Combining Multiple Sources of Knowledge in Deep CNNs for Action Recognition

Eunbyung Park, Xufeng Han, Tamara L. Berg, Alexander C. Berg
University of North Carolina at Chapel Hill

UNC DEPARTMENT OF COMPUTER SCIENCE

CV²R Consortium for Vision and Virtual Reality

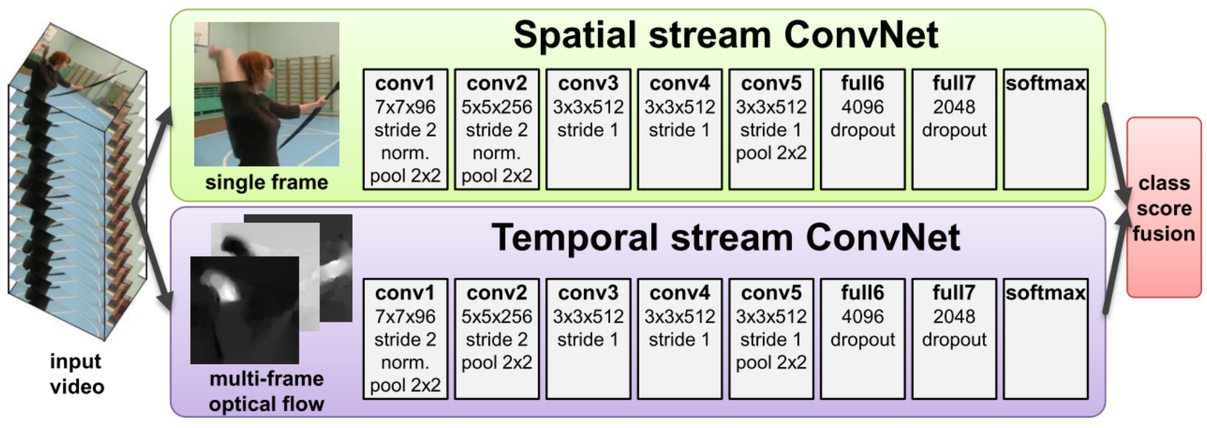
Motivation

Difficulties of learning spatio-temporal features with deep CNNs

- Existing datasets are not large enough and non-representative
- Training 3D deep CNN requires too many parameters causing high computational cost

Incorporating hand-crafted temporal features with spatial CNN features

- Two-stream CNN is promising solution, but each network was highly over-tuned to the particular input type (RGB and optical flow)



Two-stream convolutional neural network[1]

