ActionVLAD: Learning spatio-temporal aggregation for action classification

Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic and Bryan Russell

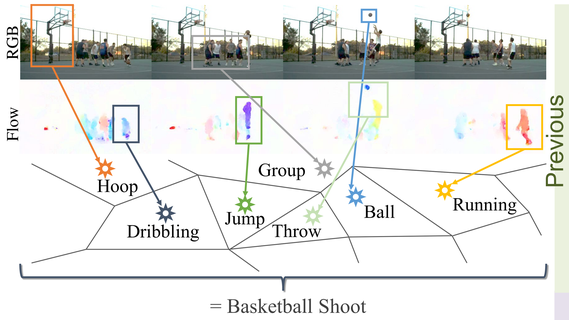
Carnegie Mellon University · Adobe · Inria

IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 HONOLULU 2017

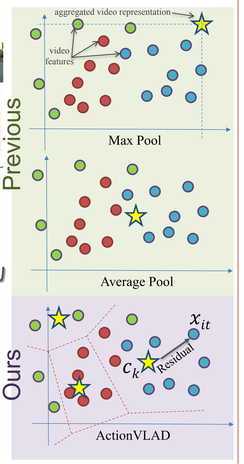
Code & Models

Goal: Action Recognition in Videos

How do we aggregate features across space and time?



= Basketball Shoot

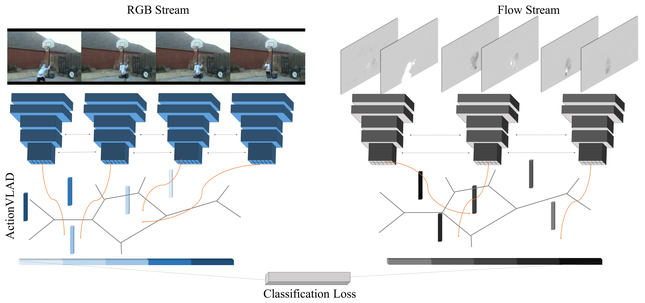


We present ActionVLAD, a differentiable pooling layer designed to aggregate features across spatio-temporal extent of videos for end-to-end trainable action classification.

$$V[\cdot,k] = \sum_{t=1}^{T}\sum_{i=1}^{N} \frac{e^{-\alpha\|x_{it}-c_k\|^2}}{\sum_{k'} e^{-\alpha\|x_{it}-c_{k'}\|^2}} (x_{it}[\cdot]-c_k[\cdot])$$

k^{th} ActionVLAD feature · soft-assignment · pool5 feature · cluster center · residual

Our two-stream ActionVLAD architecture



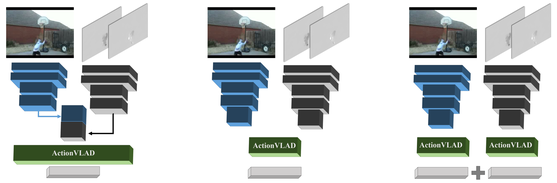
Analysis

(On HMDB-51 dataset, split 1)

Effect of training (rgb/flow)

Two-Stream	VLAD	ActionVLAD
47.1/55.2	44.9/55.6	51.2/58.4

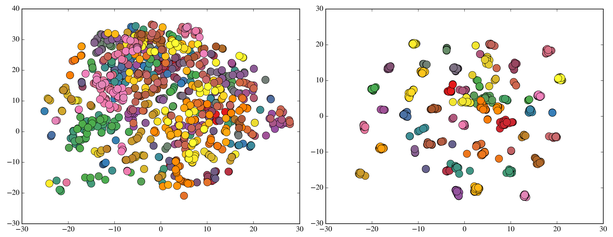
How to fuse RGB and Flow?



Concat Fuse	Early Fuse	Late Fuse
56.0	64.8	66.9

Where to ActionVLAD?

tSNE embedding of conv5 & fc7 features; same video => same color



conv5_3	fc7
51.2/58.4	43.3/53.1

(rgb/flow)

Other Pooling Strategies

Average	Max	ActionVLAD
41.6/53.4	41.5/54.6	51.2/58.4

Overall*

	RGB	Flow	Combined
2-stream	42.9	55.0	58.5
ActionVLAD	49.8	59.1	66.3

* 3-split average

Experiments and Results

UCF/HMDB 3-split average (+multi-crop)

	UCF101	HMDB51
iDT + FV	85.9	57.2
TDD + FV	90.3	63.2
RNN + FV	88.0	54.3
LRCN	82.9	
C3D + iDT	90.4	-
LSTM	88.6	-
2-Stream	92.5	65.4
LTC + iDT	92.7	67.2
2-Stream Fusion + iDT	93.5	69.2
ActionVLAD	92.7	66.9
ActionVLAD + iDT	**93.6**	**69.8**

Charades

	mAP	wAP
2 Stream	18.6	-
Inception RGB Stream	16.8	23.1
ActionVLAD (RGB)	17.6	25.1
ActionVLAD (RGB) + iDT	**21.0**	**29.9**

Regions assigned to an "Action Word"

Action word assignment tracks spatial regions over time

Action Word assignments for a sample video (brushing hair)