Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward

Kaiyang Zhou,[1,2] Yu Qiao[1]*, Tao Xiang[2]
[1] Guangdong Key Lab of Computer Vision and Virtual Reality,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
[2] Queen Mary University of London, UK
k.zhou@qmul.ac.uk, yu.qiao@siat.ac.cn, t.xiang@qmul.ac.uk

Abstract

Video summarization aims to facilitate large-scale video browsing by producing short, concise summaries that are diverse and representative of original videos. In this paper, we formulate video summarization as a sequential decision-making process and develop a deep summarization network (DSN) to summarize videos. DSN predicts for each video frame a probability, which indicates how likely a frame is selected, and then takes actions based on the probability distributions to select frames, forming video summaries. To train our DSN, we propose an end-to-end, reinforcement learning-based framework, where we design a novel reward function that *jointly* accounts for diversity and representativeness of generated summaries and does not rely on labels or user interactions at all. During training, the reward function judges how diverse and representative the generated summaries are, while DSN strives for earning higher rewards by learning to produce more diverse and more representative summaries. Since labels are not required, our method can be *fully unsupervised*. Extensive experiments on two benchmark datasets show that our unsupervised method not only outperforms other state-of-the-art unsupervised methods, but also is comparable to or even superior than most of published supervised approaches.

Introduction

Driven by the exponential growth in the amount of online videos in recent years, research in video summarization has gained increasing attention, leading to various methods proposed to facilitate large-scale video browsing (Gygli et al. 2014; Gygli, Grabner, and Van Gool 2015; Zhang et al. 2016a; Song et al. 2015; Panda and Roy-Chowdhury 2017; Mahasseni, Lam, and Todorovic 2017; Potapov et al. 2014).

Recently, recurrent neural network (RNN), especially with the long short-term memory (LSTM) cell (Hochreiter and Schmidhuber 1997), has been exploited to model the sequential patterns in video frames, as well as to tackle the end-to-end training problem. Zhang et al. (Zhang et al. 2016b) proposed a deep architecture that combines a bidirectional LSTM network with a Determinantal Point Process (DPP) module that increases diversity in summaries, referring to as DPP-LSTM. They trained DPP-LSTM with supervised learning, using both video-level summaries and frame-level importance scores. At test time, DPP-LSTM predicts importance scores and outputs feature vectors simultaneously, which are together used to construct a DPP matrix. Due to the DPP modeling, DPP-LSTM needs to be trained in a two-stage manner.

Although DPP-LSTM (Zhang et al. 2016b) has shown state-of-the-art performances on several benchmarks, we argue that supervised learning cannot fully explore the potential of deep networks for video summarization because there does not exist a single ground truth summary for a video. This is grounded by the fact that humans have subjective opinions on which parts of a video should be selected as the summary. Therefore, devising more effective summarization methods that rely less on labels is still in demand.

Mahasseni et al. (Mahasseni, Lam, and Todorovic 2017) developed an adversarial learning framework to train DPP-LSTM. During the learning process, DPP-LSTM selects keyframes and a discriminator network is used to judge whether a synthetic video constructed by the keyframes is real or not, in order to enforce DPP-LSTM to select more representative frames. Although their framework is unsupervised, the adversarial nature makes the training unstable, which may result in model collapse. In terms of increasing diversity, DPP-LSTM cannot benefit maximally from the DPP module without the help of labels. Since a RNN-based encoder-decoder network following DPP-LSTM for video reconstruction requires pretraining, their framework requires multiple training stages, which is not efficient in practice.

In this paper, we formulate video summarization as a sequential decision-making process and develop a deep summarization network (DSN) to summarize videos. DSN has an encoder-decoder architecture, where the encoder is a convolutional neural network (CNN) that performs feature extraction on video frames and the decoder is a bidirectional LSTM network that produces probabilities based on which actions are sampled to select frames. To train our DSN, we propose an end-to-end, reinforcement learning-based framework with a diversity-representativeness (DR) reward function that jointly accounts for diversity and representativeness of generated summaries, and does not rely on labels or user interactions at all.

The DR reward function is inspired by the general criteria of what properties a high-quality video summary should

*Corresponding author.

have. Specifically, the reward function consists of a diversity reward and a representativeness reward. The diversity reward measures how dissimilar the selected frames are to each other, while the representativeness reward computes distances between frames and their nearest selected frames, which is essentially the k-medoids problem. These two rewards complement to each other and work jointly to encourage DSN to produce diverse, representative summaries. The intuition behind this learning strategy is closely concerned with how humans summarize videos. To the best of our knowledge, this paper is the first to apply reinforcement learning to unsupervised video summarization.

The learning objective of DSN is to maximize the expected rewards over time. The rationale for using reinforcement learning (RL) to train DSN is two-fold. Firstly, we use RNN as part of our model and focus on the unsupervised setting. RNN needs to receive supervision signals at each temporal step but our rewards are computed over the whole video sequence, i.e., they can only be obtained after a sequence finishes. To provide supervision from a reward that is only available in the end of sequence, RL becomes a natural choice. Secondly, we conjecture that DSN can benefit more from RL because RL essentially aims to optimize the action (frame-selection) mechanism of an agent by iteratively enforcing the agent to take better and better actions. However, optimizing action mechanism is not particularly highlighted in a normal supervised/unsupervised setting.

