applied sciences

MDPI

Article

Deep Reinforcement Learning for Query-Conditioned Video Summarization

Yujia Zhang [1,2,*], Michael Kampffmeyer [3], Xiaoguang Zhao [1,2] and Min Tan [1,2]

1 Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; xiaoguang.zhao@ia.ac.cn (X.Z.); min.tan@ia.ac.cn (M.T.)
2 University of Chinese Academy of Sciences, Beijing 100190, China
3 Machine Learning Group, UiT The Arctic University of Norway, Tromsø 9019, Norway; michael.c.kampffmeyer@uit.no
* Correspondence: zhangyujia2014@ia.ac.cn

Received: 2 January 2019; Accepted: 18 February 2019; Published: 21 February 2019

Abstract: Query-conditioned video summarization requires to (1) find a diverse set of video shots/frames that are representative for the whole video, and that (2) the selected shots/frames are related to a given query. Thus it can be tailored to different user interests leading to a better personalized summary and differs from the generic video summarization which only focuses on video content. Our work targets this query-conditioned video summarization task, by first proposing a Mapping Network (MapNet) in order to express how related a shot is to a given query. MapNet helps establish the relation between the two different modalities (videos and query), which allows mapping of visual information to query space. After that, a deep reinforcement learning-based summarization network (SummNet) is developed to provide personalized summaries by integrating relatedness, representativeness and diversity rewards. These rewards jointly guide the agent to select the most representative and diversity video shots that are most related to the user query. Experimental results on a query-conditioned video summarization benchmark demonstrate the effectiveness of our proposed method, indicating the usefulness of the proposed mapping mechanism as well as the reinforcement learning approach.

Keywords: query-conditioned video summarization; deep reinforcement learning; visual-text embedding; temporal modeling; vision application

1. Introduction

It is estimated that by 2021 it will take a person about 5 million years to watch all the videos that are uploaded each month [1]. In order to process the enormous amount of videos that emerge every day, a tool is required which can significantly reduce the cumbersomeness for digesting the ever-increasing amount of videos. Video summarization helps to address this problem and aims to automatically select a small subset of the frames/shots that captures the most interesting parts in a concise manner [2–7]. Thus it reduces the time and cost required to analyze video information, and provides significant advantages for efficient video browsing, searching and understanding, and further enhances various down-streaming applications such as video-based question answering [8], robot learning [9] and surveillance data analysis [10,11].

Different from traditional video summarization which only focuses on video content, query-conditioned video summarization is tasked to generate user-oriented summaries conditioned on a given query in the form of text [4,12–15]. As shown in Figure 1, different user queries for the same video will have different summary results, so that the summary can be tailored to different user interests, leading to a better personalized summary. Given a user query that reveals a specific user

interest, query-conditioned video summarization focuses on predicting summaries that are (1) in close correspondence with the semantic meaning of the user query; (2) most representative of the original video, producing a minimal and compact set of video frames/shots. Therefore, this task can be viewed as a step towards personalized video summarization [13]. It also enables more accurate evaluation by addressing the issue that generic video summarization is a highly subjective task where users may have different video-content preferences. Moreover, it enables users to more efficiently search video contents via text with multiple key words instead of only searching for video titles [14].

Figure 1. The illustration of the query-conditioned video summarization task. Given a user query, query-conditioned video summarization aims to predict a summary that is relevant to the query in a concise manner. Summary 1 and Summary 2 are two summarization results for the same video but for different user queries.

Queries can be supplied in the form of text and can be either storyline-based [16] or keyword-based. The storyline-based summarization is proposed in order to retrieve video results based on a query that can be represented as a graph. Xiong et al. [16] generate a storyline representation consisting of the following four story elements along a time-line: *{Actors, Location, Supporting objects, Events}*. However their method assumes prior knowledge for the list of locations and events in order to train location/event specific classifiers based on web-images, which can be challenging in many situations. Keyword-based summarization requires fewer fine-grained annotations and has been explored in several previous works. In [17,18], the authors investigated an attention-based personalized summarization technique tailored to Cultural Heritage scenarios by looking for scenes that are similar to web images that match user preferences. However, it is tailored to a cultural heritage scenario, and may not generalize to wider scenarios. More recently, Sharghi et al. [13] propose a more generalized query-conditioned summarization task including datasets and evaluation metrics. They collect a set of concepts containing representative semantic information for a wide range of commonly used terms such as specific objects, people, and fine-grained entities. The annotation in each video shot is either 0 or 1 indicating the absence/presence of the shot in the summary for a given query. Thus, for each query and each video, the ground truth consists of a binary semantic vector. Due to its general applicability to a wide range of scenarios as well as its personalized query-specific evaluation, our proposed algorithm targets this task.

The key challenge for this query-conditioned video summarization task is to identify the relevance between a video and a given query, and simultaneously generate summaries that are of great interest with a minimum number of video shots. To address the above issues, we propose to use a deep reinforcement learning based approach for query-conditioned video summarization. Reinforcement learning is used in this work to guide our summarization agent using a set of rewards that encode our underlying intuition of what qualities a successful summarization result should have. Deep reinforcement learning approaches [19–21] have been extensively used in a variety of computer vision tasks such as object segmentation [22], video captioning [23], action recognition [24], and also generic video summarization [25–28]. For example, Zhou and Qiao [25] develop a deep reinforcement learning-based summarization network with a diversity-representativeness reward to generate summaries, and achieve a good performance on generic video summarization. Inspired

by their work, we hypothesize that deep reinforcement learning can also be applied to address the query-conditioned video summarization task to provide personalized summary results instead of a single generic summary for all users. We propose to use deep reinforcement learning to solve this sequential decision-making problem by learning a good policy for video shot selection in order to generate good video summaries. A recurrent neural network is used to learn a sequence of shot-selection actions and based on the complete set of actions a reward is assigned. Reinforcement learning in this scenario allows us to iteratively refine the selection process, allowing the model to take better and better actions.

Specifically, we develop a deep reinforcement learning query-conditioned video summarization network (SummNet), to learn a robust video shot selection policy by jointly modeling relatedness, diversity and representativeness to tackle the above challenges. We introduce a notion of relatedness that expresses how related a video shot is to a given query by proposing a mapping network (MapNet) that maps video shots to query space. Establishing this relation between the videos and the queries allows us to formalize a relatedness reward. Together with the diversity and representativeness rewards inspired by the work of Zhou and Qiao [25], SummNet is able to provide personalized summary results taking into consideration different user interests. The contributions of our paper include: (1) We propose SummNet, a deep reinforcement learning-based framework for predicting summaries given different user queries. To the best of our knowledge, this paper is the first to use deep reinforcement learning for this task. (2) We introduce MapNet for modeling relatedness to capture the relations between the video shots and the user queries, and jointly encourage the agent to select summary results by computing relatedness, as well as representativeness and diversity rewards. (3) We conduct comprehensive experiments on a benchmark dataset which is particularly designed for query-conditioned video summarization. The proposed method outperforms the current state-of-the-art algorithms, which demonstrates its effectiveness.

