Watch Hours in Minutes: Summarizing Video with User Intent

Saiteja Nalla[1], Mohit Agrawal[1], Vishal Kaushal[1], Ganesh Ramakrishnan[1], and Rishabh Iyer[2]

[1] Indian Institute of Technology Bombay, Mumbai, India
{saitejan, mohitagr, vkaushal, ganesh}@cse.iitb.ac.in
[2] The University of Texas at Dallas, Texas, USA
rishabh.iyer@utdallas.edu

Abstract. With the ever increasing growth of videos, automatic video summarization has become an important task which has attracted lot of interest in the research community. One of the challenges which makes it a hard problem to solve is presence of multiple 'correct answers'. Because of the highly subjective nature of the task, there can be different "ideal" summaries of a video. Modelling user intent in the form of queries has been posed in literature as a way to alleviate this problem. The query-focused summary is expected to contain shots which are relevant to the query in conjunction with other important shots. For practical deployments in which very long videos need to be summarized, this need to capture user's intent becomes all the more pronounced. In this work, we propose a simple two stage method which takes user query and video as input and generates a query-focused summary. Specifically, in the first stage, we employ attention within a segment and across all segments, combined with the query to learn the feature representation of each shot. In the second stage, such learned features are again fused with the query to learn the score of each shot by regressing through fully connected layers. We then assemble the summary by arranging the top scoring shots in chronological order. Extensive experiments on a benchmark query-focused video summarization dataset for long videos give better results as compared to the current state of the art, thereby demonstrating the effectiveness of our method even without employing computationally expensive architectures like LSTMs, variational autoencoders, GANs or reinforcement learning, as done by most past works.

Keywords: Query-focused, Video Summarization, Attention, User-intent

1 Introduction

Videos have become an indispensable medium for capturing and conveying information. The increasing availability of cheaper and better video capturing and storage devices have led to the unprecedented growth in the amount of video data available today. Most of this data, however, comes with a lot of redundancy, partly because of the inherent nature of videos (as a set of *many* images) and

partly due to the 'capture-now-process-later' mentality. Consequently this has given rise to the need of automatic video summarization techniques which essentially aim at producing shorter videos without significantly compromising the

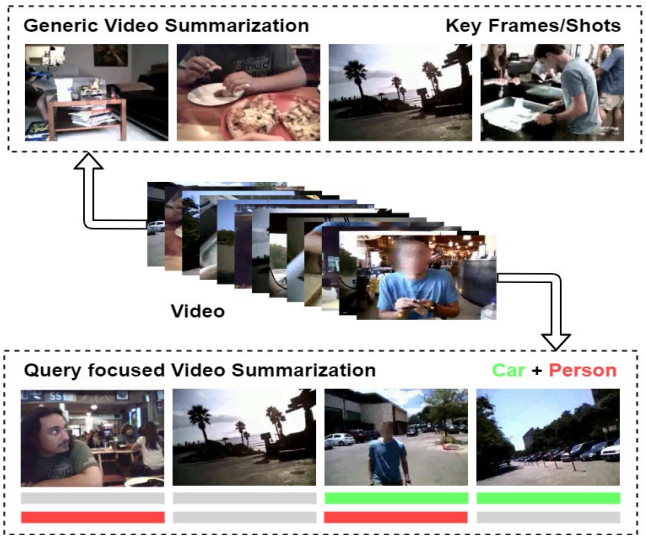

Fig. 1. Illustration of Generic vs Query focused Video Summarization for a given video

quality and quantity of information contained in them. A Video Summarization technique aims to select important, diverse (non-redundant) and representative frames (static video summarization) or shots (dynamic video summarization) from a video to enable quicker and easier consumption of information contained in the video. In this work we focus on producing summary as a sequence of shots (set of frames), i.e. dynamic video summarization. One of the characteristic challenges which make this problem hard to solve is the fact that there is no single correct answer (summary). Owing to the highly subjective nature of the task, summaries produced by different users tend to be different due to varying intents and perception. Researchers have looked at query-focused video summarization as a way to alleviate this problem. The user intent is taken as an additional input in the form of a query and the summary produced is influenced to contain more shots which are relevant to the query. The summary thus produced is semantically relevant to the query, modelling user intent and preferences. This is especially welcome in the real-world setting where very long videos need to be summarized.

Query-focused Video Summarization has attracted a lot of attention in the recent past. To the best of our knowledge, the first work in Query-focused Video Summarization was by Sharghi et. al. [34] which employed determinantal point processes [21]. This was followed by use of memory networks [35], submodular

mixtures [36, 27], adversarial networks [43] and attention [17, 40, 39] as different ways of computing the query relevance. Motivated by [17], [40] and [15] we propose a simple attention based two-stage method to address query-focused video summarization enhanced by query fusion. Let us say we are provided a video that is divided into segments of fixed size shots. In the first stage, local attention is employed to model the query agnostic importance of the shots in the segments and local attention features are learnt. In addition, for each segment, query relevant shots within a segment are identified to represent the aggregate segment level semantic features relevant to the query. These segment representatives are then combined with the visual features to learn the global attention features of all the shots, considering the query. To enhance the effectiveness of query-relevance further, in the second stage, local attention features along with the global attention features are fused with the query representation vector. This is followed by a regression through fully connected layers to obtain shot scores, indicative of the rank of shots with respect to the query. We then assemble the summary by arranging the top scoring shots in chronological order. For a fair comparison with recent techniques, we test our method on the benchmark dataset for query-focused video summarization [35] and demonstrate the effectiveness of our method both quantitatively and qualitatively.