As the training process does not require labels, our method can be fully unsupervised. To fit the case where labels are available, we further extend our unsupervised method to the supervised version by adding a supervised objective that directly maximizes the log-probability of selecting annotated keyframes. By learning the high-level concepts encoded in labels, our DSN can recognize globally important frames and produce summaries that highly align with human-annotated summaries.

We conduct extensive experiments on two datasets, SumMe (Gygli et al. 2014) and TVSum (Song et al. 2015), to quantitatively and qualitatively evaluate our method. The quantitative results show that our unsupervised method not only outperforms other state-of-the-art unsupervised alternatives, but also is comparable to or even superior than most of published supervised methods. More impressively, the qualitative results illustrate that DSN trained with our unsupervised learning algorithm can identify important frames that coincide with human selections.

The main contributions of this paper are summarized as follows: (1) We develop an end-to-end, reinforcement learning-based framework for training DSN, where we propose a label-free reward function that jointly accounts for diversity and representativeness of generated summaries. To the best of our knowledge, our work is the first to apply reinforcement learning to unsupervised video summarization. (2) We extend our unsupervised approach to the supervised version to leverage labels. (3) We conduct extensive experiments on two benchmark datasets to show that our unsupervised method not only outperforms other state-of-the-art unsupervised methods, but also is comparable to or even superior than most of published supervised approaches.

Related Work

Video summarization. Research in video summarization has been significantly advanced in recent years, leading to approaches of various characteristics. Lee et al. (Lee, Ghosh, and Grauman 2012) identified important objects and people in summarizing videos. Gygli et al. (Gygli et al. 2014) learned a linear regressor to predict the degree of interestingness of video frames and selected keyframes with the highest interestingness scores. Gygli et al. (Gygli, Grabner, and Van Gool 2015) cast video summarization as a subset selection problem and optimized submodular functions with multiple objectives. Ejaz et al. (Ejaz, Mehmood, and Baik 2013) applied an attention-modeling technique to extracting keyframes of visual saliency. Zhang et al. (Zhang et al. 2016a) developed a nonparametric approach to transfer structures of known video summaries to new videos with similar topics. Auxiliary resources have also been exploited to aid the summarization process such as web images/videos (Song et al. 2015; Khosla et al. 2013; Chu, Song, and Jaimes 2015) and category information (Potapov et al. 2014). Most of these non-deep summarization methods processed video frames independently, thus ignoring the inherent sequential patterns. Moreover, non-deep summarization methods usually do not support end-to-end training, which causes extra costs at test time. To address the aforementioned issues, we model video summarization via a deep RNN to capture long-term dependencies in video frames, and propose a reinforcement learning-based framework to train the network end to end.

Reinforcement learning (RL). RL has become an increasingly popular research area due to its effectiveness in various tasks. Mnih et al. (Mnih et al. 2013) successfully approximated Q function with a deep CNN, and enabled their agent to beat a human expert in several Atari games. Later on, many researchers have applied RL algorithms to vision-related applications such as image captioning (Xu et al. 2015) and person re-identification (Lan et al. 2017). In the domain of video summarization, our work is not the first to use RL. Previously, Song et al. (Song et al. 2016) has applied RL to training a summarization network for selecting category-specific keyframes. Their learning framework requires keyframe-labels and category information of training videos. However, our work significantly differs from the work of Song et al. and other RL-based work in the way that labels or user interactions are not required at all during the learning process, which is attributed to our novel reward function. Therefore, our summarization method can be fully unsupervised and is more practical to be deployed for large-scale video summarization.

Proposed Approach

We formulate video summarization as a sequential decision-making process. In particular, we develop a deep summarization network (DSN) to predict probabilities for video frames and make decisions on which frames to select based on the predicted probability distributions. We present an end-to-end, reinforcement learning-based framework for training our DSN, where we design a diversity-