2. Related Work

2.1. Video Summarization Using Reinforcement Learning

For generic video summarization, the first work using reinforcement learning was Masumitsu and Echigo [29] who propose a voting scheme-based method to predict the importance score for each frame and thereby generate the video summary. The voting scheme is obtained by user's action of watching (Accept) or skipping (Reject) previous similar frames. They use projected low-level feature vectors in an eigenspace to reduce the influence of correlation between different elements. However, the performance of the approach is restricted due to the limited learning abilities of the feature representations. Instead of hand-crafted feature representations for video frames, recent advances in the field of deep learning have enabled the learning of more robust features directly from data and have thus led to better reinforcement learning-based video summarization performances.

Zhou and Qiao [25] propose the DSN (Deep Summarization Network), an unsupervised video summarization technique, that makes use of a diversity-representativeness reward to mimic the way how humans summarize videos. The summaries are generated by predicting the probabilities that a given frame is a key-frame and then sampling summary frames based on this probability. Subsequently, they propose another reinforcement learning-based method that relies on video-level category labels to address the problem in a weakly supervised manner [27]. In order to do this, a summarization network with deep Q-learning (DQSN) is explored that guides the agent using a global recognizability reward based on a companion classification network, which is tasked with computing whether a summary can be recognized. Note, this task differs from the query-conditioned task in that it only requires that the summaries maintain the category label for each video and is, therefore, still a form of generic summarization and not personalized towards a specific query. Another work [28] is mainly targeting the challenge of summarizing extremely lengthy videos and implements a reinforcement learning-based approach that relies on SeqDPPs [30] to provide a guiding principle

about how to best partition a video sequence into different segments. This differs from prior work, which perform manually partitioning of the video into segments of the same length and do thereby not take the video content into account. Lei et al. [26] propose another reinforcement learning-based summarization approach that also dynamically segments videos. The video is first segmented using a trained action classifier, so that each clip contains a single action, then a deep recurrent neural network is applied to select the most distinct frames for each clip.

More recently, Lan et al. [31] propose a framework tasked with fast-forwarding a video to present a representative subset of frames. This can be regarded as another form of video summarization for producing key information. In their work, FastForwardNet (FFNet), a reinforcement learning framework using a Markov decision process is developed. It automatically fast-forwards a video and decides the number of frames to skip next while presenting the most important subsets to the users, which makes it computationally efficient as not all frames are processed. However, different from query-conditioned video summarization, the model only processes a subset of the video instead of the whole video due to the forwarding prediction.

2.2. Query-Conditioned Video Summarization

Query-conditioned video summarization aims to produce different summaries corresponding to different user interests. Oosterhuis et al. [32] present a graph-based approach to incorporate videos and queries in order to construct both relevant and visual appealing trailers for video summarization. Each video segment and query is represented by a node, and the weights of the edges between the video segments and the queries in the graph are computed based on semantic matching, while visual similarity is used to add edges between nodes corresponding to video segments. The segment is thus regarded to be related to a query if it is either semantically similar to the query, or is visually similar to other segments which are related to the query. In [14], a quality-aware relevance estimation is proposed that relies on measuring the distance using the cosine similarity between embeddings of the frames and the text queries, where the textual representation is achieved by first using the word2vec model [33], and then an Long Short-Term Memory (LSTM) [34] to encode each into a single fixed-length embedding. Summaries are then obtained by selecting key frames using a linear combination of submodular objectives which jointly consider diversity, representativeness and quality of the visual features, as well as the similarity between the query and frames. In [17,18], the authors investigate summarization of personalized videos of culture heritage scenarios with user preferences as input. They propose a 3D Convolutional Neural Network (CNN)-based approach to encode both visual apparent motion features and real motion features, that are measured by GPS sensors. Then visual semantic classifiers, one trained for each user preference, are used to assess the relatedness between the users preferences and the extracted key shots in the cultural heritage scenario.

The problem has also been addressed by making use of Determinantal Point Processes (DPP) [35] to model temporal relations. In [12], the authors propose a Sequential and Hierarchical Determinantal Point Process (SH-DPP) to generate summaries which consist of the key shots in the video that are most related to a given query, and study experimental performance on two densely annotated benchmark datasets. However, their datasets are originally collected for the generic video summarization task, which leads to restrictions when performing comprehensive evaluation due to the differences between the two tasks. In follow up work, they propose a more comprehensive dataset with an efficient evaluation metric for the query-conditioned summarization task [13]. Instead of focusing on low-level visual features or only temporal overlap, they evaluate similarity between the predicted and ground-truth video shots based on the shots' semantic information, where each video shot is annotated with a list of semantic concepts. Their proposed approach to combine the query and visual information uses a query-conditioned DPP that integrates a memory network for modeling query-related and contextual importance.

More recently, Zhang et al. [36] apply a generative adversarial network (GAN) [37] architecture and propose a three-player query-conditioned adversarial loss (*prediction loss*, *ground-truth loss*,

and random loss) to force the generator to learn better summary results. Though they achieve state-of-the-art performance, GAN training can often be unstable making the model difficult to train and making the approach highly-reliant on parameter choices. This is true even when more stable GAN modules, such as the Wasserstein GANs [38], are used as in their proposed framework. We, therefore, propose to exploit a deep reinforcement learning framework (SummNet) for the query-conditioned video summarization task to achieve more stable and robust performances, by introducing a mapping mechanism (MapNet) to measure relatedness between the given query and the video shots, and then jointly training the agent with additional diversity and representativeness rewards.

3. Our Approach

Our approach consists of two main parts: (1) MapNet maps the video shots to the user query space in order to provide a link between the two modalities and enable us to express the relatedness reward. (2) SummNet uses the trained MapNet and given a certain query, predicts the importance scores based on a deep reinforcement learning policy that is guided by the three rewards. Importance scores in this context refer to the probability that a given video shot is being included in the summary. In this section, we first introduce the architecture of MapNet and describe how it can be used to relate the visual information to the text queries. Then we present the details of SummNet and describe how the three rewards work jointly to obtain the summarization results.

3.1. Mapping Network

3.1.1. Video Embedding

As shown in Figure 2, we evenly segment each video $\mathcal{V}$ into 75-frame video shots (5-second long). This corresponds to the procedure proposed in [13] and enables us to compare our method to their collected ground-truths and use their evaluation metrics. We denote $\mathcal{V} = \{v_t\}_{t=1}^{T}$, where T is the total number of video shots in the video. We make use of both the 2D and 3D CNNs to extract the visual representation and encode the video shots more robustly. The C3D video descriptor can capture shot-level feature representation along both spatial and temporal dimensions, while the ResNet feature extractor is used to obtain frame-level video representation. By integrating both the above two feature vectors, the model can utilize the enhanced feature representation and perform better. We use the ResNet 152 model [39], which has been pretrained on the ILSVRC 2015 dataset [40], to obtain the features after the "fc7" layer, where "fc" corresponds to the fully connected layer. The 3D CNN feature extractor we apply is the C3D trained on the Sports1M dataset [41] and features are extracted after the "fc6" layer. The shots are down-sampled to a temporal length of 16 and the features extracted by the 2D and 3D CNNs for these shots are denoted as $\{f_t^{2d}\}_{t=1}^{T}$ and $\{f_t^{3d}\}_{t=1}^{T}$. Note that we average the feature representation of the 16 frames for the 2D CNN extractor to obtain a single feature vector for each shot. We down-sample $\{f_t^{2d}\}_{t=1}^{T}$ and $\{f_t^{3d}\}_{t=1}^{T}$ in order to reduce computational complexity, and concatenate them to obtain a joint visual representation. The combined feature representation is denoted as $\{f_t^{c}\}_{t=1}^{T}$. Afterwards, we feed them into a fc layer to get the encoded visual representation $\{f_t\}_{t=1}^{T}$ for each video.

3.1.2. Query Embedding

We use the Skip-gram model that has been trained on the Google News dataset [33] to encode each word into a feature vector. We sum the feature vectors of the two concepts for each query and the resulting feature vector is used as the query embedding $\{f_i^{w}\}_{i=1}^{Q}$, where Q is the total number of queries. Note that, following the setting of the dataset proposed in Sharghi et al. [13], each query contains two concepts in order to account for the fact that users often enter a query that contains more than one word. In order to obtain an embedding vector for each video shot the different user queries corresponding to that shot are averaged to get the whole-query embedding $\{f_t^{q}\}_{t=1}^{T}$.

3.1.3. Mapping Mechanism

We develop the mapping mechanism to model the mapping between the visual space and the query space. We take the visual representation $\{f_t\}_{t=1}^{T}$ as the input, and use three fc embedding layers in order to predict the query embedding $\{\dot{f}_t^q\}_{t=1}^{T}$ for the video shots. In between the fc layers we utilize leaky ReLUs to improve the nonlinearity learning ability of the model and dropout to avoid overfitting. MapNet, illustrated in Figure 2, aims to predict the query embedding given a video, so that $\{\dot{f}_t^q\}_{t=1}^{T}$ is as close to the ground-truths query embedding $\{f_t^q\}_{t=1}^{T}$ as possible. The loss $\mathcal{L}$ between $\{\dot{f}_t^q\}_{t=1}^{T}$ and $\{f_t^q\}_{t=1}^{T}$ is computed using the mean squared error:

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T} ||f_t^q - \dot{f}_t^q||_2. \tag{1}$$

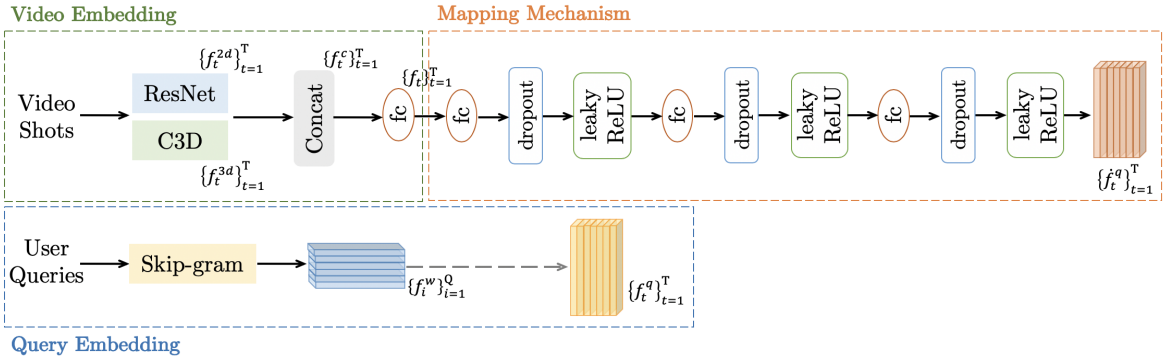

Figure 2. The architecture of our proposed Mapping Network (MapNet) has two branches. (1) Video shots are encoded using a ResNet and a C3D model to generate 2D/3D visual features $\{f_t^{2d}\}_{t=1}^{T}$ and $\{f_t^{3d}\}_{t=1}^{T}$. After down-sampling, the visual features are fused through concatenation ($\{f_t^c\}_{t=1}^{T}$) followed by a fc layer to get $\{f_t\}_{t=1}^{T}$ as a video embedding. Then three groups of fc layers, dropout layers and leaky ReLUs are used to map the video shots embeddings to the predicted query embeddings $\{\dot{f}_t^q\}_{t=1}^{T}$. (2) Different user queries each consists of two concepts are encoded via a Skip-gram model to generate query embedding $\{f_i^w\}_{t=1}^{T}$. The combined whole-query embedding $\{f_t^q\}_{t=1}^{T}$ is then achieved by averaging over all queries for each video shot. MapNet is trained to predict $\{\dot{f}_t^q\}_{t=1}^{T}$ to be close to $\{f_t^q\}_{t=1}^{T}$.

3.2. Summarization Network

As shown in Figure 3, SummNet takes both video shots and the user query as the input and first encodes them to a joint embedding space. Then the importance score for each video shot is predicted, with a higher score implying a larger probability for a shot being selected in the summary. Here we utilize the same video embedding vector $\{f_t^c\}_{t=1}^{T}$ and the word embedding f_i^w (the embedding for each query instead of the whole-query embedding) as illustrated in Section 3.1. Each time we take the query embedding for one query as the input, so here we use f_i^w, instead of $\{f_i^w\}_{i=1}^{Q}$ as the notation.

3.2.1. Importance Scores Prediction

We transform the visual information $\{f_t^c\}_{t=1}^{T}$ and the query information f_i^w using a fc layer each and then concatenate the feature representations to generate the joint video-query embedding. Note, before the concatenation, the query feature vector is repeated to match the number of shots in the batch in order to fuse the visual information at each time step with the query information. A Bidirectional-LSTM (Bi-LSTM) [42] module is then applied to capture long-term temporal dependencies in the video and model both past and future temporal relations. After that, the output is transformed by two more fc layers, with a dropout layer and a batch normalization layer in between. A sigmoid activation is used in the end to predict the importance score $\{s_t\}_{t=1}^{T}$ for each video shot, which corresponds to the probabilities that are used to generate the video summaries.

3.2.2. Video Summary Prediction

During the training phase, the three rewards are used to guide the agent to select a good set of video shots as a summary. The video shots selection is performed using a Bernoulli distribution, which is defined as:

$$p(a_t|\pi(s_t;\theta)) = Bernoulli(s_t;\theta), \tag{2}$$

where $p(\cdot)$ is the probability of taking the action a_t in the state of s_t given the learned policy π and its parameter θ. The action $a_t \in \{0, 1\}$ indicates if the current video shot is selected.

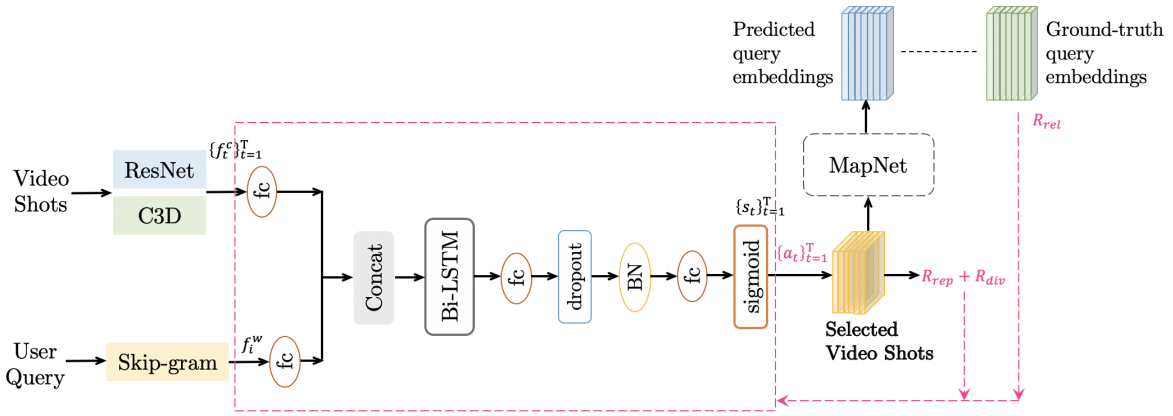

Figure 3. The architecture of our proposed Summarization Network (SummNet). Given a user query, it takes both the visual embedding $\{f_t^c\}_{t=1}^T$ and the query embedding f_i^w as the input, processes each using a fc layer and then concatenates both representations to generate the joint video-query embedding. A Bi-LSTM module is then used to model long temporal relations and the output is passed to two fc layers, with a dropout layer and a batch normalization layer in between. Afterwards, a sigmoid activation is used to predict the importance scores $\{s_t\}_{t=1}^T$ for the video shots. Based on the predicted importance scores, the reinforcement learning-based loss is used to guide the agent to take good actions. A policy π is learned based on the following three rewards to perform the right actions $\{a_t\}_{t=1}^T$ in order to select the summary shots: (1) Relatedness reward R_{rel} uses MapNet and measures the distances between the predicted and ground-truth query embedding. (2) Diversity reward R_{div} reduces the redundancies in generated summaries. (3) Representativeness reward R_{rep} aims to provide the most representative video shots of the original video as the summary.

3.2.3. Relatedness Reward Using MapNet

We take the video embedding after the concatenating layer $\{f_t^c\}_{t=1}^T$ as illustrated in Section 3.1, and feed them into the pretrained MapNet. MapNet maps each shot to the query embedding space, producing $\{\bar{f}_t^q\}_{t=1}^T$. These feature vectors are used to compute the relatedness reward in our proposed framework by considering their distance to the input query. We propose that with larger relatedness reward (smaller distance between the input query embedding and the predictions), the selected video shots are more relevant to the query making them stronger candidates for the predicted summary. We therefore compute the average squared difference between the mapping results produced by MapNet and the current user query for the selected shots as the relatedness reward and use the exponential function to ensure a positive reward. The relatedness reward is thus defined as:

$$R_{rel} = exp(-\frac{1}{|\mathcal{M}|}\sum_{\bar{f}_t^q \in \mathcal{M}} ||\bar{f}_t^q - f^q||_2), \tag{3}$$

where f^q denotes the query embedding for the current user query, and $\mathcal{M}$ denotes the selected video shots in the video.

3.2.4. Diversity and Representativeness Rewards

We apply the frame-level diversity and representativeness rewards designed by Zhou and Qiao [25] for our shot-level summarization task due to its effectiveness in measuring the quality of generated summaries in visual space. Ideally, the predicted summaries should be the most representative of the original video content and diverse in the sense that summaries should be compact and not include redundancies. The diversity reward is computed between every two video shots based on a pairwise dissimilarity measure and is defined as:

$$d(f^c_{t'}, f^c_{t''}) = 1 - \frac{f^c_{t'} f^c_{t']}}{||f^c_{t'}||_2 ||f^c_{t''}||_2}, \tag{4}$$

where $f^c_{t'}$ and $f^c_{t''}$ are the video embeddings for different video shots. We follow [25] by restricting the number of close neighbor shots that are considered, and set $d(f^c_{t'}, f^c_{t''}) = 1$ if $|t' - t''| > \lambda$. Restricting the number of neighbors is required in order to ensure that similarity between two video shots that are far from each other along the temporal dimension in the video are not penalized as this would cause problems especially for long video sequences [30]. Here, λ is a hyperparameter used to control the restriction. The diversity reward for the selected video shots $\mathcal{M}$ in the video can thus be computed based on the dissimilarity function (4):

$$R_{div} = \frac{1}{|\mathcal{M}|(|\mathcal{M}| - 1)} \sum_{f^c_{t'} \in \mathcal{M}} \sum_{\substack{f^c_{t''} \in \mathcal{M} \\ t' \neq t''}} d(f^c_{t'}, f^c_{t''}). \tag{5}$$

The representativeness reward is computed using the k-medoids algorithm as proposed in Gygli et al. [43] by first selecting a set of medoids (selected video shots) and then measuring the distances among all different video shot embeddings and their nearest medoids. It aims to minimize the mean squared error of those distances as follows:

$$R_{rep} = exp(-\frac{1}{T} \sum_{t=1}^{T} \min_{f^c_{t'} \in \mathcal{M}} ||f^c_t - f^c_{t'}||_2), \tag{6}$$

$f^c_{t'}$ denotes the video shot that is selected in the summary, and each video shot f^c_t in the video is compared to all selected ones in $\mathcal{M}$ to compute the minimal distance. We averages along the feature dimension in the exponent to avoid that the values for the representativeness reward are consistently very small for feature vectors of high dimension.

3.3. Policy Gradient Descend

Our goal is to train a summarization agent that maximizes the rewards under the policy π with the parameter θ. The expected reward $J(\theta)$ can be defined as:

$$J(\theta) = \mathbb{E}_{p(a|\pi(s;\theta))}[R_{rel} + R_{div} + R_{rep}], \tag{7}$$

where $p(a|\pi(s;\theta))$ denotes the probability distribution over the actions of sequences. We follow [25] to compute the derivative of the objective function and approximate the gradient by taking the average of N repeated episodes for each video while subtracting a constant baseline b that is computed as the moving average of the previous rewards. Introducing the constant baseline and averaging over a number of repeated episodes helps lower the variance of the gradients and thereby allows the network to converge faster to a good solution. The detailed formulation for the derivative of the objective function $J(\theta)$ is formalized as:

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{j=1}^{N} \sum_{t=1}^{T} (R^j_{rel} + R^j_{div} + R^j_{rep} - b) \nabla_\theta \log \pi(a_t|s_t;\theta), \tag{8}$$

where R^j_{rel}, R^j_{div} and R^j_{rep} are the three rewards at the j^{th} output of N episodes.

To optimize θ, we further introduce a ground-truth loss term $\mathcal{L}_{summ}$ to allow the network to predict better summary results through a direct supervised feedback signal. Here we use the output $\{s_t\}_{t=1}^T$ in Section 3.2 and compare it to the ground-truth summary $\{s_t^g\}_{t=1}^T$ ($s_t^g \in \{0,1\}$) by computing the mean squared error:

$$\mathcal{L}_{summ} = \frac{1}{T}\sum_{t=1}^{T} ||s_t - s_t^g||_2. \tag{9}$$

Thus θ can be optimized with the learning rate α as:

$$\theta = \theta - \alpha\nabla_\theta(-J(\theta) + \mathcal{L}_{summ}). \tag{10}$$

3.4. Video Summarization Inference

As shown in Figure 4, the inference phase is independent of MapNet, as well as the rewards. Given a video and a certain user query, the predicted scores $\{s_t\}_{t=1}^T$ are generated as outlined in the Importance Scores Prediction and Video Summary Prediction paragraphs in Section 3.2. A threshold ϕ is then applied to these raw scores in order to get the binarized summary prediction.

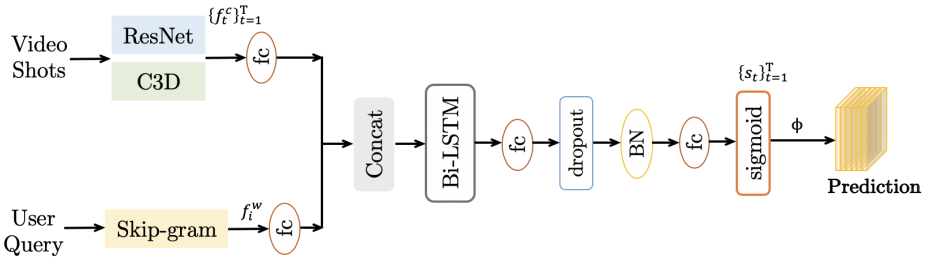

Figure 4. The pipeline for the inference phase. Given a video and a certain user query, the predicted scores are generated using SummNet without the MapNet and the reinforcement learning rewards. A threshold ϕ is applied to the raw scores to achieve binarized summary prediction.

4. Experiments

4.1. Experimental Settings

4.1.1. Videos

The proposed framework is evaluated on a query-conditioned video summarization dataset that was proposed in Sharghi et al. [13]. The dataset contains dense per-video-shot concept annotations, which are used as semantic descriptions in order to evaluate the method. The dataset contains four 3∼5 h long videos that capture uncontrolled everyday lives. The video contents contain noise due to changes in camera view point, illumination changes and motion blur [44] which makes predicting video summaries a challenging task. The queries are selected from a dictionary of 48 concepts that are constructed by Sharghi et al. [13] and that are comprehensive for the videos. Each query consists of two concepts in order to account for the fact that users often enter a query that contains more than one word. In total 46 unique queries each containing two concepts are formalized including one empty query, which aims to learn the generic video summarization that only depends on visual features. Following previous works, we use two videos for training, one for validation, and the remaining one for testing.

4.1.2. Evaluation Metrics

We follow the protocol developed by Sharghi et al. [13] to compare our framework to previous approaches. In order to match the ground-truth and predicted summaries, it uses the maximum weight matching of a bipartite graph, where the prediction and the ground-truth summaries are on two opposing sides, and the edge weights are computed by the Intersection over Union (IoU). Here the IoU for the matched shots is semantically measured based on the overlapping concepts on a shot level. The concepts that are being considered for each shot are the aggregated concepts for all queries in the ground-truth. If one shot is labelled as several concepts and is denoted as $\mathcal{A}$, and another shot's label is $\mathcal{B}$, IoU between the two shots can be defined as:

$$\text{IoU} = \frac{\text{the number of overlap concepts between } \mathcal{A} \text{ and } \mathcal{B}}{\text{the total number of concepts in } \mathcal{A} \text{ and } \mathcal{B}}. \tag{11}$$

For example, if one shot is labelled as *{Car, Street}*, another one as *{Street, Tree, Sign}*, the IoU for the two shots is $1/4 = 0.25$ [13]. Based on the number of matched video shots corresponding to the maximum weights, Precision, Recall and F-measure scores are computed to evaluate the performance of the predicted summaries. In this way, the evaluation metric focuses on the higher-level semantic information instead of lower-level visual features or temporal overlaps, which is a key requirement for measuring the quality of video summaries [45].

4.1.3. Implementation Details

We implement our framework using PyTorch [46] on a Tesla V100 GPU. We down-sample both the 2D and 3D visual features to 1024-dim. In MapNet, the three fc layers transform the learned features to 1024-, 512- and 300-dim, and a dropout rate of 0.6 and a leaky ReLU with a slope of 0.2 are used. Further, Adam [47] is used for the optimization of MapNet. In SummNet, the outputs of the fc layers for the visual feature and text feature are 2048- and 300-dim respectively. The Bi-LSTM module we use has one layer, with 1024 hidden units for each direction, followed by another fc layer to transform the features to 128-dim. After that, a dropout layer is used with a drop rate of 0.2. In the end, the last fc layer embeds the features from 128-dim to 1-dim in order to predict the importance score. We use a zero vector for the scenario where none of the video shots are related to the two concepts of the given query. Following the example of Zhou and Qiao [25], we set the temporal distance λ equal to 20, and the number of episodes N equal to 5. The constant baseline b is computed as a weighted moving average, where the previously accumulated rewards account for 90% and the current reward for 10%. We use Adam for optimizing SummNet, with a weight decay of 0.00001 and clipping of gradients with a norm larger than 5. The learning rate α, the step and the ratio for learning rate decay are optimized via cross-validation. The threshold ϕ is experimentally set by performing a grid search for the best parameter to get the binary summarization results. The training time for MapNet and SummNet are on average 2.99 ms and 408.98 ms respectively for each batch. The overall training time of the model is around 3 h. During the inference phase, given a query it takes on average 23.78 ms to obtain the prediction. This demonstrates the efficiency of our proposed method which allows small training and inference times.

4.2. Comparisons to the State-of-the-Art Methods

We compare our methods to four previous methods that have been evaluated on this query-conditioned video summarization benchmark. The results are shown in Table 1. It can be observed that our method outperforms the previous four approaches on all datasets and especially improves performance on the fourth video by large margins. The first three methods SeqDPP [30], SH-DPP [12] and QC-DPP [13] are all based on a sequential DPP, a probabilistic model, which defines a probability distribution for modeling subset selection that integrates both individual importance and collective diversity. However, DPP-based models may be influenced by an exposure bias problem as pointed out

in Sharghi et al. [4], where the model, unlike during the inference phase, is only exposed to the training data distribution and not its own predictions. While the GAN architecture of the state-of-the-art method with adversarial learning allows the learning of good summaries [36], training of GANs can be unstable making the approach more reliant on different parameter choices. Our approach, instead, is more stable and robust, and outperforms QueryGAN consistently on all videos. On average, it achieves an improvement of 1.15% in terms of F-measure, which demonstrates that our proposed deep reinforcement learning model together with the proposed mapping mechanism facilitates learning of a model that can predict more accurate summaries by jointly considering relatedness, representativeness and diversity.

Table 1. Comparison of our proposed algorithm to the state-of-the-art methods for query-conditioned video summarization in terms of Precision (Pre), Recall (Rec) and F-measure (F). The best results (F-measure) are highlighted in bold.

	SeqDPP [30]			**SH-DPP [12]**			**QC-DPP [13]**			**QueryGAN [36]**			**Ours**		
	Pre	**Rec**	**F**	**Pre**	**Rec**	**F**	**Pre**	**Rec**	**F**	**Pre**	**Rec**	**F**	**Pre**	**Rec**	**F**
Video1	53.43	29.81	36.59	50.56	29.64	35.67	49.86	53.38	48.68	49.66	50.91	48.74	50.81	51.57	**48.77**
Video2	44.05	46.65	43.67	42.13	46.81	42.72	33.71	62.09	41.66	43.02	48.73	45.30	39.99	58.51	**46.65**
Video3	49.25	17.44	25.26	51.92	29.24	36.51	55.16	62.40	56.47	58.73	56.49	56.51	63.48	54.91	**57.01**
Video4	11.14	63.49	18.15	11.51	62.88	18.62	21.39	63.12	29.96	36.70	35.96	33.64	39.03	38.37	**36.37**
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	48.33	50.84	**47.20**

4.3. Ablation Analysis

4.3.1. The Effect of the Proposed Mapping Mechanism

We analyze the effect of the mapping mechanism by dropping the MapNet network in our method and compare the performance to the full model. Note that the MapNet is used to express the relatedness reward, and that dropping MapNet leads to removing the relatedness reward in our model. The *W/O MapNet* model is therefore only trained with two rewards: representativeness and diversity rewards. The results of this comparison are shown in Table 2. We can see that the combined proposed model outperforms the model without MapNet (*W/O MapNet*) on all four videos and that the average performance decreases by 3.29% when removing the mapping mechanism of MapNet. This indicates that the mapping mechanism allows the model to incorporate the connection between the visual features and the text features by learning a transformation between the two modalities. It also shows that it is able to exploit this transformation to provide a good reward signal for the summarization model in order to learn query-conditioned video summaries. Despite of the fact that the ground-truth labels are included in the summarization model as part of the supervised loss and should enable the model to learn the relation between the visual and the query information, it appears that the proposed relatedness reward of MapNet enhances the robustness of the results.

Table 2. We compare the proposed model to two models, one that does not make use of MapNet, and one that does not make use of the ground-truth loss $\mathcal{L}_{summ}$, but that are identical otherwise. The best results (F-measure) are highlighted in bold.

	W/O MapNet			**W/O $\mathcal{L}_{summ}$**			**Ours**		
	Pre	**Rec**	**F**	**Pre**	**Rec**	**F**	**Pre**	**Rec**	**F**
Video1	49.88	44.26	43.37	30.45	64.36	40.20	50.81	51.57	**48.77**
Video2	44.17	45.49	43.54	42.01	37.78	39.50	39.99	58.51	**46.65**
Video3	61.89	52.95	55.45	39.72	74.92	51.63	63.48	54.91	**57.01**
Video4	28.19	48.13	33.28	36.28	29.40	32.12	39.03	38.37	**36.37**
Avg.	46.03	47.71	43.91	37.12	51.62	40.86	48.33	50.84	**47.20**

4.3.2. The Effect of Ground-Truth Regularization

We train the model without the ground-truth loss term (*W/O* $\mathcal{L}_{summ}$) to evaluate the effect of including the ground-truth labels. The results are shown in Table 2, and the performances on four videos decrease by 4.0%∼8.5%. From these results we can conclude that the ground-truth loss $\mathcal{L}_{summ}$ does have a large contribution to the learning process, which makes sense as the supervised feedback through the ground-truth labels provides extensive information of how the "real" summaries should look like. Especially the link between the visual information and the query will be reduced considerably when removing the ground-truth loss as it would only be supplied indirectly through the MapNet reward. This can also be seen by comparing the model without ground-truth loss (*W/O* $\mathcal{L}_{summ}$) to the model trained without MapNet (*W/O MapNet*), where the performance decreases considerably in average from 43.91% to 40.86% in terms of F-measure. This illustrates that both the ground-truth loss and the relatedness reward contribute to the overall performance, however, the ground-truth labels, which provide direct supervised feedback, contribute more to the overall performance.

4.3.3. The Effect of Rewards

We conduct another ablation experiment by dropping the three rewards (*W/O rewards*) to analyze the effect of the joint relatedness, representativeness and diversity rewards. The results are shown in Table 3. We observe that the F-measure drops 5.20% from 47.20% to 42.00%. This indicates that the model can learn a better summary by making use of the three rewards. After dropping the three rewards, the model is trained independent of MapNet as well as reinforcement learning. That is, it is only trained by the mean squared error to compute the loss between the ground-truths and the predictions in a supervised manner. The results demonstrate that while the model trained only using a supervised loss can be able to predict some good summaries, the usage of reinforcement learning can further enable the model to predict higher quality summaries by large margins.

Table 3. We compare the proposed model to two other models, one that does not make use of the three rewards, and one that does not make use of the representativeness+diversity rewards, but that are otherwise identical. Note that we repeat the results of the combined full model in this table, in order to provide a more convenient comparison. The best results (F-measure) are highlighted in bold.

	W/O Rewards			**W/O Rep+Div Rewards**			**Ours**		
	Pre	**Rec**	**F**	**Pre**	**Rec**	**F**	**Pre**	**Rec**	**F**
Video1	58.49	35.23	42.07	39.89	60.67	46.18	50.81	51.57	**48.77**
Video2	33.88	59.23	42.38	37.17	57.83	44.61	39.99	58.51	**46.65**
Video3	43.51	63.06	50.83	46.86	71.73	56.03	63.48	54.91	**57.01**
Video4	24.90	49.82	32.73	39.91	31.27	31.71	39.03	38.37	**36.37**
Avg.	40.20	51.84	42.00	40.96	55.38	44.63	48.33	50.84	**47.20**

4.3.4. The Effect of Representativeness+Diversity Rewards

We further compare our model to the one that does not make use of representativeness and diversity rewards (*W/O rep+div rewards*), and the results are shown in Table 3. The representativeness and diversity rewards are originally designed for the traditional generic video summarization task that does not take into account the user queries. However, it can be observed that the two rewards can also be utilized in this query-conditioned task. The performance decreases by 2.57% on average after dropping the two rewards, with decreases around 1.0%∼2.5% for the first three videos and 4.66% for the last one. This indicates that modeling the representativeness and diversity of the summaries in addition to the relatedness reward, which establishes the relation between the videos and the query, helps to produce a minimal and concise set of video shots.

4.4. Qualitative Results

We provide two qualitative results for Video1 of our framework. In Figure 5, the ground-truths and the predictions for user queries $\{Drink, Food\}$ and $\{Hat, Phone\}$ are presented. Figure 5a,d illustrate several video shot samples of Video1 for the two user queries, and Figure 5b,c,e,f present the selected video shots as the summary for the ground-truths (blue lines) and the predictions (green lines). Note that the evaluation metric was proposed in [13] and that we use the matched pairs in the bipartite graph based on the semantic user annotations, instead of low-level visual features. The similarity between the two shots is defined by IoU, which is computed based on the overlap of the corresponding concepts of the two shots. It aims to match semantic concepts in the video instead of unique video shots in order to provide an evaluation metric that is more robust to subjectiveness and highly similar video shots in the video. Thus the matched pairs do not necessarily map from the predictions to the ground-truths in a tight and strict sequential order.

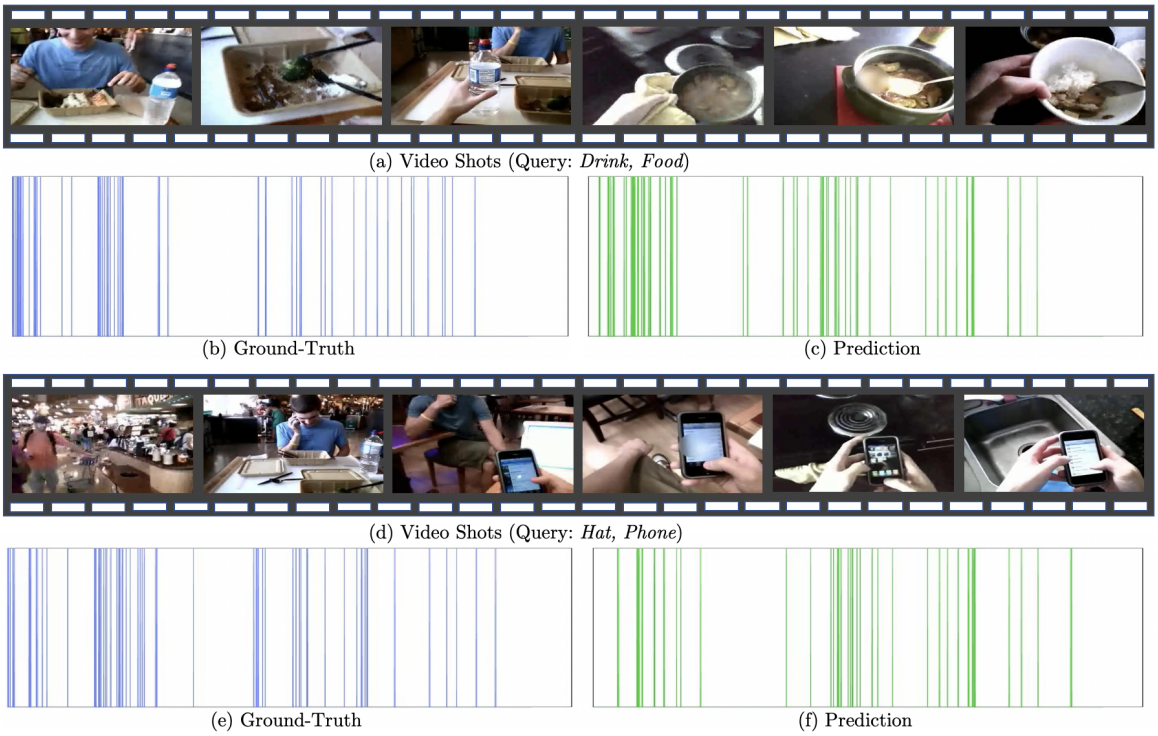

Figure 5. Visualization results of our proposed method given the query $\{Drink, Food\}$, and $\{Hat, Phone\}$ for Video1. (**a**,**d**) are several video shots samples. (**b**,**e**) and (**c**,**f**) show the ground-truths and the predicted summaries, respectively. The $x-$axis is the shot number, and the blue/green lines represent the selected video shots in the summary.

5. Conclusions

In this paper, we propose a deep reinforcement learning framework for query-conditioned video summarization, by applying relatedness, diversity and representativeness rewards to guide the agent to learn a video shots selection policy. To measure the relatedness, we design a MapNet that maps video shots from visual space to query space and based on this mapping we design a reward that encourages the agent to select the most related video shots given a certain query. Combined with additional rewards that encourage diversity and representativeness of the video shots in the summary, our reinforcement learning-based approach enables us to learn the video shots selection policy. Comprehensive experiments demonstrate the effectiveness of our approach, which outperforms the current state-of-the-art algorithm on all videos. In future work, we plan to explore the link between

the visual features and text features more in order to further improve the modelling of the relatedness, which is a key challenge in the query-conditioned video summarization task.

Author Contributions: Methodology, M.K. and Y.Z.; software, Y.Z.; writing–original draft preparation, Y.Z.; writing–review and editing, M.K. and Y.Z.; visualization, Y.Z.; supervision, M.T. and X.Z.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 61673378 and 61333016) and the Norwegian Research Council FRIPRO grant no. 239844 on developing the *Next Generation Learning Machines*.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cisco, V. Cisco Visual Networking Index: Forecast and Methodology 2016–2021. 2017. Available online: https://www.cisco.com/ (accessed on 15 February 2019).
2. Kanehira, A.; Gool, L.V.; Ushiku, Y.; Harada, T. Viewpoint-aware video summarization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7435–7444.
3. Zhang, K.; Grauman, K.; Sha, F. Retrospective encoders for video summarization. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 383–399.
4. Sharghi, A.; Borji, A.; Li, C.; Yang, T.; Gong, B. Improving sequential determinantal point processes for supervised video summarization. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 517–533.
5. Panda, R.; Roy-Chowdhury, A.K. Collaborative summarization of topic-related videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7083–7092.
6. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 202–211.
7. Amato, F.; Castiglione, A.; Moscato, V.; Picariello, A.; Sperlì, G. Multimedia summarization using social media content. *Multimed. Tools Appl.* **2018**, *77*, 17803–17827. [CrossRef]
8. Zhu, L.; Xu, Z.; Yang, Y.; Hauptmann, A.G. Uncovering the temporal context for video question answering. *Int. J. Comput. Vis.* **2017**, *124*, 409–421. [CrossRef]
9. Kalashnikov, D.; Irpan, A.; Pastor, P.; Ibarz, J.; Herzog, A.; Jang, E.; Quillen, D.; Holly, E.; Kalakrishnan, M.; Vanhoucke, V.; et al. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *arXiv* **2018**, arXiv:1806.10293.
10. Feng, W.; Ji, D.; Wang, Y.; Chang, S.; Ren, H.; Gan, W. Challenges on large scale surveillance video analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop on the AI City Challenge, Salt Lake City, UT, USA, 18–22 June 2018; pp. 69–76.
11. Wang, M.; Hong, R.; Li, G.; Zha, Z.J.; Yan, S.; Chua, T.S. Event driven web video summarization by tag localization and key-shot identification. *IEEE Trans. Multimed.* **2012**, *14*, 975–985. [CrossRef]
12. Sharghi, A.; Gong, B.; Shah, M. Query-focused extractive video summarization. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 3–19.
13. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 4788–4797.
14. Vasudevan, A.B.; Gygli, M.; Volokitin, A.; Van Gool, L. Query-adaptive video summarization via quality-aware relevance estimation. In Proceedings of the ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 582–590.
15. Ji, Z.; Ma, Y.; Pang, Y.; Li, X. Query-aware sparse coding for multi-video summarization. *arXiv* **2017**, arXiv:1707.04021.
16. Xiong, B.; Kim, G.; Sigal, L. Storyline representation of egocentric videos with an applications to story-based search. In Proceedings of the IEEE International Conference on Computer Vision, Las Condes, Chile, 11–18 December 2015; pp. 4525–4533.

17. Varini, P.; Serra, G.; Cucchiara, R. Personalized egocentric video summarization for cultural experience. In Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Shanghai, China, 23–26 June 2015; pp. 539–542.
18. Varini, P.; Serra, G.; Cucchiara, R. Personalized egocentric video summarization of cultural tour on user preferences input. *IEEE Trans. Multimed.* **2017**, *19*, 2832–2845. [CrossRef]
19. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef] [PubMed]
20. Mnih, V.; Badia, A.P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1928–1937.
21. Salimans, T.; Ho, J.; Chen, X.; Sidor, S.; Sutskever, I. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv* **2017**, arXiv:1703.03864.
22. Song, G.; Myeong, H.; Mu Lee, K. SeedNet: Automatic seed generation with deep reinforcement learning for robust interactive segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1760–1768.
23. Wang, X.; Chen, W.; Wu, J.; Wang, Y.F.; Yang Wang, W. Video captioning via hierarchical reinforcement learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4213–4222.
24. Tang, Y.; Tian, Y.; Lu, J.; Li, P.; Zhou, J. Deep progressive reinforcement learning for skeleton-based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5323–5332.
25. Zhou, K.; Qiao, Y. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. *arXiv* **2017**, arXiv:1801.00054.
26. Lei, J.; Luan, Q.; Song, X.; Liu, X.; Tao, D.; Song, M. Action parsing driven video summarization based on reinforcement learning. *IEEE Trans. Circ. Syst. Video Technol.* **2018**. [CrossRef]
27. Zhou, K.; Xiang, T.; Cavallaro, A. Video summarisation by classification with deep reinforcement learning. *arXiv* **2018**, arXiv:1807.03089.
28. 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, Munich, Germany, 8–14 September 2018; pp. 151–167.
29. Masumitsu, K.; Echigo, T. Video summarization using reinforcement learning in eigenspace. In Proceedings of the IEEE International Conference on Image Processing, Vancouver, BC, Canada, 10–13 September 2000; Volume 2, pp. 267–270.
30. Gong, B.; Chao, W.L.; Grauman, K.; Sha, F. Diverse Sequential subset selection for supervised video summarization. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2069–2077.
31. Lan, S.; Panda, R.; Zhu, Q.; Roy-Chowdhury, A.K. FFNet: Video fast-forwarding via reinforcement learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6771–6780.
32. Oosterhuis, H.; Ravi, S.; Bendersky, M. Semantic video trailers. *arXiv* **2016**, arXiv:1609.01819.
33. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 3111–3119.
34. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
35. Kulesza, A.; Taskar, B. Determinantal point processes for machine learning. *Found. Trends® Mach. Learn.* **2012**, *5*, 123–286. [CrossRef]
36. Zhang, Y.; Kampffmeyer, M.; Liang, X.; Tan, M.; Xing, E.P. Query-conditioned three-player adversarial network for video summarization. *arXiv* **2018**, arXiv:1807.06677.
37. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.

38. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein gan. *arXiv* **2017**, arXiv:1701.07875.
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas Valley, NV, USA, 26 June–1 July 2016; pp. 770–778.
40. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput.Vis.* **2015**, *115*, 211–252. [CrossRef]
41. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4489–4497.
42. Graves, A.; Mohamed, A.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 6645–6649.
43. Gygli, M.; Grabner, H.; Van Gool, L. Video summarization by learning submodular mixtures of objectives. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3090–3098.
44. Lee, Y.J.; Ghosh, J.; Grauman, K. Discovering important people and objects for egocentric video summarization. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Providence, Rhode Island, 16–21 June 2012; pp. 1346–1353.
45. Yeung, S.; Fathi, A.; Li, F.-F. Videoset: Video summary evaluation through text. *arXiv* **2014**, arXiv:1406.5824.
46. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in pytorch. In Proceedings of the Advances in Neural Information Processing Systems Workshop, Long Beach, CA, USA, 4–9 December 2017.
47. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).