The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

Convolutional Hierarchical Attention Network for Query-Focused Video Summarization

Shuwen Xiao,[1] Zhou Zhao,[1]* Zijian Zhang,[1] Xiaohui Yan,[2] Min Yang[3]
[1]College of Computer Science and Technology, Zhejiang University, Hangzhou, China
[2]CBG Intelligent Engineering Dept., Huawei Technologies, China
[3]Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences
{xiaoshuwen, zhaozhou, ckczzj}@zju.edu.com, yanxiaohui2@huawei.com, min.yang@siat.ac.cn

20210428 何嘉峻

Outline

Introduction

Related Work

Proposed Method

Experiments

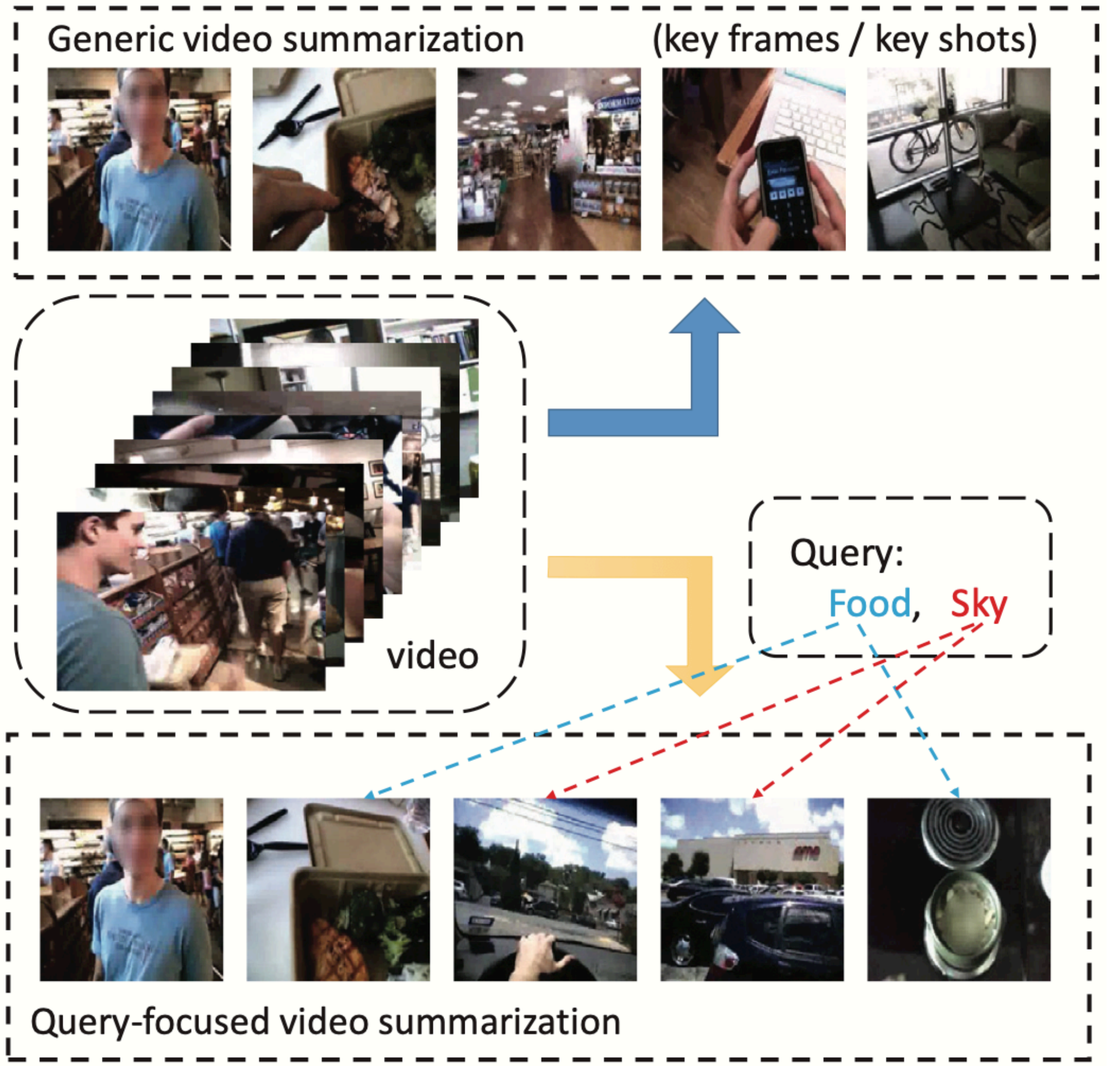
Conclusions

Comments

Introduction

Video Summarization

- Given a long video and a user query which is some concepts
  - The goal of **query-focused video summarization** is not only to remove the redundant parts of the video, to find key frame / shots in the video
  - But also to pick out those segments that **related to the user's query**.



Introduction

Differences between query-focused & generic video summarization

- Summary needs to take to the subjectivity of users into account
  - Different user queries may receive different video summaries
- Trained video summarizers can't meet all the users' preference
- Textual query will bring additional semantic information to the task

Related Work

Generic Video Summarization

- **Unsupervised:** take advantage of specific selection criteria to measure the importance or interestingness of video frames and then generate video summaries.
  - hand-crafted heuristics (2014), frame clustering (2016), GAN (2017)…
- **Supervised:** training data used in these methods contains raw videos and human-created ground-truth annotations. Combined with a large number of human annotations, models can capture the video content with more semantic information
  - Web image (2014), dppLSTM (2016), bi-LSTM (2016), RNN+LSTM (2018)…
- **Weakly supervised approaches:** readily available label, such as video categories, as additional information to improve model performance
  - 3D ConvNet (2017)

Related Work

Query-Focused Video Summarization

- These approaches are sequential models, using DPP-based algorithm or LSTM structure:
  - DPP-based algorithm (2016)
  - Memory network based model (2017)
  - Quality-aware relevance model (2017)
  - GAN (2018)

Introduction

Proposed Framework

- Consider the task of query-focused video summarization as **ranking problem**
  - First select out the **important** visual content
  - Then compute the **similarity** between visual content and the given query
- Propose method name Convolutional Hierarchical Attention Network (CHAN)