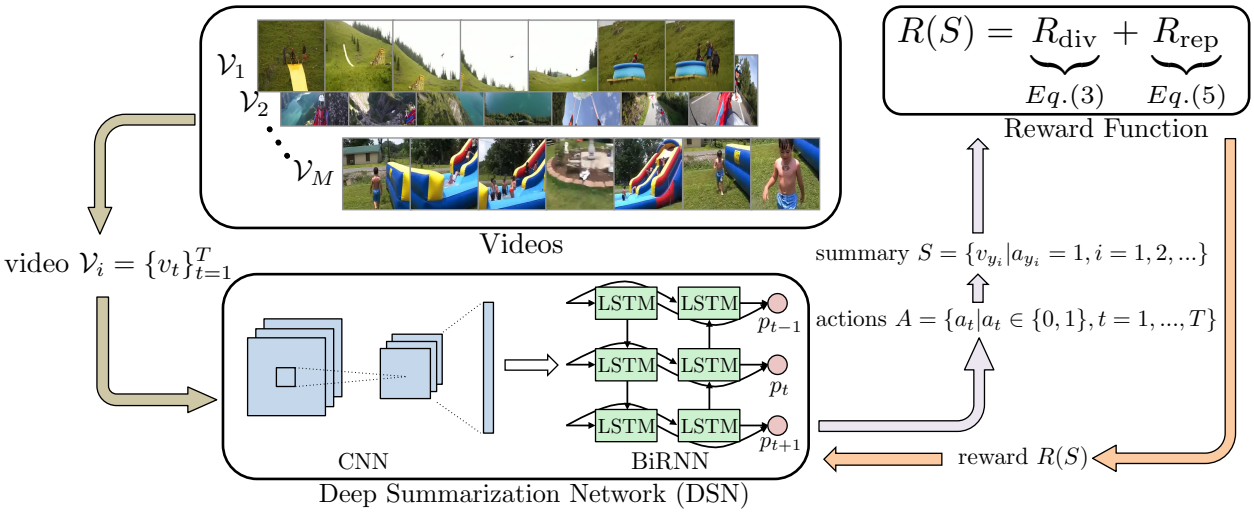


Figure 1: Training deep summarization network (DSN) via reinforcement learning. DSN receives a video $\mathcal{V}_i$ and takes actions A (i.e., a sequence of binary variables) on which parts of the video are selected as the summary S. The feedback reward $R(S)$ is computed based on the quality of the summary, i.e., diversity and representativeness.

representativeness reward function, which directly assesses how diverse and representative the generated summaries are. Figure 1 illustrates the overall learning process.

Deep Summarization Network

We adopt the encoder-decoder framework for our deep summarization network (DSN). The encoder is a convolutional neural network (CNN) that extracts visual features $\{x_t\}_{t=1}^T$ from the input video frames $\{v_t\}_{t=1}^T$ with the length T. The decoder is a bidirectional recurrent neural network (BiRNN) topped with a fully connected (FC) layer. The BiRNN takes as input the entire visual features $\{x_t\}_{t=1}^T$ and produces corresponding hidden states $\{h_t\}_{t=1}^T$. Each h_t is the concatenation of the forward hidden state h_t^f and the backward hidden state h_t^b, which encapsulate the future information and the past information with a strong emphasis on the parts surrounding the t^{th} frame. The FC layer that ends with the sigmoid function predicts for each frame a probability p_t, from which a frame-selection action a_t is sampled:

$$p_t = \sigma(W h_t), \tag{1}$$

$$a_t \sim \text{Bernoulli}(p_t), \tag{2}$$

where σ represents the sigmoid function, $a_t \in \{0, 1\}$ indicates whether the t^{th} frame is selected or not. The bias in Eq. (1) is omitted for brevity. A video summary is composed of the selected frames, $S = \{v_{y_i} | a_{y_i} = 1, i = 1, 2, ...\}$.

In practice, we use the GoogLeNet (Szegedy et al. 2015) pretrained on ImageNet (Deng et al. 2009) as the CNN model. The visual feature vectors $\{x_t\}_{t=1}^T$ are extracted from the penultimate layer of the GoogLeNet. For the RNN cell, we employ long short-term memory (LSTM) to enhance RNN's ability for capturing long-term dependencies in video frames. During training, we only update the decoder.

Diversity-Representativeness Reward Function

During training, DSN will receive a reward $R(S)$ that evaluates the quality of generated summaries, and the objective of DSN is to maximize the expected rewards over time by producing high-quality summaries. In general, a high-quality video summary is expected to be both diverse and representative of the original video so that temporal information across the entire video can be maximally preserved. To this end, we propose a novel reward that assesses the degree of diversity and representativeness of generated summaries. The proposed reward is composed of a diversity reward R_{div} and a representativeness reward R_{rep}, which we detail as follows.

Diversity reward. We evaluate the degree of diversity of a generated summary by measuring the dissimilarity among the selected frames in the feature space. Let the indices of the selected frames be $\mathcal{Y} = \{y_i | a_{y_i} = 1, i = 1, ..., |\mathcal{Y}|\}$, we compute R_{div} as the mean of the pairwise dissimilarities among the selected frames:

$$R_{\text{div}} = \frac{1}{|\mathcal{Y}|(|\mathcal{Y}| - 1)} \sum_{t \in \mathcal{Y}} \sum_{\substack{t' \in \mathcal{Y} \\ t' \neq t}} d(x_t, x_{t'}), \tag{3}$$

where $d(\cdot, \cdot)$ is the dissimilarity function calculated by

$$d(x_t, x_{t'}) = 1 - \frac{x_t^T x_{t'}}{||x_t||_2 ||x_{t'}||_2}. \tag{4}$$

Intuitively, the more diverse (or more dissimilar) the selected frames to each other, the higher the diversity reward that the agent can receive. However, Eq. (3) treats video frames as randomly permutable items which ignore the temporal structure inherent in sequential data. In fact, the similarity between two temporally distant frames should be ignored because they are essential to the storyline construction (Gong et al. 2014). To overcome this problem, we set

$d(x_t, x_{t'}) = 1$ if $|t - t'| > \lambda$, where λ controls the degree of temporal distance. We will validate this hypothesis in the Experiments section.

