Weakly-Supervised Video Summarization Using Variational Encoder-Decoder and Web Prior

Sijia Cai[1,2], Wangmeng Zuo[3], Larry S. Davis[4], and Lei Zhang[1](✉)

[1] Department of Computing, The Hong Kong Polytechnic University,
Kowloon, Hong Kong
{csscai,cslzhang}@comp.polyu.edu.hk
[2] DAMO Academy, Alibaba Group, Hangzhou, China
[3] School of Computer Science and Technology, Harbin Institute of Technology,
Harbin, China
cswmzuo@gmail.com
[4] Department of Computer Science, University of Maryland, College Park, USA
lsd@umiacs.umd.edu

Abstract. Video summarization is a challenging under-constrained problem because the underlying summary of a single video strongly depends on users' subjective understandings. Data-driven approaches, such as deep neural networks, can deal with the ambiguity inherent in this task to some extent, but it is extremely expensive to acquire the temporal annotations of a large-scale video dataset. To leverage the plentiful web-crawled videos to improve the performance of video summarization, we present a generative modelling framework to learn the latent semantic video representations to bridge the benchmark data and web data. Specifically, our framework couples two important components: a variational autoencoder for learning the latent semantics from web videos, and an encoder-attention-decoder for saliency estimation of raw video and summary generation. A loss term to learn the semantic matching between the generated summaries and web videos is presented, and the overall framework is further formulated into a unified conditional variational encoder-decoder, called variational encoder-summarizer-decoder (VESD). Experiments conducted on the challenging datasets CoSum and TVSum demonstrate the superior performance of the proposed VESD to existing state-of-the-art methods. The source code of this work can be found at https://github.com/cssjcai/vesd.

Keywords: Video summarization · Variational autoencoder

1 Introduction

Recently, it has been attracting much interest in extracting the representative visual elements from a video for sharing on social media, which aims to effectively

This research is supported by the Hong Kong RGC GRF grant (PolyU 152135/16E) and the City Brain project of DAMO Academy, Alibaba Group.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11218, pp. 193–210, 2018.
https://doi.org/10.1007/978-3-030-01264-9_12

express the semantics of the original lengthy video. However, this task, often referred to as video summarization, is laborious, subjective and challenging since videos usually exhibit very complex semantic structures, including diverse scenes, objects, actions and their complex interactions.

A noticeable trend appeared in recent years is to use the deep neural networks (DNNs) [10,44] for video summarization since DNNs have made significant progress in various video understanding tasks [2,12,19]. However, annotations used in the video summarization task are in the form of frame-wise labels or importance scores, collecting a large number of annotated videos demands tremendous effort and cost. Consequently, the widely-used benchmark datasets [1,31] only cover dozens of well-annotated videos, which becomes a prominent stumbling block that hinders the further improvement of DNNs based summarization techniques. Meanwhile, annotations for summarization task are subjective and not consistent across different annotators, potentially leading to overfitting and biased models. Therefore, the advanced studies toward taking advantage of augmented data sources such as web images [13], GIFs [10] and texts [23], which are complimentary for the summarization purpose.

To drive the techniques along with this direction, we consider an efficient weakly-supervised setting of learning summarization models from a vast number of web videos. Compared with other types of auxiliary source domain data for video summarization, the temporal dynamics in these user-edited "templates" offer rich information to locate the diverse but semantic-consistent visual contents which can be used to alleviate the ambiguities in small-size summarization. These short-form videos are readily available from web repositories (*e.g.*, YouTube) and can be easily collected using a set of topic labels as search keywords. Additionally, these web videos have been edited by a large community of users, the risk of building a biased summarization model is significantly reduced. Several existing works [1,21] have explored different strategies to exploit the semantic relatedness between web videos and benchmark videos. So motivated, we aim to effectively utilize the large collection of weakly-labelled web videos in learning more accurate and informative video representations which: (i) preserve essential information within the raw videos; (ii) contain discriminative information regarding the semantic consistency with web videos. Therefore, the desired deep generative models are necessitated to capture the underlying latent variables and make practical use of web data and benchmark data to learn abstract and high-level representations.

To this end, we present a generative framework for summarizing videos in this paper, which is illustrated in Fig. 1. The basic architecture consists of two components: a variational autoencoder (VAE) [14] model for learning the latent semantics from web videos; and a sequence encoder-decoder with attention mechanism for summarization. The role of VAE is to map the videos into a continuous latent variable, via an inference network (encoder), and then use the generative network (decoder) to reconstruct the input videos conditioned on samples from the latent variable. For the summarization component, the association is temporally ambiguous since only a subset of fragments in the raw video is relevant to its summary semantics. To filter out the irrelevant fragments and identify informative temporal regions for the better summary generation, we exploit the soft attention mechanism where the attention vectors (*i.e.*, context representations) of raw videos are obtained by integrating the latent semantics trained from web videos. Furthermore, we provide a weakly-supervised semantic matching loss instead of reconstruction loss to learn the topic-associated summaries in our generative framework. In this sense, we take advantage of potentially accurate and flexible latent variable distribution from external data thus strengthen the expressiveness of generated summary in the encoder-decoder based summarization model. To evaluate the effectiveness of the proposed method, we comprehensively conduct experiments using different training settings and demonstrate that our method with web videos achieves significantly better performance than competitive video summarization approaches.

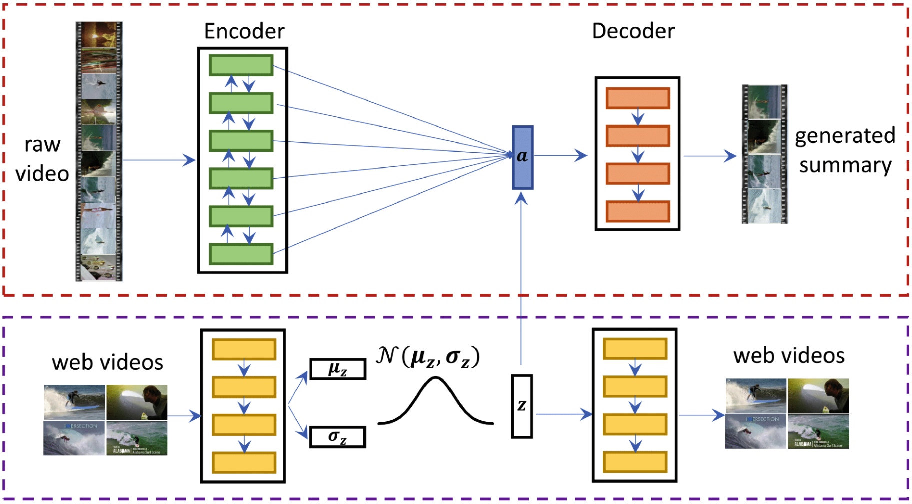


Fig. 1. An illustration of the proposed generative framework for video summarization. A VAE model is pre-trained on web videos (purple dashed rectangle area); And the summarization is implemented within an encoder-decoder paradigm by using both the attention vector and the sampled latent variable from VAE (red dashed rectangle area). (Color figure online)