Introduction

Contribution

- Propose **Convolutional Hierarchical Attention Network**
  - Based on convolution network and global-local attention mechanism
  - Able to generate video query-related summary in parallel
- Present a **feature encoding network** to learn the features of each video shot
  - Employ fully convolutional network with local self-attention and query-aware global attention mechanism to obtain features with more semantic information
- Employ a **query-relevance computing module**
  - Takes the feature of video shot and query as input and then calculate the similarity score

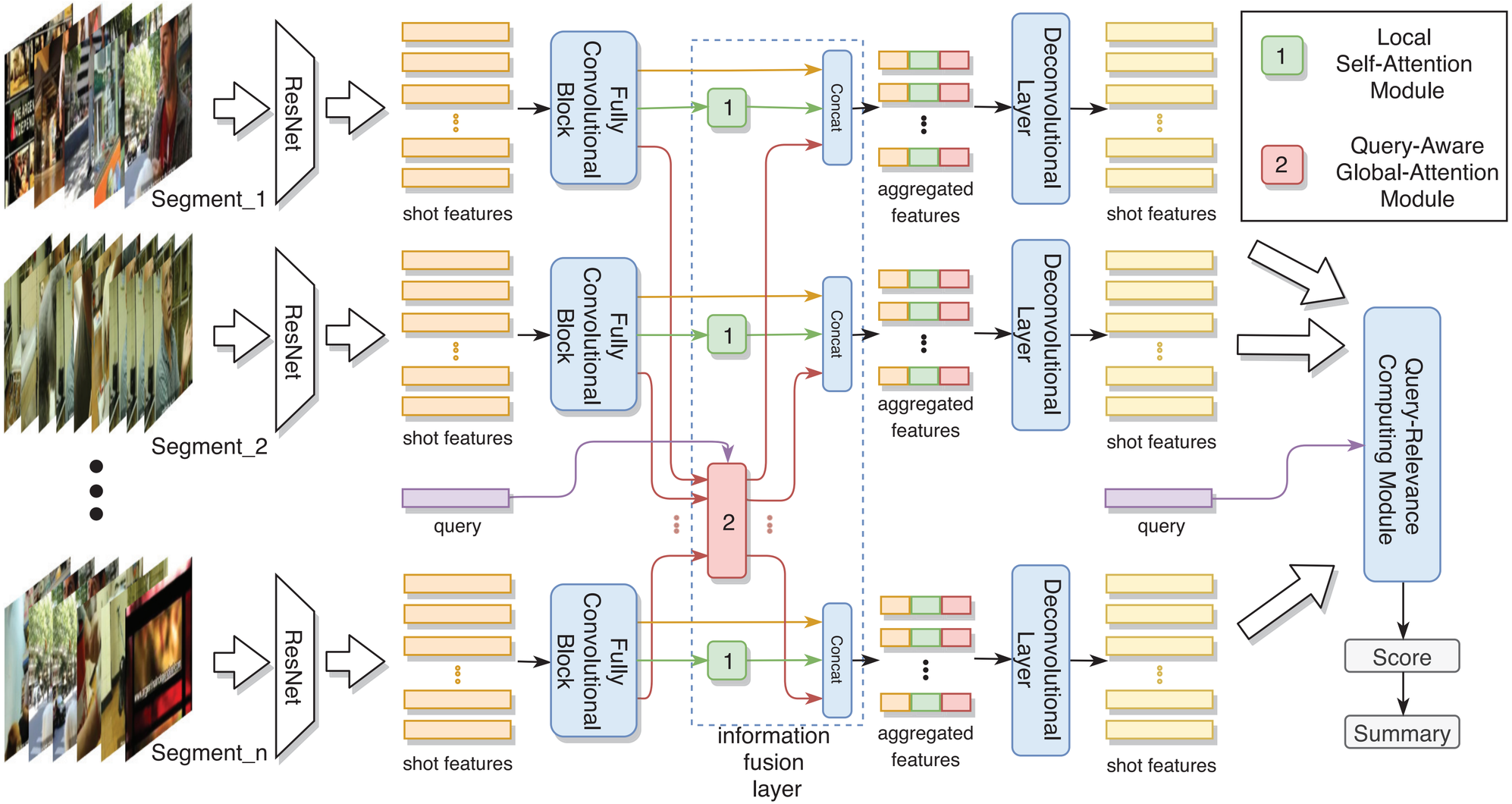
Introduction

Convolutional Hierarchical Attention Network (CHAN)

- Consists of two parts:
  - A **feature encoding network** to learn features from each video shot in parallel from a local perspective and a global perspective.
  - **Query-relevance ranking module** to calculate the similarity score with respect to a query for each shot and then select video content related to the given query.

Proposed Method

Convolutional Hierarchical Attention Network (CHAN)



Proposed Method

Problem Formalization

- Given a long video v and a query q, find out the diverse and representative subset of query-related frames or video shots.
- Denote the task as a problem of calculating the shot-query similarity.
- Denote the video as a sequence of video shots $(s_1, s_2, \ldots, s_n)$
  - n is the number of video shots
  - Each video shot is a small clip of the original video
- Denote h_q as representation of query

