurrent layer is applied to the output and iteratively proceeds for T times. This procedure makes each unit evolve over discrete time steps and aggregate larger RFs. The overall architecture of DANN has 6 alternative layers with 64, 128, 256, 256, 512 and 512 kernel response maps, followed by a VPPL and 3 fc layers of size 2048 each.

AL_1 64		Pool1	AL_2 128		Pool2	AL_3 256		Pool3	AL_4 256		Pool4	AL_5 512		Pool5	AL_6 512		VPPL	FC_1 2048		FC_2 2048		FC_3 2048		Softmax

Results

Table 4: Comparison with the state-of-the-art on HMDB51 and UCF101 (over three splits).

	Method	HMDB	UCF		Method	HMDB	UCF
CNN	Slow fusion [13]	-	65.4	Fusion	Two-stream [22]	59.4	88.0
	C3D [28]	-	85.2		CNN+deep LSTM [35]	-	88.6
	Two-Stream(spatial) [22]	40.5	73.0		TDD [31]	63.2	90.3
	Two-Stream(temporal) [22]	54.6	83.7		TDD+iDT [31]	65.9	91.5
	LTC [29]	57.9	83.3		C3D+iDT [28]	-	90.4
	Very deep (temporal) [32]	-	87.0		Very deep (two-stream) [32]	-	91.4
	Very deep (spatial) [32]	-	87.0		LTC+spatial	61.5	88.6
Hand	IDT+FV [30]	57.2	85.9	Ours	DANN	63.3	89.2
	IDT+HSV [19]	61.1	87.9		DANN+spatial	65.9	**91.6**
	IDT+MIFS [16]	65.1	89.1				
	IDT+SFV [20]	**66.8**	-				

Table 4 reports the best DANN model and state-of-the-art approaches over three splits on UCF101 and HMDB51 datasets in terms of video-level accuracy. As can be seen, trajectory-based features are still competitive in the area of deep learning, especially with the help of high-order encodings or deep architectures. Fusion strategies often outperform pure single deep networks. Note that all the other deep networks use a pre-defined temporal length to generate video clip as input such as 16-frame and 60-frame, while our DANN determines it in an adaptive manner. Combined with spatial stream, DANN achieves the accuracy of 65.9% and 91.6% on HMDB51 and UCF101, separately.

Main References

[1] K. Simonyan and A. Zisserman Two-stream Convolutional Networks for Action Recognition in Videos. *NIPS '14*

[2] M. Liang, X. Hu and B. Zhang. Convolutional Neural Networks with Intra-layer Recurrent Connections for Scene Labeling. *NIPS '15*

Future Direction

Although adaptive temporal length is used in our model, the spatial size is still chosen in ad hoc manner. A natural direction as discussed in our paper is using a more compact video clip such as *tube* as the input of our DANN.

We expect the proposed architecture DANN be a general framework used for more CV tasks especially in which contexts are essential. Also, some related domains such as NLP have potential to enjoy the advantages of our DANN.

Acknowledgement

This work was supported by Shenzhen Peacock Plan (20130408-183003656).

The complete source code and pre-trained models are available a at `http://github.com/wangjinzhuo/DANN`.