2 Related Work

Video Summarization is a challenging task which has been explored for many years [18,37] and can be grouped into two broad categories: unsupervised and supervised learning methods. Unsupervised summarization methods focus on low-level visual cues to locate the important segments of a video. Various strategies have been investigated, including clustering [7,8], sparse optimizations [3,22], and energy minimization [4,25]. A majority of recent works mainly study the summarization solutions based on the supervised learning from human annotations. For instance, to make a large-margin structured prediction, submodular functions are trained with human-annotated summaries [9]. Gygli *et al.* [8] propose a linear regression model to estimate the interestingness score of shots. Gong *et al.* [5] and Sharghi *et al.* [28] learn from user-created summaries for selecting informative video subsets. Zhang *et al.* [43] show summary structures can be transferred between videos that are semantically consistent. More recently, DNNs based methods have been applied for video summarization with the help of pairwise deep ranking model [42] or recurrent neural networks (RNNs) [44]. However, these approaches assume the availability of a large number of human-created video-summary pairs or fine-grained temporal annotations, which are in practice difficult and expensive to acquire. Alternatively, there have been attempts to leverage information from other data sources such as web images, GIFs and texts [10,13,23]. Chu *et al.* [1] propose to summarize shots that co-occur among multiple videos of the same topic. Panda *et al.* [20] present an end-to-end 3D convolutional neural network (CNN) architecture to learn summarization model with web videos. In this paper, we also consider to use the topic-specific cues in web videos for better summarization, but adopt a generative summarization framework to exploit the complementary benefits in web videos.

Video Highlight Detection is highly related to video summarization and many earlier approaches have primarily been focused on specific data scenarios such as broadcast sport videos [27,35]. Traditional methods usually adopt the mid-level and high-level audio-visual features due to the well-defined structures. For general highlight detection, Sun *et al.* [32] employ a latent SVM model detect highlights by learning from pairs of raw and edited videos. The DNNs also have achieved big performance improvement and shown great promise in highlight detection [41]. However, most of these methods treat highlight detection as a binary classification problem, while highlight labelling is usually ambiguous for humans. This also imposes heavy burden for humans to collect a huge amount of labelled data for training DNN based models.

Deep Generative Models are very powerful in learning complex data distribution and low-dimensional latent representations. Besides, the generative modelling for video summarization might provide an effective way to bring scalability and stability in training a large amount of web data. Two of the most effective approaches are VAE [14] and generative adversarial network (GAN) [6]. VAE aims at maximizing the variational lower bound of the observation while encouraging the variational posterior distribution of the latent variables to be close to the prior distribution. A GAN is composed of a generative model and a discriminative model and trained in a min-max game framework. Both VAE and GAN have already shown promising results in image/frame generation tasks [17,26,38]. To embrace the temporal structures into generative modelling, we propose a new variational sequence-to-sequence encoder-decoder framework for video summarization by capturing both the video-level topics and web semantic prior. The attention mechanism embedded in our framework can be naturally used as key shots selection for summarization. Most related to our generative summarization is the work of Mahasseni *et al.* [16], who present an unsupervised summarization in the framework of GAN. However, the attention mechanism in their approach depends solely on the raw video itself thus has the limitation in delivering diverse contents in video-summary reconstruction.

3 The Proposed Framework

As an intermediate step to leverage abundant user-edited videos on the Web to assist the training of our generative video summarization framework, in this section, we first introduce the basic building blocks of the proposed framework, called variational encoder-summarizer-decoder (VESD). The VESD consists of three components: (i) an encoder RNN for raw video; (ii) an attention-based summarizer for raw video; (iii) a decoder RNN for summary video.

Following the video summarization pipelines in previous methods [24,44], we first perform temporal segmentation and shot-level feature extraction for raw videos using CNNs. Each video $\mathcal{X}$ is then treated as a sequential set of multiple non-uniform shots, where $\boldsymbol{x}_t$ is the feature vector of the t-th shot in video representation $\boldsymbol{X}$. Most supervised summarization approaches aim to predict labels/scores which indicate whether the shots should be included in the summary, however, suffering from the drawbacks of selection of redundant visual contents. For this reason, we formulate video summarization as video generation task which allows the summary representation $\boldsymbol{Y}$ does not necessarily be restricted to a subset of $\boldsymbol{X}$. In this manner, our method centres on the semantic essence of a video and can exhibit the high tolerance for summaries with visual differences. Following the encoder-decoder paradigm [33], our summarization framework is composed of two parts: the encoder-summarizer is an inference network $q_\phi(\boldsymbol{a}|\boldsymbol{X}, \boldsymbol{z})$ that takes both the video representation $\boldsymbol{X}$ and the latent variable $\boldsymbol{z}$ (sampled from the VAE module pre-trained on web videos) as inputs. Moreover, the encoder-summarizer is supposed to generate the video content representation $\boldsymbol{a}$ that captures all the information about $\boldsymbol{Y}$. The summarizer-decoder is a generative network $p_{\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{a}, \boldsymbol{z})$ that outputs the summary representation $\boldsymbol{Y}$ based on the attention vector $\boldsymbol{a}$ and the latent representation $\boldsymbol{z}$.

3.1 Encoder-Summarizer

To date, modelling sequence data with RNNs has been proven successful in video summarization [44]. Therefore, for the encoder-summarizer component, we employ a pointer RNN, *e.g.*, a bidirectional Long Short-Term Memory (LSTM), as an encoder that processes the raw videos, and a summarizer aims to select the shots of most probably containing salient information. The summarizer is exactly the attention-based model that generates the video context representation by attending to the encoded video features.

