DeepQAMVS: Query-Aware Hierarchical Pointer Networks for Multi-Video Summarization

Safa Messaoud
University of Illinois at Urbana-Champaign
messaou2@illinois.edu

Ismini Lourentzou
Virginia Tech
ilourentzou@vt.edu

Assma Boughoula*
University of Illinois at Urbana-Champaign
boughou1@illinois.edu

Mona Zehni*
University of Illinois at Urbana-Champaign
mzehni2@illinois.edu

Zhizhen Zhao
University of Illinois at Urbana-Champaign
zhizhenz@illinois.edu

Chengxiang Zhai
University of Illinois at Urbana-Champaign
czhai@illinois.edu

Alexander G. Schwing
University of Illinois at Urbana-Champaign
aschwing@illinois.edu

ABSTRACT

The recent growth of web video sharing platforms has increased the demand for systems that can efficiently browse, retrieve and summarize video content. Query-aware multi-video summarization is a promising technique that caters to this demand. In this work, we introduce a novel Query-Aware Hierarchical Pointer Network for Multi-Video Summarization, termed DeepQAMVS, that jointly optimizes multiple criteria: (1) conciseness, (2) representativeness of important query-relevant events and (3) chronological soundness. We design a hierarchical attention model that factorizes over three distributions, each collecting evidence from a different modality, followed by a pointer network that selects frames to include in the summary. DeepQAMVS is trained with reinforcement learning, incorporating rewards that capture representativeness, diversity, query-adaptability and temporal coherence. We achieve state-of-the-art results on the MVS1K dataset, with inference time scaling linearly with the number of input video frames.

CCS CONCEPTS

• **Computing methodologies → Video summarization**; *Reinforcement learning*; *Neural networks.*

KEYWORDS

Video summarization; Multi-video Summarization

ACM Reference Format:
Safa Messaoud, Ismini Lourentzou, Assma Boughoula, Mona Zehni, Zhizhen Zhao, Chengxiang Zhai, and Alexander G. Schwing. 2021. DeepQAMVS: Query-Aware Hierarchical Pointer Networks for Multi-Video Summarization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3404835.3462959

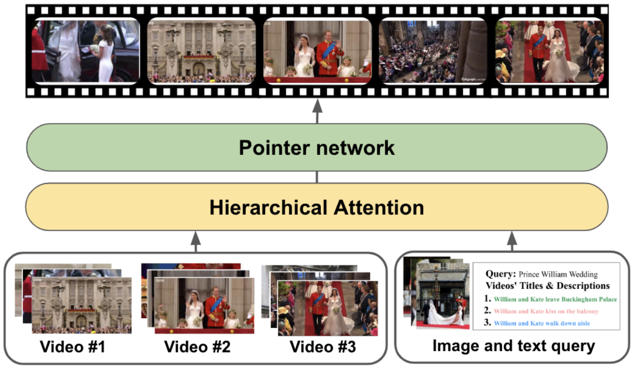

Pointer network
Hierarchical Attention
Video #1
Video #2
Video #3
Query: Prince William Wedding
Videos' Titles & Descriptions
1. William and Kate leave Buckingham Palace
2. William and Kate kiss on the balcony
3. William and Kate walk down aisle
Image and text query

Figure 1: Overview of the Deep Query-Aware Multi-Video Summarization (DeepQAMVS) model.

1 INTRODUCTION

From Snapchat and Youtube to Twitter, Facebook and ByteDance, video sharing has influenced social media significantly over the past years. Video views increased over 99% on YouTube and 258% on Facebook, in just a single year[1]. To date, more than 5 billion videos have been shared on Youtube, where users daily spend 1 billion hours watching the uploaded content[2]. Facebook also reached 100 million hours of video watching every day[3]. Given a query, current video search engines return hundreds of videos, often redundant and difficult for the user to comprehend without spending a significant amount of time and effort to find the information of interest. To effectively tackle this issue, *Query-Aware Multi-Video Summarization (QAMVS)* methods select a subset of frames from the retrieved videos and form a concise topic-related summary conditioned on the user search intent [22, 25].

A compelling summary should be (1) concise, (2) representative of the query-relevant events, and (3) chronologically sound. Naively applying traditional single video summarization (SVS) techniques

*Both authors contributed equally.

SIGIR '21, July 11–15, 2021, Virtual Event, Canada
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462959

[1] https://www.wyzowl.com/video-social-media-2020/
[2] https://www.omnicoreagency.com/youtube-statistics/
[3] https://99firms.com/blog/facebook-video-statistics/

results in suboptimal summaries, as SVS methods fail to capture all aforementioned criteria. Overall, QAMVS is more challenging than SVS. First, QAMVS needs to ensure temporal coherence, a non-trivial task since the frames are selected from multiple different videos. In contrast, for SVS the chronological order is given by the video frame order. Secondly, QAMVS methods need to filter large noisy content as videos contain a lot of query-irrelevant information. Hence, QAMVS involves modeling the interactions between two or more modalities, *i.e.*, the set of videos and the query contents. In contrast, a clustering formulation optimizing for the summary diversity yields good results for SVS.

Prior work relies on multi-stage pipelines to sequentially optimize for the aforementioned criteria. First, a set of candidate frames is selected following graph-based [7, 24, 28], decomposition-based [22, 25, 46] or learning-based [41, 65] methods. Next, the list of frames is refined to be query-adaptive by ignoring frames that are dissimilar to a set of web-images retrieved with the same query [22, 24, 25, 28]. Finally, the selected frames are ordered to form a coherent summary, either based on importance scores assigned at the video level [46, 65] or by topic-closeness [22, 24, 25]. Due to the sequential nature of these methods we observe significant shortcomings: (1) multi-stage procedures result in error propagation; (2) existing methods have polynomial complexity with respect to the size of the video set and the video lengths, and (3) the use of multi-modal meta-data is often limited to candidate frame selection instead of guiding the summarization in every step.

To address these shortcomings, in this work, we propose a *unified end-to-end trainable model* for the QAMVS task. Our architecture (summarized in Figure 1) is a hierarchical attention-based sequence-to-sequence model which significantly reduces the computational complexity from polynomial to linear compared to the current state-of-the-art methods and alleviates error propagation due to being a unified approach. We achieve this via a pointer network, which selects the frames to include in the summary, thus removing the burden of rearranging the frames in a separate subsequent step. The attention of the pointer network factorizes over three distributions, each collecting evidence from a different modality, guiding the summarization process in every step. To address the challenge of limited ground truth supervision, we train our model using reinforcement learning, incorporating representativeness, diversity, query-adaptability and temporal coherence rewards.

The key contributions of this work are summarized as follows: (1) We design a novel end-to-end Query-Aware Multi-Video Summarization (DeepQAMVS) framework that jointly optimizes multiple crucial criteria of this challenging task: (i) conciseness, (ii) chronological soundness and (iii) representativeness of all query-related events. (2) We adopt pointer networks to remove the burden of rearranging the selected frames towards forming a chronologically coherent summary and design a hierarchical attention mechanism that models the cross-modal semantic dependencies between the videos and the query, achieving state-of-the-art performance. (3) We employ reinforcement learning to avoid over-fitting to the limited ground-truth data. We introduce two novel rewards that capture query-adaptability and temporal coherence. We conduct extensive experiments on the challenging MVS1K dataset. Quantitative and qualitative analysis shows that our model achieves state-of-the-art results and generates visually coherent summaries.

2 RELATED WORK

We cover related work on single video summarization (SVS), multi-video summarization (MVS) and pointer networks (PN).

2.1 Single Video Summarization

Both *supervised* and *unsupervised* methods have been proposed for the SVS task. On the *supervised* side, methods involve category-specific classifiers for importance scoring of different video segments [51, 62], sequential determinantal point processes [16, 57, 58], LSTMs [37, 55, 71], encoder-decoder architectures [6, 23], memory networks [14] and semantic aware techniques which include video descriptors [67], vision-language embeddings [50, 63] and text-summarization metrics [69]. Instead, *unsupervised* methods rely on low-level visual features to determine the important parts of a video. Strategies include clustering [10, 17, 45], maximal bi-clique finding [7], energy minimization [52] and sparse-coding [8, 12, 13, 80]. Recently, convolutional models [54], generative adversarial networks [15, 39, 53, 75, 76] and reinforcement learning [29, 48, 74, 81] have shown compelling results on the SVS task.

Using queries to guide the summary has been explored in SVS. Proposed methods condition the summary generation on the textual query embedding [74, 76], learn common textual-visual co-embeddings for both the query and the frames [63], or enrich the visual features with textual ones obtained from dense textual shot annotations [58, 59]. As current multi-video datasets contain video-level titles/descriptions and abstract queries (*e.g.*, retirement, wedding, terror attack), the aforementioned methods are not applicable. Instead, we use the query to retrieve a set of web-images that represent its major sub-concepts/sub-events and use these images to condition the summary generation process. Note that query-adaptability is more critical in the case of MVS due to large irrelevant content across different videos.

