Weakly Supervised Deep Reinforcement Learning for Video Summarization With Semantically Meaningful Reward

Zutong Li Lei Yang
Weibo R&D Limited, USA
{zutongli0805, trilithy}@gmail.com

Abstract

Conventional unsupervised video summarization algorithms are usually developed in a frame level clustering manner. For example, frame level diversity and representativeness are two typical clustering criteria used for unsupervised reinforcement learning-based video summarization. Inspired by recent progress in video representation techniques, we further introduce the similarity of video representations to construct a semantically meaningful reward for this task. We consider that a good summarization should also be semantically identical to its original source, which means that the semantic similarity can be regarded as an additional criterion for summarization. Through combining a novel video semantic reward with other unsupervised rewards for training, we can easily upgrade an unsupervised reinforcement learning-based video summarization method to its weakly supervised version. In practice, we first train a video classification sub-network (VCSN) to extract video semantic representations based on a category-labeled video dataset. Then we fix this VCSN and train a summary generation sub-network (SGSN) using unlabeled video data in a reinforcement learning way. Experimental results demonstrate that our work significantly surpasses other unsupervised and even supervised methods. To the best of our knowledge, our method achieves state-of-the-art performance in terms of the correlation coefficients, Kendall's τ and Spearman's ρ.

1. Introduction

With the explosive growth of video data on the internet, more and more researchers have paid their attention to develop new technologies for efficient video indexing, retrieval, browsing and classification. Video summarization aims to shorten an input video into a short summary, which can help users relieve the tedious work of browsing and managing the video content of interest. Due to the extremely diverse nature of online videos, it still remains a challenging task to robustly produce a semantically meaningful video summary.

Many machine learning technology-based video summarization approaches have been proposed over the past few years. They can be roughly classified into three categories: supervised, weakly supervised and unsupervised. Zhang *et al*. [31] proposed a bidirectional LSTM network with a Determinantal Point Process module (dppLSTM) for summarization. This method directly utilizes the human annotated frame level importance scores as ground-truth to train the model. Based on the learned video semantic knowledge, an effective video summarization can be achieved by using this method. Although supervised learning-based methods look robust and easy to understand, they would be suffered from the difficulties to define which frames deserve higher scores and to label massive frame level importance scores, leading to relatively limited studies in this category. In contrast, weakly supervised and unsupervised learning-based methods attract more attention in the research community.

Otani *et al*. [20] utilized contrastive loss to map videos as well as its descriptions to a semantic space. During the test step, they extract video segment level features and apply clustering techniques to generate summary. Mahasseni *et al*. [17] designed an adversarial learning framework to train the dppLSTM model. Based on the work of Mahasseni et al, Jung *et al*. [12] introduced CSNet, which reconstructed the input sequence in a stride and chunk way, to improve summarization for long-length videos. It is very interesting to introduce the adversarial learning techniques to this task, however, the adversarial nature may incur mode collapse, leading to an unstable training procedure. A novel reinforcement learning-based deep summarization network (DR-DSN), which combines frame level diversity and representativeness of the generated summaries as unsupervised training rewards, is proposed by Zhou *et al*. in [33]. This method does not need to label frame level importance scores for training data and therefore is easy to reproduce in practice. Extended from Zhou's solution, Chen *et al*. [7] decomposed the task into several sub-tasks and proposed a hierarchical reinforcement learning method for summarization. These two methods perform superior than other unsupervised methods, however they still have some limitations. For example, DR-DSN [33] ignores the content information although the content is very important for a semantically meaningful summarization, and frame level importance score annotations are still needed for Chen's method [7] to guide the network training.