In time step t, we denote $\boldsymbol{x}_t$ as the feature vector for the t-th shot and $\boldsymbol{h}_t^e$ as the state output of the encoder. It is known that $\boldsymbol{h}_t^e$ is obtained by concatenating the hidden states from each direction:

$$\boldsymbol{h}_t^e = [\mathrm{RNN}_{\overrightarrow{enc}}(\overrightarrow{\boldsymbol{h}_{t-1}}, \boldsymbol{x}_t); \mathrm{RNN}_{\overleftarrow{enc}}(\overleftarrow{\boldsymbol{h}_{t+1}}, \boldsymbol{x}_t)]. \tag{1}$$

The attention mechanism is proposed to compute an attention vector $\boldsymbol{a}$ of input sequence by summing the sequence information $\{\boldsymbol{h}_t^e, t = 1, \ldots, |\boldsymbol{X}|\}$ with the location variable $\boldsymbol{\alpha}$ as follows:

$$\boldsymbol{a} = \sum_{t=1}^{|\boldsymbol{X}|} \alpha_t \boldsymbol{h}_t^e, \tag{2}$$

where α_t denotes the t-th value of $\boldsymbol{\alpha}$ and indicates whether the t-th shot is included in summary or not. As mentioned in [40], when using the generative modelling on the log-likelihood of the conditional distribution $p(\boldsymbol{Y}|\boldsymbol{X})$, one approach is to sample attention vector $\boldsymbol{a}$ by assigning the Bernoulli distribution to $\boldsymbol{\alpha}$. However, the resultant Monte Carlo gradient estimator of the variational lower-bound objective requires complicated variance reduction techniques and may lead to unstable training. Instead, we adopt a deterministic approximation to obtain $\boldsymbol{a}$. That is, we produce an attentive probability distribution based on $\boldsymbol{X}$ and $\boldsymbol{z}$, which is defined as $\alpha_t := p(\alpha_t|\boldsymbol{h}_t^e, \boldsymbol{z}) = \mathrm{softmax}(\varphi_t([\boldsymbol{h}_t^e; \boldsymbol{z}]))$, where $\boldsymbol{\varphi}$ is a parameterized potential typically based on a neural network, *e.g.*, multilayer perceptron (MLP). Accordingly, the attention vector in Eq. (2) turns to:

$$\boldsymbol{a} = \sum_{t=1}^{N} p(\alpha_t|\boldsymbol{h}_t^e, \boldsymbol{z}) \boldsymbol{h}_t^e, \tag{3}$$

which is fed to the decoder RNN for summary generation. The attention mechanism extracts an attention vector $\boldsymbol{a}$ by iteratively attending to the raw video features based on the latent variable $\boldsymbol{z}$ learned from web data. In doing so the model is able to adapt to the ambiguity inherent in summaries and obtain salient information of raw video through attention. Intuitively, the attention scores α_ts are used to perform shot selection for summarization.

3.2 Summarizer-Decoder

We specify the summary generation process as $p_{\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{a}, \boldsymbol{z})$ which is the conditional likelihood of the summary given the attention vector $\boldsymbol{a}$ and the latent variable $\boldsymbol{z}$. Different with the standard Gaussian prior distribution adopted in VAE, $p(\boldsymbol{z})$ in our framework is pre-trained on web videos to regularize the latent semantic representations of summaries. Therefore, the summaries generated via $p_{\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{a}, \boldsymbol{z})$ are likely to possess diverse contents. In this manner, $p_{\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{a}, \boldsymbol{z})$ is then reconstructed via a RNN decoder at each time step t: $p_{\boldsymbol{\theta}}(\boldsymbol{y}_t|\boldsymbol{a}, [\boldsymbol{\mu}_{\boldsymbol{z}}, \boldsymbol{\sigma}_{\boldsymbol{z}}^2])$, where $\boldsymbol{\mu}_{\boldsymbol{z}}$ and $\boldsymbol{\sigma}_{\boldsymbol{z}}$ are nonlinear functions of the latent variables specified by two learnable neural networks (detailed in Sect. 4).

3.3 Variational Inference

Given the proposed VESD model, the network parameters $\{\boldsymbol{\phi}, \boldsymbol{\theta}\}$ need to be updated during inference. We marginalize over the latent variables $\boldsymbol{a}$ and $\boldsymbol{z}$ by maximizing the following variational lower-bound $\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta})$

$$\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta}) = \mathbb{E}_{q_{\phi}(\boldsymbol{a},\boldsymbol{z}|\boldsymbol{X},\boldsymbol{Y})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{a},\boldsymbol{z}) - \mathrm{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{a},\boldsymbol{z}|\boldsymbol{X},\boldsymbol{Y})|p(\boldsymbol{a},\boldsymbol{z}))], \qquad (4)$$

where $\mathrm{KL}(\cdot)$ is the Kullback-Leibler divergence. We assume the joint distribution of the latent variables $\boldsymbol{a}$ and $\boldsymbol{z}$ has a factorized form, *i.e.*, $q_{\boldsymbol{\phi}}(\boldsymbol{a},\boldsymbol{z}|\boldsymbol{X},\boldsymbol{Y}) = q_{\boldsymbol{\phi}^{(z)}}(\boldsymbol{z}|\boldsymbol{X},\boldsymbol{Y})q_{\boldsymbol{\phi}^{(a)}}(\boldsymbol{a}|\boldsymbol{X},\boldsymbol{Y})$, and notice that $p(\boldsymbol{a}) = q_{\boldsymbol{\phi}^{(a)}}(\boldsymbol{a}|\boldsymbol{X},\boldsymbol{Y})$ is defined with a deterministic manner in Sect. 3.1. Therefore the variational objective in Eq. (4) can be derived as:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta}) &= \mathbb{E}_{q_{\phi^{(z)}}(\boldsymbol{z}|\boldsymbol{X},\boldsymbol{Y})}[\mathbb{E}_{q_{\phi^{(a)}}(\boldsymbol{a}|\boldsymbol{X},\boldsymbol{Y})}\log p_{\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{a},\boldsymbol{z}) \\ &\quad -\mathrm{KL}(q_{\boldsymbol{\phi}^{(a)}}(\boldsymbol{a}|\boldsymbol{X},\boldsymbol{Y})||p(\boldsymbol{a}))] + \mathrm{KL}(q_{\boldsymbol{\phi}^{(z)}}(\boldsymbol{z}|\boldsymbol{X},\boldsymbol{Y})||p(\boldsymbol{z})) \\ &= \mathbb{E}_{q_{\phi}(\boldsymbol{z}|\boldsymbol{X},\boldsymbol{Y})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{a},\boldsymbol{z})] + \mathrm{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{X},\boldsymbol{Y})||p(\boldsymbol{z})). \qquad (5)\end{aligned}$$

The above variational lower-bound offers a new perspective for exploiting the reciprocal nature of raw video and its summary. Maximizing Eq. (5) strikes a balance between minimizing generation error and minimizing the KL divergence between the approximated posterior $q_{\boldsymbol{\phi}^{(z)}}(\boldsymbol{z}|\boldsymbol{X},\boldsymbol{Y})$ and the prior $p(\boldsymbol{z})$.