Representativeness reward. This reward measures how well the generated summary can represent the original video. To this end, we formulate the degree of representativeness of a video summary as the k-medoids problem (Gygli, Grabner, and Van Gool 2015). In particular, we want the agent to select a set of medoids such that the mean of squared errors between video frames and their nearest medoids is minimal. Therefore, we define R_{rep} as

$$R_{\text{rep}} = \exp(-\frac{1}{T}\sum_{t=1}^{T}\min_{t' \in \mathcal{Y}}||x_t - x_{t'}||_2). \quad (5)$$

With this reward, the agent is encouraged to select frames that are close to the cluster centers in the feature space. An alternative formulation of R_{rep} can be the inverse reconstruction errors achieved by the selected frames, but this formulation is too computationally expensive.

Diversity-representativeness reward. R_{div} and R_{rep} complement to each other and work jointly to guide the learning of DSN:

$$R(S) = R_{\text{div}} + R_{\text{rep}}. \quad (6)$$

During training, R_{div} and R_{rep} are similar in the order of magnitude. In fact, it is non-trivial to keep R_{div} and R_{rep} at the same order of magnitude during training, thus none of them would dominate in gradient computation. We give zero reward to DSN when no frames are selected, i.e., the sampled actions are all zeros.

Training with Policy Gradient

The goal of our summarization agent is to learn a policy function π_θ with parameters θ by maximizing the expected rewards

$$J(\theta) = \mathbb{E}_{p_\theta(a_{1:T})}[R(S)], \quad (7)$$

where $p_\theta(a_{1:T})$ denotes the probability distributions over possible action sequences, and $R(S)$ is computed by Eq. (6). π_θ is defined by our DSN.

Following the REINFORCE algorithm proposed by Williams (Williams 1992), we can compute the derivative of the objective function $J(\theta)$ w.r.t. the parameters θ as

$$\nabla_\theta J(\theta) = \mathbb{E}_{p_\theta(a_{1:T})}[R(S)\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_t|h_t)], \quad (8)$$

where a_t is the action taken by DSN at time t and h_t is the hidden state from the BiRNN.

Since Eq. (8) involves the expectation over high-dimensional action sequences, which is hard to compute directly, we approximate the gradient by running the agent for N episodes on the same video and then taking the average gradient

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T} R_n \nabla_\theta \log \pi_\theta(a_t|h_t), \quad (9)$$

where R_n is the reward computed at the n^{th} episode. Eq. (9) is also known as the episodic REINFORCE algorithm.

Although the gradient in Eq. (9) is a good estimate, it may contain high variance which will make the network hard to converge. A common countermeasure is to subtract the reward by a constant baseline b, so the gradient becomes

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T}(R_n - b)\nabla_\theta \log \pi_\theta(a_t|h_t), \quad (10)$$

where b is simply computed as the moving average of rewards experienced so far for computational efficiency.

Regularization

Since selecting more frames will also increase the reward, we impose a regularization term on the probability distributions $p_{1:T}$ produced by DSN in order to constrain the percentage of frames selected for the summary. Inspired by (Mahasseni, Lam, and Todorovic 2017), we minimize the following term during training,

$$L_{\text{percentage}} = ||\frac{1}{T}\sum_{t=1}^{T} p_t - \epsilon||^2, \quad (11)$$

where ϵ determines the percentage of frames to be selected.

In addition, we also add the $\ell 2$ regularization term on the weight parameters θ to avoid overfitting

$$L_{\text{weight}} = \sum_{i,j}\theta_{i,j}^2. \quad (12)$$

Optimization

We optimize the policy function's parameters θ via stochastic gradient-based method. By combing the gradients computed from Eq. (10), Eq. (11) and Eq. (12), we update θ as

$$\theta = \theta - \alpha\nabla_\theta(-J + \beta_1 L_{\text{percentage}} + \beta_2 L_{\text{weight}}), \quad (13)$$

where α is learning rate, and β_1 and β_2 are hyperparameters that balance the weighting.

In practice, we use Adam (Kingma and Ba 2014) as the optimization algorithm. As a result of learning, the log-probability of actions taken by the network that have led to high rewards is increased, while that of actions that have resulted in low rewards is decreased.

Extension to Supervised Learning

Given the keyframe indices for a video, $\mathcal{Y}^* = \{y_i^*|i = 1, ..., |\mathcal{Y}^*|\}$, we use Maximum Likelihood Estimation (MLE) to maximize the log-probability of selecting keyframes specified by $\mathcal{Y}^*$, $\log p(t;\theta)$ where $t \in \mathcal{Y}^*$. $p(t;\theta)$ is computed from Eq. (1). The objective is formalized as

$$L_{\text{MLE}} = \sum_{t \in \mathcal{Y}^*}\log p(t;\theta). \quad (14)$$

Summary Generation

For a test video, we apply a trained DSN to predict the frame-selection probabilities as importance scores. We compute shot-level scores by averaging frame-level scores within the same shot. For temporal segmentation, we use KTS proposed by (Potapov et al. 2014). To generate a summary, we select shots by maximizing the total scores while ensuring that the summary length does not exceed a limit, which is usually 15% of the video length. The maximization step is essentially the 0/1 Knapsack problem, which is known as NP-hard. We obtain a near-optimal solution via dynamic programming (Song et al. 2015).

Besides evaluating generated summaries in the Experiments part, we also qualitatively analyze the raw predictions of DSN so as to exclude the effect of this summary generation step, by which we can better understand what DSN has learned.

