Query-Conditioned Three-Player Adversarial Network for Video Summarization

Yujia Zhang*[1][2]
zhangyujia2014@ia.ac.cn
Michael Kampffmeyer[3]
michael.c.kampffmeyer@uit.no
Xiaodan Liang[4]
xdliang328@gmail.com
Min Tan[1][2]
min.tan@ia.ac.cn
Eric P. Xing[4]
epxing@cs.cmu.edu

[1] Institute of Automation,
Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
[3] UiT The Arctic University of Norway
[4] Carnegie Mellon University

Abstract

Video summarization plays an important role in video understanding by selecting key frames/shots. Traditionally, it aims to find the most representative and diverse contents in a video as short summaries. Recently, a more generalized task, query-conditioned video summarization, has been introduced, which takes user queries into consideration to learn more user-oriented summaries. In this paper, we propose a query-conditioned three-player generative adversarial network to tackle this challenge. The generator learns the joint representation of the user query and the video content, and the discriminator takes three pairs of query-conditioned summaries as the input to discriminate the real summary from a generated and a random one. A three-player loss is introduced for joint training of the generator and the discriminator, which forces the generator to learn better summary results, and avoids the generation of random trivial summaries. Experiments on a recently proposed query-conditioned video summarization benchmark dataset show the efficiency and efficacy of our proposed method.

1 Introduction

Video summarization aims to select key frames/shots among videos to summarize the main storyline and has been widely investigated for facilitating video understanding [5, 7, 16, 20, 29, 34]. As shown in Figure 1, this task can be classified into two types: a) generic video summarization, which only takes the visual features of the video contents as the input and b) query-conditioned video summarization which conditions summarization on user queries.

The generic video summarization task has been addressed at three different levels: shot-level [14, 15], frame-level [11, 12], and object-level [17, 33] video summarization by selecting key shots/frames/objects in the videos. However, one main issue with generic video

*Work done while the first author was at CMU.

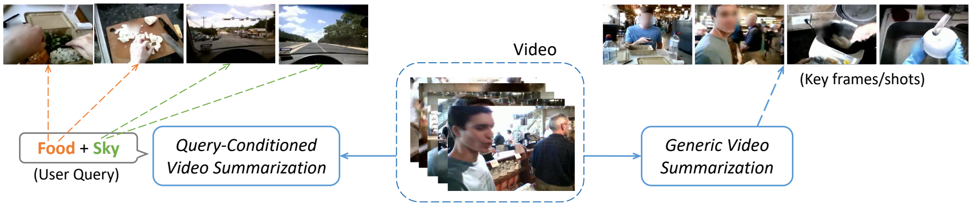

Food + Sky
(User Query)
Query-Conditioned Video Summarization
Video
Generic Video Summarization
(Key frames/shots)

Figure 1: Different video summarization tasks. Generic video summarization aims to generate key contents of a video, while query-conditioned video summarization takes the user query into consideration and generates summaries accordingly.

summarization is the fact that it does not take user preferences into account, since different users may have different preferences towards the video content, and a single evaluation metric is not robust enough for all video summaries [23].

Recently, another research direction, query-conditioned video summarization [22, 23, 27], has been explored, which takes advantage of different user queries in form of texts to learn more user-oriented summaries. It generates user-oriented summaries that have effective correlations between summaries and query, and capture the overall essence of the video. Several approaches to query-conditioned video summarization have been proposed. Sharghi et al. [22] first extend a sequential DPP (seqDPP) [4] to extract key shots. Afterwards, they develop a more comprehensive dataset for this task, and propose a memory network [25] parameterized seqDPP model. However, there is still room to learn a better summarizer due to the limitation of the memory to jointly encode video and query.