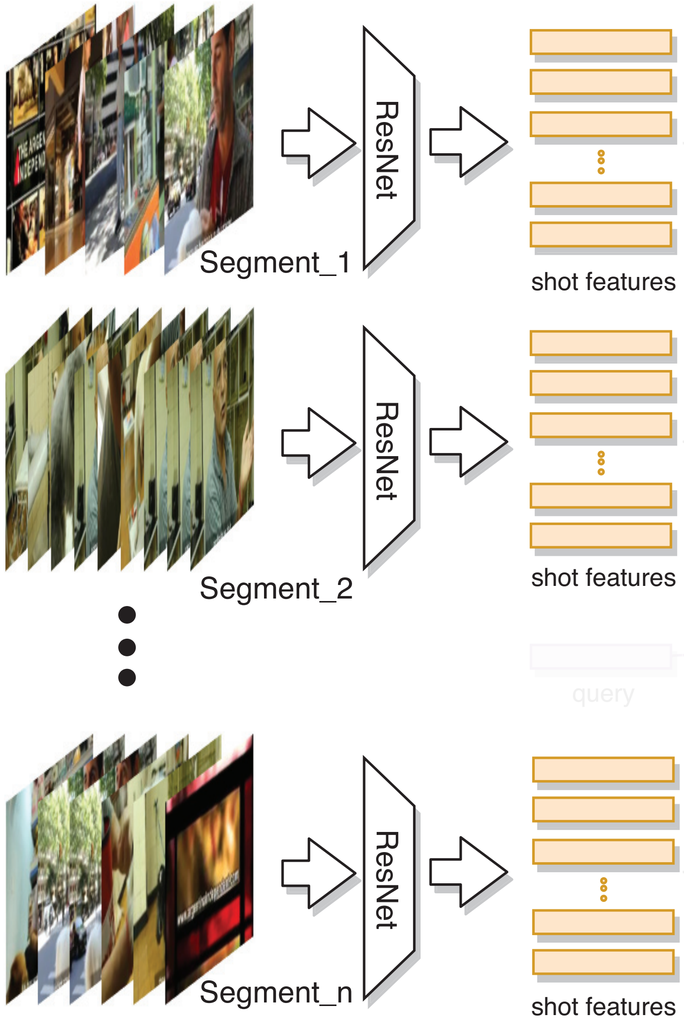
Proposed Method

Problem Formalization

- In the benchmark dataset, in where each query is consist of two concept (c_1, c_2)
  - Compute concept-related score for each shot
  - Then merge two kind of score as the query-related score
  - Finally, based on the score, can produce a diverse subset of video segments
    - Not only represent the origin video but related to the query
- **Input:** A long video v and a query q
- **Output:** A diverse subset of video shots remains original video info and related to query

Proposed Method

Feature Encoding Network

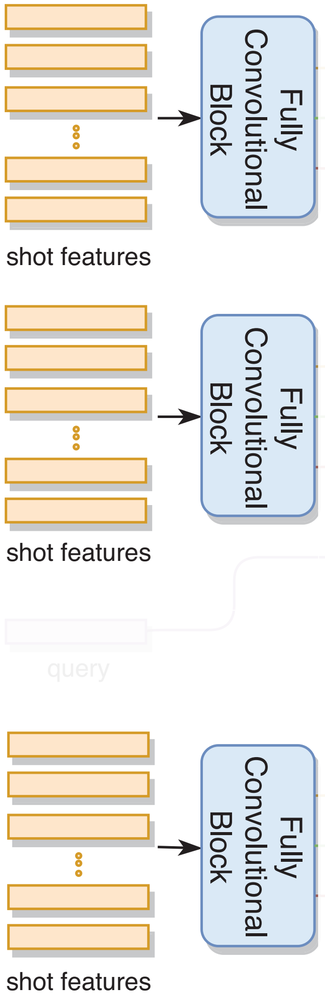


- Use the pertained ResNet to extract the visual feature of each shot after the segmentation process
- Denote the shot-level features in a specific video segment as $(v_1, v_2, \ldots, v_{s_k})$
  - s_k : the number of shots in the video segment
- Then use convolutional structure to further encode the visual feature

Proposed Method

Fully Convolutional Block

- Use 1D fully-convolutional network architecture to encode the shot-level visual feature
- Utilize **dilated convolutions** to obtain larger receptive field for handling long distance among the video segment

