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
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


Dong Wang, Yuan Yuan, Qi Wang, *Cross-Modal Message Passing for Two-stream Fusion*
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.
Benefits

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

Suppose \( x_a, x_m \in \mathbb{R}^{T \times D} \) denote the convolutional features from spatial and temporal stream respectively. Therefore, the two message generator networks can be formatted as follows:

\[
m_a = \text{lstm}_2(x_a; w_a); \quad m_m = \text{lstm}_2(x_m; w_m)
\]  

(1)

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

\[
x_a^f = x_a + m_m; \quad x_m^f = x_m + m_a
\]  

(2)
The standard cross-entropy loss is utilized as loss function for each ConvNets, which is formed as

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

\[ AL_a = L_a(y, s_a) + \max(L_a(y, s_a) - L_m(y, s_m), 0) \]  \hspace{1cm} (4)

while the adversarial objective function for temporal ConvNet is:

\[ AL_m = L_m(y, s_m) + \max(L_m(y, s_m) - L_a(y, s_a), 0) \]  \hspace{1cm} (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 categorical 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Spatial</th>
<th>Temporal</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUM</td>
<td>53.01</td>
<td>54.05</td>
<td>53.79</td>
</tr>
<tr>
<td>MAX</td>
<td>52.61</td>
<td>52.29</td>
<td>52.68</td>
</tr>
<tr>
<td>CMMP+noAL</td>
<td>46.99</td>
<td>47.71</td>
<td>60.13</td>
</tr>
<tr>
<td>SUM+AL</td>
<td>51.70</td>
<td>52.29</td>
<td>51.96</td>
</tr>
<tr>
<td>MAX+AL</td>
<td>53.79</td>
<td>53.66</td>
<td>53.88</td>
</tr>
<tr>
<td>CMMP</td>
<td>50.07</td>
<td>65.23</td>
<td>66.67</td>
</tr>
</tbody>
</table>

- 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>UCF-101</th>
<th>HMDB-51</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional</td>
<td>iDT+FV</td>
<td>85.9</td>
<td>57.2</td>
</tr>
<tr>
<td></td>
<td>iDT+HSV</td>
<td>87.9</td>
<td>61.6</td>
</tr>
<tr>
<td>Deep</td>
<td>EMV-CNN</td>
<td>86.4</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Two Stream</td>
<td>88.0</td>
<td>59.4</td>
</tr>
<tr>
<td></td>
<td>FST-CN</td>
<td>88.1</td>
<td>59.1</td>
</tr>
<tr>
<td></td>
<td>C3D</td>
<td>85.2</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>VideoLSTM</td>
<td>89.2</td>
<td>56.4</td>
</tr>
<tr>
<td></td>
<td>TDD+FV</td>
<td>90.3</td>
<td>63.2</td>
</tr>
<tr>
<td></td>
<td>Fusion</td>
<td>91.8</td>
<td>64.6</td>
</tr>
<tr>
<td>Ours</td>
<td>CMMP</td>
<td><strong>91.3</strong></td>
<td><strong>65.9</strong></td>
</tr>
</tbody>
</table>
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!