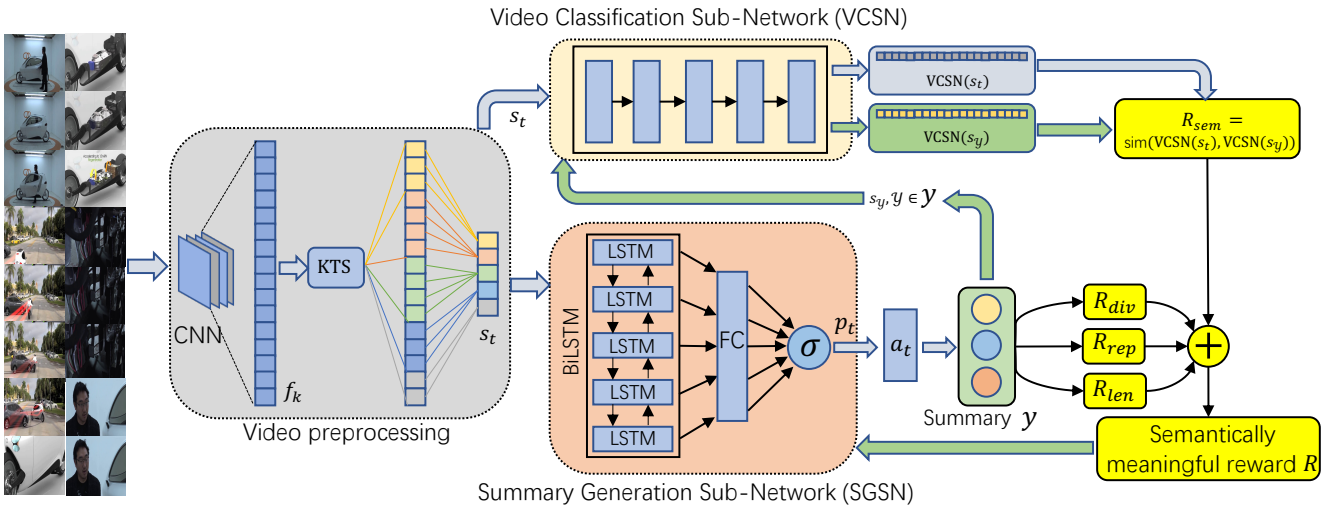

Figure 1. Training of our proposal. A pre-trained CNN converts the raw input video into a sequence of frame level feature representations. Kernel Temporal Segmentation (KTS) based shot segmentation is proceeded to cluster the frame level feature representations $\{f_k\}_{k=1}^{K}$ into its shot level feature representations $\{s_t\}_{t=1}^{T}$, where K and T denote the frame numbers of the raw input video and clustered shot level feature representations, respectively. A summary generation sub-network (SGSN) is subsequently used to predict the importance score $\{p_t\}_{t=1}^{T}$ for the segmented video shots, which will then be applied to generate the video summary $\mathcal{Y}$. The shot level feature representations of the input video s_t and those of its summary $s_y, y \in \mathcal{Y}$ are fed into a video classification sub-network (VCSN) to obtain their semantic representations $\mathrm{VCSN}(\mathrm{s_t})$ and $\mathrm{VCSN}(s_y)$. A new semantic reward R_{sem} is proposed to measure the similarity between these two video representations, where $\mathrm{sim}(\cdot,\cdot)$ is a similarity function (here we use Cosine similarity). A semantically meaningful reward R, designed as a summation of a video semantic reward term R_{sem}, a summary length reward term R_{len} and two unsupervised reward terms R_{div} and R_{rep}, are used to guide the RL procedure of the SGSN for video summarization.

In this paper, we propose a weakly supervised reinforcement learning method for video summarization. Our proposal consists of two sub-networks: video classification sub-network (VCSN) and summary generation sub-network (SGSN), where the former sub-network plays a supervisor role to guide the learning of the latter one. We first train the VCSN based on a large-scale video dataset in which each video has been classified into some specific semantic categories (based on its content), such as concert, animal, boxing, cooking show, and so on. Commonly, video level semantic category annotation is much easier and less ambiguous than frame level importance score labeling, which indicates that less efforts would be required to train this VCSN, compared with the workload for training a supervised summarization network directly. Then, regarding the input of the last fully connected layer in the frozen VCSN as feature representation of the raw input video, a video semantic reward can be evaluated by measuring the similarity between the summary video representation and the raw input video representation. The training step of our proposal is illustrated in Figure 1. As we can seen from this figure, in order to remove redundant footage in the raw video sequence, we first apply a video preprocessing step to cluster the consecutive similar frames into a sequence of video shots. Each video shot will be regarded as a basic summary element for following processes. Both the preprocessed input video and its summary are fed into the VCSN to obtain their semantic representations, respectively. A new training reward term R_{sem}, defined as the similarity measurement between the two video representations, is proposed to guide the reinforcement learning of the SGSN. By doing so, the learning procedure of our SGSN can also be considered as a weakly supervised upgrade from its original unsupervised version given in [33]. In addition, here we note that only the video preprocessing step and the trained SGSN are needed for inference, as shown in flow chart Figure 2.

We conduct extensive experience on four benchmark datasets: TVSum [27], SumMe [10], OVP [1] and YouTube [2], and evaluate algorithm performance based on three metrics: Kendall's τ and Spearman's ρ correlation coefficients [19] and F-Score [31]. Experimental results confirm that our proposed method outperforms other leading methods in video summarization.

We summarize our contributions as follows: (1) we present a new weakly supervised reinforcement learning solution for video summarization. In our proposal, the VCSN is introduced to guide the unsupervised reinforcement learning procedure of the SGSN; (2) a new semantic reward term is proposed to guide the unsupervised reinforcement learning procedure for summarization. This improvement can effectively help to generate a semantically meaningful summary from its original; (3) we introduce an efficient preprocessing step to reduce the redundant video content and shorten the input sequence for the following processes. It also makes the training converge faster; (4) we conduct extensive experiments on four benchmark datasets and confirm