4 Weakly-Supervised VESD

In practice, as only a few video-summary pairs are available, the latent variable $\boldsymbol{z}$ cannot characterize the inherent semantic in video and summary accurately. Motivated by the VAE/GAN model [15], we explore a weakly-supervised learning framework and endow our VESD the ability to make use of rich web videos for the latent semantic inference. The VAE/GAN model extends VAE with the discriminator network in GAN, which provides a method that constructs the latent space from inference network of data rather than random noises and implicitly learns a rich similarity metric for data. The similar idea has also been investigated in [16] for unsupervised video summarization. Recall that the discriminator in GAN tries to distinguish the generated examples from real examples; Following the same spirit, we apply the discriminator in the proposed VESD which naturally results in minimizing the following adversarial loss function:

$$\mathcal{L}(\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}) = -\mathbb{E}_{\hat{\boldsymbol{Y}}}[\log \mathrm{D}_{\psi}(\hat{\boldsymbol{Y}})] - \mathbb{E}_{\boldsymbol{X},\boldsymbol{z}}[\log(1-\mathrm{D}_{\psi}(\boldsymbol{Y}))], \qquad (6)$$

where $\hat{\boldsymbol{Y}}$ refers to the representation of web video. Unfortunately, the above loss function suffers from the unstable training in standard GAN models and cannot be directly extended into supervised scenario. To address these problems, we propose to employ a semantic feature matching loss for the weakly-supervised setting of VESD framework. The objective requires the representation of generated summary to match the representation of web videos under a similarity function. For the prediction of the semantic similarity, we replace $p_{\boldsymbol{\theta}}(\boldsymbol{Y}|\boldsymbol{a}, \boldsymbol{z})$ with the following sigmoid function:

$$p_{\boldsymbol{\theta}}(c|\boldsymbol{a}, \boldsymbol{h}^d(\hat{\boldsymbol{Y}})) = \sigma(\boldsymbol{a}^T \boldsymbol{M} \boldsymbol{h}^d(\hat{\boldsymbol{Y}})), \tag{7}$$

where $\boldsymbol{h}^d(\hat{\boldsymbol{Y}})$ is the last output state of $\hat{\boldsymbol{Y}}$ in the decoder RNN and $\boldsymbol{M}$ is the sigmoid parameter. We randomly pick $\hat{\boldsymbol{Y}}$ in web videos and c is the pair relatedness label, *i.e.*, $c = 1$ if $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ are semantically matched. We can also generalize the above matching loss to multi-label case by replacing c with one-hot vector $\boldsymbol{c}$ whose nonzero position corresponds the matched label. Therefore, the objective (5) can be rewritten as:

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\psi}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{c}|\boldsymbol{a}, \boldsymbol{h}^d(\hat{\boldsymbol{Y}}))] + \mathrm{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z})||p(\boldsymbol{z}|\hat{\boldsymbol{Y}})). \tag{8}$$

It is found that the above variational objective shares the similarity with conditional VAE (CVAE) [30] which is able to produce diverse outputs for a single input. For example, Walker *et al.* [39] use a fully convolutional CVAE for diverse motion prediction from a static image. Zhou and Berg [45] generate diverse timelapse videos by incorporating conditional, twostack and recurrent architecture modifications to standard generative models. Therefore, our weakly-supervised VESD naturally embeds the diversity in video summary generation.

4.1 Learnable Prior and Posterior

In contrast to the standard VAE prior that assumes the latent variable $\boldsymbol{z}$ to be drawn from latent Gaussian (*e.g.*, $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$), we impose the prior distribution learned from web videos which infers the topic-specific semantics more accurately. Thus we impose $\boldsymbol{z}$ to be drawn from the Gaussian with $p(\boldsymbol{z}|\hat{\boldsymbol{Y}}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}(\hat{\boldsymbol{Y}}), \boldsymbol{\sigma}^2(\hat{\boldsymbol{Y}})\boldsymbol{I})$ whose mean and variance are defined as:

$$\boldsymbol{\mu}(\hat{\boldsymbol{Y}}) = f_{\boldsymbol{\mu}}(\hat{\boldsymbol{Y}}), \log \boldsymbol{\sigma}^2(\hat{\boldsymbol{Y}}) = f_{\boldsymbol{\sigma}}(\hat{\boldsymbol{Y}}), \tag{9}$$

where $f_{\boldsymbol{\mu}}(\cdot)$ and $f_{\boldsymbol{\sigma}}(\cdot)$ denote any type of neural networks that are suitable for the observed data. We adopt two-layer MLPs with ReLU activation in our implementation.

Likewise, we model the posterior of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\cdot) := q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{X}, \hat{\boldsymbol{Y}}, \boldsymbol{c})$ with the Gaussian distribution $\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}(\boldsymbol{X}, \hat{\boldsymbol{Y}}, \boldsymbol{c}), \boldsymbol{\sigma}^2(\boldsymbol{X}, \hat{\boldsymbol{Y}}, \boldsymbol{c})$ whose mean and variance are also characterized by two-layer MLPs with ReLU activation:

$$\boldsymbol{\mu} = f_{\boldsymbol{\mu}}([\boldsymbol{a}; \boldsymbol{h}^d(\hat{\boldsymbol{Y}}); \boldsymbol{c}]), \log \boldsymbol{\sigma}^2 = f_{\boldsymbol{\sigma}}([\boldsymbol{a}; \boldsymbol{h}^d(\hat{\boldsymbol{Y}}); \boldsymbol{c}]). \tag{10}$$