In the following sections, we begin by talking about the related work in this area. In Section 3 we then describe our proposed method in details. This is followed by the details of the experiments and results in Section 4. We finally conclude by reporting our proposed method as a simple, yet effective improvement over the current state of the art Query-focused Video Summarization techniques as tested on the benchmark dataset.

2 Related Work

2.1 Generic Video Summarization

In terms of the generated output, broadly speaking, video summarization can be categorized as compositional video summarization, which aims at producing spatio-temporal synopsis or mosaic composed of more than one frames [30, 32, 29, 31] and extractive video summarization which aims at selecting key frames or key shots. Extractive video summarization with key frame selection is also often referred to as key frame extraction, static story board creation or static video summarization [5] while it is referred to as dynamic video summarization or dynamic video skimming [13]) in case of shots. In this work, we focus on dynamic extractive video summarization for single video summarization, as against multi-video summarization [26]. Video summarization, at least from what appears at the surface, boils down to identify important, representative portions of a video while eliminating redundancy. Early approaches were mainly unsupervised and summarized a video using low level cues [38, 24]. More advanced approaches looked into better indicators of 'important' portions of a video through presence of people or objects in egocentric videos and more recently, actionness [22, 14, 6]. First truly supervised approach, in terms of learning directly from a ground

truth summary, was presented by [12] who adapted determinantal point processes (DPP) [21] to videos. Motivated by the fact that video is a form of sequence data where LSTMs have demonstrated superior performance [42] was the first work to use LSTMs for video summarization. They also proposed an additional DPP layer on top to ensure diversity. Another body of work looks at using external clues as an aide to summarization [19, 4, 20, 46]. [3, 23] explicitly focus on enhancing the diversity and representativeness of the generated summary. [23] for example employed sequential DPP to learn the time span of a video segment upon which the local diversity is imposed to guarantee the diversity of a long video. The absence of a large annotated dataset and the fact that there are multiple ground truth summaries possible for a video, has led to a recent rise of unsupervised techniques [25, 45, 18, 41, 1]. [45] was the first to apply reinforcement learning to unsupervised video summarization motivated by the fact that reward is available only at the end of the sequence. [25] used a generative adversarial framework, consisting of the summarizer and discriminator in an unsupervised setting to achieve comparable performance to supervised techniques. There is also a lot of recent work combining adversarial and attention based networks with an aim to produce better video summaries [25, 7, 16, 8, 41, 1]. [16] was the first to use attentive encoder decoder based network to video summarization.

2.2 Query-Focused Video Summarization

SeqDPP introduced in [12] was used to model the problem of video summarization as diverse sequential subset selection. Based on this idea, in [34] Sharghi et al proposed sequential hierarchical DPP (SH-DPP) where the first layer modeled query relevance and the second layer modeled importance conditioned on first. Diversity was naturally modelled by DPP [21]. In [4], topic-based summary is generated by finding shots which co-occur mostly across videos collected using the given topic and a MBF (Maximal Biclique Finding) algorithm is optimized to find sparsely co-occuring pattern. In [35] Sharghi et al introduced QC-DPP (Query Conditioned DPP) where a memory network was used to model query importance as well as contextual importance of a shot. This is then fed into the seqDPP. They also, for the first time, introduced a dataset specifically prepared for the task of Query-focused Video Summarization. They also introduced a new evaluation metric which focuses on the semantic relationship between the shots in predicted and ground truth summary. This has emerged as a benchmark dataset for Query-focused Video Summarization with several recent techniques reporting their results on it. In [36] Vasudevan *et. al.*, model Query-focused Video Summarization as a subset selection problem where the best subset is found by maximizing a mixture of different submodular terms which capture (i) query similarity between frame and query in a common semantic embedding space, (i) quality score, (iii) diversity and (iv) representativeness. A similar approach is adopted by [27] where they demonstrate the importance of using joint vision-langauge embedding in addition to visual features. [43] use adversarial networks where the generator learns the joint representation of the user

query and the video content, and the discriminator takes three pairs of query-conditioned summaries (generator, ground truth and random) as the input to discriminate the real summary from a generated and a random one, trained via a three-player loss. In [17] Jiang et al use a query-focused attention module to combine the semantic information of the query and a multilevel self-attention variational block to obtain context-important information and add user-oriented diversity and stochasticity. Reinforcement Learning is used in [44] to target this problem where a Mapping Network (MapNet) is used to map video shot and query in same space and after that a deep RL-based summarization network is used to provide query based summary by including parameters like relatedness, representativeness and diversity as rewards. Xiao et al in [40] employ a hierarchical attention network and demonstrate the effectiveness of local and global attention. In [39], Xiao et al extended their work [40] and used a pre-trained RL caption generator to generate captions for the video shots for textual information and along with semantic information generated from self-attentive module which helped to decide the important shots and then use a query-aware scoring module to generate query-focused summary. Huang et al [15] addressed query-focused video summarization by learning the query relevance scores using a combination of visual features and a vector representation of input query.

3 Proposed Method

3.1 Problem Formulation

The objective of the Query-focused Video Summarization is to output a video summary which is a sequence of diverse, representative and query relevant video shots given a long video and a query. We extract visual and textual features from shots, compute shot scores using the proposed method and finally construct the summary based on the shot scores. We denote a video as a sequence of non-overlapping shots of fixed length. Let there be n shots in the video denoted by $\{s_1, s_2, \cdots, s_n\}$ These shots are further grouped together into fixed-sized non overlapping segments. This can easily be extended to using variable sized segments formed using Kernel Temporal Segmentation [28] or other alternate techniques for shot detection. We use the lexicon of concepts constructed by [35] and represent the textual query t_q as a collection of two concepts $\{c_1, c_2\}$. Each concept is a noun like 'SKY', 'LADY', 'FLOWER', 'COMPUTER' etc. More complex queries can easily be supported by using appropriate embedding for them as in some of the video localization works [9, 33, 10, 37] or following the approach in [27]. We leave that to future work in this area. For a textual query t_q shot scores are calculated corresponding to visual features of each shot to construct a query-focused video summary.

3.2 Feature Embedding

Motivated by the success of I3D features [2] in better modelling of temporal resolution and in capturing long spatial and temporal dependencies, which is

especially important for long videos, we extract p dimensional I3D features for every shot in the video. These features are reduced from p to d dimensions using a fully connected layer. The output of the fully connected layer corresponding to the extracted visual features $\{v_1^{'}, v_2^{'}, \cdots, v_n^{'}\}$ is $\{v_1, v_2, \cdots, v_n\}$, $v_i \in R^d$ is referred to as 'visual features' here after in this paper. For representing queries, we use one hot feature encoding of the concepts to form a query representation vector of 48 dimensions (corresponding to the 48 concepts introduced in [35]). The query representation vector is passed through another fully connected layer and the resultant d dimensional features are referred to as 'textual features' $f_q \in R^d$ here after in this paper. In order to generate the video summary of a long video, it is important to look at both local context as well as global context in determining query-relevance and importance of shots. To facilitate this, we define fixed-size windows, called segments, which are non-overlapping groups of shots. For local context, we consider shots within a segment and for global context we consider query-relevance and importance across segments. Visual features are used for the computation of local attention feature vectors within a segment in the Local Attention Module (LAM). Visual features along with textual features are used to compute query relevant segment representatives in the Query Relevant Segment Representation Module (QSRM). Query relevant features along with the local features are used to compute the global attention features across the segments in the Global Attention Module(GAM) which are further regressed to compute the shot scores. We illustrate this end to end pipeline in 2. In what follows, we give details of each of these modules.

3.3 Local Attention Module

The Local Attention Module (LAM) computes the attentive features of the shots within a segment for all segments of a video. It captures the semantic relations among the shots within a segment. We represent the visual features corresponding to shots $\{s_1, s_2, \cdots, s_n\}$ as $\{v_1, v_2, \cdots, v_n\}$ respectively. LAM takes these visual features as input and outputs attention vectors corresponding to each shot. Semantic similarity matrix is calculated from the visual features corresponding to each segment. Semantic similarity score of shots (v_i, v_j) i.e., $(i, j)^{th}$ element of semantic similarity matrix for a segment is computed as

$$\phi(v_i, v_j) = \mathbf{Z_l} \tanh(\mathbf{W_1^l} v_i + \mathbf{W_2^l} v_j + \mathbf{b}) \tag{1}$$

$W_1^l, W_2^l, Z_l \in R^{d \times d}$ and $b \in R^d$ are parameters to be learnt. The semantic score vector $\phi(v_i, v_j) \in R^d$ and the shape of semantic similarity score matrix Φ is $k \times k \times d$, where k is the segment length. Semantic similarity matrix captures the semantic relations between the shots within a segment (also representing the temporal features) and this is done for all the segments. The semantic relations of one segment are interacted with other segments temporally by sharing the trainable parameters across the segments and thereby reducing the number of trainable parameters in the model. Softmax interactions within a segment are then calculated from the semantic similarity matrix and the values are computed corresponding to visual features (v_i, v_j) as,

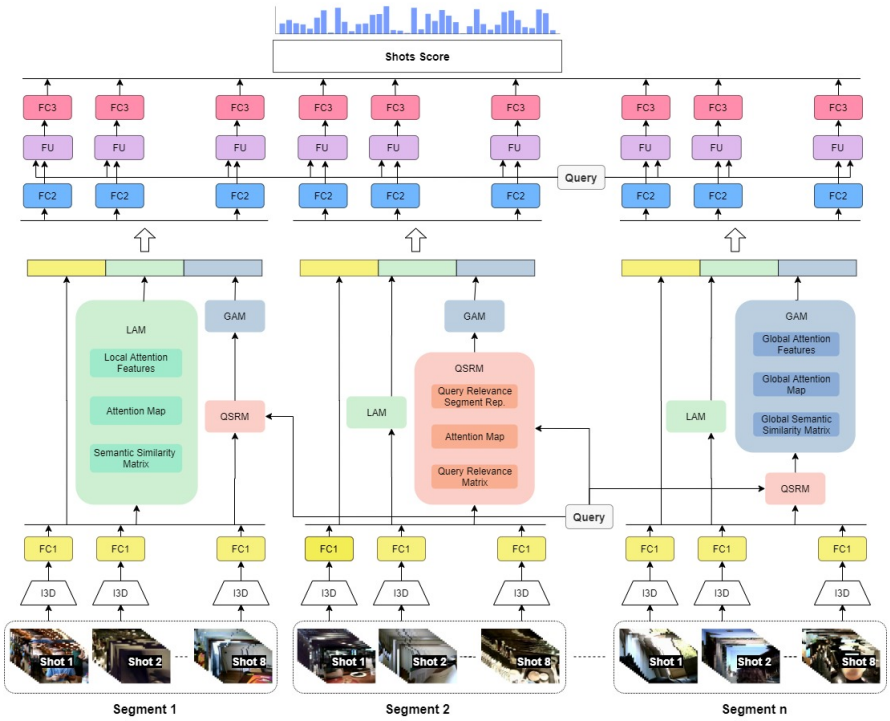

Fig. 2. Architecture of the proposed method. We split a video into non-overlapping shots and group them into non-overlapping segments. We extract the visual features using a pretrained model and compute the local attention features in LAM using visual features. We encode the textual query and compute the segment representatives in QSRM using textual and visual features. We pass the outputs of QSRM and visual features to GAM to compute the global attention vectors. Local and global attention vectors are concatenated with visual features, down-sampled and are passed to the Fusion Unit followed by FC layer to generate the shot scores which are then used to construct the summary

$$\chi_{ij} = \frac{\exp(\phi(v_i, v_j))}{\sum_{t=0}^{k} \exp(\phi(v_i, v_t))} \tag{2}$$

The shape of the softmax interaction matrix is $k \times k \times d$ corresponding to every combination of the shots with the segment.

The local attentive features for i^{th} shot are then calculated from the softmax interactions and visual features as,

$$v_i^l = \sum_{j=0}^{k} \chi_{ij} v_j \tag{3}$$

The shape of the local attention feature matrix of a segment is $k \times d$ corresponding to every shot of the segment.

3.4 Query relevant Segment Representation Module

Query relevant Segment Representation Module (QSRM) captures the semantic relations of the shots with the textual query and outputs the query-relevant representative features for each segment which are further used in the computation of global attention vectors in the GAM. Visual features $\{v_1, v_2, \cdots, v_n\}$ along with the textual features of a query are fed as input to the QSRM. Query relevant semantic scores are calculated as,

$$r_i = \mathbf{Z_g} \tanh(\mathbf{W_1^g} v_i + \mathbf{W_2^g} f_q + \mathbf{b})) \tag{4}$$

$W_1^g, W_2^g, Z_g \in R^{d \times d}$ and $b \in R^d$ are trainable parameters. Query relevant semantic matrix is of shape $k \times d$ for each segment. Query relevant softmax interactions are then computed from query relevant semantic scores as,

$$\chi_i = \frac{\exp(r_i)}{\sum_{t=0}^{k} \exp(r_t)} \tag{5}$$

Query relevant softmax interaction matrix is of shape $k \times d$ corresponding to every shot of a segment.

Query relevant segment representations for a segment are computed from the visual features $\{v_1, v_2, ..v_n\}$ of the shots and their corresponding query-relevant softmax interactions as,

$$v^{(s)} = \sum_{i=0}^{k} \chi_i v_i \tag{6}$$

Query relevant representation vectors for m segments are represented as $\{v_1^{(s)}, v_2^{(s)}, ..v_m^{(s)}\}$ and have a shape of $m \times d$ representing the aggregated attention representations for a query.

3.5 Global Attention Module

Global attention features are computed from the visual features and query relevant segment representations. These capture semantic interactions between the intra segment semantic features and query-relevant inter segment attention features. Global semantic similarity scores are computed from the visual features $\{v_1, v_2, \cdots, v_n\}$ and query relevant semantic representation features $\{v_1^{(s)}, v_2^{(s)}, \cdots, v_m^{(s)}\}$ as

$$r_j^g = \mathbf{Z_g} \tanh(\mathbf{W_1^g} v_i + \mathbf{W_2^g} v_j^{(s)} + \mathbf{b}) \tag{7}$$

Global semantic similarity score matrix has a shape of $n \times d$. $W_1^g, W_2^g, Z_g \in R^{d \times d}$ and $b \in R^d$ are trainable parameters. The softmax interaction scores are calculated from the semantic segment scores as,

$$\chi_j^g = \frac{\exp(r_j^g)}{\sum_{k=0}^{m} \exp(r_k^g)} \tag{8}$$

The shape of the softmax interaction matrix is $n \times d$. The global attentive features are computed from the softmax interaction scores and query relevant segment representatives as,

$$v_i^g = \sum_{j=0}^{m} \chi_j^g v_j^{(s)} \tag{9}$$

v^g is the global attention vector corresponding to each shot which captures the query relevant global semantic attention features of all the segments in a video. Shape of v^g is $n \times d$

3.6 Fusion Unit

To better learn the shot scores, we make use of visual features v_i, local attention vectors v_i^l and global attention vectors v_i^g for each shot, these features are concatenated to form a single shot feature vector $v_i^{c'}$, $v_i^{c'} = [v_i, v_i^l, v_i^g], v_i^{c'} \in R^{3d}$. The concatenated feature vector $v_i^{c'}$ is reduced to d dimensions using a fully connected layer and the output of the fully connected layer is represented as v_i^c. To enhance the effectiveness of query-relevance with respect to the given query, the condensed features vector v_i^c along with the textual features $f_q \in R^d$ of the query t_q are fed as inputs to the Fusion Unit. The Fusion Unit (FU) aggregates the features by performing point wise additions, multiplications and concatenating the features and outputs a feature vector $v_i^f \in R^{4d}$. These features are used to finally predict the shot scores through a fully connected layer and the top ranked shots are used as predicted selections.

We use Adam optimizer to train the model end-to-end based on the Binary Cross Entropy (BCE) loss between the predicted shots and ground truth shots.

4 Experiments and Results

4.1 Dataset

For a fair comparison with other techniques, we evaluate our model's performance on the benchmarking dataset introduced by [35] which was built upon UTE dataset [22]. It contains four egocentric consumer grade videos captured in uncontrolled everyday scenarios and each video is 3 to 5 hours long containing a diverse set of events. A set of 48 concepts is defined by [35] and every query is made up of two concepts. The queries are so defined by [35] as to cover four different scenarios: 1) all concepts in the query appear in the same video shots together 2) all concepts appear in video occur but never jointly in a shot 3) only one of the concepts in the query appears in some shots and 4) none of the concept in the query are present in the video. We follow the same convention in our work. The dataset provides four ground truth query-focused summaries for each video and query pair, 1 oracle summary and 3 user summaries.

4.2 End to End Pipeline

Preprocessing The videos in the UTE dataset are divided into non-overlapping shots of 5 seconds each. We sample the frames at 3 fps and compute features of each shot (15 frames per shot) as follows: we use a pre trained I3D model as a feature extractor. For each shot, 15 frames each of size 224 x 224 is given as input to the I3D Model and it generates a 512 dimension feature vector (output from the temporal layer in the I3D Model) to be further used in our pipeline. We define segments as non-overlapping groups of 8 shots. As far as queries are concerned, we deal with bi-concept queries. There are a total of 48 concepts defined in the dataset. We represent the query by the addition of the one-hot vectors corresponding to each concept resulting in a 48-dimensional vector which is then used in our pipeline.

Training We performed 4 experiments by using one video for test and rest for training and validation in turn. We train the model for 25 epochs. In each epoch, for each training video, I3D features are passed through a FC layer to create a 300-dimensional visual feature representation for each shot. These features are also used to create the local and global representation for each shot. The visual feature of each shot of a segment is given as input to the LAM (Local Attention Module). This module uses attention (refer section 3.3 in the main paper) to generate a 300-dimensional local feature representation of each shot of the segment. Since these are non-overlapping segments, we improve the efficiency by doing this in parallel. The visual feature of each shot of a segment is also given as input to QSRM (Query relevant Segment Representative Module) in parallel along with the textual feature of the query. QSRM uses attention (refer section 3.4 in the main paper) to generate a 300-dimensional segment representative vector. Again, since we have non overlapping segments, QSRM operates on all segments in parallel. For each segment, visual features of the shots are given as input to the GAM (Global Attention Module) which uses all the 300-dimensional segment representatives to generate a global representation of the shots using attention (refer section 3.5 in the main paper). The visual features, local and global attention vector for each shot thus generated are concatenated together to generate a 900-dimensional representation which is passed through a fully connected layer to create a 300-dimensional embedding. This 300-dimensional embedding is fused with the 300-dimensional textual feature vector of the query using Fusion Unit (refer section 3.6 in the main paper) which performs concatenation, pointwise addition and multiplication to generate a 1200-dimensional representation for each shot. This is regressed through a fully-connected layer followed by a sigmoid function to generate a score between 0 and 1 (inclusive) which is the predicted shot importance. We initialize all learnable parameters with Xavier Uniform initialization and learn the parameters end-to-end by Adam optimizer with a learning rate of 1e04 and weight decay of 1e01 using Binary Cross Entropy loss.

Inference Inference involves taking a video and a query as input and generate a query-focussed summary for this video. We generate the visual features, local and global attention vector for each shot of the test video as described above along with the textual features of the input query. The trained model predicts the scores for each shot. The top ranking shots based on a threshold (empirically found using validation video) are assembled chronologically to construct the query-focussed summary.

4.3 Evaluation

Authors in [35] have also defined an evaluation metric which first finds a mapping between the ground truth shots and the generated summary shots by doing the maximum weight matching on a bipartite graph where weights are based on intersection-over-union (IoU) between the shots using the dense concept annotations of the shots provided in the dataset. This notion of distance or similarity takes the semantics into account and has been shown to be better than matching in visual domain or matching based on shot numbers [35]. Standard Precision, Recall and F1 scores are than calculated based on the number of matches.

	Video 1			Video 2			Video 3			Video 4			Average		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
SeqDPP [12]	53.43	29.81	36.59	44.05	46.65	43.67	49.25	17.44	25.26	11.14	63.49	18.15	39.47	39.35	30.92
SH-DPP [34]	50.56	29.64	35.67	42.13	46.81	42.72	51.92	29.24	36.51	11.51	62.88	18.62	39.03	42.14	33.38
QC-DPP [35]	49.86	**53.38**	48.68	33.71	62.09	41.66	55.16	62.40	56.47	21.39	**63.12**	29.96	40.03	**60.25**	44.19
TPAN [43]	49.66	50.91	48.74	43.02	48.73	45.30	58.73	56.49	56.51	36.70	35.96	33.64	47.03	48.02	46.05
CHAN [40]	54.73	46.57	49.14	45.92	50.26	46.53	59.75	**64.53**	58.65	25.23	51.16	33.42	46.40	53.13	46.94
HVN [17]	52.55	52.91	**51.45**	38.66	**62.70**	47.49	**60.28**	62.58	**61.08**	26.79	54.21	35.47	44.57	58.10	48.87
QSAN [39]	48.41	52.34	48.52	46.51	51.36	46.64	56.78	61.14	56.93	30.54	46.90	34.25	45.56	52.94	46.59
Ours	**54.58**	52.51	50.96	**48.12**	52.15	**48.28**	58.48	61.66	58.41	**37.40**	43.90	**39.18**	**49.64**	52.55	**49.20**

Table 1. Quantitative results comparing our method against some existing Query-focused Video Summarization techniques

4.4 Implementation Details

As defined by the UTE dataset used by [35] each shot is 5 seconds long. In this work we use fixed size segments, and empirically chose the size of each segment to be 8 shots. Since the dataset contains long ego-centric videos which do not contain fast changing events, this choice is neither too small (hence retaining sufficient local context) nor too big (hence not missing out on event changes). We leave out one video for testing and use remaining 3 videos for training and validation. We report the results when each video is used as a test video, retaining the remaining for train and validation and we also report the average performance over the four experiments. We use MLP with fully connected layers to increase and reduce the features dimensionality as required by the architecture. In each MLP, we use 3 fully connected (FC) layers and all the FC layers has 300 hidden units followed by ReLU activation functions. We initialized the

weights using Xavier Uniform initialization [11], used Adam Optimizer with a learning rate $1e-04$ and weight decay $1e-01$. We use binary cross entropy loss to train the model. We compare the results of our method with some of the existing methods for Query-focused Video Summarization that have reported their results on the UTE benchmark dataset. Specifically, we chose SeqDPP [12], SH-DPP [34], QC-DPP [35], TPAN [43], CHAN [40], HVN [17] and QSAN [39] for comparison.

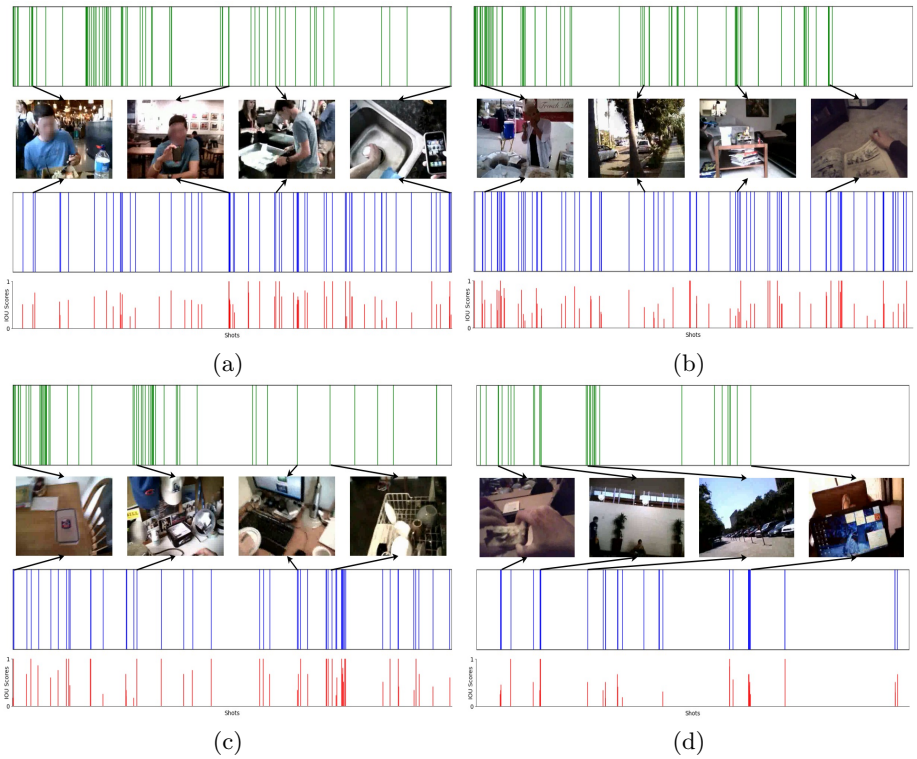

Fig. 3. Qualitative Analysis of summaries generated using our method. In each subfigure, the x-axis represents the video shot numbers, the green lines represents the ground truth shots, blue lines represents the predicted shots for a query and red bars shows the IoU score between predicted shot and the matching ground truth shot. Subfigure (a), (b), (c) and (d) correspond to video 1, 2, 3 and 4 and queries (Face and Phone), (Book and Garden), (Cupglass and Desk) and (Car and Food) respectively

4.5 Results and Analysis

Table 1 shows the Precision, Recall and F1 score of our method as compared to other methods. On an average across all four videos and in three out of four

videos, our method scores higher than all other methods on Precision without significant compromise in Recall. With regards to F1, which considers both precision and recall, we perform better than all techniques on an average. This is because of the use of better temporal and spatial representation using I3D features and enhancing the effect of query in selecting shots by fusing the textual features of query with the local and global attentive visual features.

In Figure 3(a), (b), (c) and (d) we plot the shots selected by machine generated summary against the ground truth shot selections for queries (Face and Phone), (Book and Garden), (Cupglass and Desk) and (Car and Food) on video numbers 1,2,3, and 4 respectively (video index same as in the dataset). We observe following cases - i) for some shots, there are exact matches (as can be seen by the matching green and blue lines in each sub-figure and through sample frame visualizations), ii) there are some shots in ground truth which are not in our summary, and iii) there are some shots in our summary which are not in ground truth. With regards to ii) and iii), it is important to note, that since the evaluation is based on matching using semantic similarity and not using shot numbers, exact match based on shot numbers is not expected. As long as there is a semantic match between the generated summary shots and ground truth shots, the generated summary is still considered good. To validate this, we generated one more visualization. We plot the IoU values between the predicted shots and matching ground truth shot. We see that even those shots which are not in ground truth, have a considerably high value of IoU (> 0.5) with a matching ground truth shot. With regards to IoU values, it may be noted that when the shot numbers exactly match, the IoU need not be 1. This is because of the maximum weight bipartite matching algorithm which is not greedy. For example, in the case of the first visualized frame in 3(a) though it is in both ground truth as well as prediction, the IoU is seen to be less than 1.

4.6 Analysis of Model Complexity

Extensive experiments on a benchmark query-focused video summarization dataset for long videos give better results as compared to the current state of the art, thereby demonstrating the effectiveness of our method even without employing computationally expensive architectures like LSTMs, variational autoencoders, GANs or reinforcement learning, as done by most past works. To better understand the simplicity of our model as compared to some of the previous methods, we estimate a rough lower bound of the number of learnable parameters used in those methods based on the information published in the respective works. Wherever details are not mentioned, we have made assumptions, if required. Methods presented in [39], [44], [43] and [17] use LSTMs or BiLSTMs as a key component in their architecture with input dimensions ranging from 512 to 4096 and number of hidden units in these LSTMs/BiLSTMs ranging from 512 to 1024. This makes the number of learnable parameters in these models to be greater than $\sim$1e7. On the other hand the number of parameters in our proposed method is of the order of $\sim$1e5. In our method we have 3 fully connected layers with the maximum number of parameters for one of them being 900(input)$\times$300(output).

The local and global attention weight matrices are 300×300 each. In addition, for mapping the features from one dimensionality to another, we have 4 weight matrices of max dimensionality for one of them being 512×300. Hence the total number of learnable parameters in our method is ∼1e6 is less than the lower bound estimate of other methods by an order of magnitude.

5 Conclusion

Query-focused video summarization is an important step forward in addressing the challenges associated with automatic video summarization. Past work has employed DPPs, memory networks, adversarial networks, submodular mixtures and attention networks in coming up with better techniques. In this work we proposed a simple architecture based on attention networks and query fusion and used I3D features to further the sate of the art. Extensive quantitative and qualitative evaluation of our method on the currently available benchmark data set of long videos especially made for this task establishes the effectiveness of our method.

References

1. Apostolidis, E., Adamantidou, E., Metsai, A.I., Mezaris, V., Patras, I.: Unsupervised video summarization via attention-driven adversarial learning. In: International Conference on Multimedia Modeling. pp. 492–504. Springer (2020)
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
3. Chen, X., Li, X., Lu, X.: Representative and diverse video summarization. In: Signal and Information Processing (ChinaSIP), 2015 IEEE China Summit and International Conference on. pp. 142–146. IEEE (2015)
4. Chu, W.S., Song, Y., Jaimes, A.: Video co-summarization: Video summarization by visual co-occurrence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3584–3592 (2015)
5. De Avila, S.E.F., Lopes, A.P.B., da Luz Jr, A., de Albuquerque Araújo, A.: Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. Pattern Recognition Letters **32**(1), 56–68 (2011)
6. Elfeki, M., Borji, A.: Video summarization via actionness ranking. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 754–763. IEEE (2019)
7. Fajtl, J., Sokeh, H.S., Argyriou, V., Monekosso, D., Remagnino, P.: Summarizing videos with attention. In: Asian Conference on Computer Vision. pp. 39–54. Springer (2018)
8. Fu, T.J., Tai, S.H., Chen, H.T.: Attentive and adversarial learning for video summarization. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1579–1587. IEEE (2019)
9. Gao, J., Sun, C., Yang, Z., Nevatia, R.: Tall: Temporal activity localization via language query. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5267–5275 (2017)

10. Ge, R., Gao, J., Chen, K., Nevatia, R.: Mac: Mining activity concepts for language-based temporal localization. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 245–253. IEEE (2019)
11. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 249–256 (2010)
12. Gong, B., Chao, W.L., Grauman, K., Sha, F.: Diverse sequential subset selection for supervised video summarization. In: Advances in neural information processing systems. pp. 2069–2077 (2014)
13. Gygli, M., Grabner, H., Riemenschneider, H., Van Gool, L.: Creating summaries from user videos. In: ECCV (2014)
14. Gygli, M., Grabner, H., Riemenschneider, H., Van Gool, L.: Creating summaries from user videos. In: European conference on computer vision. pp. 505–520. Springer (2014)
15. Huang, J.H., Worring, M.: Query-controllable video summarization. arXiv preprint arXiv:2004.03661 (2020)
16. Ji, Z., Xiong, K., Pang, Y., Li, X.: Video summarization with attention-based encoder-decoder networks. IEEE Transactions on Circuits and Systems for Video Technology (2019)
17. Jiang, P., Han, Y.: Hierarchical variational network for user-diversified & query-focused video summarization. In: Proceedings of the 2019 on International Conference on Multimedia Retrieval. pp. 202–206 (2019)
18. Jung, Y., Cho, D., Kim, D., Woo, S., Kweon, I.S.: Discriminative feature learning for unsupervised video summarization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8537–8544 (2019)
19. Khosla, A., Hamid, R., Lin, C.J., Sundaresan, N.: Large-scale video summarization using web-image priors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2698–2705 (2013)
20. Kim, G., Sigal, L., Xing, E.P.: Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4225–4232 (2014)
21. Kulesza, A., Taskar, B., et al.: Determinantal point processes for machine learning. Foundations and Trends® in Machine Learning **5**(2–3), 123–286 (2012)
22. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 1346–1353. IEEE (2012)
23. Li, Y., Wang, L., Yang, T., Gong, B.: How local is the local diversity? reinforcing sequential determinantal point processes with dynamic ground sets for supervised video summarization. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 151–167 (2018)
24. Ma, Y.F., Lu, L., Zhang, H.J., Li, M.: A user attention model for video summarization. In: Proceedings of the tenth ACM international conference on Multimedia. pp. 533–542. ACM (2002)
25. Mahasseni, B., Lam, M., Todorovic, S.: Unsupervised video summarization with adversarial lstm networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
26. Panda, R., Mithun, N.C., Roy-Chowdhury, A.K.: Diversity-aware multi-video summarization. IEEE Transactions on Image Processing **26**(10), 4712–4724 (2017)
27. Plummer, B.A., Brown, M., Lazebnik, S.: Enhancing video summarization via vision-language embedding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5781–5789 (2017)

28. Potapov, D., Douze, M., Harchaoui, Z., Schmid, C.: Category-specific video summarization. In: European conference on computer vision. pp. 540–555. Springer (2014)
29. Pritch, Y., Ratovitch, S., Hendel, A., Peleg, S.: Clustered synopsis of surveillance video. In: 2009 Advanced Video and Signal Based Surveillance. pp. 195–200. IEEE (2009)
30. Pritch, Y., Rav-Acha, A., Gutman, A., Peleg, S.: Webcam synopsis: Peeking around the world. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. pp. 1–8. IEEE (2007)
31. Pritch, Y., Rav-Acha, A., Peleg, S.: Nonchronological video synopsis and indexing. IEEE transactions on pattern analysis and machine intelligence **30**(11), 1971–1984 (2008)
32. Rav-Acha, A., Pritch, Y., Peleg, S.: Making a long video short: Dynamic video synopsis. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. vol. 1, pp. 435–441. IEEE (2006)
33. Shao, D., Xiong, Y., Zhao, Y., Huang, Q., Qiao, Y., Lin, D.: Find and focus: Retrieve and localize video events with natural language queries. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 200–216 (2018)
34. Sharghi, A., Gong, B., Shah, M.: Query-focused extractive video summarization. In: European Conference on Computer Vision. pp. 3–19. Springer (2016)
35. Sharghi, A., Laurel, J.S., Gong, B.: Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4788–4797 (2017)
36. Vasudevan, A.B., Gygli, M., Volokitin, A., Van Gool, L.: Query-adaptive video summarization via quality-aware relevance estimation. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 582–590 (2017)
37. Wang, W., Huang, Y., Wang, L.: Language-driven temporal activity localization: A semantic matching reinforcement learning model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 334–343 (2019)
38. Wolf, W.: Key frame selection by motion analysis. In: Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on. vol. 2, pp. 1228–1231. IEEE (1996)
39. Xiao, S., Zhao, Z., Zhang, Z., Guan, Z., Cai, D.: Query-biased self-attentive network for query-focused video summarization. IEEE Transactions on Image Processing **29**, 5889–5899 (2020)
40. Xiao, S., Zhao, Z., Zhang, Z., Yan, X., Yang, M.: Convolutional hierarchical attention network for query-focused video summarization. arXiv preprint arXiv:2002.03740 (2020)
41. Yuan, L., Tay, F.E., Li, P., Zhou, L., Feng, J.: Cycle-sum: Cycle-consistent adversarial lstm networks for unsupervised video summarization. arXiv preprint arXiv:1904.08265 (2019)
42. Zhang, K., Chao, W.L., Sha, F., Grauman, K.: Video summarization with long short-term memory. In: European Conference on Computer Vision. pp. 766–782. Springer (2016)
43. Zhang, Y., Kampffmeyer, M., Liang, X., Tan, M., Xing, E.P.: Query-conditioned three-player adversarial network for video summarization. arXiv preprint arXiv:1807.06677 (2018)
44. Zhang, Y., Kampffmeyer, M., Zhao, X., Tan, M.: Deep reinforcement learning for query-conditioned video summarization. Applied Sciences **9**(4), 750 (2019)

45. Zhou, K., Qiao, Y., Xiang, T.: Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
46. Zhu, X., Loy, C.C., Gong, S.: Learning from multiple sources for video summarisation. International Journal of Computer Vision **117**(3), 247–268 (2016)