In general, SVS methods that operate on a single long video obtained by concatenating all videos to be summarized, such as k-means [10] and Dominant Set Clustering (DSC) [3], also result in lower performance than methods designed specifically for the QAMVS problem. These methods first form clusters of frames, select centroids as candidate frames and then compute diversity to eliminate similar keyframes before generating the final summary. Due to the lack of an ordering mechanism, SVS methods result in low consistency across selected frames that reduces readability and smoothness of the overall summary, affecting significantly the user viewing experience [11]. Nevertheless, to emphasize the importance of designing techniques that tackle QAMVS specifically, we also report results for SVS approaches in our evaluation.

2.2 Multi-Video Summarization

Applications range from multi-view summarization aiming at summarizing videos captured for the same scene with several dynamically moving cameras (*e.g.*, in surveillance) [21, 38, 43, 44, 82], and summarizing of user-devices' videos [1, 41, 72, 73, 77–79] (*e.g.*, for cities hotspot preview [78] or city navigation [77]) to topic-related MVS (QAMVS) [22, 25, 28, 41, 46, 66]. Early attempts to solve the QAMVS task applied techniques optimizing for diversity [9, 20, 26, 31–35, 46, 47, 65, 66]. However, methods that advocate for these metrics cannot solve the QAMVS task satisfactorily, as

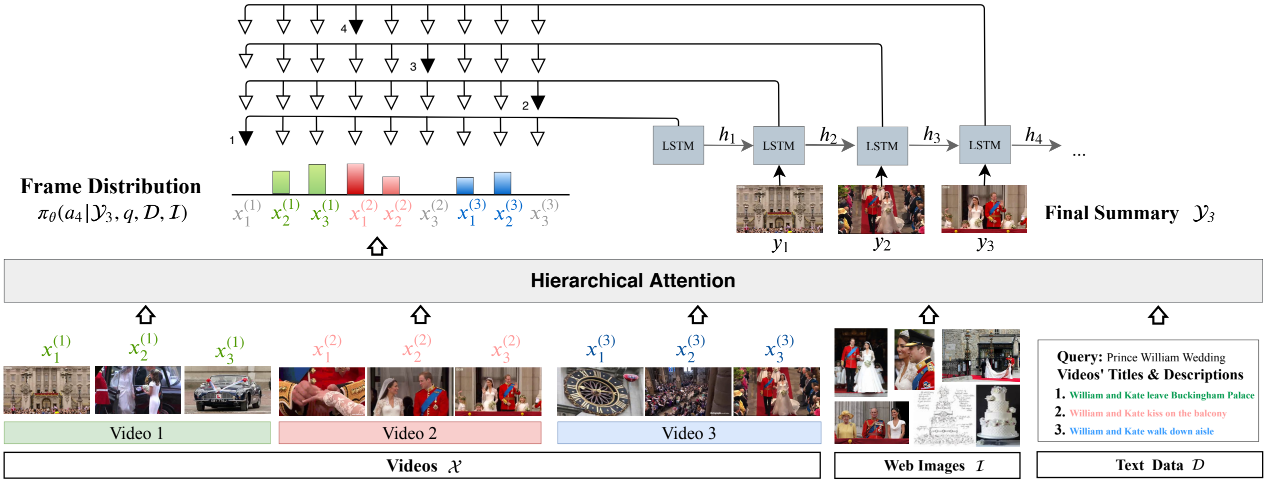

1
2
3
4
LSTM
h_1
LSTM
h_2
LSTM
h_3
LSTM
h_4
...
Frame Distribution
$\pi_\theta(a_4|\mathcal{Y}_3, q, \mathcal{D}, \mathcal{I})$
$x_1^{(1)}$ $x_2^{(1)}$ $x_3^{(1)}$ $x_1^{(2)}$ $x_2^{(2)}$ $x_3^{(2)}$ $x_1^{(3)}$ $x_2^{(3)}$ $x_3^{(3)}$
y_1
y_2
y_3
Final Summary $\mathcal{Y}_3$
Hierarchical Attention
Video 1
Video 2
Video 3
Videos $\mathcal{X}$
Web Images $\mathcal{I}$
Query: Prince William Wedding
Videos' Titles & Descriptions
1. William and Kate leave Buckingham Palace
2. William and Kate kiss on the balcony
3. William and Kate walk down aisle
Text Data $\mathcal{D}$

Figure 2: Overview of the policy network. DeepQAMVS is modeled as a Pointer Network with Hierarchical Attention (Figure 3). The policy $\pi_\theta(a_t|\mathcal{Y}_{t-1}, q, \mathcal{D}, \mathcal{I})$ is constructed by gathering evidence from the videos, the query images and the textual data. During inference, the frame with the highest probability from the video collection is copied into the final summary $\mathcal{Y}_L$.

(1) unimportant yet diverse frames are selected due to the high amount of irrelevant information across the different videos and (2) frames are not ordered chronologically to make a coherent story. Nevertheless, to emphasize the importance of designing techniques that tackle QAMVS challenges specifically, we also report results for diversity-oriented approaches in our evaluation. More recent QAMVS methods can be divided into three categories: (1) graph-based, (2) decomposition-based, and (3) learning-based.

Graph-based methods construct a graph of relationships between frames of different videos, from which the most representative ones are selected. For example, Kim *et al.* [28] summarized query related videos by performing diversity ranking on top of the similarity graphs between query web-images and video frames, in order to reconstruct a storyline graph of query-relevant events. Ji *et al.* [24] proposed a clustering-based procedure using a hyper-graph dominant set, followed by a refinement step to filter frames that are most dissimilar to the query web-images, and a final step where the remaining candidates are ordered based on topic closeness.

Decomposition-based approaches subsume weighted archetypal analysis and sparse-coding. Ji *et al.* [25] proposed a two-stage approach, where the frames are first extracted using multimodal Weighted Archetypal Analysis (MWAA). Here, the weights are obtained from a graph fusing information from video frames, textual meta-data and query-dependent web-images. Next, the frames are chronologically ordered based on upload time and topic-closeness. Panda *et al.* [46] formulated QAMVS as a sparse coding program regularized with *interestingness* and *diversity* metrics, followed by ordering the frames using a video-relevance score. While Panda *et al.* [46] did not account for query-adaptability, Ji *et al.* [22] extended the latter with an additional regularization term enforcing the selected frames to be similar to the query web-images. To form the final summary, frames are then ordered chronologically by grouping them into events based on textual and visual similarity.

For *learning-based* methods, Wang *et al.* [66] proposed a multiple-instance learning approach to localize the tags into video shots and select the query-aware frames in accordance with the tags. Nie *et al.* [41] selected frames from semantically important regions and then use a probabilistic model to jointly optimize for multiple attributes such as aesthetics, coherence, and stability.

In contrast to previous approaches that propose modularized solutions, we design a unified end-to-end model for QAMVS to generate visually coherent summaries in an end-to-end fashion.

2.3 Pointer Networks

Pointer Networks (PNs) have been applied to solve combinatorial optimization problems, *e.g.*, traveling-salesman [2] and language modeling tasks [64]. At every time step, the output is constructed by iteratively copying an input item that is chosen by the pointer. This property is uniquely convenient for the QAMVS task. Our model is the first to use a Pointer Network for QAMVS. PNs, unlike other Seq2Seq models (*e.g.*, LSTM [71] or seqDPP [16]), enable attending to any frame in any video at any time point. Hence, they naturally generate an ordered sequence of frames, while the attention mechanism fuses the multi-modal information to select the next best frame satisfying diversity, query-relevance and visual coherence (Figure 3). We train the Pointer Network in our model using reinforcement learning, as it is useful for tasks with limited labeled data [4, 5, 19, 30, 36, 40, 56], as in the case of QAMVS.

