Cross-Modal Interaction Networks for Query-Based Moment Retrieval in Videos

Zhu Zhang, Zhijie Lin, Zhou Zhao and Zhenxin Xiao

College of Computer Science, Zhejiang University, China
{zhangzhu, zhaozhou, linzhijie}@zju.edu.cn, alanshawzju@gmail.com
Video Retrieval

- Video retrieval task searches the most relevant video from large collections according to a given natural language query.

Database of video clips
Query-Based Moment Retrieval

- Moment Retrieval task aims to localize the most relevant moment in an untrimmed video according to the given natural language query.

Language Query:
A person runs to the window and then look out

9.3 s | 14.4 s
The query describes **three successive actions**, corresponding to **complex object interactions** within the video.

Challenges:
- sufficient understanding of both video and query contents.
- cross-modal interactions to capture the matching clues.
Motivation

Query representation learning

- Previous method: RNN networks.
- Weakness: ignore the syntactic structure of queries.

- Our method: develop a syntactic GCN to exploit the syntactic structure of queries.
Motivation

Video representation learning

- Previous method: CNN+RNN or R-C3D networks.
- Weakness: fail to capture long-range semantic dependencies.
- Our method: propose a multi-head self-attention mechanism.
Motivation

Cross-modal interaction

- Previous method: widely-used attention mechanism.
- Weakness: remain in the rough one-stage interaction.
- Our method: adopt a multi-stage cross-modal interaction method.
In summary, we consider multiple crucial factors for this challenging task, including

- the syntactic structure of natural language queries;
- long-range semantic dependencies in video context;
- the sufficient cross-modal interaction.
Our proposed model consists of 4 modules:
1. Syntactic GCN module for query modeling
2. Multi-head self-attention module for video modeling
3. Multi-stage interaction module for cross-modal interaction
4. Moment retrieval module for moment localization
Model

**Syntactic GCN module**

**Input:** query

**Output:** syntactic-aware query representation

- Build syntactic dependency graph.
- By passing information along the dependency edges between relevant words, we learn syntactic-aware query representations.
The original GCN regard the syntactic dependency graph as an undirected graph.

We consider a syntactic GCN to exploit the directional and labeled dependency edges between nodes.

\[ g_i^1 = \text{ReLU} \left( \sum_{j \in \mathcal{N}(i)} W^g h_j^q + b^g \right) \]

\[ g_i^1 = \text{ReLU} \left( \sum_{j \in \mathcal{N}(i)} W_{dir(i,j)}^g h_j^q + b_{lab(i,j)}^g \right) \]

dir(i,j) represents the direction from i to j.
lab(i,j) represents the label of the edge.
Model

Multi-Head Self-Attention Module

Input: video

Output: video semantic representation

- By the self-attention method, each frame is able to interact not only with adjacent frames but also with distant ones.

\[
\text{Attention}(\overline{Q}, \overline{K}, \overline{V}) = \text{Softmax}\left(\frac{\overline{Q}^T \overline{K}}{\sqrt{d_k}}\right)\overline{V}^T,
\]

\[
\text{MultiHead}(\overline{Q}, \overline{K}, \overline{V}) = W^O \text{Concat}(\text{head}_1, \ldots, \text{head}_H)
\]

where \(\text{head}_i = \text{Attention}(W_i^Q \overline{Q}, W_i^K \overline{K}, W_i^V \overline{V})\)

Model

Multi-Stage Interaction Module

**Input:** query and video semantic representation

**Output:** cross-modal representation

- **Attentive aggregation** to capture the query clues for each frame.
  \[
  M(h^v_i, o^l_j) = w^\top \tanh(W_1^m h^v_i + W_2^m o^l_j + b^m),
  \]
  \[
  M^{row}_{ij} = \frac{\exp(M_{ij})}{\sum_{k=1}^{m} \exp(M_{ik})}, \quad h^s_i = \sum_{j=1}^{m} M^{row}_{ij} o^l_j,
  \]

- **Cross gate** to emphasize crucial contents and weaken inessential parts.
  \[
  g^v_i = \sigma(W^v h^v_i + b^v),
  \]
  \[
  \tilde{h}^s_i = h^s_i \odot g^v_i,
  \]
  \[
  g^s = \sigma(W^s \tilde{h}^s_i + b^s),
  \]
  \[
  \tilde{h}^v_i = h^v_i \odot g^s_i,
  \]

- **Bilinear fusion** to obtain cross-modal representations.
  \[
  f_i = P^f (\sigma(W^v \tilde{h}^v_i) \odot \sigma(W^s \tilde{h}^s_i)) + b^f,
  \]
Moment Retrieval Module

**Input:** cross-modal representations  
**Output:** final moments

- We pre-define a set of candidate moments with multi-scale windows at each time step $i$.
- We can simultaneously produce the confidence scores for these moments.
- We produce the predicted offsets for these moments.

**Loss**

- **Alignment loss**
  \[
  \mathcal{L}_{\text{align}} = -\frac{1}{nk} \sum_{i=1}^{n} \sum_{j=1}^{k} \mathcal{L}_{ij}
  \]

- **Regression loss**
  \[
  \mathcal{L}_{\text{reg}} = \frac{1}{N} \sum_{C_h} (R(\delta_s \hat{\delta}_s) + R(\delta_e \hat{\delta}_e))
  \]
Dataset: ActivityCaption and TACoS

- The contents of ActivityCaption dataset are **diverse and open**.

- The contents of TACoS are **limited to cooking scenes**, thus lack the diversity.
• Performance Criteria $R@n$, $IoU=m$: the percentage of at least one of top-$n$ selected moments having IoU larger than $m$.

<table>
<thead>
<tr>
<th>Method</th>
<th>$R@1$ IoU=0.3</th>
<th>$R@1$ IoU=0.5</th>
<th>$R@1$ IoU=0.7</th>
<th>$R@5$ IoU=0.3</th>
<th>$R@5$ IoU=0.5</th>
<th>$R@5$ IoU=0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCN</td>
<td>39.35</td>
<td>21.36</td>
<td>6.43</td>
<td>68.12</td>
<td>53.23</td>
<td>29.70</td>
</tr>
<tr>
<td>VSA-RNN</td>
<td>39.28</td>
<td>23.43</td>
<td>9.01</td>
<td>70.84</td>
<td>55.52</td>
<td>32.12</td>
</tr>
<tr>
<td>VSA-STV</td>
<td>41.71</td>
<td>24.01</td>
<td>8.92</td>
<td>71.05</td>
<td>56.62</td>
<td>34.52</td>
</tr>
<tr>
<td>CTRL</td>
<td>47.43</td>
<td>29.01</td>
<td>10.34</td>
<td>75.32</td>
<td>59.17</td>
<td>37.54</td>
</tr>
<tr>
<td>ACRN</td>
<td>49.70</td>
<td>31.67</td>
<td>11.25</td>
<td>76.50</td>
<td>60.34</td>
<td>38.57</td>
</tr>
<tr>
<td>QSPN</td>
<td>52.13</td>
<td>33.26</td>
<td>13.43</td>
<td>77.72</td>
<td>62.39</td>
<td>40.78</td>
</tr>
<tr>
<td>CMIN</td>
<td><strong>63.61</strong></td>
<td><strong>43.40</strong></td>
<td><strong>23.88</strong></td>
<td><strong>80.54</strong></td>
<td><strong>67.95</strong></td>
<td><strong>50.73</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>$R@1$ IoU=0.1</th>
<th>$R@1$ IoU=0.3</th>
<th>$R@1$ IoU=0.5</th>
<th>$R@5$ IoU=0.1</th>
<th>$R@5$ IoU=0.3</th>
<th>$R@5$ IoU=0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCN</td>
<td>3.11</td>
<td>1.64</td>
<td>1.25</td>
<td>3.11</td>
<td>2.03</td>
<td>1.25</td>
</tr>
<tr>
<td>VSA-RNN</td>
<td>8.84</td>
<td>10.77</td>
<td>4.78</td>
<td>19.05</td>
<td>13.90</td>
<td>9.10</td>
</tr>
<tr>
<td>VSA-STV</td>
<td>15.01</td>
<td>10.77</td>
<td>7.56</td>
<td>32.82</td>
<td>23.92</td>
<td>15.50</td>
</tr>
<tr>
<td>CTRL</td>
<td>24.32</td>
<td>18.32</td>
<td>13.30</td>
<td>48.73</td>
<td>36.69</td>
<td>25.42</td>
</tr>
<tr>
<td>ACRN</td>
<td>24.22</td>
<td>19.52</td>
<td>14.62</td>
<td>47.42</td>
<td>34.97</td>
<td>24.88</td>
</tr>
<tr>
<td>QSPN</td>
<td>25.31</td>
<td>20.15</td>
<td>15.23</td>
<td>53.21</td>
<td>36.72</td>
<td>25.30</td>
</tr>
<tr>
<td>CMIN</td>
<td><strong>32.48</strong></td>
<td><strong>24.64</strong></td>
<td><strong>18.05</strong></td>
<td><strong>62.13</strong></td>
<td><strong>38.46</strong></td>
<td><strong>27.02</strong></td>
</tr>
</tbody>
</table>

• On all the criteria of two datasets, our method outperforms all previous state-of-the-art baselines, especially on ActivityCaption.
Ablation Study

- **CMIN(w/o. GCN):** We remove the syntactic GCN layer from the query representation learning.

- **CMIN(w/o. SA):** We discard the multi-head self-attention from the video representation learning.

- **CMIN(w/o. CG):** We remove the cross gate in the multi-stage cross-modal interaction module.

- **CMIN(w/o. BF):** We replace the bilinear fusion method with a simple concatenation for query and video features.

<table>
<thead>
<tr>
<th>Method</th>
<th>R@1 IoU=0.1</th>
<th>R@1 IoU=0.3</th>
<th>R@1 IoU=0.5</th>
<th>R@5 IoU=0.1</th>
<th>R@5 IoU=0.3</th>
<th>R@5 IoU=0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o. GCN</td>
<td>60.12</td>
<td>40.84</td>
<td>21.79</td>
<td>78.23</td>
<td>65.67</td>
<td>45.43</td>
</tr>
<tr>
<td>w/o. SA</td>
<td>61.22</td>
<td>41.56</td>
<td>22.36</td>
<td>79.43</td>
<td>66.91</td>
<td>48.12</td>
</tr>
<tr>
<td>w/o. CG</td>
<td>60.57</td>
<td>41.21</td>
<td>22.01</td>
<td>78.62</td>
<td>65.99</td>
<td>46.89</td>
</tr>
<tr>
<td>w/o. BF</td>
<td>61.32</td>
<td>41.89</td>
<td>22.12</td>
<td>79.27</td>
<td>66.21</td>
<td>47.92</td>
</tr>
<tr>
<td>full</td>
<td>63.61</td>
<td>43.40</td>
<td>23.88</td>
<td>80.54</td>
<td>67.95</td>
<td>50.73</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>R@1 IoU=0.1</th>
<th>R@1 IoU=0.3</th>
<th>R@1 IoU=0.5</th>
<th>R@5 IoU=0.1</th>
<th>R@5 IoU=0.3</th>
<th>R@5 IoU=0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o. GCN</td>
<td>30.54</td>
<td>23.22</td>
<td>10.03</td>
<td>57.69</td>
<td>37.12</td>
<td>26.16</td>
</tr>
<tr>
<td>w/o. SA</td>
<td>30.21</td>
<td>23.02</td>
<td>16.87</td>
<td>55.51</td>
<td>36.6</td>
<td>25.37</td>
</tr>
<tr>
<td>w/o. CG</td>
<td>31.96</td>
<td>23.59</td>
<td>17.47</td>
<td>61.87</td>
<td>38.11</td>
<td>26.79</td>
</tr>
<tr>
<td>w/o. BF</td>
<td>32.01</td>
<td>24.79</td>
<td>17.61</td>
<td>61.59</td>
<td>38.23</td>
<td>26.75</td>
</tr>
<tr>
<td>full</td>
<td>32.48</td>
<td>24.64</td>
<td>18.05</td>
<td>62.13</td>
<td>38.16</td>
<td>27.02</td>
</tr>
</tbody>
</table>
Hyper-Parameter Analysis:

- For the syntactic GCN module, the number of stacked layers is a crucial hyper-parameter. Therefore, we further explore the effect of this hyper-parameter by varying the number of layers from 1 to 5.

![Figure 4: Effect of the Number of Stacked Syntactic GCN layers on the ActivityCaption Dataset.](image1)

![Figure 5: Effect of the Number of Stacked Syntactic GCN layers on the TACoS Dataset.](image2)
Attention Visualization

- The word "line" has the highest attention score over the query for the fourth frame.

Figure 8: The Video-to-Query Attention Results in the Multi-Stage Cross-Modal Interaction Module
Experiment

Example

Query: The boy drops the cloths and takes the iron away before the baby can pick it up.

Ground Truth: 114.12s | QSPN: 112.42s | CMIN: 113.56s

Query: The female athlete jumped over the pole and wave at everyone.

Ground Truth: 25.01s | QSPN: 21.56s | CMIN: 26.98s

Figure 6: Examples on the ActivityCaption dataset.

Query: After getting out the juicer, he juices the first orange half.

Ground Truth: 74.56s | QSPN: 79.1s | CMIN: 72.32s

Query: She washes herb stems in the sink before placing them on the cuttingboard.

Ground Truth: 126.29s | QSPN: 130.1s | CMIN: 123.04s

Figure 7: Examples on the TACoS dataset.
Thank you!

Zhu Zhang, Zhijie Lin, Zhou Zhao and Zhenxin Xiao

College of Computer Science, Zhejiang University, China
{zhangzhu, zhaozhou, linzhijie}@zju.edu.cn, alanshawzju@gmail.com