Proposed Techniques

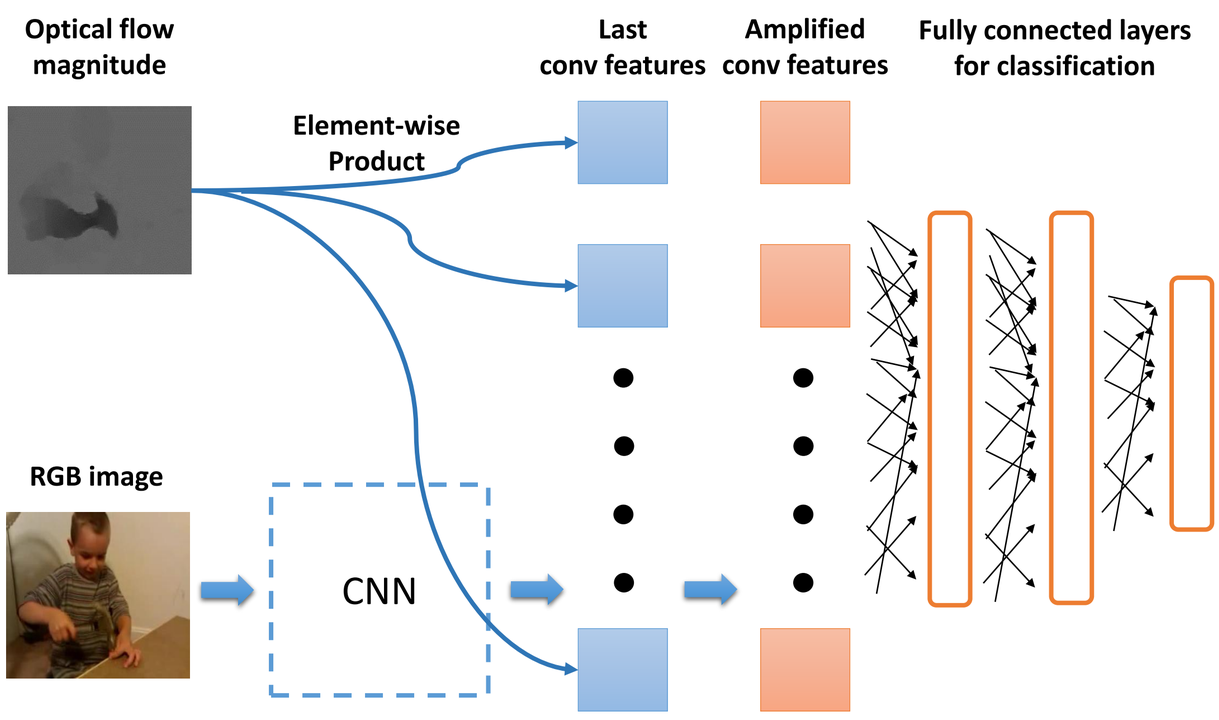
Feature amplification

- Amplifying spatial CNN features based on hand-crafted temporal features

Multiplicative fusion of two-stream networks

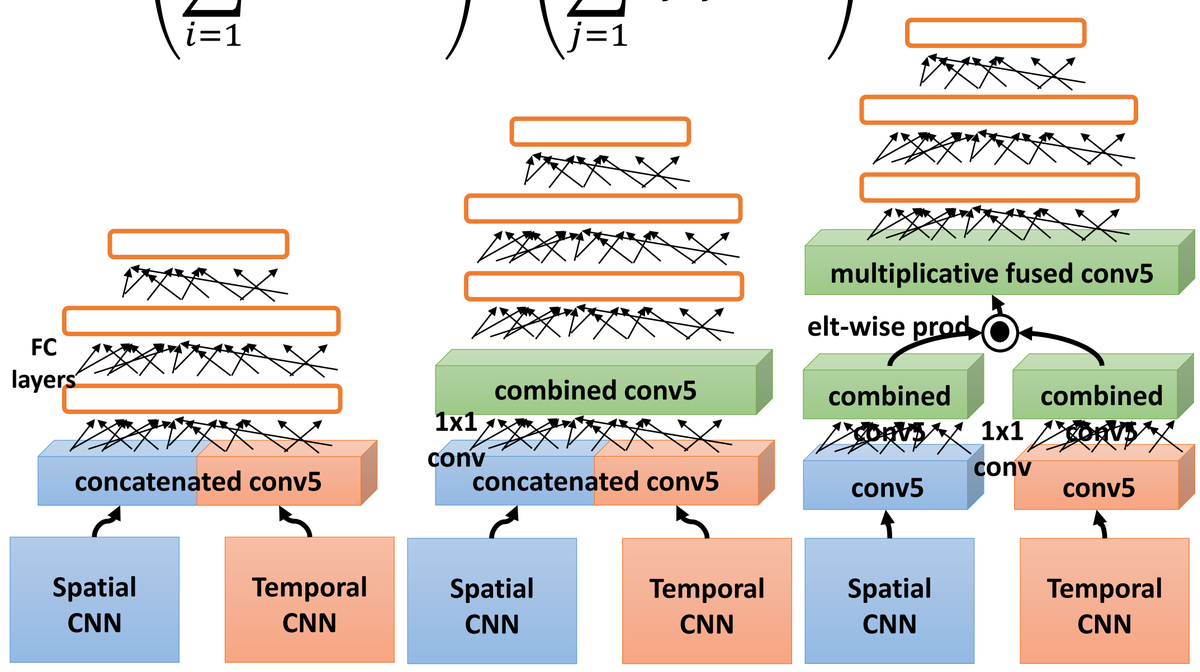
- Amplifying or suppressing the feature activations of each network based on their agreement
- Knowledge exchange between two networks

Feature Amplification



Multiplicative Fusion

$$c_k = \left(\sum_{i=1}^{M} \alpha_i a_i + \gamma_k\right) \odot \left(\sum_{j=1}^{N} \beta_j b_j + \delta_k\right)$$