Experiments

Experimental Setup

Datasets. We evaluate our methods on SumMe (Gygli et al. 2014) and TVSum (Song et al. 2015). SumMe consists of 25 user videos covering various topics such as holidays and sports. Each video in SumMe ranges from 1 to 6 minutes and is annotated by 15 to 18 persons, thus there are multiple ground truth summaries for each video. TVSum contains 50 videos, which include the topics of news, documentaries, etc. The duration of each video varies from 2 to 10 minutes. Similar to SumMe, each video in TVSum has 20 annotators that provide frame-level importance scores. Following (Song et al. 2015; Zhang et al. 2016b), we convert importance scores to shot-based summaries for evaluation. In addition to these two datasets, we exploit two other datasets, OVP[1] that has 50 videos and YouTube (De Avila et al. 2011) that has 39 videos excluding cartoon videos, to evaluate our method in the settings where training data is augmented (Zhang et al. 2016b; Mahasseni, Lam, and Todorovic 2017).

Evaluation metric. For fair comparison with other approaches, we follow the commonly used protocol from (Zhang et al. 2016b) to compute F-score as the metric to assess the similarity between automatic summaries and ground truth summaries. We also follow (Zhang et al. 2016b) to deal with multiple ground truth summaries.

Evaluation settings. We use three settings as suggested in (Zhang et al. 2016b) to evaluate our method. (1) Canonical: we use the standard 5-fold cross validation (5FCV), i.e., 80% of videos for training and the rest for testing. (2) Augmented: we still use the 5FCV but we augment the training data in each fold with OVP and YouTube. (3) Transfer: for a target dataset, e.g. SumMe or TVSum, we use the other three datasets as the training data to test the transfer ability of our model.

Implementation details. We downsample videos by 2 fps as did in (Zhang et al. 2016b). We set the temporal distance λ to 20, the ϵ in Eq. 11 to 0.5, and the number of episodes N to 5. The other hyperparameters α, β_1 and β_2 in Eq. (13) are optimized via cross-validation. We set the dimension of hidden state in the RNN cell to 256 throughout this paper. Training is stopped when it reaches a maximum number of epochs (60 in our case). Early stopping is executed when reward creases to increase for a period of time (10 epochs in our experiments). We implement our method based on Theano (Al-Rfou et al. 2016)[2].

Comparison. To compare with other approaches, we implement Uniform sampling, K-medoids and Dictionary selection (Elhamifar, Sapiro, and Vidal 2012) by ourselves. We retrieve results of other approaches including Video-MMR (Li and Merialdo 2010), Vsumm (De Avila et al. 2011), Web image (Khosla et al. 2013), Online sparse coding (Zhao and Xing 2014), Co-archetypal (Song et al. 2015), Interestingness (Gygli et al. 2014), Submodularity (Gygli, Grabner, and Van Gool 2015), Summary transfer (Zhang et al. 2016a), Bi-LSTM and DPP-LSTM (Zhang et al. 2016b), GAN_{dpp} and GAN_{sup} (Mahasseni, Lam, and Todorovic 2017) from published papers. Due to space limit, we do not include these citations in tables.

Quantitative Evaluation

We first compare our method with several baselines that differ in learning objectives. Then, we compare our methods with current state-of-the-art unsupervised/supervised approaches in the three evaluation settings.

Comparison with baselines. We set the baseline models as the ones trained with R_{div} only and R_{rep} only, which are denoted by D-DSN and R-DSN, respectively. We represent the model trained with the two rewards jointly as DR-DSN. The model that is extended to the supervised version is denoted by $\text{DR-DSN}_{\text{sup}}$. We also validate the effectiveness of the proposed technique (we call this λ-technique from now on) that ignores the distant similarity when computing R_{div}. We represent the D-DSN trained without the λ-technique as $\text{D-DSN}_{\text{w/o }\lambda}$. To verify that DSN can benefit more from reinforcement learning than from supervised learning, we add another baseline as the DSN trained with the cross entropy loss using keyframe annotations, where a confidence penalty (Pereyra et al. 2017) is imposed on the output distributions as a regularization term. This model is denoted by DSN_{sup}.

Table 1: Results (%) of different variants of our method on SumMe and TVSum.

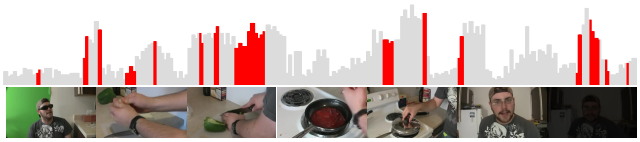
Method	SumMe	TVSum
DSN_{sup}	38.2	54.5
$\text{D-DSN}_{\text{w/o }\lambda}$	39.3	55.7
D-DSN	40.5	56.2
R-DSN	40.7	56.9
DR-DSN	41.4	57.6
$\text{DR-DSN}_{\text{sup}}$	**42.1**	**58.1**

Table 1 reports the results of different variants of our method on SumMe and TVSum. We can see that DR-DSN

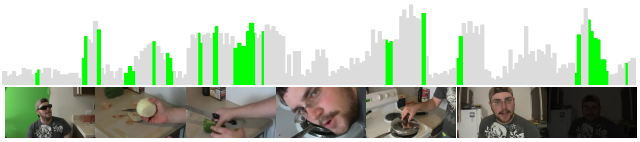
[1]Open video project: https://open-video.org/.