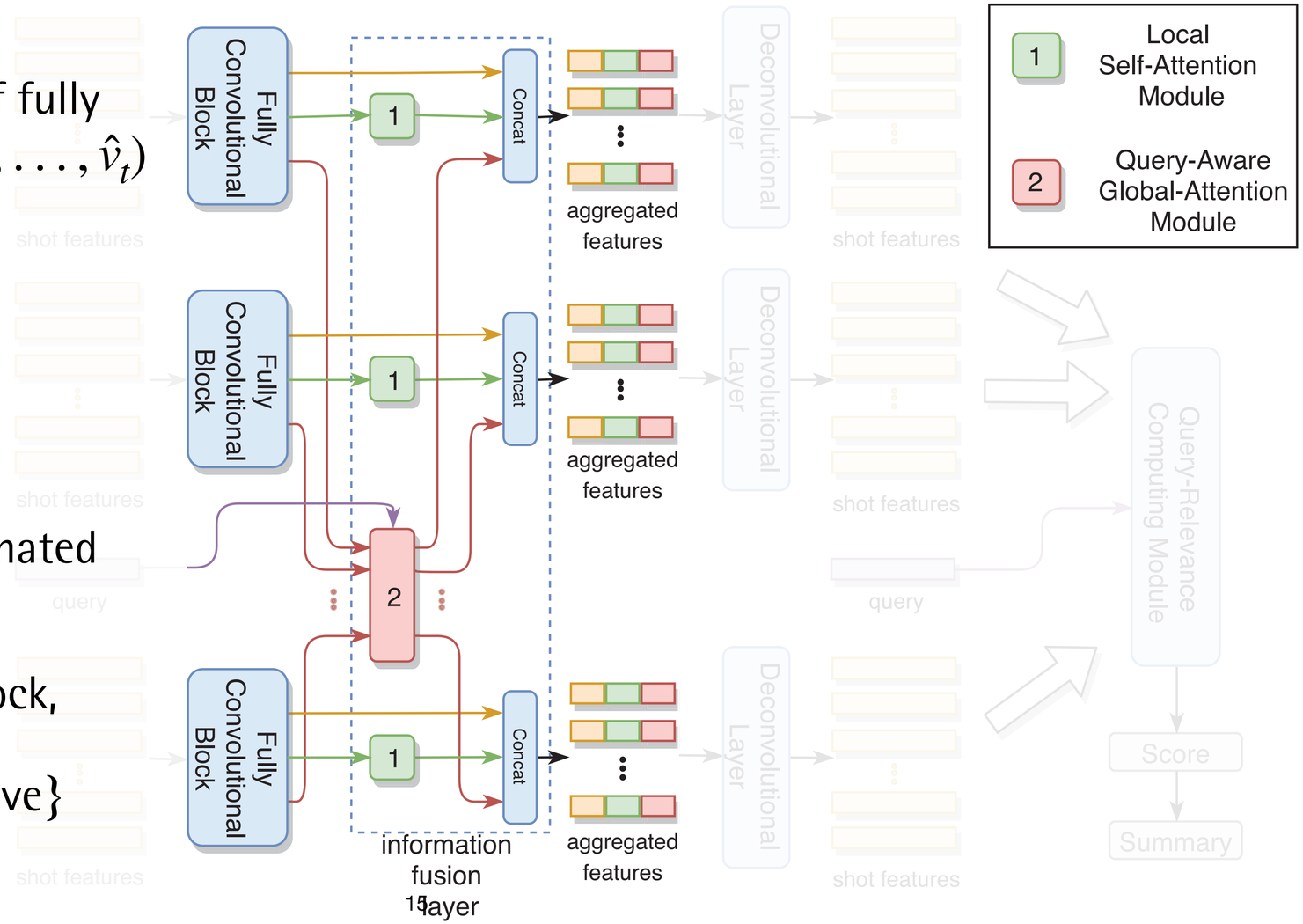


- First propose convolutional networks with different filter size and then concatenate their outputs which enables the model to receive more information
- The dilated convolution operation on i-th shot in a video segment: $o_i = \sum_{t=-k}^{k} f(t) \cdot v_{i+d \cdot t}$, where $2k+1$ is the filter size, f is the filter and d is dilation factor
- Then employ a pooling layer on the temporal axis of the video, can reduce computing time and also decrease the running memory of the model
- Connect different fully convolutional block and construct a multi-layer block to extracted representative features

Proposed Method

Information Fusion Layer

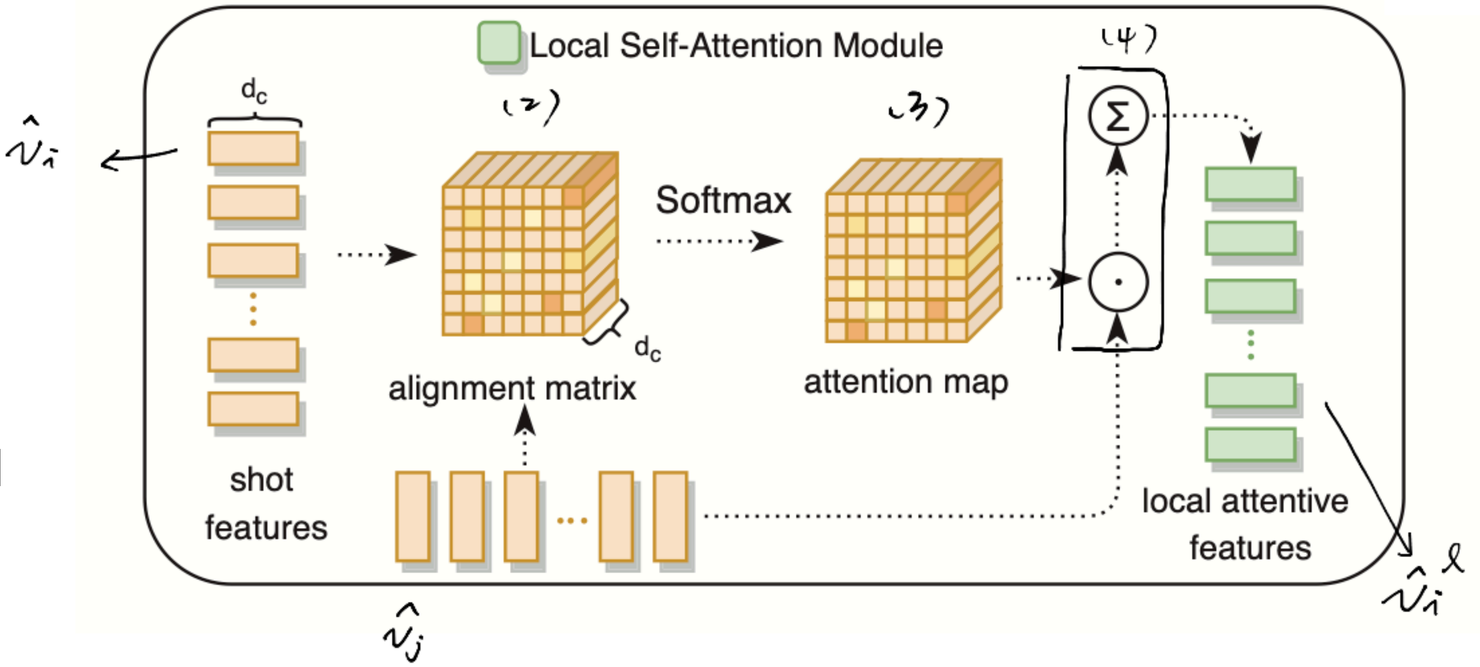
- Denote the output features of fully convolutional block as $(\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_t)$
  - t: length of output feature sequence
- **Input**: features from fully convolutional block
- **Output**: Sequence of concatenated vectors
  - {outputs from previous block, local attentive, query-aware global attentive}



Proposed Method

Local self-attention module

- Capture the semantic relations between all shots among a video segment.
- Given $(\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_t)$ to compute the alignment matrix. (shape: $t \times t \times d_c$)
- Module can learn the relative semantic relationship of different frames in the same segments.
- For different segments, the relation structure should be similar. Therefore, modules share all the trainable parameters, also reduces the amounts of parameters in our model.