Additive baseline fusion vs multiplicative fusion

Experiments on Video Classification

	UCF101		HMDB51	
	Base	Amp	Base	amp
T	80.3		47.3	
S	78.8	**81.0**	40.3	**44.9**
S + T	87.8	**88.5**	50.1	**54.5**

Feature amplification result

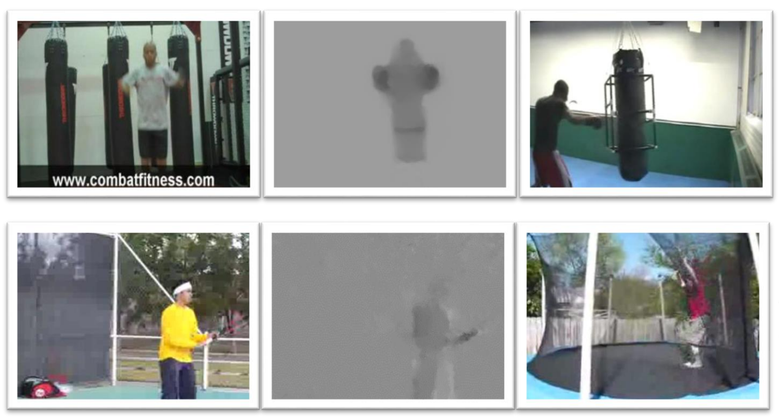
	UCF101	HMDB51
Base A	82.2	36.9
Base B	81.0	38.7
M-fuse(conv5)	**84.4**	**52.7**
M-fuse(fc7)	**87.6**	**53.3**

Multiplicative fusion result

	UCF101	HMDB51
S + T	87.8	50.1
S + T + m-fuse	88.3	54.4
S(amp) + T + m-fusion(conv5)	**88.9**	**56.2**
S(amp) + T + m-fusion(fc7)	**89.1**	**54.9**

Feature amplification and multiplicative fusion results

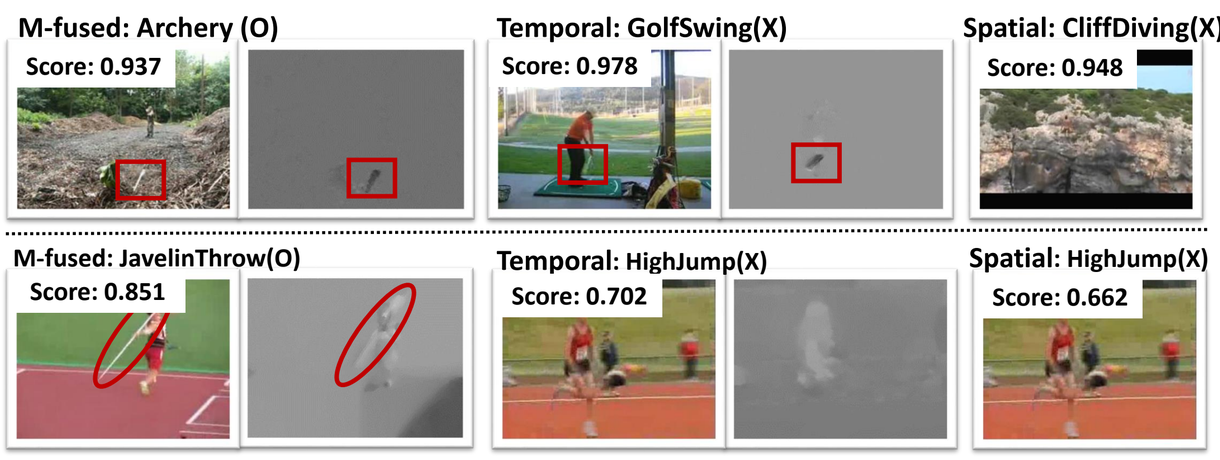
Examples of Classification



Amplified net: Jumping (O)
Spatial net: Boxing (X)

Amplified net: Tennis (O)
Spatial net: Trampoline jumping (X)

Effects of feature amplification



Effects of multiplicative fusion

Reference

K. Simonyan and A. Zisserman, Two-stream convolutional networks for action recognition in Videos, NIPS, 2014