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

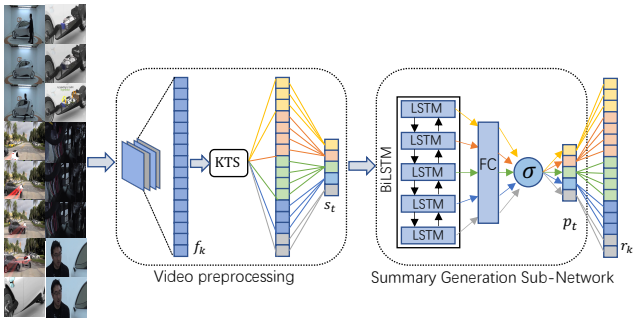

Figure 2. During inference, the video preprocessing step is first applied to obtain the shot level feature representation of the raw input video s_t, then the SGSN is used to predict the corresponding shot level importance score p_t. The final frame level importance scores for summarization r_k can be recovered based on the segmentation boundaries of the frame level feature representations f_k.

that our weakly supervised reinforcement learning method can reach a state-of-the-art performance for video summarization in terms of Kendall's τ and Spearman's ρ correlation coefficients.

2. Related Work

Video Summarization: Machine learning technology-based video summarization techniques have achieved significant improvement in recent years. As mentioned above, they can be classified into three categories. Supervised methods are straightforward and provide a strong baseline for reference. Zhang *et al*. [31] trained a dppLSTM using training data with frame level importance score annotations. Due to the difficulty to label frame level importance scores for a large amount of training data, more researchers paid their attention to develop weakly supervised or unsupervised learning-based methods. Instead of annotating training data, different implementation rules like frame level clustering or specially designed learning rewards, are proposed to solve the summarization problems in an unsupervised way, as proposed by Zhou *et al*. in [33]. In contrast, some high-level semantic knowledge, even a small amount of annotated frame importance scores data, are involved in the weakly supervised training procedures for better model learning. For example, Cai *et al*. [5] presented a generative model with weakly supervised semantic constraint to generate topic-associated summaries. A variational autoencoder (VAE) was first trained to learn the latent semantic video representations from web videos, then a simple encoder-decoder with attention as well as sampled latent variable was presented for summarization. In this paper, we also treat video level semantic information as an additional constraint condition to enhance the summarization quality.

Video Classification: Recently, with the availability of large-scale video datasets, such as YouTube-8M [1], automatic video classification has attracted more and more attention. Commonly, video classification needs massive computational power and takes temporal information into account. Recurrent Neural Networks [3, 4] like LSTM and GRU are usually applied here to learn temporal dependencies from frame-level feature space. These methods first employ sophisticated image representation techniques to convert video streams into frame level feature sequences, then use the RNNs to learn spatiotemporal relationships in the feature space. The great success of 2D CNNs in image classification also triggered many researchers to upgrade 2D CNNs to their corresponding 3D cases [6, 8, 14]. The introduction of an additional temporal dimension to 2D convolution networks makes the training of these networks more challenging. Some researchers therefore proposed pseudo 3D [23] and "R(2+1)D" [28] solutions to alleviate computational cost. Some local frame descriptors are aggregated into a global compact vector for video representation and classification in BOW [25], FV [21], NetVlad [13] and NeXtVlad [16]. These methods demonstrated a great balance of computational efficiency and algorithm performance for this task. In our work, we apply NeXtVlad method to construct our VCSN, based on its outstanding performance in large-scale video classification [11, 32, 15].

Reinforcement Learning (RL): RL is well known for its superior capability of solving decision-making problems. It also demonstrates a great availability in computer vision applications. Sahba *et al*. [24] trained an opposition-based Q-learning model for image segmentation. Mnih *et al*. [18] proposed a variant of the Q-learning algorithm to learn game control policies directly from raw video data in complex RL environments. Xu *et al*. [30] applied RL technique to propose an encoder-decoder with "hard" attention to solve image captioning problems. Furuta *et al*. [9] applied a new pixel-wise reward to extend the application of deep RL to various low-level image processing applications, such as image denoising, image restoration, and local color enhancement *etc*.

Video summarization, aiming to select important key frames from the input frame sequence, can be also considered as a decision-making problem [26, 33, 7]. Based on the key frame labels and category information of the training video, Song *et al*. [26] proposed a RL model to select category-specific key frames. Limited by the number of annotated summary data, Zhou *et al*. [33] introduced a combined diversity-representativeness reward to guide the learning of an unsupervised RL model. To solve the sparse reward problem in RL, Chen *et al*. [7] decomposed the whole task into several subtasks and presented a hierarchical RL framework for summarization. Though this method achieves the state-of-the-art results, human annotated importance scores are necessary to train the model. Different from their work, in our paper, we introduce an additional video level semantic similarity reward to guide the unsupervised RL procedure, which can avoid the tedious frame level importance score annotation work. We also introduce an effective video segmentation method to reduce redundant content and shorten the input sequence. This process can help to alleviate the sparse reward problem, especially for long-length

input videos.