4.2 Mixed Training Objective Function

One potential issue of purely weakly-supervised VESD training objective (8) is that the semantic matching loss usually results in summaries focusing on very few shots in raw video. To ensure the diversity and fidelity of the generated

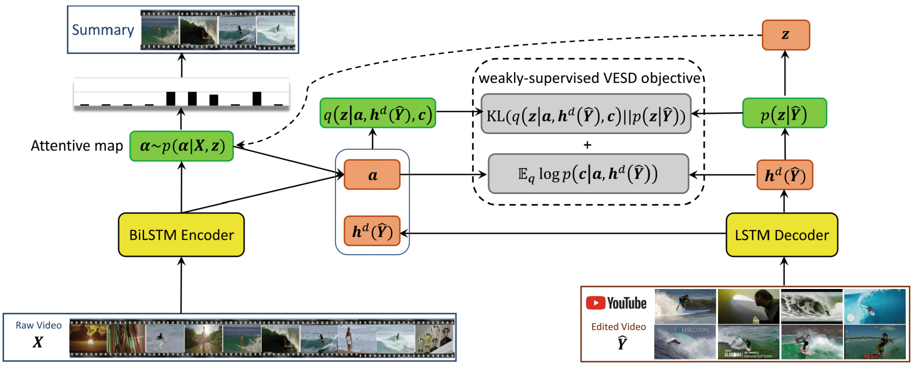


Fig. 2. The variational formulation of our weakly-supervised VESD framework.

summaries, we can also make use of the importance scores on partially finely-annotated benchmark datasets to consistently improves performance. For those detailed annotations in benchmark datasets, we adopt the same keyframe regularizer in [16] to measure the cross-entropy loss between the normalized ground-truth importance scores $\boldsymbol{\alpha}_{\boldsymbol{X}}^{gt}$ and the output attention scores $\boldsymbol{\alpha}_{\boldsymbol{X}}$ as below:

$$\mathcal{L}_{\text{score}} = \text{cross-entropy}(\boldsymbol{\alpha}_{\boldsymbol{X}}^{gt}, \boldsymbol{\alpha}_{\boldsymbol{X}}). \tag{11}$$

Accordingly, we train the regularized VESD using the following objective function to utilize different levels of annotations:

$$\mathcal{L}_{\text{mixed}} = \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\omega}) + \lambda \mathcal{L}_{\text{score}}. \tag{12}$$

The overall objective can be trained using back-propagation efficiently and is illustrated in Fig. 2. After training, we calculate the salience score $\boldsymbol{\alpha}$ for each new video by forward passing the summarization model in VESD.

5 Experimental Results

Datasets and Evaluation. We test our VESD framework on two publicly available video summarization benchmark datasets CoSum [1] and TVSum [31]. The CoSum [1] dataset consists of 51 videos covering 10 topics including Base Jumping (BJ), Bike Polo (BP), Eiffel Tower (ET), Excavators River Cross (ERC), Kids Playing in leaves (KP), MLB, NFL, Notre Dame Cathedral (NDC), Statue of Liberty (SL) and SurFing (SF). The TVSum [31] dataset contains 50 videos organized into 10 topics from the TRECVid Multimedia Event Detection task [29], including changing Vehicle Tire (VT), getting Vehicle Unstuck (VU), Grooming an Animal (GA), Making Sandwich (MS), ParKour (PK), PaRade (PR), Flash Mob gathering (FM), BeeKeeping (BK), attempting Bike Tricks (BT), and Dog Show (DS). Following the literature [9,44], we randomly choose 80% of the videos for training and use the remaining 20% for testing on both datasets.

As recommended by [1,20,21], we evaluate the quality of a generated summary by comparing it to multiple user-annotated summaries provided in benchmarks. Specifically, we compute the pairwise average precision (AP) for a proposed summary and all its corresponding human-annotated summaries, and then report the mean value. Furthermore, we average over the number of videos to achieve the overall performance on a dataset. For the CoSum dataset, we follow [20,21] and compare each generated summary with three human-created summaries. For the TVSum dataset, we first average the frame-level importance scores to compute the shot-level scores, and then select the top 50% shots for each video as the human-created summary. Finally, each generated summary is compared with twenty human-created summaries. The top-5 and top-15 mAP performances on both datasets are presented in evaluation.

Web Video Collection. This section describes the details of web video collection for our approach. We treat the topic labels in both datasets as the query keywords and retrieve videos from YouTube for all the twenty topic categories. We limit the videos by time duration (less than 4 min) and rank by relevance to constructing a set of weakly-annotated videos. However, these downloaded videos are still very lengthy and noisy in general since they contain a proportion of frames that are irrelevant to search keywords. Therefore, we introduce a simple but efficient strategy to filter out the noisy parts of these web videos: (1) we first adopt the existing temporal segmentation technique KTS [24] to segment both the benchmark videos and web videos into non-overlapping shots, and utilize CNNs to extract feature within each shot; (2) the corresponding features in benchmark videos are then used to train a MLP with their topic labels (the shots do not belong to any topic label are set with background label) and perform prediction for the shots in web videos; (3) we further truncate web videos based on the relevant shots whose topic-related probability is larger than a threshold. In this way, we observe that the trimmed videos are sufficiently clean and informative for learning the latent semantics in our VAE module.

Architecture and Implementation Details. For the fair comparison with state-of-the-art methods [16,44], we choose to use the output of pool5 layer of the GoogLeNet [34] for the frame-level feature. The shot-level feature is then obtained by averaging all the frame features within a shot. We first use the features of segmented shots on web videos to pre-train a VAE module whose dimension of the latent variable is set to 256. To build encoder-summarizer-decoder, we use a two-layer bidirectional LSTM with 1024 hidden units, a two-layer MLP with [256, 256] hidden units and a two-layer LSTM with 1024 hidden units for the encoder RNN, attention MLP and decoder RNNs, respectively. For the parameter initialization, we train our framework from scratch using stochastic gradient descent with a minibatch size of 20, a momentum of 0.9, and a weight decay of 0.005. The learning rate is initialized to 0.01 and is reduced to its 1/10 after every 20 epochs (100 epochs in total). The trade-off parameter λ is set to 0.2 in the mixed training objective.

5.1 Quantitative Results

Exploration Study. To better understand the impact of using web videos and different types of annotations in our method, we analyzed the performances under the following six training settings: (1) benchmark datasets with weak supervision (topic labels); (2) benchmark datasets with weak supervision and extra 30 downloaded videos per topic; (3) benchmark datasets with weak supervision and extra 60 downloaded videos per topic; (4) benchmark datasets with strong supervision (topic labels and importance scores); (5) benchmark datasets with strong supervision and extra 30 downloaded videos per topic; and (6) benchmark datasets with strong supervision and extra 60 downloaded videos per topic. We have the following key observations from Table 1: (1) Training on the benchmark data with only weak topic labels in our VESD framework performs much worse than either that of training using extra web videos or that of training using detailed importance scores, which demonstrates our generative summarization model demands a larger amount of annotated data to perform well. (2) We notice that the more web videos give better results, which clearly demonstrates the benefits of using web videos and proves the scalability of our generative framework. (3) This big improvements with strong supervision illustrate the positive impact of incorporating available importance scores for mixed training of our VESD. That is not surprising since the attention scores should be imposed to focus on different fragments of raw videos in order to be consistent with ground-truths, resulting in the summarizer with the diverse property which is an important metric in generating good summaries. We use the training setting (5) in the following experimental comparisons.

Table 1. Exploration study on training settings. Numbers show top-5 mAP scores.

Training settings	CoSum	TVSum
Benchmark with weak supervision	0.616	0.352
Benchmark with weak supervision + 30 web videos/topic	0.684	0.407
Benchmark with weak supervision + 60 web videos/topic	0.701	0.423
Benchmark with strong supervision	0.712	0.437
Benchmark with strong supervision + 30 web videos/topic	0.755	0.481
Benchmark with strong supervision + 60 web videos/topic	0.764	0.498

Effect of Deep Feature. We also investigate the effect of using different types of deep features as shot representation in VESD framework, including 2D deep features extracted from GoogLeNet [34] and ResNet101 [11], and 3D deep features extracted from C3D [36]. In Table 2, we have following observations: (1) ResNet produces better results than GoogLeNet, with a top-5 mAP score improvement of 0.012 on the CoSum dataset, which indicates more powerful visual features still lead improvement for our method. We also compare

Table 2. Performance comparison using different types of features on CoSum dataset. Numbers show top-5 mAP scores averaged over all the videos of the same topic.

Feature	BJ	BP	ET	ERC	KP	MLB	NFL	NDC	SL	SF	Top-5
GoogLeNet	0.715	0.746	0.813	0.756	0.772	0.727	0.737	0.782	0.794	0.709	0.755
ResNet101	0.727	0.755	0.827	0.766	0.783	0.741	0.752	0.790	0.807	0.722	0.767
C3D	0.729	0.754	0.831	0.761	0.779	0.740	0.747	0.785	0.805	0.718	0.765

2D GoogLeNet features with C3D features. Results show that the C3D features achieve better performance over GoogLeNet features (0.765 vs 0.755) and comparable performance with ResNet101 features. We believe this is because C3D features exploit the temporal information of videos thus are also suitable for summarization.

Table 3. Experimental results on CoSum dataset. Numbers show top-5/15 mAP scores averaged over all the videos of the same topic.

Topic	Unsupervised methods					Supervised methods					VESD
	SMRS	Quasi	MBF	CVS	SG	KVS	DPP	sLstm	SM	DSN	
BJ	0.504	0.561	0.631	0.658	0.698	0.662	0.672	0.683	0.692	0.685	**0.715**
BP	0.492	0.625	0.592	0.675	0.713	0.674	0.682	0.701	0.722	0.714	**0.746**
ET	0.556	0.575	0.618	0.722	0.759	0.731	0.744	0.749	0.789	0.783	**0.813**
ERC	0.525	0.563	0.575	0.693	0.729	0.685	0.694	0.717	0.728	0.721	**0.756**
KP	0.521	0.557	0.594	0.707	0.729	0.701	0.705	0.714	0.745	0.742	**0.772**
MLB	0.543	0.563	0.624	0.679	0.721	0.668	0.677	0.714	0.693	0.687	**0.727**
NFL	0.558	0.587	0.603	0.674	0.693	0.671	0.681	0.681	0.727	0.724	**0.737**
NDC	0.496	0.617	0.595	0.702	0.738	0.698	0.704	0.722	0.759	0.751	**0.782**
SL	0.525	0.551	0.602	0.715	0.743	0.713	0.722	0.721	0.766	0.763	**0.794**
SF	0.533	0.562	0.594	0.647	0.681	0.642	0.648	0.653	0.683	0.674	**0.709**
Top-5	**0.525**	**0.576**	**0.602**	**0.687**	**0.720**	**0.684**	**0.692**	**0.705**	**0.735**	**0.721**	**0.755**
Top-15	**0.547**	**0.591**	**0.617**	**0.699**	**0.731**	**0.702**	**0.711**	**0.717**	**0.746**	**0.736**	**0.764**

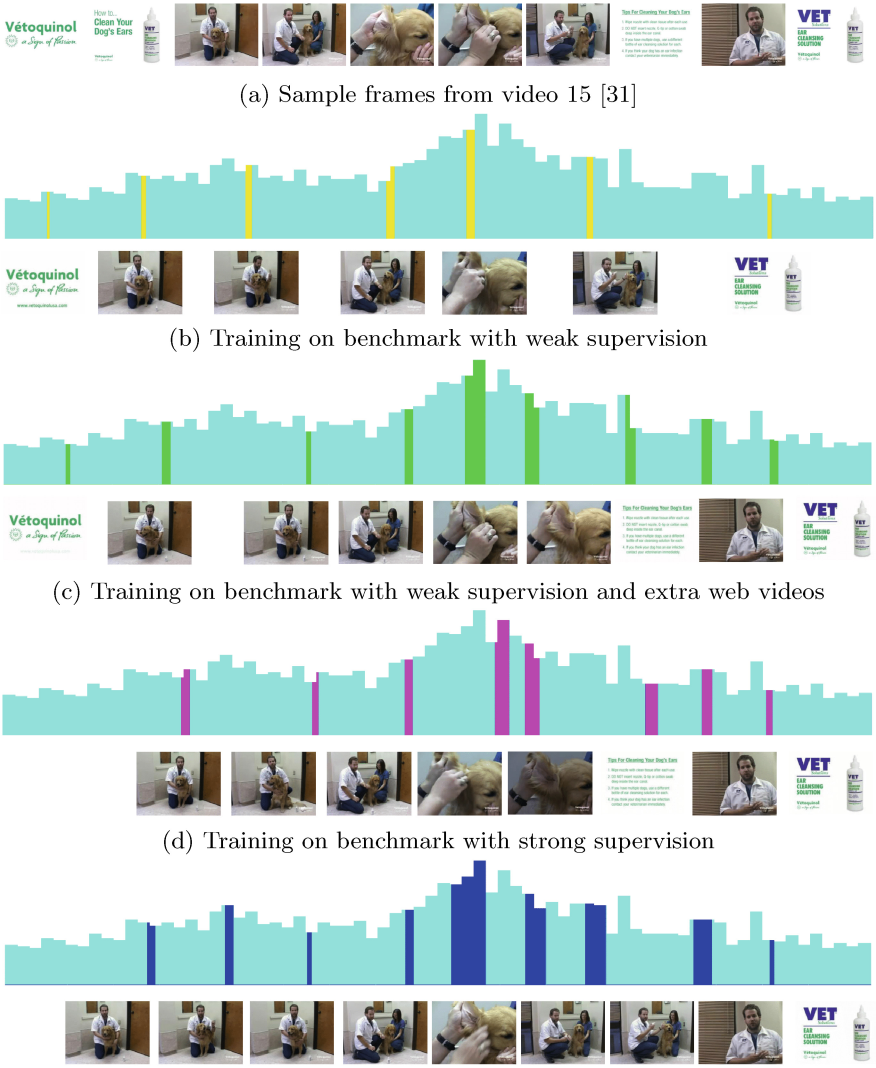
Comparison with Unsupervised Methods. We first compare VESD with several unsupervised methods including SMRS [3], Quasi [13], MBF [1], CVS [21] and SG [16]. Table 3 shows the mean AP on both top 5 and 15 shots included in the summaries for the CoSum dataset, whereas Table 4 shows the results on TVSum dataset. We can observe that: (1) Our weakly supervised approach obtains the highest overall mAP and outperforms traditional non-DNN based methods SMRS, Quasi, MBF and CVS by large margins. (2) The most competing DNN based method, SG [16] gives top-5 mAP that is 3.5% and 1.9% less than

Table 4. Experimental results on TVSum dataset. Numbers show top-5/15 mAP scores averaged over all the videos of the same topic.

Topic	Unsupervised methods					Supervised methods					VESD
	SMRS	Quasi	MBF	CVS	SG	KVS	DPP	sLstm	SM	DSN	
VT	0.272	0.336	0.295	0.328	0.423	0.353	0.399	0.411	0.415	0.373	**0.447**
VU	0.324	0.369	0.357	0.413	0.472	0.441	0.453	0.462	0.467	0.441	**0.493**
GA	0.331	0.342	0.325	0.379	0.475	0.402	0.457	0.463	0.469	0.428	**0.496**
MS	0.362	0.375	0.412	0.398	0.489	0.417	0.462	0.477	0.478	0.436	**0.503**
PK	0.289	0.324	0.318	0.354	0.456	0.382	0.437	0.448	0.445	0.411	**0.478**
PR	0.276	0.301	0.334	0.381	0.473	0.403	0.446	0.461	0.458	0.417	**0.485**
FM	0.302	0.318	0.365	0.365	0.464	0.397	0.442	0.452	0.451	0.412	**0.487**
BK	0.297	0.295	0.313	0.326	0.417	0.342	0.395	0.406	0.407	0.368	**0.441**
BT	0.314	0.327	0.365	0.402	0.483	0.419	0.464	0.471	0.473	0.435	**0.492**
DS	0.295	0.309	0.357	0.378	0.466	0.394	0.449	0.455	0.453	0.416	**0.488**
Top-5	**0.306**	**0.329**	**0.345**	**0.372**	**0.462**	**0.398**	**0.447**	**0.451**	**0.461**	**0.424**	0.481
Top-15	**0.328**	**0.347**	**0.361**	**0.385**	**0.475**	**0.412**	**0.462**	**0.464**	**0.483**	**0.438**	0.503

ours on the CoSum and TVSum dataset, respectively. Note that with web videos only is better than training with multiple handcrafted regularizations proposed in SG. This confirms the effectiveness of incorporating a large number of web videos in our framework and learning the topic-specific semantics using a weakly-supervised matching loss function. (3) Since the CoSum dataset contains videos that have visual concepts shared with other videos from different topics, our approach using generative modelling naturally yields better results than that on the TVSum dataset. (4) It's worth noticing that TVSum is a quite challenging summarization dataset because topics on this dataset are very ambiguous and difficult to understand well with very few videos. By accessing the similar web videos to eliminate ambiguity for a specific topic, our approach works much better than all the unsupervised methods by achieving a top-5 mAP of 48.1%, showing that the accurate and user-interested video contents can be directly learned from more diverse data rather than complex summarization criteria.

Comparison with Supervised Methods. We then conduct comparison with some supervised alternatives including KVS [24], DPP [5], sLstm [44], SM [9] and DSN [20] (weakly-supervised), we have the following key observations from Tables 3 and 4: (1) VESD outperforms KVS on both datasets by a big margin (maximum improvement of 7.1% in top-5 mAP on CoSum), showing the advantage of our generative modelling and more powerful representation learning with web videos. (2) On the Cosum dataset, VESD outperforms SM [9] and DSN [20] by a margin of 2.0% and 3.4% in top-5 mAP, respectively. The results suggest that our method is still better than the fully-supervised methods and the weakly-supervised method. (3) On the TVSum dataset, a similar performance gain of 2.0% can be achieved compared with all other supervised methods.



(a) Sample frames from video 15 [31]

(b) Training on benchmark with weak supervision

(c) Training on benchmark with weak supervision and extra web videos

(d) Training on benchmark with strong supervision

(e) Training on benchmark with strong supervision and extra web videos

Fig. 3. Qualitative comparison of video summaries using different training settings, along with the ground-truth importance scores (cyan background). In the last subfigure, we can easily see that weakly-supervised VESD with web videos and available importance scores produces more reliable summaries than training on benchmark videos with only weak labels. (Best viewed in colors) (Color figure online)

5.2 Qualitative Results

To get some intuition about the different training settings for VESD and their effects on the temporal selection pattern, we visualize some selected frames on an example video in Fig. 3. The cyan background shows the frame-level importance scores. The coloured regions are the selected subset of frames using the specific training setting. The visualized keyframes for different setting supports the results presented in Table 1. We notice that all four settings cover the temporal regions with the high frame-level score. By leveraging both the web videos and importance scores in datasets, VESD framework will shift towards the highly topic-specific temporal regions.