3 PROPOSED DEEPQAMVS MODEL

Given a collection of videos and images retrieved by searching with a common text query that encodes user preferences, the goal is to generate a topic-related summary for the videos. DeepQAMVS utilizes both web-images and textual meta-data. Web-images are particularly useful as they guide the summarization towards discarding irrelevant information (**image attention**). However, they

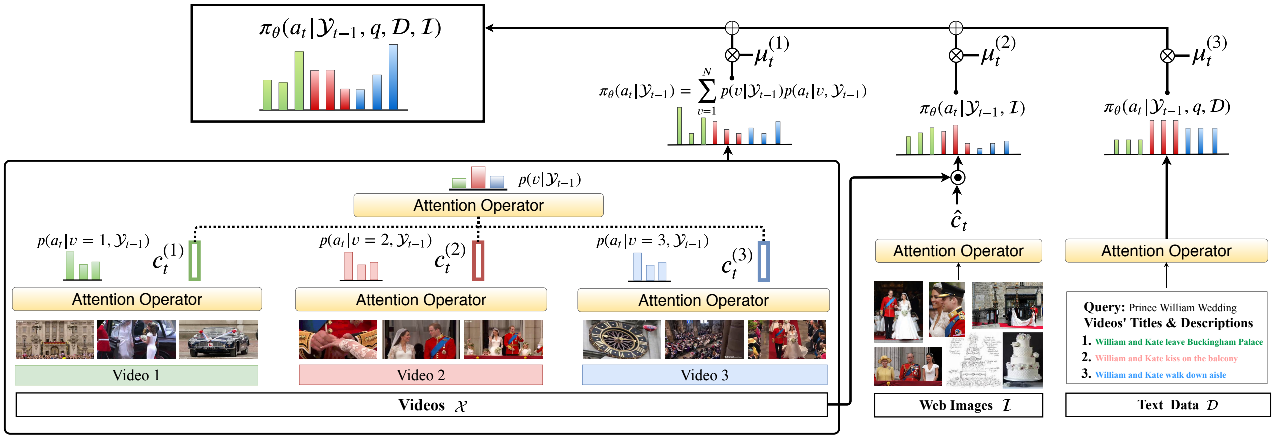

$\pi_\theta(a_t|\mathcal{Y}_{t-1}, q, \mathcal{D}, \mathcal{I})$
$\mu_t^{(1)}$
$\mu_t^{(2)}$
$\mu_t^{(3)}$
$\pi_\theta(a_t|\mathcal{Y}_{t-1}) = \sum_{v=1}^{N} p(v|\mathcal{Y}_{t-1})p(a_t|v, \mathcal{Y}_{t-1})$
$\pi_\theta(a_t|\mathcal{Y}_{t-1}, \mathcal{I})$
$\pi_\theta(a_t|\mathcal{Y}_{t-1}, q, \mathcal{D})$
$p(v|\mathcal{Y}_{t-1})$
Attention Operator
$p(a_t|v = 1, \mathcal{Y}_{t-1})$
$c_t^{(1)}$
$p(a_t|v = 2, \mathcal{Y}_{t-1})$
$c_t^{(2)}$
$p(a_t|v = 3, \mathcal{Y}_{t-1})$
$c_t^{(3)}$
Attention Operator
Attention Operator
Attention Operator
Video 1
Video 2
Video 3
Videos $\mathcal{X}$
$\hat{c}_t$
Attention Operator
Attention Operator
Web Images $\mathcal{I}$
Query: Prince William Wedding
Videos' Titles & Descriptions
1. William and Kate leave Buckingham Palace
2. William and Kate kiss on the balcony
3. William and Kate walk down aisle
Text Data D

Figure 3: Illustration of DeepQAMVS's Hierarchical Attention $\pi_\theta(a_t|\mathcal{Y}_{t-1}, q, \mathcal{D}, \mathcal{I})$.

might contain irrelevant information. To robustly ensure query-relevance, DeepQAMVS leverages multi-modal attention which enables the web-images and textual meta-data (**query attention**) to act as complementary information that guides the summarization process. In the following, we first formally define the problem, then introduce the proposed DeepQAMVS model.

3.1 Problem Formulation

Let q be the semantic embedding of the textual query and let $\mathcal{I} = \{I_1, \cdots, I_{|\mathcal{I}|}\}$ refer to the set of web-image embeddings. We denote by $\mathcal{X}^{(v)} = \{x_1^{(v)}, \cdots, x_{|\mathcal{X}^{(v)}|}^{(v)}\}$ the set of frame embeddings from video $v \in \{1, \cdots, N\}$. Let $\mathcal{D} = \{d^{(1)}, \cdots, d^{(N)}\}$ be the text embeddings of the videos' textual data, constructed by averaging the embeddings of the title and description for every video. The goal is to generate a summary $\mathcal{Y}_L = \{y_1, \cdots, y_L\}$ of L frames selected from the input video frames, *i.e.*, $\mathcal{Y}_L \subset \mathcal{X} = \bigcup_v \mathcal{X}^{(v)}$.

Due to the sequential nature of the problem, *i.e.*, selecting the next candidate frame based on what has been selected so far, we formulate the QAMVS problem as a Markov Decision Process (MDP). Specifically, an agent operates in $t \in \{1, \cdots, L\}$ time-steps according to a policy $\pi_\theta(a_t|\mathcal{Y}_{t-1}, q, \mathcal{D}, \mathcal{I})$ with trainable parameters θ. The policy encodes the probability of selecting an action a_t given the state $\mathcal{Y}_{t-1}$, the query q, the text meta-data $\mathcal{D}$ and the web-images $\mathcal{I}$. The state $\mathcal{Y}_{t-1}$ denotes the set of frames that are already selected in the summary up to time step t. Note that $\mathcal{Y}_0 = \emptyset$. The set of possible actions is the set of input frames after eliminating the ones that have already been selected in the summary, *i.e.*, $a_t \in \mathcal{A}_t = \mathcal{X} \setminus \mathcal{Y}_{t-1}$. We denote by $\mathcal{A}_t^{(v)}$ the set of valid actions corresponding to frames from video v.

We model the policy function $\pi_\theta(a_t|\mathcal{Y}_{t-1}, q, \mathcal{D}, \mathcal{I})$ as a pointer network with hierarchical attention, as illustrated in Figure 2. **At inference** step t, the inputs ($\mathcal{X}$, $\mathcal{D}$ and $\mathcal{I}$), together with the state $\mathcal{Y}_{t-1}$, are used to compute the distribution $\pi_\theta(a_t|\mathcal{Y}_{t-1}, q, \mathcal{D}, \mathcal{I})$ over possible actions a_t, *i.e.*, over possible frames. The frame with the highest probability is then copied to the summary $\mathcal{Y}_t$. The process continues until a summary of length L is reached. Next, we describe the policy $\pi_\theta(a_t|\mathcal{Y}_{t-1}, q, \mathcal{D}, \mathcal{I})$.

3.2 DeepQAMVS Policy Network

Our proposed policy function models the cross-modal semantic dependencies between the videos, the text query and the web-images. More specifically, the policy network $\pi_\theta(a_t|\mathcal{Y}_{t-1}, q, \mathcal{D}, \mathcal{I})$ is the weighted combination of three distributions, video frame attention $\pi_\theta(a_t|\mathcal{Y}_{t-1})$, image attention $\pi_\theta(a_t|\mathcal{Y}_{t-1}, \mathcal{I})$, and query attention $\pi_\theta(a_t|\mathcal{Y}_{t-1}, q, \mathcal{D})$. Formally,

$$\begin{aligned} \pi_\theta(a_t|\mathcal{Y}_{t-1}, q, \mathcal{D}, \mathcal{I}) = \mu_t^{(1)} \pi_\theta(a_t|\mathcal{Y}_{t-1}) + \\ \mu_t^{(2)} \pi_\theta(a_t|\mathcal{Y}_{t-1}, \mathcal{I}) + \mu_t^{(3)} \pi_\theta(a_t|\mathcal{Y}_{t-1}, q, \mathcal{D}), \end{aligned} \tag{1}$$

where $\mu_t^{(1)}$, $\mu_t^{(2)}$ and $\mu_t^{(3)}$ are learnable interpolation terms satisfying $\mu_t^{(1)} + \mu_t^{(2)} + \mu_t^{(3)} = 1$. An illustration of the hierarchical attention is provided in Figure 3. In the following, we introduce each of these three distributions.

The **video frame attention** $\pi_\theta(a_t|\mathcal{Y}_{t-1})$ is modeled as a two-level attention, *i.e.*, at each time step t, video attention selects video v and then selects a frame a_t from video v:

$$\pi_\theta(a_t|\mathcal{Y}_{t-1}) = \sum_{v=1}^{N} p(v|\mathcal{Y}_{t-1}) p(a_t|v, \mathcal{Y}_{t-1}), \tag{2}$$

where $p(a_t|v, \mathcal{Y}_{t-1})$ is the probability of selecting a frame a_t from video v, and $p(v|\mathcal{Y}_{t-1})$ is the distribution over the collection of videos. We compute both probabilities via

$$p(a_t|v, \mathcal{Y}_{t-1}), c_t^{(v)} = \text{Attention}(\mathcal{A}_t^{(v)}, \mathcal{Y}_{t-1}), \tag{3}$$

$$p(v|\mathcal{Y}_{t-1}), c_t = \text{Attention}\left(\{c_t^{(1)}, \cdots, c_t^{(N)}\}, \mathcal{Y}_{t-1}\right). \tag{4}$$

The *Attention* operator, as well as the context vectors c_t and $\{c_t^{(v)}\}_{v=1}^{N}$ are defined below. Intuitively, the two-level attention enables scaling to a large number of videos and video lengths since it decomposes a joint distribution into the product of two conditional distributions.