[2]Codes are available on https://github.com/KaiyangZhou/vsumm-reinforce

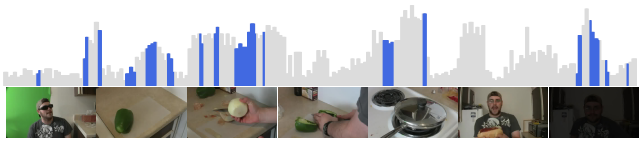
(a) Example frames from video 18 in TVSum (indexed as in (Song et al. 2015)).



(b) DR-DSN$_{sup}$



(c) DR-DSN



(d) R-DSN

(e) D-DSN

Figure 2: Video summaries generated by different variants of our approach for video 18 in TVSum. The light-gray bars in (b) to (e) correspond to ground truth importance scores, while the colored areas correspond to the selected parts by different models.

clearly outperforms D-DSN and R-DSN on both datasets, which demonstrates that by using R_{div} and R_{rep} collaboratively, we can better teach DSN to produce high-quality summaries that are diverse and representative. Comparing the unsupervised model with the supervised one, we see that DR-DSN significantly outperforms DSN$_{sup}$ on the two datasets (41.4 vs. 38.2 on SumMe and 57.6 vs. 54.5 on TVSum), which justifies our assumption that DSN can benefit more from reinforcement learning than from supervised learning.

By adding the supervision signals of L_{MLE} (Eq. (14)) to DR-DSN, the summarization performances are further improved (1.7% improvements on SumMe and 0.9% improvements on TVSum). This is because labels encode the high-level understanding of the video content, which is exploited by DR-DSN$_{sup}$ to learn more useful patterns.

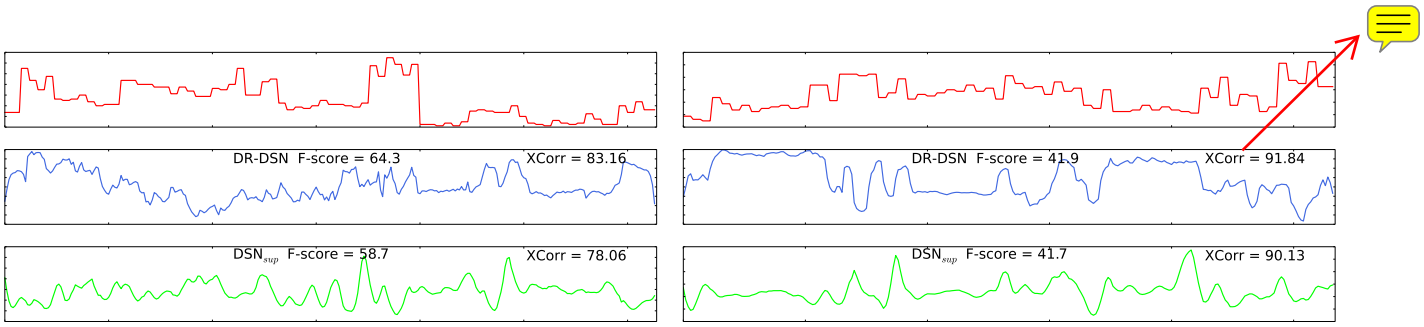
The performances of R-DSN are slightly better than those of D-DSN on the two datasets, which is because diverse summaries usually contain redundant information that are irrelevant to the video subject. We observe that the performances of D-DSN are better than those of D-DSN$_{w/o\ \lambda}$ that does not consider temporally distant frames. When using the λ-technique in training, around $50\% \sim 70\%$ of the distance matrix was set to 1 (varying across different videos) at the early stage. As the training epochs increased, the percentage went up too, eventually staying around $80\% \sim 90\%$. This makes sense because selecting temporally distant frames can lead to higher rewards and DSN is encouraged to do so with the diversity reward function.

Comparison with unsupervised approaches. Table 2 shows the results of DR-DSN against other unsupervised approaches on SumMe and TVSum. It can be seen that DR-DSN outperforms the other unsupervised approaches on both datasets by large margins. On SumMe, DR-DSN is 5.9% better than the current state-of-the-art, GAN$_{dpp}$. On TVSum, DR-DSN substantially beats GAN$_{dpp}$ by 11.4%. Although our reward functions are analogous to the objectives of GAN$_{dpp}$ in concepts, ours directly model diversity and representativeness of selected frames in the feature space, which is more useful to guide DSN to find good solutions. In addition, the training performances of DR-DSN are 40.2% on SumMe and 57.2% on TVSum, which suggest that the model did not overfit to the training data (note that we do not explicitly optimize the F-score metric in the training objective function).

Table 2: Results (%) of unsupervised approaches on SumMe and TVSum. Our DR-DSN performs the best, especially in TVSum where it exhibits a huge advantage over others.

Method	SumMe	TVSum
Video-MMR	26.6	-
Uniform sampling	29.3	15.5
K-medoids	33.4	28.8
Vsumm	33.7	-
Web image	-	36.0
Dictionary selection	37.8	42.0
Online sparse coding	-	46.0
Co-archetypal	-	50.0
GAN$_{dpp}$	39.1	51.7
DR-DSN	**41.4**	**57.6**

Comparison with supervised approaches. Table 3 reports the results of our supervised model, DR-DSN$_{sup}$, and other supervised approaches. In terms of LSTM-based methods, our DR-DSN$_{sup}$ beats the others, i.e., Bi-LSTM, DPP-LSTM and GAN$_{sup}$, by $1.0\% \sim 12.0\%$ on SumMe and $3.2\% \sim 7.2\%$ on TVSum, respectively. It is also interesting to see that the summarization performance of our unsupervised method, DR-DSN, is even superior than the state-of-the-art supervised approach on TVSum (57.6 vs. 56.3), and is better than most of the supervised approaches on SumMe.



(a) Video 11 in TVSum

(b) Video 10 in TVSum

Figure 3: Ground truth (top) and importance scores predicted by DR-DSN (middle) and DSN$_{sup}$ (bottom). Besides the F-score for each prediction, we also compute cross-correlation (XCorr) for each pair of prediction and ground truth to give a quantitative measure of similarity over two series of 1D arrays. The higher the XCorr, the more similar two arrays are to each other.

These results strongly prove the efficacy of our learning framework.

Table 3: Results (%) of supervised approaches on SumMe and TVSum. Our DR-DSN$_{sup}$ performs the best.

Method	SumMe	TVSum
Interestingness	39.4	-
Submodularity	39.7	-
Summary transfer	40.9	-
Bi-LSTM	37.6	54.2
DPP-LSTM	38.6	54.7
GAN$_{sup}$	41.7	56.3
DR-DSN$_{sup}$	**42.1**	**58.1**

Comparison in the Augmented (A) and Transfer (T) settings. Table 4 compares our methods with current state-of-the-art LSTM-based methods in the A and T settings. The results in the Canonical setting are also provided to exhibit the improvements obtained by increased training data. In the A setting, DR-DSN$_{sup}$ performs marginally better than GAN$_{sup}$ on SumMe (43.9 vs. 43.6), whereas it is defeated by GAN$_{sup}$ on TVSum (59.8 vs. 61.2). This may be because the LSTM model in GAN$_{sup}$ has more hidden units (1024 vs. our 256). In the T setting, DR-DSN$_{sup}$ performs the best on both datasets, suggesting that our model is able to transfer knowledge between datasets. Furthermore, it is interesting to see that our unsupervised model, DR-DSN, is superior or comparable with other methods in both settings. Overall, we firmly believe that by using a larger model and/or designing a better network architecture, we can obtain better summarization performances with our learning framework.

We also experiment with different gated RNN units, i.e., LSTM vs. GRU (Cho et al. 2014), and find that LSTM-based models consistently beat GRU-based models (see Table 5). This may be interpreted as that the memory mechanism in LSTM has a higher degree of complexity, thus allowing more complex patterns to be learned.

Qualitative Evaluation

Video summaries. We provide qualitative results for an exemplar video that talks about a man making a spicy sausage sandwich in Figure 2. In general, all four methods produce high-quality summaries that span the temporal structure, with only small variations observed in some frames. The peak regions of ground truth are almost captured. Nevertheless, the summary produced by the supervised model, DR-DSN$_{sup}$, is much closer to the complete storyline conveyed by the original video i.e., from food preparation to cooking. This is because DR-DSN$_{sup}$ benefits from labels that allow high-level concepts to be better captured.

Table 4: Results (%) of the LSTM-based approaches on SumMe and TVSum in the Canonical (C), Augmented (A) and Transfer (T) settings, respectively.

Method	SumMe			TVSum		
	C	A	T	C	A	T
Bi-LSTM	37.6	41.6	40.7	54.2	57.9	56.9
DPP-LSTM	38.6	42.9	41.8	54.7	59.6	58.7
GAN$_{dpp}$	39.1	43.4	-	51.7	59.5	-
GAN$_{sup}$	41.7	43.6	-	56.3	**61.2**	-
DR-DSN	41.4	42.8	42.4	57.6	58.4	57.8
DR-DSN$_{sup}$	**42.1**	**43.9**	**42.6**	**58.1**	59.8	**58.9**

Table 5: Results (%) of using different gated recurrent units.

Method	SumMe		TVSum	
	LSTM	GRU	LSTM	GRU
DR-DSN	41.4	41.2	57.6	56.7
DR-DSN$_{sup}$	42.1	41.5	58.1	57.8

Predicted importance scores. We visualize the raw predictions by DR-DSN and DSN$_{sup}$ in Figure 3. By comparing predictions with ground truth, we can better understand in more depth how well DSN has learned. It is worth highlighting that the curves of importance scores predicted by the unsupervised model resemble those predicted by the supervised model in several parts. More importantly, these parts coincide with the ones also considered as important by humans. This strongly demonstrates that reinforcement learning with our diversity-representativeness reward function can well imitate the human-learning process and effectively teach DSN to recognize important frames.

Conclusion

In this paper, we proposed a label-free reinforcement learning algorithm to tackle unsupervised video summarization. Extensive experiments on two benchmark datasets showed that using reinforcement learning with our unsupervised reward function outperformed other state-of-the-art unsupervised alternatives, and produced results comparable to or even superior than most supervised methods.