3. Proposed Method

As defined in [33, 7], we formulate video summarization as a sequential decision-making problem in which frame level importance scores are predicted for summary frame selection. In [33], Zhou *et al*. combined two frame level clustering rewards, diversity reward and representativeness reward to guide an unsupervised RL process for the task. Inspired by recent progress in video representation techniques, here we further introduce video level semantic similarity as an additional reward to weakly supervise the RL procedure. The rationality of this idea stems from our observation that a good video summarization should also be semantically identical to its original source. The semantic similarity measurement can therefore play a supervisor role in our task. In this paper, we employ a VCSN to extract video semantic representations. The similarity between the representation of the raw input video and that of its summary will then be considered as an additional constraint condition to construct a semantically meaningful reward to guide the learning of our SGSN. In practice, we find that the training process is sometimes hard to converge due to the inherent sparse reward problem of RL. We therefore apply a KTS algorithm module to first cluster the original video sequence into a sequence of video shots. Each video shot will be regarded as a basic summary element for summarization. We find this preprocessing can effectively help to improve the training of our model. We will describe our work in detail in the following sections.

3.1. Video Preprocessing

Commonly, reinforcement learning-based video summarization approaches may face a sparse reward problem, which is inherently caused by the learning mechanism of RL that the agents can only receive the reward after the whole summary is generated. This problem becomes more serious when the inputs are long-length videos, even sometimes makes RL hard to converge. Here we apply a Kernel Temporal Segmentation (KTS) algorithm [22] to segment the consecutive similar frames into T video shots, as shown in Figure 1. This KTS algorithm calculates shot boundaries based on frame feature similarity measurement, so different shots may have different numbers of covered frames. Since a video shot can also be considered as a content segment captured by a temporal sliding window, it is quite similar to the fact that human annotators always like to scroll forward and backward to review the video content in adjacent frames for frame level importance score annotation. In practice, referring to the preprocessing step of the famous Youtube-8M challenge [1], we first feed each frame image of the raw input video into an Inception-V3 feature extractor and apply Principal Component Analysis (PCA) transformation to obtain the frame level feature representations. For each shot clustered by applying KTS algorithm to the input video, a shot level feature representation is then calculated as the mean of all the frame level feature representation vectors covered by the boundary of this video shot. It can be formulated as:

$$s_t = \frac{\sum_{k=i_t}^{i_{t+1}-1} f_k}{i_{t+1} - i_t}, \tag{1}$$

where s_t stands for t^{th} shot level feature representation, i_t denotes the index of the first frame in the t^{th} video shot, f_k represents the feature representation vector of the k^{th} frame extracted by Inception-V3 feature extractor and the followed PCA transformation. After this video preprocessing step, an input video can be converted into a sequence of shot level feature representations. It can significantly benefit our training, particularly on long-length video sequences.

3.2. Video Classification Sub-Network (VCSN)

The video representation can be seen as a by-product of video classification tasks. In our work, we introduce NeXtVlad model [16], which has shown promising performance in large-scale video classification task, to train the VCSN. This network will be used to generate video level semantic representation of the input video. Any video dataset with category annotations can be used to train NeXtVlad network. Here we use Youtube-8M dataset, which contains 6 million videos with 3,862 class labels. Since each video sample in Youtube-8M dataset may contain multiple labels, we define our task as a multi-class multi-label video classification problem. The training loss can be written as:

$$\text{loss}_{bce} = -\frac{1}{M}\sum_{i=0}^{M} t_i \log\left(o_i\right) + \left(1 - t_i\right)\log\left(1 - o_i\right), \tag{2}$$

where the subscript bce means that it is a binary cross entropy loss for solving this multi-class multi-label classification problem, M denotes the total number of categories, t_i represents the i^{th} target category, and o_i stands for the i^{th} output prediction. We follow the parameter settings given in [16] to train this network. After training, the network structure and weights will be fixed. We consider the input of the last fully connected (FC) layer of VCSN as the video level semantic representation of the input video/frames.

3.3. Summary Generation Sub-Network (SGSN)

The backbone of our SGSN is constructed as a bidirectional LSTM (BiLSTM) topped with a FC layer (see Figure 1). The input sequence of this network is the shot level feature representations $\{s_t\}_{t=1}^{T}$ obtained by the video preprocessing step. A sigmoid function is applied after the FC layer. We regard the output of the sigmoid function as the importance score of the corresponding input video shot, which indicates the probability that this video shot should be selected as a part of the final summary. This process can be formulated as Eq. 3. Bernoulli sampling is subsequently applied to select video shots.

$$p_t = \text{sigmoid}\left(Wh_t\right), \tag{3}$$

$$a_t \sim \text{ Bernoulli}\left(p_t\right), \tag{4}$$

In Eq. 3, $\{p_t\}_{t=1}^{T}$ represents the estimated importance score for the input video shot, $a_t \in \{0, 1\}$ represents if a t^{th} video shot is selected or not, h_t is the hidden state of BiLSTM, W is the learnable parameters.