The **image attention** $\pi_\theta(a_t|\mathcal{Y}_{t-1}, \mathcal{I})$ reflects the correlation between video frames and web-images. We first generate a context

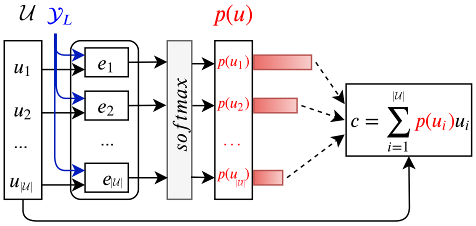


Figure 4: The *Attention* operator.

vector $\hat{c}_t$ encoding the most relevant information in the web-images at time t given the current summary $\mathcal{Y}_{t-1}$:

$$p(I|\mathcal{Y}_{t-1}), \hat{c}_t = \text{Attention}(\mathcal{I}, \mathcal{Y}_{t-1}). \tag{5}$$

$\pi_\theta(a_t|\mathcal{Y}_{t-1}, \mathcal{I})$ is then obtained by transforming the dot product between $\hat{c}_t$ and the action representations, *i.e.*, representations from not previously selected frames, into a distribution via a softmax.

The **query attention** $\pi_\theta(a_t|\mathcal{Y}_{t-1}, q, \mathcal{D})$ captures the correlation between the query q, the text data $\mathcal{D}$ and the summary at time t. For this, we first weigh every video's text embedding by its similarity to the query. Next, we compute an attention over the weighted embeddings, given the current summary $\mathcal{Y}_{t-1}$, via

$$\pi(a_t|\mathcal{Y}_{t-1}, q, \mathcal{D}), \tilde{c_t} = \text{Attention}\left((q^T d^{(v)})d^{(v)}, \mathcal{Y}_{t-1}\right). \tag{6}$$

The interpolation weights in Eq. (1) can be obtained by attending over the modalities' context vectors, c_t, $\hat{c}_t$ and $\tilde{c_t}$:

$$[\mu_t^{(1)}, \mu_t^{(2)}, \mu_t^{(3)}], \cdot = \text{Attention}(\{c_t, \hat{c}_t, \text{MLP}(\tilde{c_t})\}, \mathcal{Y}_{t-1}), \tag{7}$$

where MLP is a multi-layer perceptron used to unify the dimensions of the three context vectors. We observe that if c_t and $\hat{c}_t$ are similar, their weights $\mu_t^{(1)}$ and $\mu_t^{(2)}$ are close, else more weight is given to *video attention.*

The **Attention operator**, illustrated in Figure 4 and used multiple times above, takes as input a sequence of vectors $\mathcal{U} = \{u_i\}_{i=1}^{|\mathcal{U}|}$ with $u_i \in \mathbb{R}^m$ and the summary $\mathcal{Y}_{t-1}$, embedded by an LSTM into a hidden state $h_t \in \mathbb{R}^n$. The Attention operator provides as output a distribution $p(u_i)$ over the vectors $\{u_i\}_{i=1}^{|\mathcal{U}|}$ and a context vector c as a linear combination of elements in $\mathcal{U}$ by conditioning them on h_t:

$$e_i = w_1^T \tanh(W_2[u_i; h_t]),$$
$$p(u_1), \cdots, p(u_{|\mathcal{U}|}) = \text{Softmax}([e_1, \cdots, e_{|\mathcal{U}|}]), \tag{8}$$
$$c = \sum_{i=1}^{|\mathcal{U}|} p(u_i)u_i, \tag{9}$$

where $w_1 \in \mathbb{R}^n$ and $W_2 \in \mathbb{R}^{n \times (n+m)}$ are trainable weight parameters. The outputs of the Attention operator are the probabilities given in Eq. (8) and the context vector c given in Eq. (9).

Embeddings: The video frames $\mathcal{X}$ are embedded with a pre-trained CNN followed by a BiLSTM network. Web-images $\mathcal{I}$ are encoded with the same CNN. Textual embeddings $\mathcal{D}$ are computed for every video by averaging Glove word embeddings [49] from its associated title and description. Note that we normalize all embeddings.

3.3 Training with Policy Gradient

Due to the limited annotated data and the subjectivity of the ground truth summaries, we train our model via reinforcement learning. The goal is to learn the policy $\pi_\theta(a_t|\mathcal{Y}_{t-1}, q, \mathcal{D}, \mathcal{I})$ by maximizing the expected reward $J(\theta) = \mathbb{E}_{\pi_\theta}[R(\mathcal{Y}_L)]$ during training, where $R(\mathcal{Y}_L)$ denotes the reward function computed for a summary $\mathcal{Y}_L$. Following REINFORCE [68], we approximate the expectation by running the agent for M episodes for a batch of videos and then taking the average gradient. To reduce variance, we use a moving average of the rewards as a computationally efficient baseline.

The reward $R = \beta_1 R_{\text{div}} + \beta_2 R_{\text{rep}} + \beta_3 R_{\text{query}} + \beta_4 R_{\text{coh}}$ is composed of four terms, measuring the diversity (R_{div}), representativeness (R_{rep}), query-adaptability (R_{query}) and temporal coherence (R_{coh}). Hyperparameters $\{\beta_i\}_{i=1}^4$ are weights associated to different rewards. Note that we use the same diversity and representativeness rewards as Zhou et al. [81]. In addition, we introduce two novel rewards, query-adaptability and temporal coherence, to accommodate the QAMVS task. To keep the rewards in the same range, we use (1) *dot product* as a similarity metric in R_{coh} to balance out R_{div} and (2) a similar form to R_{rep} for R_{query}.

The **Diversity Reward** measures the dissimilarity among the selected frames in the feature space via

$$R_{\text{div}}(\mathcal{Y}_L) = \frac{1}{L(L-1)} \sum_{\substack{y_t, y_{t'} \in \mathcal{Y}_L \\ t \neq t'}} \left(1 - y_t^T y_{t'}\right). \tag{10}$$

Intuitively, the more dissimilar the selected frames to each other, the higher the diversity reward the agent receives.

The **Representativeness Reward** measures how well the generated summary represents the main events occurring in the collection of videos. Thus, the reward is higher when the selected frames are closer to the cluster centers. Formally,

$$R_{\text{rep}}(\mathcal{Y}_L) = \exp(-\frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \min_{y_t \in \mathcal{Y}_L} \|x - y_t\|^2). \tag{11}$$

The **Query-Adaptability Reward**[4] encourages the model to select the summary frames to be similar to the web-images $\mathcal{I}$ via

$$R_{\text{query}}(\mathcal{Y}_L) = \exp\left(-\frac{1}{L} \sum_{y_t \in \mathcal{Y}_L} \min_{I \in \mathcal{I}} \|y_t - I\|^2\right). \tag{12}$$

The **Temporal Coherence Reward** encourages the visual coherence of the generated summary via

$$R_{\text{coh}}(\mathcal{Y}_L) = \frac{1}{L} \sum_{y_t \in \mathcal{Y}_L} \rho(y_t), \tag{13}$$

[4] Other explored forms include $R_{\text{query}}(\mathcal{Y}_L) = -\frac{1}{|\mathcal{I}|} \sum_{I \in \mathcal{I}} \min_{y_t \in \mathcal{Y}_L} \|y_t - I\|^2$ and $R_{\text{query}}(\mathcal{Y}_L) = -\frac{1}{L} \sum_{y_t \in \mathcal{Y}_L} \|y_t - \frac{1}{|\mathcal{I}|} \sum_{I \in \mathcal{I}} I\|^2$. We found the formulation in Eq. (12) to work best.

where $\rho(y_t)$ is calculated by adding up the correlation between two consecutive frames:

$$\rho(y_t) = \frac{1}{2} \sum_{k \in \{\pm 1\}} y_t^T y_{t+k}. \tag{14}$$

Hence, the more correlated the neighboring frames, the higher the temporal coherence reward. Note that optimizing for visual/temporal coherence, *i.e.*, smoothness of the transitions, is just a proxy for chronological soundness, which is a much harder problem.

4 EXPERIMENTAL SETUP

We describe experimental details, such as the evaluation dataset and metrics, and present quantitative and qualitative results, comparing the proposed DeepQAMVS model with several baselines. Our experiments aim to show that (1) SVS methods cannot properly address QAMVS and multi-stage MVS procedures result in lower performance than a unified system (sections 4.2 and 4.5), (2) the use of multi-modal information is crucial in guiding the summarization process (section 4.3), (3) the introduced novel temporal coherence reward generates more visually coherent summaries (section 4.4) and (4) our method reduces the computational complexity compared to the current state-of-the-art methods (section 4.6).

4.1 Experimental Settings

Dataset: We perform our experiments on the MVS1K dataset [22]. MVS1K is a collection of 1000 videos on 10 queries (events), with associated web-images, video titles, and their text descriptions. Each query has 4 different user summaries, serving as a ground truth. Table 1 lists the events, the query used to retrieve them, the number of videos and query web-images for each event as well as the total number of input frames across all videos. Each video is associated with a title and a text description. We use the features introduced by Ji *et al.* [22]. The dimensionality of the video frame and web-image embeddings is 4352. The embeddings are composed of a 4096 dimensional VGGNet-19 [60] (trained on ImageNet) CNN feature vector concatenated with a 256 dimensional HSV color histogram feature vector. These embeddings are reduced to a vector of length 256 through a fully connected layer. The input frames to the model are selected such that they represent the segment centers obtained using the shot boundary detection algorithm [70]. The textual features (titles and descriptions) and the query are Glove embeddings [49] of dimension 100. We set the hidden state dimension of the LSTM and the pointer network to be 256 and 32, respectively.