(2) $f(\hat{v}_i, \hat{v}_j) = P \tanh(W_1 \hat{v}_i + W_2 \hat{v}_j + b) \in R^{d_c}$

- $P, W_1, W_2 \in R^{d_c \times d_c}$: trainable parameters
- $b \in R^{d_c}$: bias vector , d_c : dimension of $\hat{v}_i$

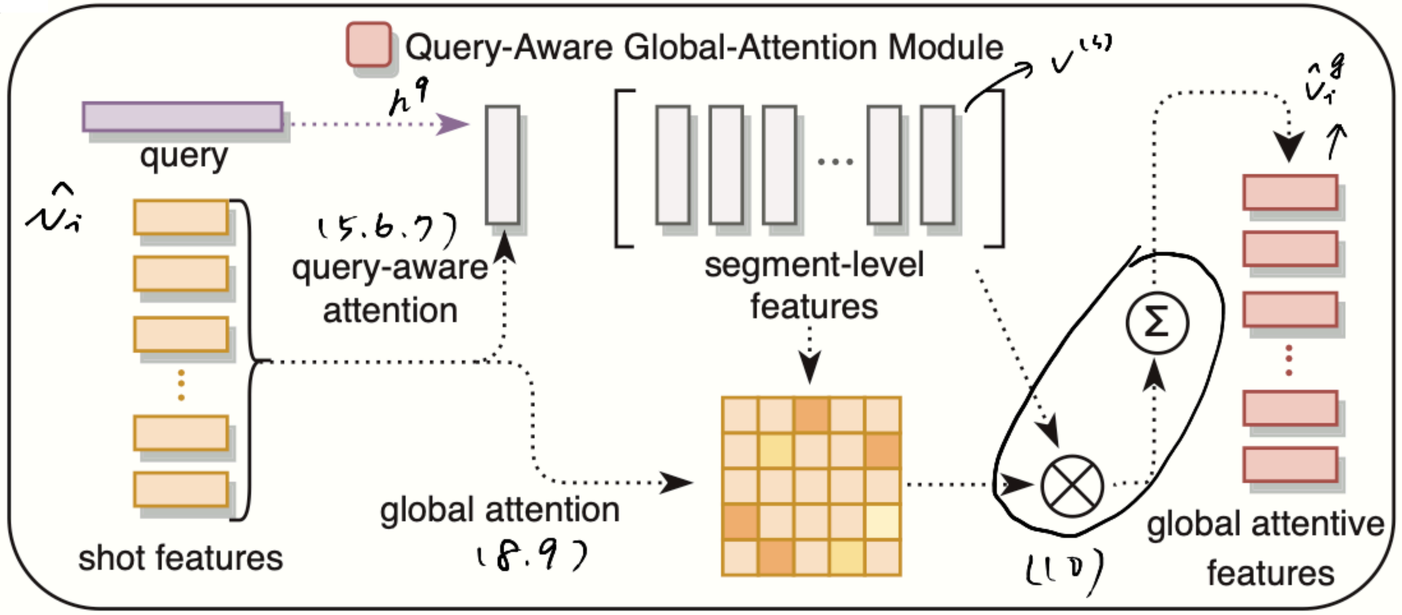
(3) $r_{ij} = \dfrac{\exp(f(\hat{v}_i, \hat{v}_j))}{\sum_{k=0}^{t} \exp(f(\hat{v}_i, \hat{v}_k))}$

(4) Local attentive video feature for i-th: $\hat{v}_i^l = \sum_{j=0}^{t} r_{ij} \odot \hat{v}_j$

Proposed Method

Query global-attention module

- Model the relationship of different video segments among the video and to generate query-focused visual representation.
- Given $(\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_t)$ and query q (composed of two concept (c_1, c_2))



(5) $e_i = v^T \tanh(W_1 \hat{v}_i + W_2 h^q + b)$

- v^T, W_1, W_2 : trainable parameters, b : bias vector
- h^q : average of representation of concepts

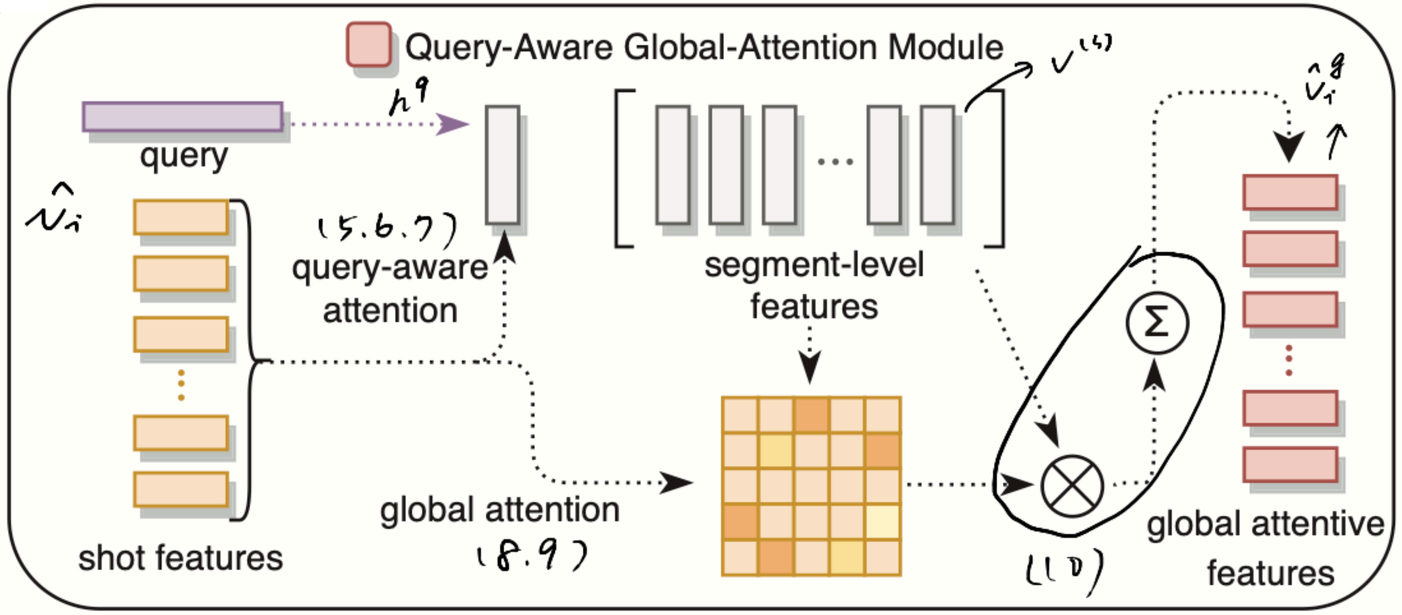
(6) $r_i = \dfrac{\exp(e_i)}{\sum_{k=0}^{t} \exp(e_k)}$

(7) Segment-level visual feature: $v^{(s)} = \sum_{i=0}^{t} r_i \hat{v}_i$

Proposed Method

Query global-attention module

- Compute the query-aware global -attentive representation for each shot.
- Given visual feature $\hat{v}_i$ & all segment-level visual representation $(v_1^{(s)}, v_2^{(s)}, \ldots, v_m^{(s)})$
  - m: number of video segments