3.4. Reward Functions

Unsupervised reward: In [33], an unsupervised diversity-representativeness reward R_{div} + R_{rep} is defined to jointly guide the RL for video summarization. In this composed reward, R_{div} represents the degree of diversity of the generated summaries, it measures the mean of the pairwise dissimilarities among the selected shot features. While R_{rep} measures how well the summary frames can represent the input video, it calculates the minimum distance between each selected shot features and input shot features. We also apply these two rewards to our task. The definitions of these two rewards can be written as:

$$R_{div} = \frac{1}{Y(Y-1)} \sum_{t \in \mathcal{Y}} \sum_{\substack{t' \in \mathcal{Y} \\ t' \neq t}} d\left(s_t, s_{t'}\right), \tag{5}$$

$$d\left(s_t, s_{t'}\right) = 1 - \frac{s_t^T s_{t'}}{\|s_t\|_2 \|s_{t'}\|_2}, \tag{6}$$

$$R_{rep} = \exp\left(-\frac{1}{T} \sum_{t=1}^{T} \min_{t' \in \mathcal{Y}} \|s_t - s_{t'}\|_2\right), \tag{7}$$

where $d(\cdot, \cdot)$ in Eq. 6 is the dissimilarity function, the indices of the selected shot level feature representations are $\mathcal{Y} = \{y_i | a_{y_i} = 1, i = 1, \ldots, Y\}$.

Supervised reward: In this paper, we propose a new semantic reward R_{sem} to measure how well the summary is semantically identical to its original source. This reward will play a supervisor role to guide the training. Through applying the proposed VCSN to the input video s_t and its summary s_y respectively, two corresponding video representations $\mathrm{VCSN}(s_t)$ and $\mathrm{VCSN}(s_y)$ can be obtained. A supervised reward will then be calculated as the similarity measurement between these two representations by,

$$R_{sem} = \mathrm{sim}\left(\mathrm{VCSN}(s_t), \mathrm{VCSN}(s_y)\right), \tag{8}$$

where $\mathrm{VCSN}(\cdot)$ denotes the process to extract the video level semantic representation vector by using the proposed VCSN model; $\mathrm{sim}(\cdot, \cdot)$ is a similarity function, in practice, we use Cosine similarity measurement.

Weakly supervised reward: We combine the supervised semantic reward R_{sem} with the unsupervised rewards R_{div} and R_{rep} to jointly train the SGSN. Therefore, we can easily upgrade an unsupervised RL method-based summarization approach to its weakly supervised version. A new semantically meaningful reward for the weakly supervised RL can therefore be formulated as:

$$R = R_{div} + R_{rep} + R_{sem}, \tag{9}$$

As mentioned in [33] and [17], due to the nature of video summarization, selecting more or even all frames will increase the rewards for the learning of RL. A regularization term is therefore imposed to constrain the percentage of frames selected for the summary in these two papers. Different from these two methods, here we further put forward a new summary length reward R_{len} that helps to constrain the length of the generated summaries. Its definition is:

$$R_{len} = 1 - \left(\frac{p_{len} - \varepsilon}{\max(\varepsilon, 1 - \varepsilon)}\right)^2, \quad p_{len} = \frac{Y}{T}, \tag{10}$$

where the reward term R_{len} represents the ratio of the number of selected video shots to the total number of shots, ε is an expected length percentage factor. With this summary length reward term, our semantically meaningful reward Eq. 9 can be updated to:

$$R = R_{div} + R_{rep} + R_{sem} + R_{len}, \tag{11}$$

As will be seen in the Experiments Section, this full combination of reward terms performs more robust than other combination cases. During training, we assign zero reward to R if none of the frames are selected.

3.5. Optimization

The SGSN is trained with REINFORCE algorithm [29], aiming to learn a policy function π_θ to maximize the expected rewards.

$$J(\theta) = E_{p_\theta(a_{1:T})}[R], \tag{12}$$

where θ denotes the trainable parameters of the summary generated sub-network, a_t is the action taken by time t, $p_\theta(a_{1:T})$ is the probability of the action sequence. R is the weakly supervised reward defined by Eq. 11. Following the REINFORCE algorithm, the derivative of the objective function Eq. 12 can be computed as:

$$\nabla_\theta J(\theta) = E_{p_\theta(a_{1:T})}\left[R \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta\left(a_t | h_t\right)\right], \tag{13}$$

where h_t is the hidden state of the BiLSTM. We introduce Monte-Carlo policy gradient method to solve this equation. By running the agent for N episodes for each input sequence, Eq. 13 can be approximately computed as:

$$\nabla J(\theta) \approx \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T} \left(R_n - b\right) \nabla_\theta \log \pi_\theta\left(a_t | p_t\right), \tag{14}$$

here b is defined as the moving average of reward R, n represents the n^{th} episode. p_t is the probability of a_t.

The pseudo code of the proposed SGSN RL is given in Algorithm 1.