Acknowledgments

We thank Ke Zhang and Wei-Lun Chao for discussions of details of their paper (Zhang et al. 2016b). This work was supported in part by National Key Research and Development Program of China (2016YFC1400704) and National Natural Science Foundation of China (U1613211, 61633021).

References

[Al-Rfou et al. 2016] Al-Rfou, R.; Alain, G.; Almahairi, A.; Angermueller, C.; Bahdanau, D.; Ballas, N.; Bastien, F.; Bayer, J.; Belikov, A.; Belopolsky, A.; et al. 2016. Theano: A python framework for fast computation of mathematical expressions. *arXiv preprint arXiv:1605.02688*.

[Cho et al. 2014] Cho, K.; Van Merriënboer, B.; Bahdanau, D.; and Bengio, Y. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

[Chu, Song, and Jaimes 2015] Chu, W.-S.; Song, Y.; and Jaimes, A. 2015. Video co-summarization: Video summarization by visual co-occurrence. In *CVPR*, 3584–3592.

[De Avila et al. 2011] De Avila, S. E. F.; Lopes, A. P. B.; da Luz, A.; and de Albuquerque Araújo, A. 2011. Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters* 32(1):56–68.

[Deng et al. 2009] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255. IEEE.

[Ejaz, Mehmood, and Baik 2013] Ejaz, N.; Mehmood, I.; and Baik, S. W. 2013. Efficient visual attention based framework for extracting key frames from videos. *Signal Processing: Image Communication* 28(1):34–44.

[Elhamifar, Sapiro, and Vidal 2012] Elhamifar, E.; Sapiro, G.; and Vidal, R. 2012. See all by looking at a few: Sparse modeling for finding representative objects. In *CVPR*, 1600–1607. IEEE.

[Gong et al. 2014] Gong, B.; Chao, W.-L.; Grauman, K.; and Sha, F. 2014. Diverse sequential subset selection for supervised video summarization. In *NIPS*, 2069–2077.

[Gygli et al. 2014] Gygli, M.; Grabner, H.; Riemenschneider, H.; and Van Gool, L. 2014. Creating summaries from user videos. In *ECCV*, 505–520. Springer.

[Gygli, Grabner, and Van Gool 2015] Gygli, M.; Grabner, H.; and Van Gool, L. 2015. Video summarization by learning submodular mixtures of objectives. In *CVPR*, 3090–3098.

[Hochreiter and Schmidhuber 1997] Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

[Khosla et al. 2013] Khosla, A.; Hamid, R.; Lin, C.-J.; and Sundaresan, N. 2013. Large-scale video summarization using web-image priors. In *CVPR*, 2698–2705.

[Kingma and Ba 2014] Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization. In *ICLR*.

[Lan et al. 2017] Lan, X.; Wang, H.; Gong, S.; and Zhu, X. 2017. Deep reinforcement learning attention selection for person re-identification. In *BMVC*.

[Lee, Ghosh, and Grauman 2012] Lee, Y. J.; Ghosh, J.; and Grauman, K. 2012. Discovering important people and objects for egocentric video summarization. In *CVPR*, 1346–1353. IEEE.

[Li and Merialdo 2010] Li, Y., and Merialdo, B. 2010. Multi-video summarization based on video-mmr. In *WIAMIS*, 1–4. IEEE.

[Mahasseni, Lam, and Todorovic 2017] Mahasseni, B.; Lam, M.; and Todorovic, S. 2017. Unsupervised video summarization with adversarial lstm networks. In *CVPR*.

[Mnih et al. 2013] Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

[Panda and Roy-Chowdhury 2017] Panda, R., and Roy-Chowdhury, A. K. 2017. Collaborative summarization of topic-related videos. In *CVPR*.

[Pereyra et al. 2017] Pereyra, G.; Tucker, G.; Chorowski, J.; Kaiser, Ł.; and Hinton, G. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.

[Potapov et al. 2014] Potapov, D.; Douze, M.; Harchaoui, Z.; and Schmid, C. 2014. Category-specific video summarization. In *ECCV*, 540–555. Springer.

[Song et al. 2015] Song, Y.; Vallmitjana, J.; Stent, A.; and Jaimes, A. 2015. Tvsum: Summarizing web videos using titles. In *CVPR*, 5179–5187.

[Song et al. 2016] Song, X.; Chen, K.; Lei, J.; Sun, L.; Wang, Z.; Xie, L.; and Song, M. 2016. Category driven deep recurrent neural network for video summarization. In *ICMEW*, 1–6. IEEE.

[Szegedy et al. 2015] Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *CVPR*, 1–9.

[Williams 1992] Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8(3-4):229–256.

[Xu et al. 2015] Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2048–2057.

[Zhang et al. 2016a] Zhang, K.; Chao, W.-L.; Sha, F.; and Grauman, K. 2016a. Summary transfer: Exemplar-based

subset selection for video summarization. In *CVPR*, 1059–1067.

[Zhang et al. 2016b] Zhang, K.; Chao, W.-L.; Sha, F.; and Grauman, K. 2016b. Video summarization with long short-term memory. In *ECCV*, 766–782. Springer.

[Zhao and Xing 2014] Zhao, B., and Xing, E. P. 2014. Quasi real-time summarization for consumer videos. In *CVPR*, 2513–2520.