(8) $e_j^g = v^T \tanh(W_1^g \hat{v}_i + W_2^g v_j^{(s)} + b)$

(9) $r_j^g = \dfrac{\exp(e_j^g)}{\sum_{k=0}^{m} \exp(e_k^g)}$

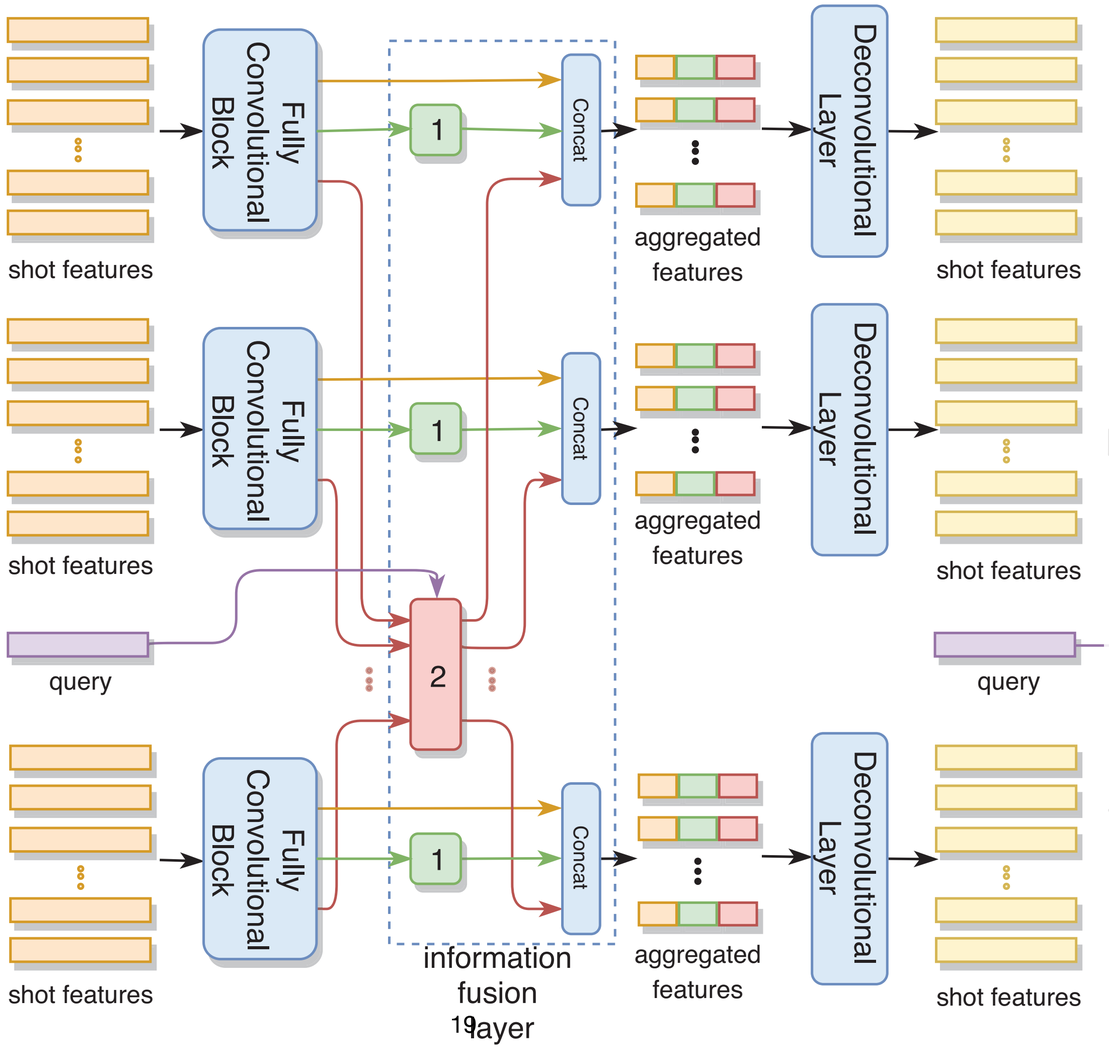
(10) Global-attentive representation for i-shot:

$$\hat{v}_j^g = \sum_{j=0}^{m} r_j^g v_j^s$$

Proposed Method

Deconvolutional Layer

- Use fully convolutional block to decrease the number on temporal axis of video features.

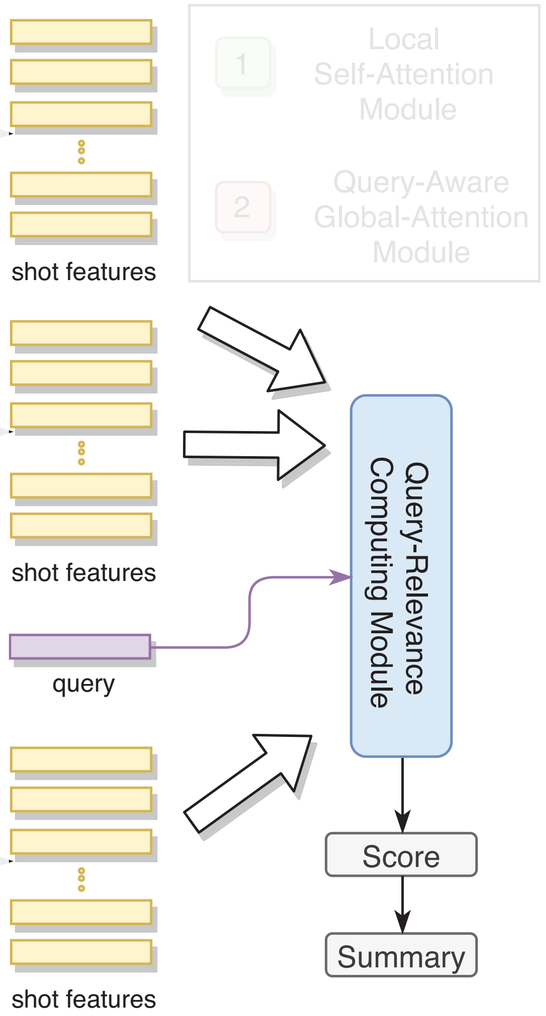


- Propose several 1D deconvolutional layers to recover the original number of video shots.