4. Experiments

4.1. Dataset

As mentioned above, we apply the large-scale dataset Youtube-8M to train our VCSN. The evaluations of algorithm performances are based on two benchmark datasets: TVSum [27] and SumMe [10]. TVSum contains 50 videos

Algorithm 1 Proposed SGSN REINFORCE Learning pseudo code

for e : 1, 2, ... to $numEpoch$ **do**
  for each $\{s_t\}_{t=1}^{T}$ in input videos **do**
    p_t = SGSN($\{s_t\}_{t=1}^{T}$)
    $avgEpisodeCost$ = 0.0
    for n : 1, 2, ... to $numEpisode$ **do**
      $a_t \sim$ Bernoulli(p_t) (Eq. 4)
      $log_m = -\frac{1}{T}\sum_{i=i}^{T} a_i \log(p_i) + (1-a_i)\log(1-p_i)$
      $R = R_{div} + R_{rep} + R_{sem} + R_{len}$ (Eq. 11)
      $avgEpisodeCost$ += $(R - baseline) * log_m$ (Eq. 14)
    end for
    $avgEpisodeCost = \frac{avgEpisodeCost}{numEpisode}$
    $avgEpisodeCost$.backward()
  end for
end for

annotated by 20 persons, ranging from 2 to 10 minutes; SumMe includes 25 videos annotated by 15 to 18 persons, ranging from 1 to 6 minutes. Refer to paper [31], we also consider OVP and YouTube [2] data to construct the "Augmented" and "Transfer" sets for evaluations. More details about these two settings will be described in Section 4.3.

4.2. Evaluation metric

Rank correlation coefficient: Two well established metrics, Kendall's τ and Spearman's ρ correlation coefficients are recently introduced to measure the strength of relationship between the predicted rankings of video summaries and human annotated frame level importance scores. They have been verified as the robust criteria to evaluate performance of video summarization [19]. We also apply these two metrics for performance evaluation in this paper.

F-score: F-score is a commonly used metric to evaluate the performance of video summarization. However, as Otani *et al*. pointed out in [19], F-score may not be stable for this task because it is highly determined by the distribution of video segment lengths. For instance, F-score tends to get higher as summary length gets longer. Even so, here we still consider F-score as a complementary metric for our comparison. Considering the intersection of two videos, the definition of F-score can be given as:

$$P = \frac{\text{overlapped}(A,B)}{A}, R = \frac{\text{overlapped}(A,B)}{B}, \quad (15)$$

$$F = \frac{2PR}{P+R} \times 100\%, \quad (16)$$

where A is the ground truth summary and B is the generated summary. There we can see that this metric describes the overlapped duration of the generated summary and its ground truth. For fair comparisons with previous methods, we strictly follow Zhou's method [33] to deal with the multiple ground truth summary problem.

4.3. Evaluation settings

Following Zhang's suggestion [31], we study three settings in performance evaluation: (1) Canonical: we use standard 5-fold cross validation (5FCV), which means 80% of the selected dataset as training set and remaining 20% as testing data; (2) Augmented: based on the Canonical result, we complement the other three datasets to the training set; (3) Transfer: pick one dataset (TVSum or SumMe) as training set, the other three as testing set.

SGSN	Reward combinations	τ	ρ
Unsupervised [33]	$R_{len} + R_{rep} + R_{div}$	0.058	0.076
Weakly supervised by LSTM-based VCSN	$R_{len} + R_{sem}$	0.063	0.082
	$R_{len} + R_{sem} + R_{div}$	0.063	0.082
	$R_{len} + R_{sem} + R_{rep}$	0.063	0.083
	$R_{len} + R_{sem} + R_{rep} + R_{div}$	0.064	0.084
Weakly supervised by NeXtVlad-based VCSN	$R_{len} + R_{sem}$	0.083	0.108
	$R_{len} + R_{sem} + R_{div}$	0.085	0.111
	$R_{len} + R_{sem} + R_{rep}$	0.090	0.117
	$R_{len} + R_{sem} + R_{rep} + R_{div}$	**0.094**	**0.122**

Table 1. Kendall's τ and Spearman's ρ correlation coefficient-based SGSN performance evaluation on TVSum dataset.

Method	τ	ρ	labels
dppLSTM [31]	0.042	0.055	470
DR-DSN [33]	0.02	0.026	-
Hierarchical RL [7]	0.078	0.116	24
Our Proposal	**0.094**	**0.122**	-
Human Annotations	0.177	0.204	-

Table 2. Kendall's τ and Spearman's ρ correlation coefficients-based comparisons on TVSum dataset in the Augmented setting.

4.4. Implementation details

Video preprocessing: In order to reduce computational time, we sample one frame per second for all the training and testing video samples. The frame level feature representations are obtained by the Inception-V3 and PCA transformation, as presented in [1]. KTS algorithm is then applied to segment the consecutive similar frames into a sequence of video shots. Here we set the maximum number of segmented video shots to 50. For those videos shorter than 50 frames, KTS segmentation will not be applied, i.e. the frame level feature representations will be directly regarded as the shot level feature representations. **VCSN**: we try two basic architectures, LSTM and NeXtVlad, to construct the VCSN. Following the parameter setting given in [16], we set 2 LSTM layers with hidden size 1024 and learning rate 2e-4 for the LSTM case, and set 8 groups, 2 expansions with cluster size 128, hidden size 2048 and learning rate 2e-4 for the NeXtVlad case. All these two VCSN cases are trained on Youtube-8M training set. **SGSN**: The input of SGSN is the shot level feature representations. We set dimension of the hidden state of the BiLSTM to 256, length ratio ε 50%, learning rate 2e-5, and baseline equals to the moving average of the learning reward.