To address the above issue we develop a query-conditioned three-player generative adversarial network architecture. We encode the query and the video sequence to learn a joint representation combining visual and text information, and take this query-conditioned representation as the input to the generative adversarial network. A three-player structure is applied during joint training, in order to achieve superior regularization. The contribution in our work can be summarized as follows: first, we propose a query-conditioned three-player adversarial network, which jointly encodes query and visual information and learns in an adversarial manner. Second, we introduce a three-player structure for the adversarial training. The discriminator regularizes the model via the three-player loss, which facilitates the generator to generate more related and meaningful video summaries. Two supervised losses are applied to ensure a more compact summary. One loss regularizes the length and the other aligns prediction and ground truth. Experimental results on a public dataset [23] demonstrate the superiority of our proposed approach against the state-of-the-art method.

2 Related Work

2.1 Generic Video Summarization

Generic video summarization [11, 12, 14, 30, 31], has been widely studied for efficient video analysis and video understanding. For shot-level video summarization [14, 15, 24, 28], Song et al. [24] propose to learn the canonical visual concepts which are shared between videos and images to find important shots. In [28], a pairwise deep ranking model is proposed to distinguish highlight segments from non-highlight ones. For frame-level video summarization [4, 11, 12, 32], Khosla et al. [11] use web-images as a prior to facilitate video summarization. In [4], a probabilistic model is proposed for learning sequential structures to generate summaries. Approaches to object-level video summarization [17, 33] aim to obtain representative objects to perform fine-grained summarization. Currently there are two existing GAN-based works [16, 32] that include regularization using adversarial training. However, they do not consider user preferences, so the summaries may not be robust and may not generalize well to different users. Therefore, we investigate the query-conditioned video summarization task to provide more personalized summarization results by relying on user queries.

2.2 Query-conditioned Video Summarization

Query-conditioned video summarization [10, 19, 22, 23, 27] takes user queries in the form of texts into consideration in order to learn more user-oriented summaries. In [22], a Sequential and Hierarchical DPP (SH-DPP) is developed to tackle this challenge. In [27], the authors adopt a quality-aware relevance model and submodular mixtures to pick relevant and representative frames. There are two other works related to query-conditioned video summarization. One is used to generate visual trailers, while the other obtains web images conditioned on user queries, and then produces video summaries from both images and videos. Specifically, Oosterhuis et al. [19] propose a graph-based method to generate visual trailers by selecting frames that are most relevant to a given user's query. Ji et al. [10] formulate the task by incorporating web images, which are obtained from user query searches. Thus the video summarization is indirectly conditioned on the query through the web images.

Recently, Sharghi et al. [23] explore a more thoroughly query-conditioned video summarization approach. Instead of using datasets which are originally collected for the generic task, they propose a new dataset and an evaluation metric towards this task. Our work is developed based on this new dataset and the evaluation metric. We propose a novel query-conditioned adversarial network which does not rely on external knowledge, such as web images, and can effectively summarize videos based on user queries by integrating a three-player adversarial training structure.

3 Proposed Algorithm

3.1 Generator Network

Our proposed network facilitates query-conditioned video summarization by applying an adversarial network that takes the query into consideration with a three-player discriminator loss. Figure 2 illustrates the whole framework of our approach. The generator is mainly tasked with embedding visual information and text jointly, in order to provide comprehensive query-conditioned representations. The discriminator aims to distinguish the real summaries, i.e., ground-truth summary from random and generated summaries.

In the following sections, we first introduce the query-conditioned generator network in our model to select key shots with regards to different queries. We then present the proposed query-conditioned discriminator with three-player loss, which distinguishes the ground-truth summary from the random and the generated summaries. Finally, we introduce the details of adversarial training with two supervision losses in our model.

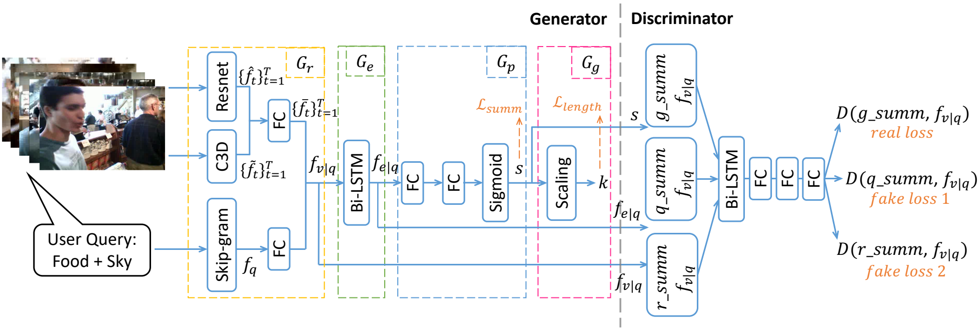


Figure 2: The network architecture of our proposed method for query-conditioned video summarization. In the generator, the video is fed into a query-conditioned feature representation module G_r, that integrates query and visual information. After a compact video encoding process in module G_e, we can predict shot scores in module G_p. The summary is then generated using video summary generator G_g, with the final results. We further introduce two regularizations: the summary regularization $\mathcal{L}_{summ}$ and the length regularization $\mathcal{L}_{length}$ to enhance the generator's ability to learn superior summaries. The discriminator uses three query-conditioned representations as the input, and is tasked to distinguish the real summary from two fake summaries in an adversarial learning manner.

3.1.1 Query-Conditioned Feature Representation Module G_r

Frame-level visual representation. We denote an input video as $\mathcal{V} = \{\mathcal{V}_t\}_{t=1}^T$, where T denotes the total number of shots in a video. Each video shot $\{\mathcal{V}_t\}$ is partitioned into 75 frames (5-second long) for fair comparison with related work [23]. As shown in Figure 2, the model G_r aims to generate feature representations which are conditioned on user queries. We first apply the ResNet-152 feature extractor [8] to encode frame-level visual features. In order to do this, we downsample each shot to 16 frames per segment. The "fc7" layer of the ResNet-152 model trained on the ILSVRC 2015 dataset [21] is used to obtain features for frames within each shot and followed by an average pooling layer. This frame-level feature vector is denoted as $\{\hat{f}_t\}_{t=1}^T$.

Shot-level visual representation. We apply the C3D video descriptor [26], a network trained on the Sports1M dataset [26], to extract shot-level feature representation. We use the output of the "fc6" layer of the C3D and uniformly split the video into 75 frames before downsampling it to 16 frames per segment, aligning it with the extracted ResNet features, to get the shot-level visual features. The features we extracted from the C3D are denoted as $\{\tilde{f}_t\}_{t=1}^T$.

Textual representation. To obtain textual feature representations, we use the Skip-gram model [18], a word2vec model pretrained on the GoogleNews dataset, to project each word into a semantic feature space. We define each user query as q. Each q contains two concepts (words), and we generate the concept embedding by summing up the two feature vectors of the two concepts. After that, we encode the textual representation f_q by applying a fully connected layer.

Query-Conditioned Feature representation. We first combine frame-level and shot-level feature vectors $\{\hat{f}_t\}_{t=1}^T$ and $\{\tilde{f}_t\}_{t=1}^T$ by concatenation, followed by a fully connected layer to get the encoded visual feature $\{\bar{f}_t\}_{t=1}^T$. After that, we use another concatenation to combine $\{\bar{f}_t\}_{t=1}^T$ and the textual representation f_q. Thus, we obtain a query-conditioned feature encoding for the video, which can be denoted as $f_{v|q} = \{f_1^v, f_2^v, \dots, f_T^v|q\}$, i.e., $f_{v|q} = G_r(\mathcal{V}|q)$.

3.1.2 Video Summarization Prediction

Compact Video Encoding Module G_e. Given the generated query-conditioned feature representation f_v from model G_r, we introduce the compact video encoding module G_e to learn the temporal dependencies among video shots. Thus the output of the compact video encoding can be produced as $f_{e|q} = G_e(f_{v|q})$, where $f_{e|q} = \{f_1^e, f_2^e, \dots, f_T^e|q\}$. The model G_e consists of a Bidirectional LSTM (Bi-LSTM) layer [6] to model the temporal representation, followed by a batch normalization layer [9] and Rectified Linear Unit (ReLU) activation [3] to learn the compact video encoding.

Shot Score Prediction Module G_p. In order to predict a confidence score for each video shot, we propose the shot score prediction module G_p. We define the confidence score as $s = \{s_t\}_{t=1}^T$, where $s_t = G_p(f_t^e|q)$. In our setting, we use two fully connected layers with a batch normalization and a ReLU activation in the middle. After that, we apply a sigmoid layer in order to generate a confidence score for each video shot being a key shot.

Video Summary Generator G_g. Given the confidence score of each video shot from model G_p, we introduce the video summary generator G_g to generate the final results for selected key shots by means of scaling. To get the video summary results, we apply $k = G_g(s)$ by passing the shot score into the video summary generator, where k is the summary result for the shot, and $k = \{k_t\}_{t=1}^T$. $k_t = 0$ means that the shot is a trivial one, while $k_t = 1$ represents that it is a key shot which will be included in the generated summary. G_g is a softmax function with a temperature parameter τ to get the result of each k_t:

$$k_t = \frac{e^{s_t/\tau}}{e^{s_t/\tau} + e^{(1-s_t)/\tau}}. \tag{1}$$

3.2 Discriminator Network

We use three pairs of different summaries together with the feature representation of the video as the input to the discriminator. For simplicity, we use q_{summ}, g_{summ} and r_{summ} to denote *generated query-conditioned summary*, *ground-truth query-conditioned summary*, and *randomly generated query-conditioned summary*, respectively. The three pairs are: *(q_{summ}, video shots)*, *(g_{summ}, video shots)*, and *(r_{summ}, video shots)*. The *video shots* we use are the learn joint embedding of visual and query information. We use r_{summ} to enhance the ability of the generator to learn a more robust summary as well as avoid the generation of random trivial short sequences.

As shown in Figure 2, we use query-conditioned feature representation $f_{v|q}$ generated from model G_R as the input of *video shots* in the three pairs as the learned feature for the video. r_{summ} is obtained using the random summary score s_r and by generating random values of 0 and 1. The length of s_r is the same as the one of predicted summary s from the video summary generator G_g. g_{summ} is produced using a ground-truth summary score s_g, where $s_g = \{s_1^g, s_2^g, \dots, s_T^g\}$. Note, s_r,$s_g \in [0,1]$. In order to get $(q_{summ}, f_{v|q})$, $(g_{summ}, f_{v|q})$ and $(r_{summ}, f_{v|q})$, the three summary representations can be defined as:

$$\begin{aligned} q_{summ} &= f_{e|q} \cdot s, \\ g_{summ} &= f_{e|q} \cdot s_g, \\ r_{summ} &= f_{e|q} \cdot s_r. \end{aligned} \tag{2}$$

After that, we take $f_{v|q}$ as the input to a Bi-LSTM layer, followed by a batch normalization layer and ReLU activation, and pass q_{summ}, g_{summ} and r_{summ} to another Bi-LSTM and a batch normalization layer with ReLU activation to learn a temporal representation. Then we concatenate them in pairs and apply three fully connected layers and jointly train the discriminator to distinguish the true summary from fake ones.

3.3 Adversarial Learning

We first introduce the summary regularization $\mathcal{L}_{summ}$ to optimize the generator by enforcing the selection of key shots to align with the ground-truth. It aligns a predicted shot score s_t from the model G_p with the corresponding ground-truth summary score s_t^g:

$$\mathcal{L}_{summ} = \frac{1}{T} \sum_{t=1}^{T} (s_t - s_t^g)^2. \tag{3}$$

We further incorporate the length regularization $\mathcal{L}_{length}$, which is computed between the number of generated summary shots and the ground-truth summary during the adversarial training to control the length of summaries:

$$\mathcal{L}_{length} = \left| \frac{1}{T} \sum_{t=1}^{T} k_t - \gamma \right|, \tag{4}$$

where γ is the percentage of the key shots in the the video based on ground-truth summary, and k_t is the summary result for each video shot generated from the model G_g.

Our adversarial objective function is based on Wasserstein GANs [2], due to its good convergence property. Note that this does not exclude the use of other GAN-based objectives, as our model is flexible enough to be combined with other GAN structures.

Instead of using the commonly used two-player learning mode, we optimize it with the three-player loss as shown in Figure 2: the real loss $D(g\_summ, f\_v|q)$ and the two fake losses $D(q\_summ, f\_v|q)$ and $D(r\_summ, f\_v|q)$. The thee-player loss can not only force the model to generate good summaries, but can also avoid the learning of a trivial summary of randomly selected shots.

The generator G and the discriminator D conditioned on query q are jointly optimized by the use of a min-max adversarial objective:

$$\begin{aligned} \min_G \max_D \mathcal{L}(G,D|q) = & \mathbb{E}_g[D(g_{summ}, f_{v|q})] \\ & - \omega \mathbb{E}_q[D(q_{summ}, f_{v|q})] - (1-\omega)\mathbb{E}_r[D(r_{summ}, f_{v|q})], \end{aligned} \tag{5}$$

where ω is the balancing parameter for the two fake losses. Here we use $\omega = 0.5$ for treating the two fake losses equally. We replace the generator G with models G_r, G_e, G_p in Section 3.1 and the three summary representation q_{summ}, g_{summ} and r_{summ} in Section 3.2, so that the objective function Eq. (5) can be reformulated as:

$$\begin{aligned}
&\min_{G_r,G_e,G_p} \max_{D} \mathcal{L}(G_r,G_e,G_p,D|q) = \\
&\mathbb{E}_g[D(G_e(G_r(\mathcal{V}|q)) \cdot s_g, G_r(\mathcal{V}|q)] \\
&-0.5\mathbb{E}_q[D(G_e(G_r(\mathcal{V}|q)) \cdot G_p(G_e(G_r(\mathcal{V}|q)), G_r(\mathcal{V}|q)] \\
&-0.5\mathbb{E}_r[D(G_e(G_r(\mathcal{V}|q)) \cdot s_r,), G_r(\mathcal{V}|q)].
\end{aligned} \tag{6}$$

Thus, the final objective function conditioned on query q including the two supervised losses, $\mathcal{L}_{summ}$ and $\mathcal{L}_{length}$, can be denoted as G^*:

$$G^* = \arg\min_{G_r,G_e,G_p} \max_{D} \mathcal{L}(G_r,G_e,G_p,D|q) + \mathcal{L}_{summ} + \mathcal{L}_{length} \ . \tag{7}$$

4 Experimental Results

4.1 Datasets and Settings

Datasets. We evaluate our approach on the query-conditioned dataset proposed in [23], which is built upon the existing UT Egocentric (UTE) dataset [13]. The dataset has 4 videos in total, containing different uncontrolled daily lives scenarios and each being 3∼5 hours long. A dictionary of concepts for user queries is supplied, which is a concise and diverse set of 48 concepts, which are deemed to be comprehensive of daily life for the query-conditioned video summarization. As for the queries, four different scenarios are included to formalize comprehensive queries [23]. Note that among different queries, one scenario is introduced where none of the concepts in the query are presented in the video. The three remaining scenarios are 1) queries, where all concepts appear together in the same video shot, 2) queries, where all concepts appear but not in the same shot, and 3) queries, where only one of the concepts appears. For fair comparison, we follow [23] and randomly select two videos for training, leaving one for testing and one for validation. Four experiments are performed to test all four videos.

Evaluation Metrics. In [23], the authors propose to find the ideal mapping between the generated query and the ground-truths summary by the maximum weight matching of a bipartite graph, based on a similarity function between two video shots. The similarity function uses the intersection-over-onion (IoU) on corresponding concepts to evaluate the performance. The IoU is defined as the edge weights, and the generated and ground-truths summaries are on opposing sides of the graph. Precision, recall, and F1-score are computed based on the number of matched summary pairs.

4.2 Implementation Details

We implement our work using TensorFlow [1], with 1 GTX TITAN X 12GB card on a single server. In the generator G, the output for frame- and shot-level visual representation are 2048- and 4096-dimensional vectors, and the textual representation is a 300-dimensional vector. The learned temporal representation after the Bi-LSTM is 2048-dimension. In the module G_p, the output of the two fully connected layers are 128- and 1-dimensional vectors, respectively. We use a low-temperature softmax function in module G_g in order to get approximately binary results. The dropout between the two fully connected layers is 0.5.

Table 1: Results obtained by our method compared to other approaches for query-conditioned video summarization in terms of Precision (Pre), Recall (Rec) and F1-score(F1).

	SeqDPP [4]			SH-DPP [22]			QC-DPP [23]			**Ours**		
	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**
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**
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**
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**
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**

In the discriminator D, the Bi-LSTM that we use encodes the features to a 512-dimensional vector. The output dimensions for the three fully connected layers are 512-, 256-, and 128-dimensions. During the training phase, we randomly select a set of 1000 successive video shots for each batch, with one user query for all shots. For the generic scenario where none of the concepts in the query are presented in the video, we use a zero vector of 300-dimension for the query embedding. During the testing phase, we obtain the predicted shot score in module G_p for each video shot.

The inference times for the 4 videos are $1472ms$, $1948ms$, $1141ms$ and $1893ms$, respectively, so on average, it takes $1.614s$ for each video to generate query-conditioned key video shots.

4.3 Quantitative Results

4.3.1 Comparison Analysis

We compare our approach with all other frameworks which have been applied to this query-conditioned video summarization dataset. The precision, recall and F1-score comparison for the four videos are shown in Table 1. It can be observed that our approach outperforms the existing state-of-the-art by 1.86%. Especially for Video 2 and Video 4, we achieve 3.64% and 3.68% better performance than [23] in terms of F1-score. Such substantial performance improvements indicate the superiority of our proposed method by using a three-player adversarial network on the joint embedding of visual information and the user query. The rest three works are all based on a DPP architecture, which can learn long time temporal relations among video shots. However, our work adopts the adversarial learning objective, which facilitates both temporal and query-conditioned joint learning. The two regularizations on summary and length also help obtaining better query-conditioned summary.

4.3.2 Ablation Analysis

We conduct experiments on different components of our model. As shown in Table 2, we use $w/o-\mathcal{L}_{length}$, $w/o-\mathcal{L}_{summ}$ and $two-player$ to denote our model when trained without the length regularization loss, the ground-truth summary regularization loss, and the random summary loss respectively. We can observe that the performance is reduced slightly after dropping the length regularization and the random summary as a form of two-player structure. Thus it demonstrates the effects of the length regularization and the three-player manner. Besides, there is a large decline after dropping the ground-truth summary regular-

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

Method	Pre	Rec	F1
Ours	43.02	48.73	**45.30**
w/o-$\mathcal{L}_{length}$	34.78	61.80	44.08
w/o-$\mathcal{L}_{summ}$	28.30	47.58	35.19
two-player	43.39	51.28	44.37

Table 3: Query length analysis on query-conditioned video summarization in terms of Precision (Pre), Recall (Rec) and F1-score (F1).

	w/o-$\mathcal{L}_{length}$				**Ours** (*w-*$\mathcal{L}_{length}$)			
	$d_{w/o-len}$	Pre	Rec	F1	d_{w-len}	Pre	Rec	F1
Vid1	50.72	45.42	57.09	47.45	**26.09**	49.66	50.91	**48.74**
Vid2	50.23	34.78	61.80	44.08	**11.59**	43.02	48.73	**45.30**
Vid3	31.15	48.50	63.59	52.98	**13.61**	58.73	56.49	**56.51**
Vid4	44.67	23.46	51.96	31.69	**17.98**	36.70	35.96	**33.64**
Avg.	40.88	40.19	55.98	44.12	**17.32**	47.03	48.02	**46.05**

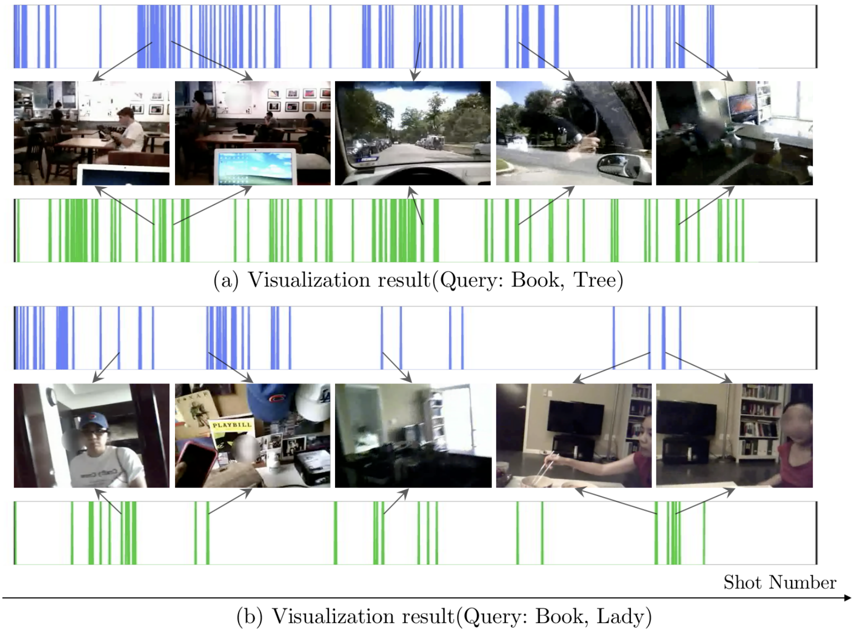
ization, which complies with the fact that additional supervised information tends to improve learning considerably.

We further conduct an experiment to more thoroughly analyze the ability of our proposed summary length regularization approach to generate summaries of suitable length. Here we define the summary length distance between the generated summary and the ground-truth summary as: $d_{w-len} = \left| \frac{1}{Q} \sum_{q=1}^{Q} \sum_{t=1}^{T} (k_{t,q} - s^{g}_{t,q}) \right|$, Q is the total number of queries. We use $k_{t,q}$ and $s^{g}_{t,q}$ to denote the summary result and the ground-truth summary given a certain query q with the length regularization $\mathcal{L}_{length}$. Similarly, the summary distance between the generated summary after dropping the length regularization and the ground-truth summary is defined as: $d_{w/o-len} = \left| \frac{1}{Q} \sum_{q=1}^{Q} \sum_{t=1}^{T} (k'_{t,q} - s^{g}_{t,q}) \right|$, $k'_{t,q}$ denotes the summary result given a certain query q after dropping the length regularization $\mathcal{L}_{length}$.

As shown in Table 3, the F1-score of the model without length regularization is reduced by 1.93% in average, compared with the result of our proposed framework, and the length distance increases from 17.32 to 40.88. This demonstrates the effect of the length regularization. Moreover, we can also observe that the differences between precision and recall values for $w/o - \mathcal{L}_{length}$ tend to be larger than the ones in our proposed approach. Note that the smaller the distance between precision and recall values is, the closer between the length of the summary and the length of the ground-truth is. This indicates the effects of our introduced query length regularization $\mathcal{L}_{length}$.

4.4 Qualitative Results

We provide some visualization results of our method in Figure. 3. We use the two user queries as examples: "Book Tree" and "Book Lady" (each user query contains two concepts). The x-axis in the figure is the shot number given a certain video. The upper blue lines denote the ground-truth key shots which are related to the user query, while the bottom



(a) Visualization result(Query: Book, Tree)

(b) Visualization result(Query: Book, Lady)

Figure 3: Some visualization results of our proposed method. The x-axis is the shot number given a certain video. The blue lines show the key shots of the ground-truths, and the green lines represent predicted key shots using our method. (a) The results for the query “Book Tree”. (b) The results for the query “Book Lady”.

green lines represent predicted key shots using our proposed method. Note that the selected summaries can be either related to one or two concepts given a user query. We can observe that our proposed method can find compact and representative summaries.

5 Conclusions

In this paper, we proposed a query-conditioned three-player generative adversarial network for query-conditioned video summarization. In the generator, video representations conditioned on user queries are obtained by jointly encoding visual information together with the text of user queries. Given these embeddings, confidence scores are predicted for each video shot in order to generate key shots based on these predicted scores. In the discriminator, we defined a three-player loss by introducing a randomly generated summary to prevent the model from generating trivial and short sequences. Experiments on videos of uncontrolled daily lives demonstrate the superiority of our proposed method.

Acknowledgment This project is supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center. This work is also partially funded by the National Natural Science Foundation of China (Grant No. 61673378 and 61333016), and Norwegian Research Council FRIPRO grant no. 239844 on developing the *Next Generation Learning Machines*.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[3] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2011.

[4] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. In *Advances in Neural Information Processing Systems*, pages 2069–2077, 2014.

[5] Prasoon Goyal, Zhiting Hu, Xiaodan Liang, Chenyu Wang, and Eric Xing. Nonparametric variational auto-encoders for hierarchical representation learning. *ICCV*, 2017.

[6] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

[7] Junwei Han, Le Yang, Dingwen Zhang, Xiaojun Chang, and Xiaodan Liang. Reinforcement cutting-agent learning for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9080–9089, 2018.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[9] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[10] Zhong Ji, Yaru Ma, Yanwei Pang, and Xuelong Li. Query-aware sparse coding for multi-video summarization. *arXiv preprint arXiv:1707.04021*, 2017.

[11] Aditya Khosla, Raffay Hamid, Chih-Jen Lin, and Neel Sundaresan. Large-scale video summarization using web-image priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2698–2705, 2013.

[12] Gunhee Kim, Leonid Sigal, and Eric P Xing. Joint summarization of large-scale collections of web images and videos for storyline reconstruction. 2014.

[13] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1346–1353. IEEE, 2012.

[14] Yen-Liang Lin, Vlad I Morariu, and Winston Hsu. Summarizing while recording: Context-based highlight detection for egocentric videos. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 51–59, 2015.

[15] Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2714–2721. IEEE, 2013.

[16] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[17] Jingjing Meng, Hongxing Wang, Junsong Yuan, and Yap-Peng Tan. From keyframes to key objects: Video summarization by representative object proposal selection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1039–1048, 2016.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[19] Harrie Oosterhuis, Sujith Ravi, and Michael Bendersky. Semantic video trailers. *arXiv preprint arXiv:1609.01819*, 2016.

[20] Bryan A Plummer, Matthew Brown, and Svetlana Lazebnik. Enhancing video summarization via vision-language embedding. In *Computer Vision and Pattern Recognition*, 2017.

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015.

[22] Aidean Sharghi, Boqing Gong, and Mubarak Shah. Query-focused extractive video summarization. In *European Conference on Computer Vision*, pages 3–19. Springer, 2016.

[23] Aidean Sharghi, Jacob S Laurel, and Boqing Gong. Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2127–2136. IEEE, 2017.

[24] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5179–5187, 2015.

[25] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448, 2015.

[26] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4489–4497, 2015.

[27] Arun Balajee Vasudevan, Michael Gygli, Anna Volokitin, and Luc Van Gool. Query-adaptive video summarization via quality-aware relevance estimation. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 582–590. ACM, 2017.

[28] Ting Yao, Tao Mei, and Yong Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 982–990, 2016.

[29] Yuan Yuan, Xiaodan Liang, Xiaolong Wang, Dit Yan Yeung, and Abhinav Gupta. Temporal dynamic graph lstm for action-driven video object detection. *ICCV*, 2017.

[30] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Summary transfer: Exemplar-based subset selection for video summarization. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 1059–1067. IEEE, 2016.

[31] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *European conference on computer vision*, pages 766–782. Springer, 2016.

[32] Yujia Zhang, Michael Kampffmeyer, Xiaodan Liang, Dingwen Zhang, Min Tan, and Eric P Xing. Dtr-gan: Dilated temporal relational adversarial network for video summarization. *arXiv preprint arXiv:1804.11228*, 2018.

[33] Yujia Zhang, Xiaodan Liang, Dingwen Zhang, Min Tan, and Eric P Xing. Unsupervised object-level video summarization with online motion auto-encoder. *arXiv preprint arXiv:1801.00543*, 2018.

[34] Kaiyang Zhou and Yu Qiao. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. *arXiv preprint arXiv:1801.00054*, 2017.