Proposed Method

Query-Relevance Computing Module

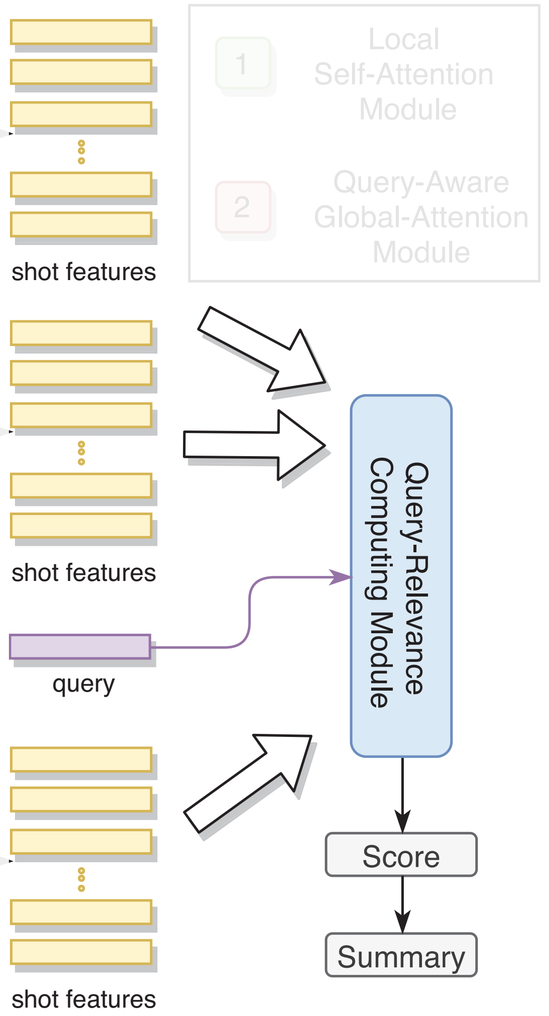
- Compute the similarity score between video shot and user's query.
- The module takes the shot-level features $(\hat{v}_1^f, \hat{v}_2^f, \ldots, \hat{v}_n^f)$ generate by feature encoding network and concepts as inputs.
- Given a specific concept c, using pretrained language model to obtain its embedding feature f_c.
- Calculate f_c and shot-level features $\hat{v}_i^f$ of i-th shot distance-based similarity: $d_i = W_f \hat{v}_i^f \odot W_c f_c$
  - W_f, W_c: parameter matrices project the visual & textual features into same vector space



Proposed Method

Query-Relevance Computing Module

- Then let the output pass a MLP and get the concept-relevant score between i-th segment and concept c.
- The average of two concept-relevant score is taken as the query-relevant score $s = \{s_1, s_2, \ldots, s_n\}$
- Given the ground truth annotations $\hat{s} = \{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_n\} \in [0,1]$
- Loss: $L_{summ} = \frac{1}{T}\sum_{t=1}^{T} \hat{s}_t \log s_t + (1-\hat{s}_t)\log(1-s_t)$
- By minimizing the loss, module can focus on the most concept-related video shots.



Experiments

Datasets

- Sharghi, A.; Laurel, J. S.; and Gong, B. (CVPR 2017): *Query-Focused Video Summarization: Dataset, Evaluation, and A Memory Network Based Approach*
- Contains videos taken from the first person perspective.
- The dataset has 4 videos containing different daily life scenarios, each of which lasts 3~5 hours.
- Provide a set of concepts for user's queries (total 48), the concepts are concise and diverse, related to some common objects in our daily life.
- Each query is composed of two concepts and there are 46 queries in the dataset.

Experiments
Datasets

- As for queries, there are four different scenarios:
    1. All concepts in the query appear in the same video shot
    2. All concepts in the query appear in the video but not in the same shot
    3. Some of the concepts in the query appear in the video
    4. None of the concepts in the query appear in the video
        - Is to some extend the same as general form video summarization
- The dataset provide per-shot annotation, from which each shot labeled with several concepts.

Experiments

Compared Models

- **SeqDPP (2014):** formulates video summarization as a subset selection problem and use sub-modular maximization to found summary. (dose not consider user queries)
- **SH-DPP (2016):** extension of SeqDPP, add a extra layer in the process of SeqDPP to judge whether a video shot is related to a given query.
- **QC-DPP (2017):** another extension of SeqDPP, introduces memory network to parameterize the kernel matrix.
- **TPAN (2018):** the three-player adversarial network, uses GAN to tackle with the task and introduce a random summary as an extra adversarial sample.

Experiments

Experimental Results

Table 1: Comparison results on the query-focused video summarization dataset in terms of Precision, Recall and F1-score.

	SeqDPP			SH-DPP			QC-DPP			TPAN			CHAN		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Vid1	53.43	29.81	36.59	50.56	29.64	35.67	49.86	53.38	48.68	49.66	50.91	48.74	54.73	46.57	**49.14**
Vid2	44.05	46.65	43.67	42.13	46.81	42.72	33.71	62.09	41.66	43.02	48.73	45.30	45.92	50.26	**46.53**
Vid3	49.25	17.44	25.26	51.92	29.24	36.51	55.16	62.40	56.47	58.73	56.49	56.51	59.75	64.53	**58.65**
Vid4	11.14	63.49	18.15	11.51	62.88	18.62	21.39	63.12	29.96	36.70	35.96	**33.64**	25.23	51.16	33.42
Avg.	39.47	39.35	30.92	39.03	42.14	33.38	40.03	60.25	44.19	47.03	48.02	46.05	46.40	53.13	**46.94**

- CHAN is outperforms the state-of-the-art approach (TPAN) by 1.9%
  - Specifically video 2 & 3, CHAN can have a better performance than TPAN (2.64%, 3.6%)
- The improvements of performance identify the effectiveness of our approaches to learn the relevance between the video shots and user's query.
  - The average running time of each video 134.4ms, shorter than TPAN 1.614s by 91.6%.

Experiments

Ablation Study

- F1-score of CHAN without local self-attention module is reduced by 7.84%
- The performance without query-aware global attention module decreases by 18.8%
- Local self-attention & query-aware global attention module can capture visual information inside a video segment and between segments
  - Helpful to improve performance

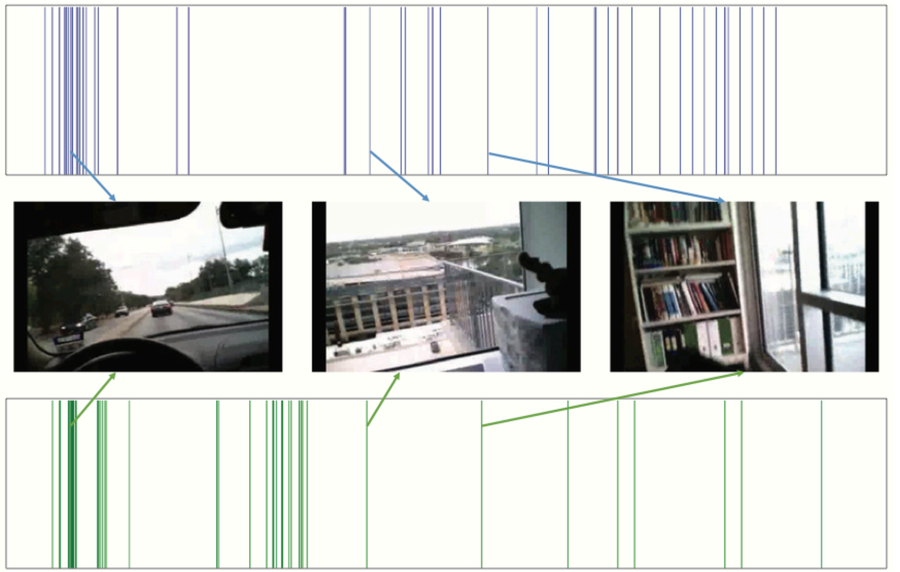
Table 2: Ablation analysis on query-conditioned video summarization in terms of Precision , Recall and F1-score.

Model	Pre	Rec	F1
CHAN w/o Local Att	42.72	49.04	43.26
CHAN w/o Global Att	37.62	43.17	38.09
CHAN	46.40	53.13	46.94

Experiments

Qualitative Results

- Observed that the selected summaries are related to one or both concepts in the given query, demonstrates that CHAN is able to find diverse, representative and query-related summaries.



(a) Visualization Result (Book, Street)

Figure 3: The visualization results of our approach. The x-axis denotes the video shot number. Green lines represent the ground truth annotations and blue lines represent the predicted key shots from our method. Figures shows the summary for the query "Book Street"

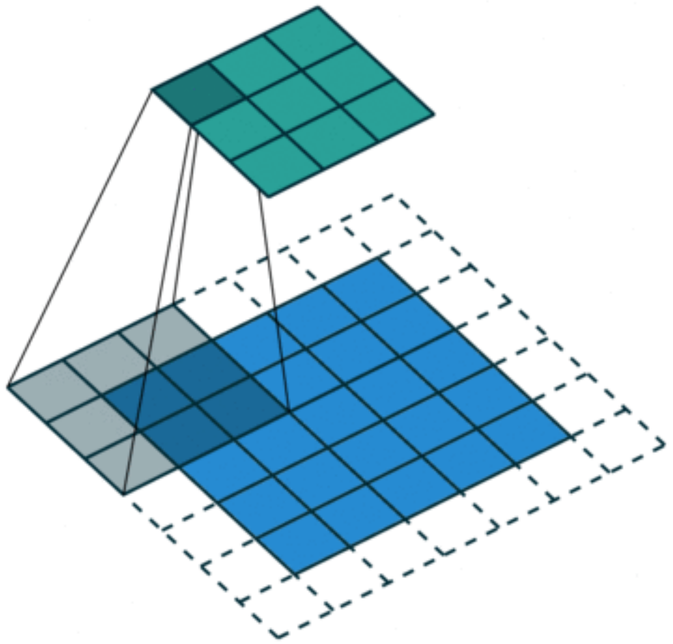
Conclusions
and Further works

- Formulate query-focused video summarization as ranking task
- CHAN is firstly encodes the video shots in parallel, then computes relevance scores between each shot and query and finally generate query-related video summary
- Extensive experiments on the benchmark dataset show the effectiveness and efficiency of CHAN
- The authors are going to design a more general model which can generate video summary for new queries.

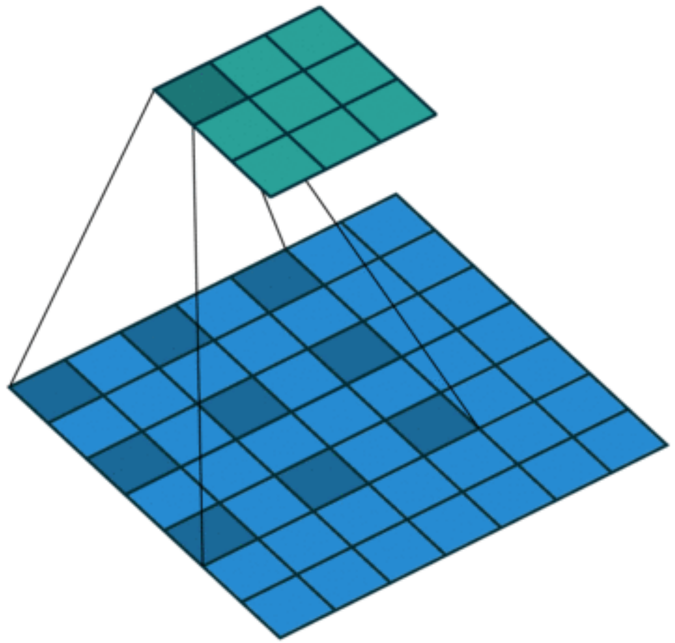
Comments

of Convolutional Hierarchical Attention Network for Query-Focused Video Summarization

- Encode video shots in parallel



Standard Convolution with a 3 x 3 kernel (and padding)



Dilated Convolution with a 3 x 3 kernel and dilation rate 2