	Method	SumMe			TVSum		
		Canonical	Augmented	Transfer	Canonical	Augmented	Transfer
Supervised	vsLSTM [31]	37.6	41.6	40.7	54.2	57.9	56.9
	dppLSTM [31]	38.6	42.9	41.8	54.7	59.6	58.7
	SUM-GAN$_{sup}$ [17]	41.7	43.6	-	56.3	**61.2**	-
	DR-DSN$_{sup}$ [33]	42.1	43.9	42.6	58.1	59.8	**58.9**
Unsupervised	SUM-GAN [17]	39.1	43.4	-	51.7	59.5	-
	DR-DSN [33]	41.4	42.8	42.4	57.6	58.4	57.8
Weakly Supervised	Hierarchical RL [7]	**43.6**	44.5	42.4	**58.4**	58.5	58.3
	Our Proposal	41.5	**44.9**	**43.8**	55.7	59.1	58.7

Table 3. Evaluation by F-score on SumMe and TVSum in the Canonical, Augmented and Transfer Settings, respectively.

4.5. Comparisons

Refer to [16], we first introduce the Global Average Precision (GAP) metric to evaluate the performance of the two VCSN cases for video classification. Our experiments show that the NeXtVlad-based VCSN can reach higher GAP (0.856) than the LSTM-based VCSN (0.830), which are consistent with the results reported in [16]. Based on the trained VCSN, we can extract the video semantic representations to guide the RL of the SGSN. To demonstrate how the new video semantic representations can help to improve video summarization, we apply Kendall's τ and Spearman's ρ correlation coefficients to evaluate performance of the trained SGSNs with different combinations of VCSN cases as well as reward terms. The results tested on augmented setting of TVSum dataset are listed in Table 1, where the second column shows the reward combination methods. For example, $R_{len} + R_{rep} + R_{div}$ represents that the learning reward for RL is a summation of summary length, representativeness and diversity reward terms, similar to the method proposed in [33]; $R_{len} + R_{sem}$ denotes that the learning reward is a summation of summary length and our new semantic reward terms, and so on.

From Table 1 we can see, our new video semantic reward can help to improve the summarization performance apparently, even if only the video semantic and summary length reward terms are applied (compare row "Unsupervised" to two cases of reward combination "$R_{len} + R_{sem}$" supervised by different VCSNs). By comparison, we can see the algorithm performance can be steadily improved while more reward terms are added to supervise the RL procedure. Unsurprisingly, all the SGSN cases supervised by the NeXtVlad-based VCSN perform significantly better than those cases supervised by the LSTM-based VCSN. In all these solutions, we find the SGSN case with the full reward combination $R_{len} + R_{sem} + R_{rep} + R_{div}$ and supervised by NeXtVlad-based VCSN can reach the best performance (see the last row in Table 1). We therefore choose this case as our preferred solution for video summarization.

We compare our method with other leading video summarization methods on TVSum dataset in the Augmented setting. The experimental results are listed in Table 2, where dppLSTM, DR-DSN and Hierarchical RL are supervised, unsupervised and weakly supervised methods, respectively. From this table, we can see that our proposal works considerably better than others. What's more, compared with the state-of-the-art Hierarchical RL method [7], our method is more practical because it does not require any frame level or shot level importance score annotations whereas shot level importance score annotation is necessary for the Hierarchical RL method.

We also apply F-score to compare the algorithm performance on two benchmark datasets, SumMe and TVSum. The results are revealed in Table 3. From this table, we can see that our method performs a bit worse than some others in the Canonical settings of the two datasets. We note that it is a reasonable result because the introduction of our new semantic reward R_{sem} increases the RL complexity of the SGSN, however, the limited training data in the Canonical setting maybe insufficient to support the RL procedure to fit the model very well. In contrast, more training data in the Augmented setting can benefit our RL obviously, as shown in Table 3 where our proposal outperforms the two RL baseline methods DR-DSN and Hierarchical RL on both two benchmark datasets. The same situation can also be identified in the Transfer settings. In addition, compared with supervised methods, our method performs better than vsLSTM method on two datasets but worse than SUM-GAN$_{sup}$ and DR-DSN$_{sup}$ methods on TVSum dataset. For SumMe dataset, our method even outperforms the majority of the listed supervised methods, particularly in the Augmented and Transfer settings.