Evaluation Metrics: To compare with previous work, generated summaries are assessed using F1-score, averaged over the ground truth user summaries. Following prior work [22, 24, 25], two frames are considered to match when the pixel-level euclidean distance is smaller than a predefined threshold of 0.6.

Table 1: Dataset Characteristics

Query ID	Query	# Videos	# Frames	# Images
1	Britains Prince William wedding 2011	90	1124	324
2	Prince death 2016	104	1549	142
3	NASA discovers Earth-like planet	100	1349	226
4	American government shut-down 2013	82	962	177
5	Malaysia Airline MH370	109	1330	435
6	FIFA corruption scandal 2015	90	785	177
7	Obama re-election 2012	85	1263	207
8	Alpha go vs Lee Sedo	84	976	118
9	Kobe Bryant retirement	109	1140	221
10	Paris terror attacks	83	857	651
Total	-	936	-	2678

Training Details: We train using a 10-fold cross-validation scheme. Specifically, for evaluating each event, we use the remaining 9 events as training data. During training, we use a batch size of 32, where each sample consists of 10 randomly sampled videos per event. We limit the number of video combinations to 4000 for every event. This large number of random combinations allowed us to avoid overfitting despite the small number of events[5]. We optimize with Adam, 0.01 learning rate and ℓ_2 regularization. During testing, we use all the videos associated with an event in the test set. Since the diversity R_{div} and representativeness R_{rep} reward on one side, and the coherence reward R_{coh} on the other side are contradictory, *i.e.*, R_{div} and R_{rep} encourage the selection of diverse frames while R_{coh} is high when the summary is smooth as measured by the similarity of the neighboring frames, we use a training schedule: (1) We set $\beta_1 = \beta_2 = \beta_3 = 1/3$ and $\beta_4 = 0$ for 60 epochs. (2) Then, set $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1/4$ for 30 additional epochs. We also experiment with different summary lengths $L \in \{30, 50, 60\}$.

4.2 Experimental Results

We compare DeepQAMVS to five SVS baselines operating on the concatenated videos. We chose the SVS baselines such that they represent the main trends in unsupervised summarization:

- ***K*-means** [10]: SVS method that clusters all video frames and then selects the one closest to the cluster centers as summary frames ($k = 9$).
- **DSC** [3]: Dominant Set Clustering (DSC) is a graph-based clustering method where a dominant set algorithm is used to extract the summary frames.
- **MSR** [3]: Minimum Sparse Reconstruction (MSR) is a decomposition based approach, which formulates video summarization as a minimum sparse reconstruction.
- **SUM-GAN** [39]: An adversarial LSTM model, where the generator is an autoencoder LSTM aiming at first selecting the summary frames then reconstructing the original video based on them, and the discriminator is trained to distinguish between the reconstructed video and the original one.
- **DSN** [81]: uses a RNN trained with deep reinforcement learning with diversity and representativeness rewards.

Moreover, we compare with four state-of-the-art QAMVS baselines:

- **QUASC** [22]: QUASC is a sparse coding program regularized with interestingness, diversity and query-relevance metrics, followed by ordering the frames chronologically by grouping them into events based on textual and visual similarity.
- **MWAA** [25]: MWAA is a two-stage approach, where the frames are first extracted using multi-modal Weighted Archetypal Analysis (MWAA), and then are chronologically ordered based on upload time and topic-closeness.

[5] Besides the proposed reinforcement learning framework, we experimented with training in a supervised fashion, however, we could not avoid over-fitting.

Table 2: Comparison of our approach against baselines (F1 score).

	Query ID	1	2	3	4	5	6	7	8	9	10	AVG
SVS	**k-means**	.576	.552	**.568**	.336	.457	.525	.651	.278	.384	.337	.466
	DSC	.578	.472	.399	.530	.407	.494	.533	.485	.529	.471	.490
	MSR	.472	.391	.370	.414	.396	.355	.418	.234	.384	.288	.372
	SUM-GAN	.620±.035	.481±.028	.519±.034	.501±.038	.413±.022	.455±.048	.458±.059	.459±.041	.510±.021	.395±.056	.486± .075
	DSN	.529±.019	.327±.062	.478±.036	.407±.026	.325±.042	.453±.033	.616±.028	.375±.022	.469±.021	.384±.016	.436±.093
MVS	**QUASC**	.520	.513	.400	**.570**	.513	**.538**	.623	.439	**.709**	**.588**	.544
	MVS-HDS	.660	.552	.475	.526	.495	.520	.642	.469	.633	.581	.555
	MWAA	.705	**.610**	.553	.511	**.563**	.466	**.664**	.483	.611	.379	.555
	Random-50	.600±.070	.349±.088	.288±.047	.492±.131	.255±.074	.352±.096	.265±.099	.429±.109	.326±.109	.284±.064	.364±.089
	Ours-30	.570±.013	.491±.037	.421±.084	.519±.017	.458±.054	.476±.030	.369±.036	.372±.014	.403±.017	.368±.041	.446±.022
	Ours-50	.706±.018	.563±.035	.525±.017	.553±.026	.549±.014	.486±.032	.524±.015	**.486±.022**	.690±.015	.542±.022	.561±.005
	Ours-60	**.722±.019**	.530±.046	.495±.009	.508±.015	.541±.036	.487±.014	.614±.026	.474±.015	.674±.025	.573±.019	.562±.004
	Ours-best	**.722±.019**	.563±.035	.525±.017	.553±.026	.549±.014	.487±.014	.614±.026	**.486±.022**	.690±.015	.573±.019	**.576±.017**

Table 3: Summary length (# frames) across methods.

	Query ID	1	2	3	4	5	6	7	8	9	10	AVG
SVS	**k-means**	48	51	59	51	63	47	48	36	39	28	47.0
	DSC	42	47	34	39	52	46	55	41	41	41	43.8
	MSR	48	51	59	51	63	47	48	36	39	28	47.0
	SUM-GAN	60	60	60	60	60	60	60	60	60	60	60.0
	DSN	60	60	60	60	60	60	60	60	60	60	60.0
MVS	**QUASC**	33	57	21	55	48	41	59	52	51	56	47.3
	MVS-HDS	51	60	48	48	60	44	58	54	60	50	53.3
	MWAA	49	75	46	39	49	37	60	36	39	39	46.9
	Ours-best	60	50	50	50	50	60	60	50	50	60	53.0

- **MVS-HDS** [24]: MVS-HDS is a clustering-based procedure using a Hyper-graph Dominant Set, followed by a refinement step to filter frames that are most dissimilar to the query web-images, and a final step where the remaining candidates are ordered based on topic closeness.
- **Random-50**: We also compare our method against a randomly generated summary with length 50.

We present quantitative results of our approach in Table 2 and the number of summary frames selected by each approach in Table 3. More specifically, in Table 2, the reported numbers represent the mean and standard deviation obtained from 5 rounds of experiments. We report the F1-scores for summaries of length 30 (*ours-30*), 50 (*ours-50*) and 60 (*ours-60*), as well as the best obtained score (*ours-best*) when selecting the best summary length for every event. We observe that SVS methods have in general lower performance than MVS methods. In addition, our proposed end-to-end DeepQAMVS model, on average, outperforms all baselines.

4.3 Ablation Study

We present an ablation study, examining the effect of different rewards and attention mechanisms in Table 4. We evaluate the average F1-score across all the events for the following combinations of the attention modalities: (1) only video frame attention ($\mu_t^{(2)} = \mu_t^{(3)} = 0$); (2) video frame and image attention ($\mu_t^{(3)} = 0$); (3) video frame and query attention ($\mu_t^{(2)} = 0$); and (4) video frame, image and query attention ($\mu_t^{(i)} \neq 0, \forall\, i \in \{1, 2, 3\}$). We also investigate the effect of the different rewards by incrementally adding the reward terms including, (1) R_{div} ($\beta_2 = \beta_3 = \beta_4 = 0$); (2) R_{div} and R_{query} ($\beta_3 = \beta_4 = 0$); (3) R_{div}, R_{query} and R_{coh} ($\beta_4 = 0$); and (4) all the rewards, *i.e.*, $\beta_i \neq 0, \forall i \in \{1, \cdots, 4\}$. Note that we do not add the query reward R_{query} when testing with attention terms that do not include the image attention (– in Table 4).

Table 4: Ablation study on attention and rewards ($L = 60$).

Attention / Reward	$\mu_t^{(2)} = 0$ $\mu_t^{(3)} = 0$	$\mu_t^{(3)} = 0$	$\mu_t^{(2)} = 0$	$\mu_t^{(i)} \neq 0$
$\beta_2 = \beta_3 = \beta_4 = 0$	.323± .013	.559±.008	.374± .015	.560± .008
$\beta_3 = \beta_4 = 0$	.321± .011	.557±.002	.373± .002	.559± .001
$\beta_4 = 0$	–	.561±.003	–	.562± .006
$\beta_i \neq 0$	.330± .020	.559±.005	.375± .017	.562± .004

When considering all forms of attention (last column), we found that R_{div} has barely improved the F1-score. In contrast, including R_{query}, helped improve the quality of the summary while adding the coherency reward R_{coh} did not lead to a consistent increase of the F1-score. This is expected as the ground truth summary consists of an unordered set of frames. However, as demonstrated by the user study below, R_{coh} helped generating more visually coherent summaries. Across the different combinations of rewards, we observe that the combination of video frame attention and image attention (column 3) yields overall a higher F1-score than the combination of video frame attention and query attention (column 4). This is due to video descriptions being noisy and associated with the whole video, unlike the web-images, which are embedded in the same space as the frames and hence better capture query-adaptability. The best results are obtained by using all the attention terms (last column), demonstrating the complementary properties of the multimodal information.

4.4 Temporal Coherence User Study

Since the provided ground truth summaries are composed of an unordered collection of frames, we resort to a user study to assess the visual coherence of our generated summaries. In total 21 participants are presented with 3 summaries generated from (1) DeepQAMVS, (2) random permutation of the video segments in

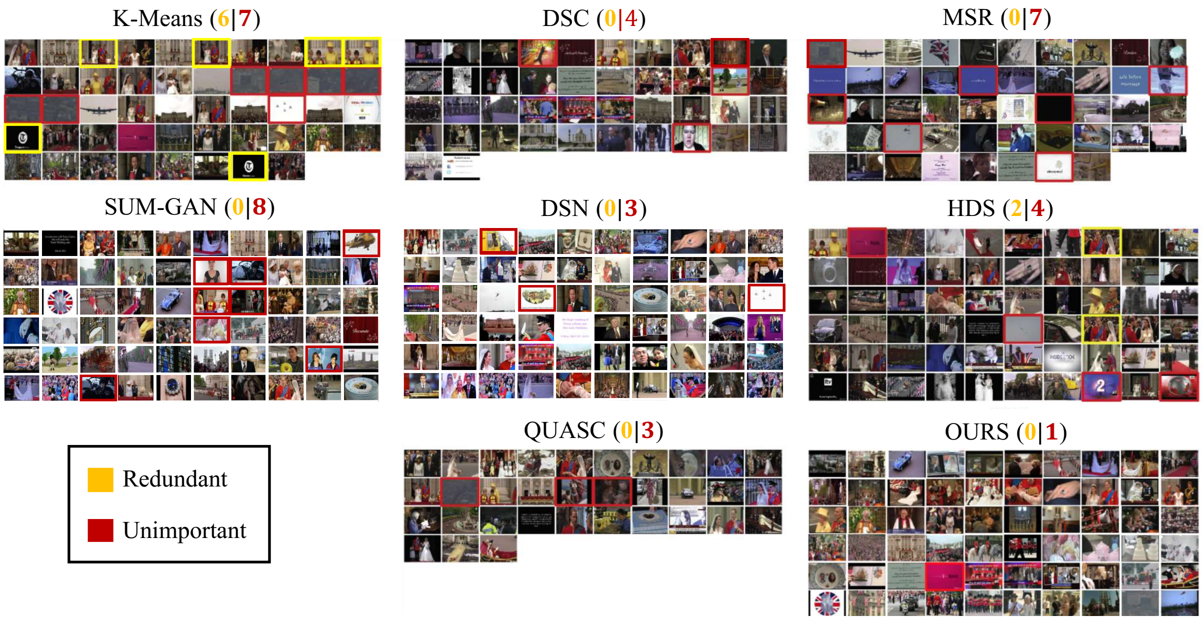


Figure 5: Qualitative results for event 1 (Prince William Wedding) by K-means [10], DSC [3], MSR [3], QUASC [22], MVS-HDS [24] and DeepQAMVS, respectively. Frames outlined in red indicate unimportant keyframes, while yellow ones show redundant ones. The number of unimportant and redundant frames are reported on top of every summary.

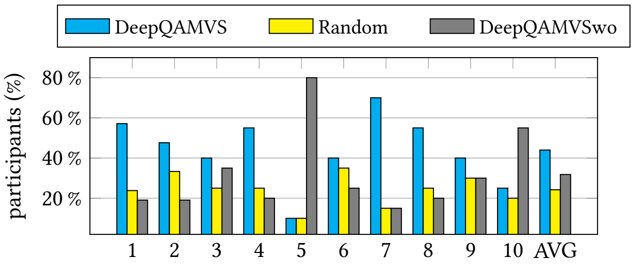


Figure 6: Temporal coherence user study for Query IDs (x-axis).

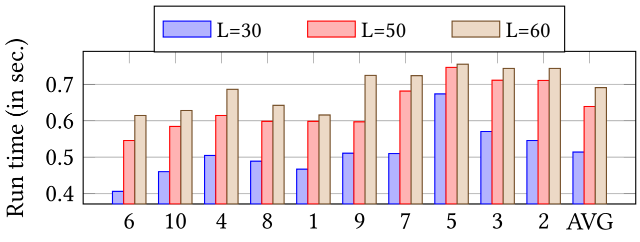


Figure 7: Run Time Analysis in seconds. Query IDs (x-axis) ordered by total number of input frames.

the DeepQAMVS summary (Random), and (3) DeepQAMVS trained without the temporal coherence reward (DeepQAMVSwo). The participants are asked to select the most coherent summary, paying special attention to transitions between different segments in each video.

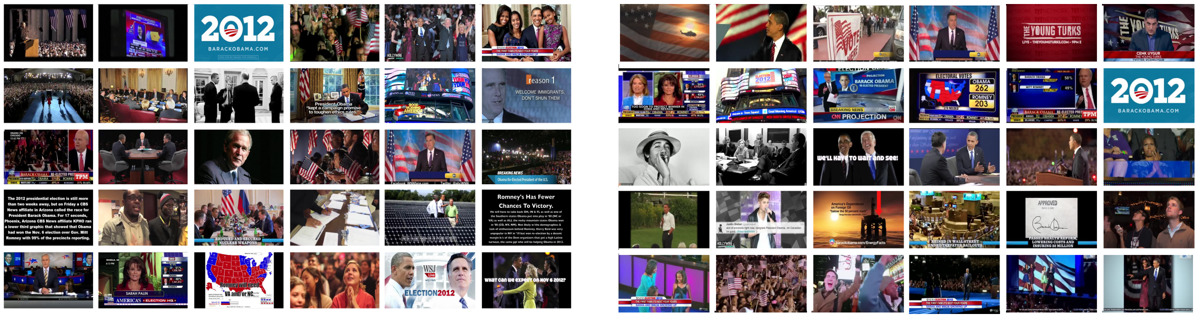
From Figure 6, we can see that users preferred our DeepQAMVS summary in 8 out of 10 events. For events 5 '*Malaysian Airline MH370*' and 10 '*Paris Attack*', users preferred the summaries generated by DeepQAMVSwo. Note that these two events deal with major news incidents and consequently mostly consist of visually similar newscaster segments. In this case, users most likely prefer the resulting summaries from DeepQAMVSwo, as it produces more visually varied summaries due to the higher importance of the diversity reward.

4.5 Qualitative Results

Figure 5 illustrates the summaries generated by different methods for the query *Prince William Wedding* (event 1). Visually, we observe that SVS methods choose many irrelevant frames. This is expected as these methods just optimize for diversity and do not take query information into account. QUASC, MWAA and HDS on the other hand have fewer irrelevant frames as they use the web-images to further guide the summarization. Compared to baselines, our method generates summaries with high diversity and selects less unimportant (red bounding box) or redundant frames (yellow bounding box).

4.6 Run-Time Analysis

For completeness, we report the run-time of our model in Figure 7 for summary lengths 30, 50 and 60. We observe that we scale linearly with the number of input frames and summary length. We do not have access to any QAMVS baseline implementations to measure



(a) $F_1 = .39$, $R_{div} = .68$, $R_{rep} = .60$, $R_{query} = .65$, $R_{coh} = .29$

(b) $F_1 = .31$, $R_{div} = .67$, $R_{rep} = .63$, $R_{query} = .67$, $R_{coh} = .43$

Figure 8: Failure case from event 7 (Obama Re-election). Although, the summary constructed from the ground-truth (left) and the DeepQAMVS generated one (right) are visually and reward-wise comparable, yet there is a remarkable difference in their corresponding F1-scores.

run-times, but complexity-wise, they all scale polynomially with the number of input frames.

4.7 Limitations and Future Work

Figure 8 presents a comparison of two summaries, the ground truth summary (left, (a)) and the summary generated by our DeepQAMVS (right, (b)). While both summaries have high diversity, representativeness and query-adaptability rewards, (b) has a lower F1-score compared to (a). This showcases the limitations of (1) the F1-score as a metric to assess the summary and (2) the subjectivity of the ground truth summaries. The F1-score relies solely on the *visual overlap* between the selected frames and the ground truth using pixel-level distances, which are highly sensitive to zooming, shifting and camera angle.

In fact, Otani *et al.* [42] showed that randomly generated summaries achieve comparable or better performance to the state-of-the-art methods when evaluated using the F1-score on two SVS datasets, SumMe [18] and TVSum [61]. Note that the ground truth in their case consists of importance scores associated with every frame. Otani *et al.* [42] proposed a new evaluation protocol based on the correlation between the ranking of the estimated scores and the human-annotated ones (Kendall [27] and Spearman [83] correlation coefficients). This metric shows the expected intuition, *i.e.*, across human-annotated summaries, the correlation metric is high. In contrast, the correlation between the randomly generated and state-of-the-art summarization methods is small.

Unfortunately, this metric is not applicable to QAMVS. To see this consider the following: if the ground truth consists of importance scores, redundant frames representing an important event will have high scores across videos. Hence, a ranked list of ground truth scores contains redundant frames, which leads to a sub-optimal summary resulting in high Spearman/Kendall scores. To fix this, we believe that a metric combining *visual*, *textual* and *temporal order* overlap would lead to a better evaluation protocol. Few papers have proposed metrics based on the *textual overlap* in the past. In particular, Yeung *et al.* [69] annotated segments in videos with sentences. The ground truth and selected segments are compared using a similarity metric for text summarization (ROUGE). Textual annotation could be very expensive for QAMVS. However, recent advances in image captioning could be leveraged to automate the process. In this paper, we design a user study to assess the coherence of the produced summaries. However, user-studies are expensive, subjective and not reproducible. Instead, a ranking correlation measure between a list of textual concepts from the ordered ground truth frames and the ones from the proposed summary may serve as a better metric, similar to [42].

Beyond the evaluation metric, training to optimize for the temporal coherence still has room for improvement. Although using the proposed reward results in visually smoother transitions, it did not lead to an overall clear story in the final summary. Embedding frames/web-images in a shared vision-language domain [50] could permit to leverage advances in text summarization. Also, the field could benefit from new benchmarks with more events and shot-level text annotations to enable a wider range of techniques and evaluation metrics.

5 CONCLUSION

In this work, we present DeepQAMVS, the first end-to-end trainable model for query-aware multi-video summarization. DeepQAMVS leverages a pointer network with hierarchical attention to fuse information from video frames, web images and textual meta-data. In addition, we introduce two novel rewards that capture query-adaptability and temporal coherence. Quantitative comparisons with an extensive set of SVS and MVS baselines and thorough qualitative analysis showcase that our model can generate a temporally coherent, query-adaptive, diverse and representative summary from a collection of retrieved videos, achieving state-of-the-art results on the MVS1K dataset. QAMVS needs more community attention and research efforts to tackle the discussed limitations and therefore provide an efficient and robust technology to leverage the exponentially growing online video content.

6 ACKNOWLEDGEMENTS

This work is supported in part by the IBM-Illinois C3SR and by the National Science Foundation under Grant No. 1718221, 1801652, 2008387, 2045586.

REFERENCES

[1] Ido Arev, Hyun Soo Park, Yaser Sheikh, Jessica Hodgins, and Ariel Shamir. 2014. Automatic editing of footage from multiple social cameras. *TOG*.

[2] Irwan Bello, Hieu Pham, Quoc V. Le, Mohammad Norouzi, and Samy Bengio. 2017. Neural Combinatorial Optimization with Reinforcement Learning. In *Procs. ICLR*.

[3] Dimitrios Besiris, Andrew Makedonas, George Economou, and Spiros Fotopoulos. 2009. Combining graph connectivity & dominant set clustering for video summarization. *Multimedia Tools Appl.* (2009).

[4] J. Boyan and A. W. Moore. 2000. Learning evaluation functions to improve optimization by local search. *JMLR* (2000).

[5] R. Bunel, M. Hausknecht, J. Devlin, R. Singh, and P. Kohli. 2018. Leveraging grammar and reinforcement learning for neural program synthesis. *In Procs. ICLR* (2018).

[6] Sijia Cai, Wangmeng Zuo, Larry S. Davis, and Lei Zhang. 2018. Weakly-supervised Video Summarization using Variational Encoder-Decoder and Web Prior. In *Procs. ECCV*.

[7] Wen-Sheng Chu, Yale Song, and Alejandro Jaimes. 2015. Video co-summarization: Video summarization by visual co-occurrence. In *Procs. CVPR*.

[8] Yang Cong, Junsong Yuan, and Jiebo Luo. 2011. Towards scalable summarization of consumer videos via sparse dictionary selection. *IEEE TMM* (2011).

[9] Kevin Dale, Eli Shechtman, Shai Avidan, and Hanspeter Pfister. 2012. Multi-video browsing and summarization. In *Procs. CVPR Workshops*.

[10] Sandra Eliza Fontes de Avila, Ana Paula Brandão Lopes, Antonio da Luz Jr., and Arnaldo de Albuquerque Araújo. 2011. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters* (2011).

[11] Tamires Tessarolli de Souza Barbieri and Rudinei Goularte. 2020. Content selection criteria for news multi-video summarization based on human strategies. *IJDL* (2020).

[12] Fadi Dornaika and I Kamal Aldine. 2015. Decremental sparse modeling representative selection for prototype selection. *Pattern Recognition* (2015).

[13] Ehsan Elhamifar, Guillermo Sapiro, and Rene Vidal. 2012. See all by looking at a few: Sparse modeling for finding representative objects. In *Procs. CVPR*.

[14] Litong Feng, Ziyin Li, Zhanghui Kuang, and Wei Zhang. 2018. Extractive Video Summarizer with Memory Augmented Neural Networks. In *Procs. ACM Multimedia*.

[15] Tsu-Jui Fu, Shao-Heng Tai, and Hwann-Tzong Chen. 2019. Attentive and Adversarial Learning for Video Summarization. In *Procs. WACV*.

[16] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. 2014. Diverse Sequential Subset Selection for Supervised Video Summarization. In *Procs. NeurIPS*.

[17] Genliang Guan, Zhiyong Wang, Shaohui Mei, Max Ott, Mingyi He, and David Dagan Feng. 2014. A top-down approach for video summarization. *TOMM* (2014).

[18] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating summaries from user videos. In *Procs. ECCV*.

[19] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T.-Y. Liu, and W.-Y. Ma. 2016. Dual learning for machine translation. In *Procs. NeurIPS*.

[20] Richang Hong, Jinhui Tang, Hung-Khoon Tan, Shuicheng Yan, Chongwah Ngo, and Tat-Seng Chua. 2009. Event driven summarization for web videos. In *Procs. SIGMM workshop on Social media*.

[21] Yedid Hoshen, Gil Ben-Artzi, and Shmuel Peleg. 2014. Wisdom of the crowd in egocentric video curation. In *Procs. CVPR Workshops*.

[22] Zhong Ji, Yaru Ma, Yanwei Pang, and Xuelong Li. 2019. Query-aware sparse coding for web multi-video summarization. *Information Sciences* (2019).

[23] Zhong Ji, Kailin Xiong, Yanwei Pang, and Xuelong Li. 2019. Video summarization with attention-based encoder-decoder networks. *IEEE TCSVT* (2019).

[24] Zhong Ji, Yuanyuan Zhang, Yanwei Pang, and Xuelong Li. 2018. Hypergraph Dominant Set Based Multi-Video Summarization. *Signal Processing* (2018).

[25] Zhong Ji, Yuanyuan Zhang, Yanwei Pang, Xuelong Li, and Jing Pan. 2019. Multi-Video Summarization with Query-Dependent Weighted Archetypal Analysis. *Neurocomputing* (2019).

[26] Atsushi Kanehira, Luc Van Gool, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Viewpoint-aware Video Summarization. In *Procs. CVPR*.

[27] Maurice G Kendall. 1945. The treatment of ties in ranking problems. *Biometrika* (1945).

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

[29] Shuyue Lan, Rameswar Panda, Qi Zhu, and Amit K. Roy-Chowdhury. 2018. FFNet: Video Fast-Forwarding via Reinforcement Learning. In *Procs. CVPR*.

[30] J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky. 2016. Deep reinforcement learning for dialogue generation. *In Procs. EMNLP* (2016).

[31] Yingbo Li and Bernard Merialdo. 2010. Multi-video summarization based on AV-MMR. In *CBMI*.

[32] Yingbo Li and Bernard Merialdo. 2010. Multi-video summarization based on video-mmr. In *WIAMIS*.