6 Conclusion

One key problem in video summarization is how to model the latent semantic representation, which has not been adequately resolved under the "single video understanding" framework in prior works. To address this issue, we introduced a generative summarization framework called VESD to leverage the web videos for better latent semantic modelling and to reduce the ambiguity of video summarization in a principled way. We incorporated flexible web prior distribution into a variational framework and presented a simple encoder-decoder with attention for summarization. The potentials of our VESD framework for large-scale video summarization were validated, and extensive experiments on benchmarks showed that VESD outperforms state-of-the-art video summarization methods significantly.

References

1. Chu, W.S., Song, Y., Jaimes, A.: Video co-summarization: video summarization by visual co-occurrence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3584–3592 (2015)
2. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2625–2634 (2015)
3. Elhamifar, E., Sapiro, G., Vidal, R.: See all by looking at a few: sparse modeling for finding representative objects. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1600–1607. IEEE (2012)
4. Feng, S., Lei, Z., Yi, D., Li, S.Z.: Online content-aware video condensation. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2082–2087. IEEE (2012)
5. Gong, B., Chao, W.L., Grauman, K., Sha, F.: Diverse sequential subset selection for supervised video summarization. In: Advances in Neural Information Processing Systems, pp. 2069–2077 (2014)
6. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)

7. Guan, G., Wang, Z., Mei, S., Ott, M., He, M., Feng, D.D.: A top-down approach for video summarization. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **11**(1), 4 (2014)
8. Gygli, M., Grabner, H., Riemenschneider, H., Van Gool, L.: Creating summaries from user videos. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 505–520. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10584-0_33
9. Gygli, M., Grabner, H., Van Gool, L.: Video summarization by learning submodular mixtures of objectives. Proc. CVPR **2015**, 3090–3098 (2015)
10. Gygli, M., Song, Y., Cao, L.: Video2gif: automatic generation of animated gifs from video. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1001–1009. IEEE (2016)
11. 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, pp. 770–778 (2016)
12. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1725–1732 (2014)
13. Kim, G., Sigal, L., Xing, E.P.: Joint summarization of large-scale collections of web images and videos for storyline reconstruction (2014)
14. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114 (2013)
15. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. arXiv preprint arXiv:1512.09300 (2015)
16. Mahasseni, B., Lam, M., Todorovic, S.: Unsupervised video summarization with adversarial LSTM networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
17. Mathieu, M., Couprie, C., LeCun, Y.: Deep multi-scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440 (2015)
18. Money, A.G., Agius, H.: Video summarisation: a conceptual framework and survey of the state of the art. J. Vis. Commun. Image Represent. **19**(2), 121–143 (2008)
19. Ng, J.Y.H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: deep networks for video classification. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4694–4702. IEEE (2015)
20. Panda, R., Das, A., Wu, Z., Ernst, J., Roy-Chowdhury, A.K.: Weakly supervised summarization of web videos. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3677–3686. IEEE (2017)
21. Panda, R., Roy-Chowdhury, A.K.: Collaborative summarization of topic-related videos. In: CVPR, vol. 2, p. 5 (2017)
22. Panda, R., Roy-Chowdhury, A.K.: Sparse modeling for topic-oriented video summarization. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1388–1392. IEEE (2017)
23. Plummer, B.A., Brown, M., Lazebnik, S.: Enhancing video summarization via vision-language embedding. In: Computer Vision and Pattern Recognition (2017)
24. Potapov, D., Douze, M., Harchaoui, Z., Schmid, C.: Category-specific video summarization. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 540–555. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_35

25. Pritch, Y., Rav-Acha, A., Gutman, A., Peleg, S.: Webcam synopsis: peeking around the world. In: IEEE 11th International Conference on Computer Vision, ICCV 2007, pp. 1–8. IEEE (2007)
26. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396 (2016)
27. Rui, Y., Gupta, A., Acero, A.: Automatically extracting highlights for TV baseball programs. In: Proceedings of the Eighth ACM International Conference on Multimedia, pp. 105–115. ACM (2000)
28. Sharghi, A., Gong, B., Shah, M.: Query-focused extractive video summarization. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 3–19. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_1
29. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and TRECVid. In: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, pp. 321–330. ACM (2006)
30. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Advances in Neural Information Processing Systems, pp. 3483–3491 (2015)
31. Song, Y., Vallmitjana, J., Stent, A., Jaimes, A.: TVSUM: summarizing web videos using titles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5179–5187 (2015)
32. Sun, M., Farhadi, A., Seitz, S.: Ranking domain-specific highlights by analyzing edited videos. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 787–802. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_51
33. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems, pp. 3104–3112 (2014)
34. Szegedy, C., et al.: Going deeper with convolutions. In: CVPR (2015)
35. Tang, H., Kwatra, V., Sargin, M.E., Gargi, U.: Detecting highlights in sports videos: cricket as a test case. In: 2011 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2011)
36. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 4489–4497. IEEE (2015)
37. Truong, B.T., Venkatesh, S.: Video abstraction: a systematic review and classification. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **3**(1), 3 (2007)
38. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: Advances in Neural Information Processing Systems, pp. 613–621 (2016)
39. Walker, J., Doersch, C., Gupta, A., Hebert, M.: An uncertain future: forecasting from static images using variational autoencoders. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 835–851. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_51
40. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015)
41. Yang, H., Wang, B., Lin, S., Wipf, D., Guo, M., Guo, B.: Unsupervised extraction of video highlights via robust recurrent auto-encoders. arXiv preprint arXiv:1510.01442 (2015)
42. Yao, T., Mei, T., Rui, Y.: Highlight detection with pairwise deep ranking for first-person video summarization (2016)

43. Zhang, K., Chao, W.L., Sha, F., Grauman, K.: Summary transfer: exemplar-based subset selection for video summarization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1059–1067. IEEE (2016)
44. Zhang, K., Chao, W.-L., Sha, F., Grauman, K.: Video summarization with long short-term memory. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 766–782. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_47
45. Zhou, Y., Berg, T.L.: Learning temporal transformations from time-lapse videos. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 262–277. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_16