Cross-Modal Message Passing for Two-stream Fusion

Dong Wang   Yuan Yuan   Qi Wang

School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University

International Conference on Acoustics, Speech and Signal Processing, 2018

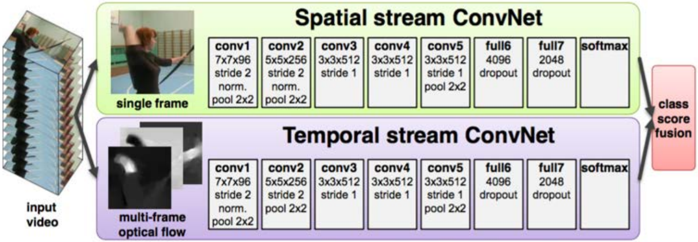
Outline

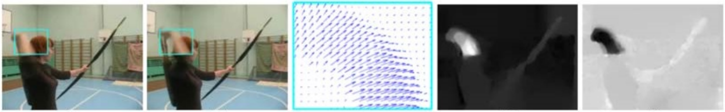
Action recognition in videos

Action Recognition: classify the short clip or untrimmed video into pre-defined class.

- Action recognition “in the lab”: KTH, Weizmann etc.
- Action recognition “in TV, Movies”: UCF Sports, Holloywood etc.
- Action recognition “in Web Videos”: HMDB, UCF101, THUMOS, ActivityNet etc.

Two Stream CNN





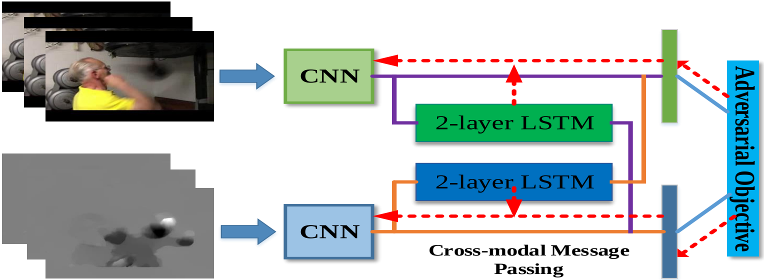
(a) (b) (c) (d) (e)

Karen Simonyan and Andrew Zisserman, *Two-Stream Convolutional Networks for Action Recognition in Videos*, in NIPS, 2014.

Contributions of this work

- Two Stream CNN
  - The spatial stream ConvNet and temporal stream ConvNet are trained independently.
  - The two stream architecture cannot exploit the spatial and temporal information simultaneously.
- Contributions
  - Presenting a novel cross-modal message passing mechanism for two-stream fusion.
  - An adversarial objective is proposed to train the two-stream network end-to-end.

Proposed Network Architecture



Benefits

- End-to-End trainable two stream action recognition network.
- The proposed frameworks explores the coupling property of appearance and motion information.

Proposed Network Architecture

Cross-Modal Message Passing Generator

Suppose $\mathrm{x}_a, \mathrm{x}_m \in R^{T \times D}$ denote the convolutional features from spatial and temporal stream respectively. Therefore, the two message generator networks can be formatted as follows:

$$\mathrm{m_a} = lstm_2(\mathrm{x_a}; \mathrm{w_a}); \quad \mathrm{m_m} = lstm_2(\mathrm{x_m}; \mathrm{w_m}) \qquad (1)$$

Then those messages are fused with convolutional features from another modal as follows:

$$\mathrm{x}_a^f = x_a + m_m; \quad \mathrm{x}_m^f = x_m + m_a \qquad (2)$$

Proposed Loss Function

The standard cross-entropy loss is utilized as loss function for each ConvNets, which is formed as

$$L(y,s) = -\sum_{i=1}^{C} y_i (s_i - log \sum_{j=1}^{C} \exp \ s_j) \tag{3}$$

where C is the number of action classes, y_i is the groundtruth label concerning class i and s_j is the classification score concerning class j.

Proposed Loss Function

Based on standard cross-entropy loss, the adversarial objective function of spatial ConvNet is defined as follows:

$$AL_a = L_a(y, s_a) + max(L_a(y, s_a) - L_m(y, s_m), 0) \tag{4}$$

while the adversarial objective function for temporal ConvNet is:

$$AL_m = L_m(y, s_m) + max(L_m(y, s_m) - L_a(y, s_a), 0) \tag{5}$$

where L_a, L_m represent the cross-entropy loss of spatial and temporal ConvNets.

Two Stage Training

- First, two-stream ConvNets is pretrained using standard categorial cross-entropy loss without updating the cross-modal message passing network.
- Second, the proposed adversarial objective loss function is utilized to train the whole two-stream network jointly.

Exploration Study

Table: CMMP components analysis on split 1 of HMDB-51.

Method	Spatial	Temporal	Fusion
SUM	53.01	54.05	53.79
MAX	52.61	52.29	52.68
CMMP+noAL	46.99	47.71	60.13
SUM+AL	51.70	52.29	51.96
MAX+AL	53.79	53.66	53.88
CMMP	**50.07**	**65.23**	**66.67**

- Fusion with Cross-Modal Message Passing Generator is better than SUM and MAX.
- The proposed adversarial objective and two-stage training strategy boost the performance.

Comparison with the-state-of-the-art

Table: Mean accuracy on the UCF-101 and HMDB-51.

Model	Method	UCF-101	HMDB-51
Traditional	iDT+FV	85.9	57.2
	iDT+HSV	87.9	61.6
Deep	EMV-CNN	86.4	-
	Two Stream	88.0	59.4
	F_{ST}CN	88.1	59.1
	C3D	85.2	-
	VideoLSTM	89.2	56.4
	TDD+FV	90.3	63.2
	Fusion	91.8	64.6
Ours	CMMP	**91.3**	**65.9**

Summary

- The message generator network is utilized to transfer the discriminative message from one modal to another, which is better than SUM and MAX.
- a novel adversarial objective to fine-tune the whole network, and boosts the performance even further.
- Comparison with the-state-of-the-arts shown the efficiency of the proposed method.

Thank you!