[33] Yingbo Li and Bernard Merialdo. 2012. Video summarization based on balanced AV-MMR. In *Procs. ICMM*.

[34] Yingbo Li and Bernard Merialdo. 2016. Multimedia maximal marginal relevance for multi-video summarization. *Multimedia Tools and Applications* (2016).

[35] Yingbo Li, Bernard Merialdo, Mickael Rouvier, and Georges Linares. 2011. Static and dynamic video summaries. In *Procs. ICM*.

[36] C. Liang, M. Norouzi, J. Berant, Q. Le, and N. Lao. 2017. Memory augmented policy optimization for program synthesis with generalization. *In Procs. NeurIPS* (2017).

[37] Y. Liu, Y. Li, F. Yang, S. Chen, and Y. F. Wang. 2019. Learning Hierarchical Self-Attention for Video Summarization. In *Procs. ICIP*.

[38] Ansuman Mahapatra, Pankaj K Sa, Banshidhar Majhi, and Sudarshan Padhy. 2016. MVS: A multi-view video synopsis framework. *SPIC* (2016).

[39] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. 2017. Unsupervised video summarization with adversarial lstm networks. In *Procs. CVPR*.

[40] Safa Messaoud, Maghav Kumar, and Alexander G Schwing. 2020. Can We Learn Heuristics for Graphical Model Inference Using Reinforcement Learning?. In *Procs. CVPR*.

[41] Liqiang Nie, Richang Hong, Luming Zhang, Yingjie Xia, Dacheng Tao, and Nicu Sebe. 2015. Perceptual attributes optimization for multivideo summarization. *Trans. on Cybernetics* (2015).

[42] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkila. 2019. Rethinking the Evaluation of Video Summaries. In *Procs. CVPR*.

[43] Shun-Hsing Ou, Chia-Han Lee, V Srinivasa Somayazulu, Yen-Kuang Chen, and Shao-Yi Chien. 2014. On-line multi-view video summarization for wireless video sensor network. *STSP* (2014).

[44] Shun-Hsing Ou, Yu-Chen Lu, Jui-Pin Wang, Shao-Yi Chien, Shou-De Lin, Mi-Yen Yeti, Chia-Han Lee, Phillip B Gibbons, V Srinivasa Somayazulu, and Yen-Kuang Chen. 2014. Communication-efficient multi-view keyframe extraction in distributed video sensors. In *Procs. VCIP*.

[45] Gang Pan, Yaoxian Zheng, Rufei Zhang, Zhenjun Han, Di Sun, and Xingming Qu. 2019. A bottom-up summarization algorithm for videos in the wild. *EURASIP Journal on Advances in Signal Processing* (2019).

[46] Rameswar Panda, Niluthpol Chowdhury Mithun, and Amit K. Roy-Chowdhury. 2017. Diversity-Aware Multi-Video Summarization. *Trans. Image Processing* (2017).

[47] Rameswar Panda and Amit K Roy-Chowdhury. 2017. Collaborative summarization of topic-related videos. In *Procs. CVPR*.

[48] Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A Deep Reinforced Model for Abstractive Summarization. In *Procs. ICLR*.

[49] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Procs. EMNLP*.

[50] Bryan A Plummer, Matthew Brown, and Svetlana Lazebnik. 2017. Enhancing video summarization via vision-language embedding. In *Procs. ICCV*.

[51] Danila Potapov, Matthijs Douze, Zaïd Harchaoui, and Cordelia Schmid. 2014. Category-Specific Video Summarization. In *Procs. ECCV*.

[52] Yael Pritch, Alex Rav-Acha, Avital Gutman, and Shmuel Peleg. 2007. Webcam synopsis: Peeking around the world. In *Procs. ICCV*.

[53] Mrigank Rochan and Yang Wang. 2019. Video Summarization by Learning From Unpaired Data. In *Procs. CVPR*.

[54] Mrigank Rochan, Linwei Ye, and Yang Wang. 2018. Video Summarization Using Fully Convolutional Sequence Networks. In *Procs. ECCV*.

[55] Dhruva Sahrawat, Mohit Agarwal, Sanchit Sinha, Aditya Adhikary, Mansi Agarwal, Rajiv Ratn Shah, and Roger Zimmermann. 2019. Video Summarization using Global Attention with Memory Network and LSTM. In *Procs. BigMM*.

[56] A. Sharaf and H. Daumé III. 2017. Structured prediction via learning to search under bandit feedback. In *Procs. Workshop on Structured Prediction for NLP ACL*.

[57] Aidean Sharghi, Ali Borji, Chengtao Li, Tianbao Yang, and Boqing Gong. 2018. Improving sequential determinantal point processes for supervised video summarization. In *Procs. ECCV*.

[58] Aidean Sharghi, Boqing Gong, and Mubarak Shah. 2016. Query-Focused Extractive Video Summarization. In *Procs. ECCV*.

[59] Aidean Sharghi, Jacob S Laurel, and Boqing Gong. 2017. Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In *Procs. CVPR*.

[60] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Procs. ICLR*.

[61] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. Tvsum: Summarizing web videos using titles. In *Procs. CVPR*.

[62] Min Sun, Ali Farhadi, and Steven M. Seitz. 2014. Ranking Domain-Specific Highlights by Analyzing Edited Videos. In *Procs. ECCV*.

[63] Arun Balajee Vasudevan, Michael Gygli, Anna Volokitin, and Luc Van Gool. 2017. Query-adaptive Video Summarization via Quality-aware Relevance Estimation. In *Procs. MM*.

[64] Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. 2016. Order Matters: Sequence to sequence for sets. In *Procs. ICLR*.

[65] Feng Wang and Bernard Mérialdo. 2009. Multi-document video summarization. In *Procs. ICME*.

[66] Meng Wang, Richang Hong, Guangda Li, Zheng-Jun Zha, Shuicheng Yan, and Tat-Seng Chua. 2012. Event Driven Web Video Summarization by Tag Localization and Key-Shot Identification. *TMM* (2012).

[67] Huawei Wei, Bingbing Ni, Yichao Yan, Huanyu Yu, Xiaokang Yang, and Chen Yao. 2018. Video Summarization via Semantic Attended Networks. In *Procs. AAAI*.

[68] Ronald J. Williams. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine Learning* (1992).

[69] Serena Yeung, Alireza Fathi, and Li Fei-Fei. 2014. Videoset: Video summary evaluation through text. In *Procs. CVPR Egocentric Vision Workshop*.

[70] Jinhui Yuan, Huiyi Wang, Lan Xiao, Wujie Zheng, Jianmin Li, Fuzong Lin, and Bo Zhang. 2007. A formal study of shot boundary detection. *Trans. Circuits Syst. Video Techn.* (2007).

[71] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Video summarization with long short-term memory. In *Procs. ECCV*.

[72] Luming Zhang, Yue Gao, Richang Hong, Yuxing Hu, Rongrong Ji, and Qionghai Dai. 2014. Probabilistic skimlets fusion for summarizing multiple consumer landmark videos. *TM*.

[73] Luming Zhang, Yingjie Xia, Kuang Mao, He Ma, and Zhenyu Shan. 2014. An effective video summarization framework toward handheld devices. *TIE* (2014).

[74] Yujia Zhang, Michael Kampffmeyer, Xiaoguang Zhao, and Min Tan. 2019. Deep Reinforcement Learning for Query-Conditioned Video Summarization. *Applied Sciences* (2019).

[75] Yujia Zhang, Michael Kampffmeyer, Xiaoguang Zhao, and Min Tan. 2019. Dtr-gan: Dilated temporal relational adversarial network for video summarization. In *Procs. ACM TUR-C*.

[76] Yujia Zhang, Michael C. Kampffmeyer, Xiaodan Liang, Min Tan, and Eric P. Xing. 2018. Query-Conditioned Three-Player Adversarial Network for Video Summarization. In *Procs. BMVC*.

[77] Ying Zhang, He Ma, and Roger Zimmermann. 2013. Dynamic multi-video summarization of sensor-rich videos in geo-space. In *In Procs. ICMM*.

[78] Ying Zhang, Guanfeng Wang, Beomjoo Seo, and Roger Zimmermann. 2012. Multi-video summary and skim generation of sensor-rich videos in geo-space. In *Procs. MMSys*.

[79] Ying Zhang and Roger Zimmermann. 2016. Efficient summarization from multiple georeferenced user-generated videos. *TM* (2016).

[80] Bin Zhao and Eric P Xing. 2014. Quasi real-time summarization for consumer videos. In *Procs. CVPR*.

[81] Kaiyang Zhou, Yu Qiao, and Tao Xiang. 2018. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Procs. AAAI*.

[82] Jianqing Zhu, Shengcai Liao, and Stan Z Li. 2015. Multicamera joint video synopsis. *TCSVT* (2015).

[83] Daniel Zwillinger and Stephen Kokoska. 1999. *CRC standard probability and statistics tables and formulae.* CRC Press.