Finally, a visual comparison on a real TVSum test video is given in Figure 3. The original video is about pet grooming. We compare the summary frames picked by using our SGSN and DR-DSN method. As seen, thanks to the newly introduced semantic constraint, our result skips many irrelevant frames, including title frames and people walking footage *etc.*, and selects most of the frames that show the details of pet grooming, comparing to DR-DSN also picks some irrelevant frames at the beginning of this video. A higher F-score is therefore achieved by using our method.

5. Conclusion

In this paper, we propose a weakly supervised reinforcement learning method for video summarization. Our proposal consists of two sub-networks: video classification sub-

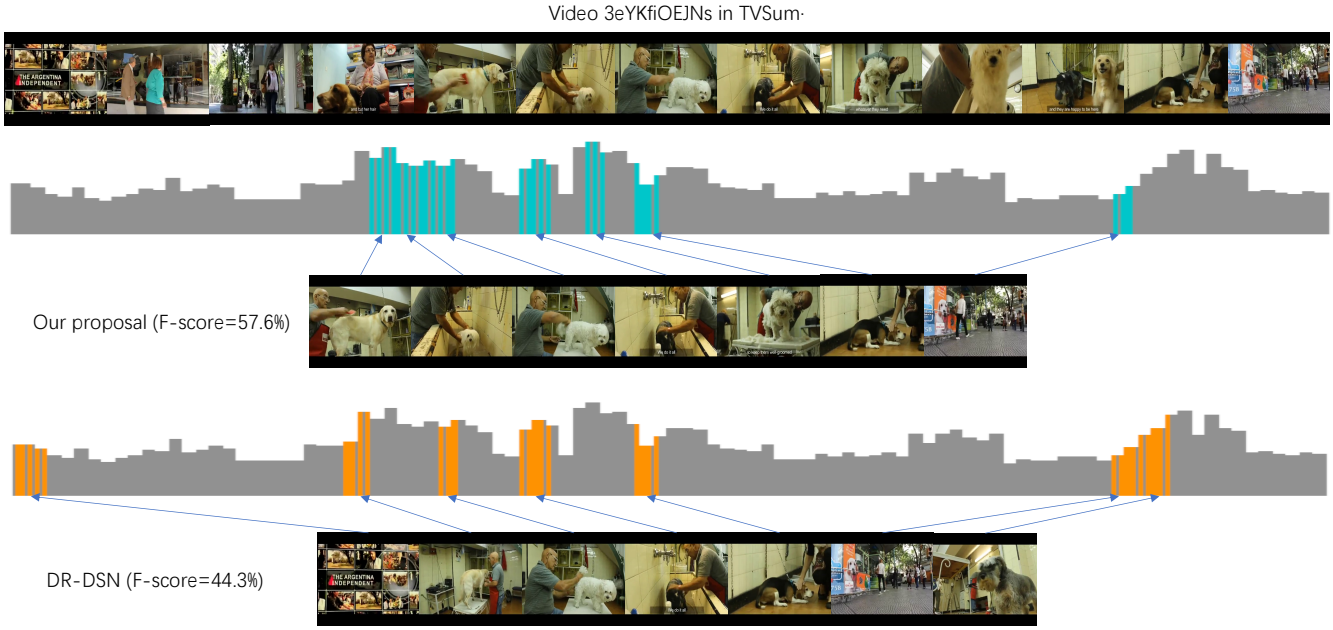

Figure 3. Examplar video summary generated by our proposal and DR-DSN, along with the ground-truth importance scores (gray background). Cyan bars represent the summaries selected by our method, while orange bars stand for the summaries selected by DR-DSN.

network and video summary generation sub-network, where the former sub-network plays a supervisor role to train the latter sub-network. A semantically meaningful reward, formulated as a combination of a new semantic reward term, a summary length reward term and the other two unsupervised reward terms, is proposed to guide the learning of our reinforcement learning model. By doing so, an unsupervised reinforcement learning-based video summarization method can be easily upgraded to its weakly supervised version, leading to the dramatically enhanced performance of summarization. Experimental results revealed that our proposed method significantly surpasses other unsupervised and even supervised methods for video summarization, and achieves state-of-the-art performance in terms of Kendall's τ and Spearman's ρ correlation coefficients.

References

[1] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv*, abs/1609.08675, 2016.

[2] S. Avila, A. Lopes, A. da Luz, and A. Araújo. Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32:56–68, 2011.

[3] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep learning for human action recognition. In *LNCS*, 2011.

[4] N. Ballas, L. Yao, C. Pal, and A. Courville. Delving deeper into convolutional networks for learning video representations. In *ICLR*, 2015.

[5] S. Cai, W. Zuo, L. S. Davis, and L. Zhang. Weakly-supervised video summarization using variational encoder-decoder and web prior. In *ECCV*, 2018.

[6] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[7] Y. Chen, L. Tao, X. Wang, and T. Yamasaki. Weakly supervised video summarization by hierarchical reinforcement learning. *Proceedings of the ACM Multimedia Asia*, 2019.

[8] F. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In *ICCV*, 2018.

[9] R. Furuta, N. Inoue, and T. