Attentive Moment Retrieval in Videos

Meng Liu[1], Xiang Wang[2], Liqiang Nie[1], Xiangnan He[2], Baoquan Chen[1], and Tat-Seng Chua[2]

[1]Shandong University, China

[2]National University of Singapore, Singapore

Pipeline

- Background
- Learning Model
- Experiment
- Conclusion

Pipeline

- Background
- Learning Model
- Experiment
- Conclusion

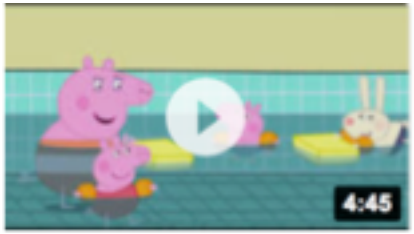
Background

- Inter-video Retrieval

Query: FIFA World Cup

… …

Query: Pig Peggy



… …

Background

- Intra-video Retrieval

Retrieving a segment from the untrimmed videos, which contain complex scenes and involve a large number of objects, attributes, actions, and interactions.

Messi's penalty/Football shot

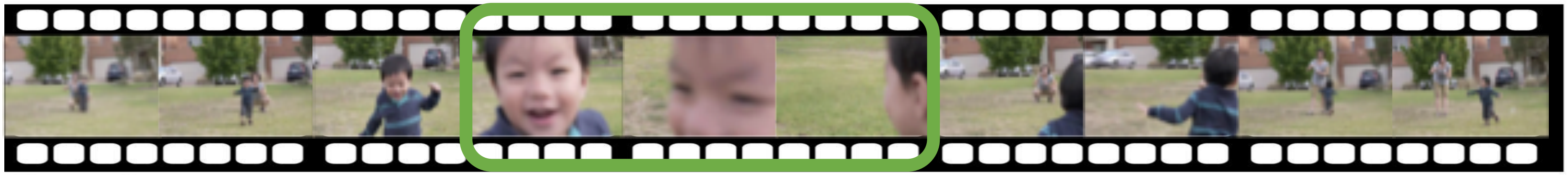
Background

- Surveillance Videos: Finding missing children or pets and suspects

Query: ***A girl in orange first walks by the camera.***

- Home videos: Recalling the desired moment

Query: ***Baby's face gets very close to the camera.***



- Online Videos: Quickly Jumping to the specific moment

Background

Reality: Dragging progress bar to locate the desired moment.

Boring and time consuming

Research: Densely segment the long video into different scale moments, and then match each moment with the query.

Expensive computational costs and the exponential search space

Problem Formulation
-Temporal Moment Localization

Input: a video and a language query

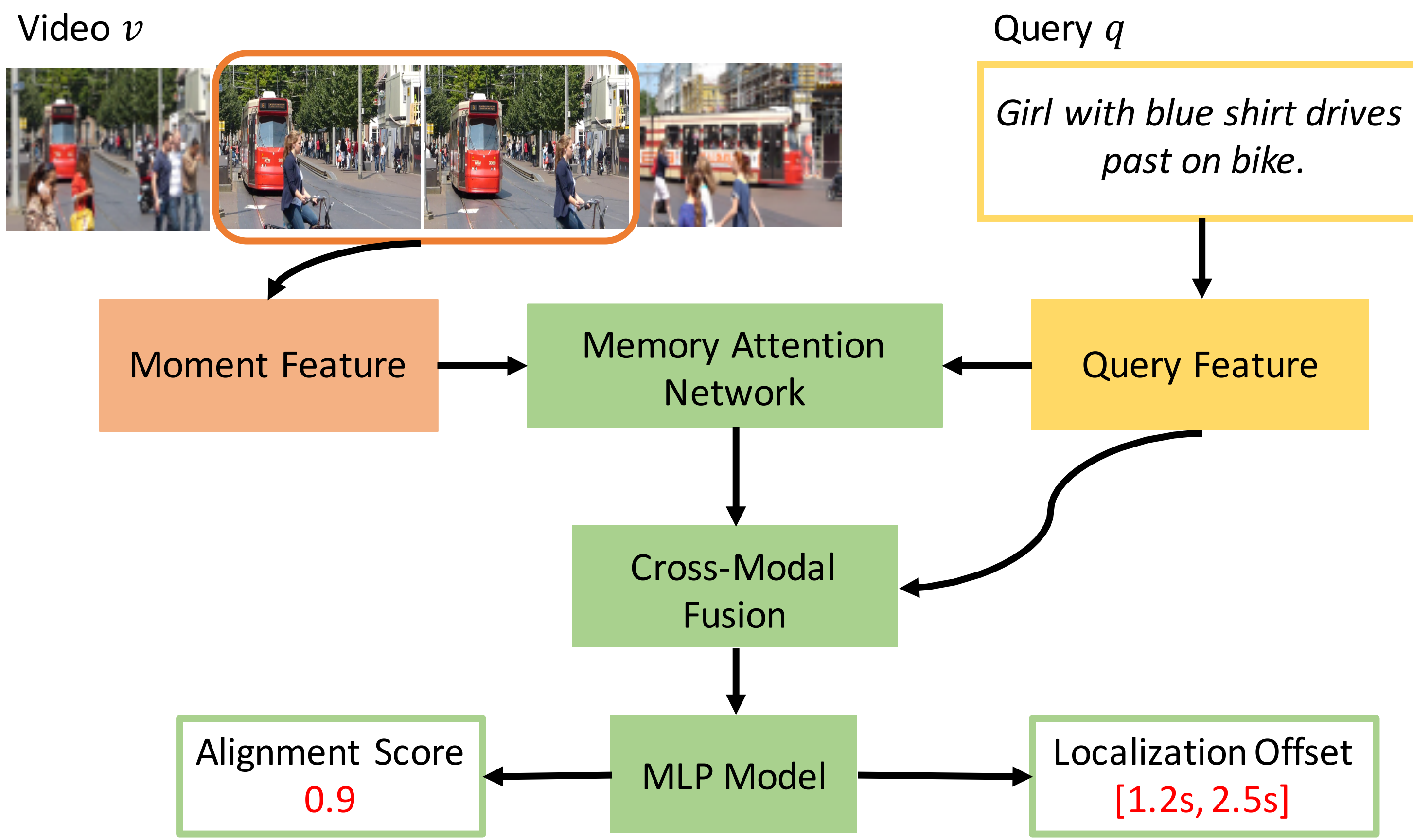
Query: ***a girl in orange walks by the camera.***

Output: Temporal moment corresponding to the given query (green box) with time points [24s,30,s]

Pipeline

- Background
- Learning Model
- Experiment
- Conclusion

Learning Model-Pipeline



Learning Model-Feature Extraction

- Video
  1. Segmentation:

     segment video into moments with sliding window, each moment c has a time location $[\tau_s, \tau_e]$
  2. Computing location offset:

     $[\delta_s, \delta_e]$=$[t_s, t_e]$-$[\tau_s, \tau_e]$, $[t_s, t_e]$ is the temporal interval of the given query
  3. Computing temporal-spatio feature $\boldsymbol{x}_c$:

     C3D feature for each moment
- Query

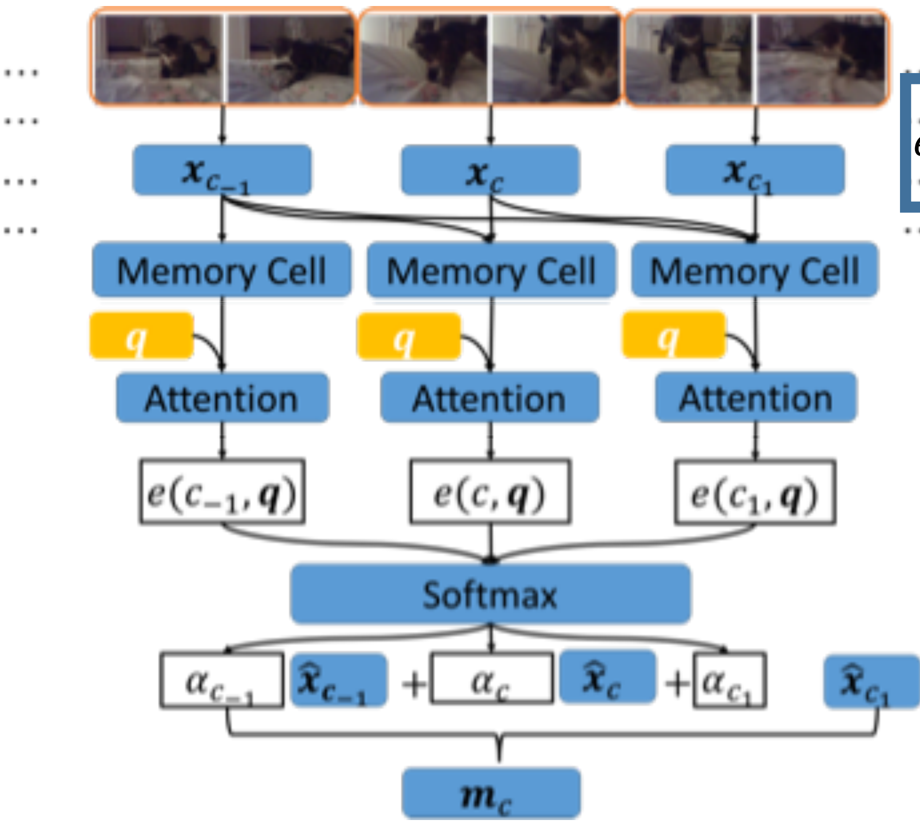
  $\boldsymbol{q}$: Skip-thoughts feature

Learning Model-Moment Attention Network

- There are many temporal constraint words in the given query, such as the term “first”, “second”, and “closer to”, therefore temporal context are useful to the localization.

- Not all the context have the same influence on the localization, the near context are more important than the far ones.

Learning Model-Memory Attention Network



Memory cell

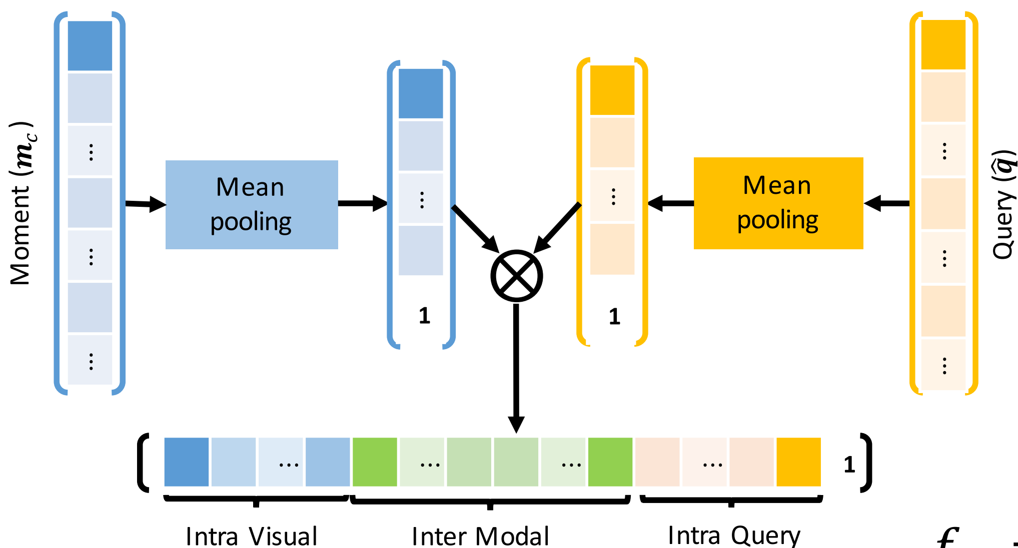
$$e(c_j, q) = \sigma\Big(\boxed{\sum_{i=-n_c}^{j} W_c x_{c_i}} + b_c\Big)^T \cdot \sigma(W_q q + b_q)$$

$$\alpha_{c_j} = \frac{e(c_j, q)}{\sum_{k=-n_c}^{n_c} e(c_k, q)}, j \in [-n_c, n_c]$$

$$\begin{cases} \hat{x}_{c_j} = W_c x_{c_j} + b_c \\ m_c = \sum_{j \in [-n_c, n_c]} \alpha_{c_j} \hat{x}_{c_j} \end{cases}$$

Learning Model- Cross-modal Fusion

The output of this fusion procedure explores the intra- modal and the inter-modal feature interactions to generate the moment-query representations.



$$f_{cq} = \begin{bmatrix} \widetilde{m}_c \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \tilde{q} \\ 1 \end{bmatrix} = [\widetilde{m}_c, \widetilde{m}_c \otimes \tilde{q}, \tilde{q}, 1]$$

Learning Model- Loss Function

Given the output of the fusion model into a two Layer MLP model, and the output of the MLP model is a three dimension vector $e_L = [s_{cq}, \delta_s, \delta_e]$.

$$L = L_{align} + \lambda L_{loc}$$

$$L_{align} = \alpha_1 \sum_{(c,q)\in\mathcal{P}} \log\left(1 + \exp\left(-s_{cq}\right)\right) + \alpha_2 \sum_{(c,q)\in\mathcal{N}} \log\left(1 + \exp\left(s_{cq}\right)\right)$$

$$L_{loc} = \sum_{(c,q)\in\mathcal{P}} [R(\delta_s^* - \delta_s) + R(\delta_e^* - \delta_e)]$$

Pipeline

- Background
- Learning Model
- Experiment
- Conclusion

Experiment - Dataset

- TACoS and DiDeMo

Table 1: The summary of the TACoS and DiDeMo datasets.

Dataset	# Videos	# Queries	# Moments	Domain	Video Source
TACoS	100	14,229	2,326	Cooking	Lab Kitchen
DiDeMo	10,464	40,543	26,892	Open	Flickr

- Evaluation:

  R(n,m)="R@n,IoU=m"

Experiment – Performance Comparison

Table 2: Performance comparison between our proposed model and the state-of-the-art baselines on TACoS. (p-value*: p-value over $R(1, 0.5)$)

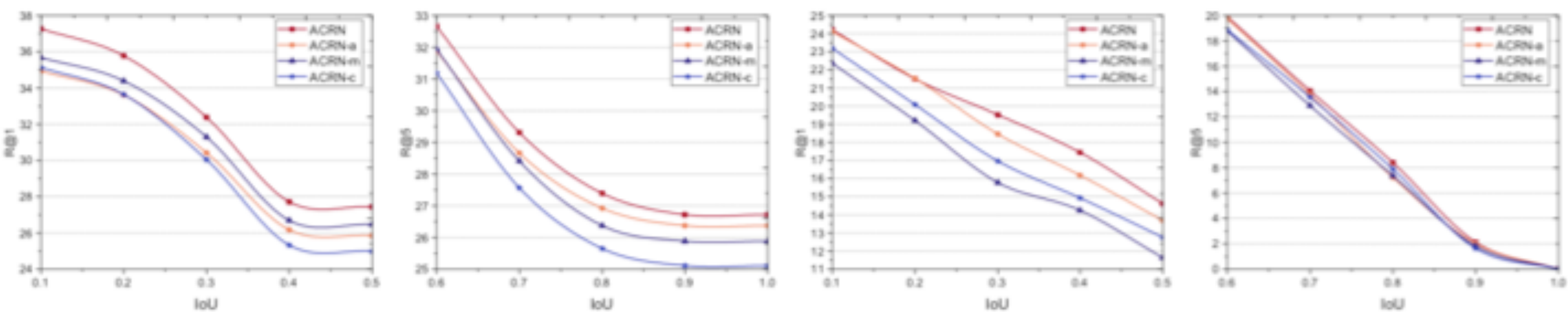
Method	R@1 IoU=0.5	R@1 IoU=0.3	R@1 IoU=0.1	R@5 IoU=0.5	R@5 IoU=0.3	R@5 IoU=0.1	p-value*
MCN	1.25%	1.64%	3.11%	1.25%	2.03%	3.11%	3.62E-10
VSA-STV	8.84%	13.59%	17.58%	16.41%	26.40%	35.86%	2.16E-06
VSA-RNN	9.96%	16.16%	20.92%	18.32%	29.19%	40.66%	1.82E-05
TALL	11.22%	15.50%	20.21%	23.46%	31.37%	44.40%	5.71E-05
ACRN	**14.62%**	**19.52%**	**24.22%**	**24.88%**	**34.97%**	**47.42%**	-

Table 3: Performance comparison between our proposed model and the state-of-the-art baselines on DiDeMo. (p-value*: p-value over $R(1, 0.5)$)

Method	R@1 IoU=0.5	R@1 IoU=0.7	R@1 IoU=0.9	R@5 IoU=0.5	R@5 IoU=0.7	R@5 IoU=0.9	p-value*
MCN	23.33%	15.37%	15.32%	41.03%	20.37%	19.77%	6.14E-09
VSA-STV	25.38%	14.49%	14.39%	68.56%	26.92%	24.24%	1.98E-03
VSA-RNN	24.94%	14.52%	14.44%	68.39%	26.10%	23.95%	3.31E-06
TALL	26.45%	15.36%	15.31%	68.78%	28.43%	26.15%	2.32E-02
ACRN	**27.44%**	**16.65%**	**16.53%**	**69.43%**	**29.45%**	**26.82%**	-

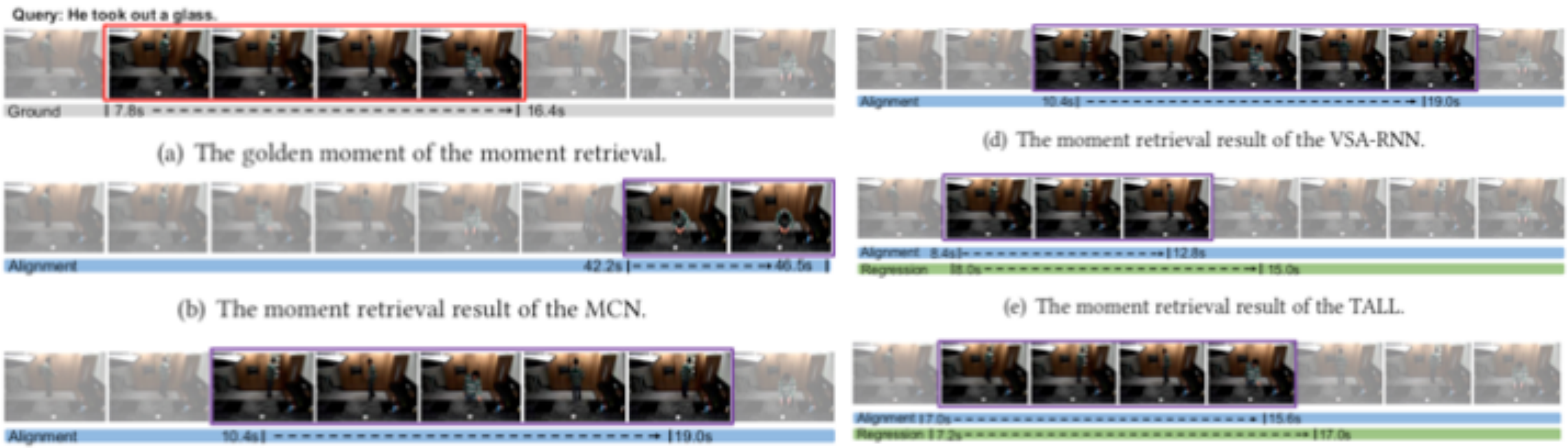
Experiment – Model Variants

- ACRN-a (pink): Mean pooling context feature as moment feature
- ACRN-m (purple): Attention model without memory part
- ACRN-c (blue): Concatenating multi-modal features



(a) R@1 vs IoU on DiDeMo (b) R@5 vs IoU on DiDeMo (c) R@1 vs IoU on TACoS (d) R@5 vs IoU on TACoS

Experiment – Qualitative Result



(a) The golden moment of the moment retrieval.

(b) The moment retrieval result of the MCN.

(c) The moment retrieval result of the VSA-STV.

(d) The moment retrieval result of the VSA-RNN.

(e) The moment retrieval result of the TALL.

(f) The moment retrieval result of the ACRN.

Pipeline

- Background
- Learning Model
- Experiment
- Conclusion

Conclusion

- We present a novel Attentive Cross-Modal Retrieval Network, which jointly characterizes the attentive contextual visual feature and the cross-modal feature representation.

- We introduce a temporal memory attention network to memorize the contextual information for each moment, and treat the natural language query as the input of an attention network to adaptively assign weights to the memory representation.

- We perform extensive experiments on two benchmark datasets to demonstrate the performance improvement.

Thank you
Q&A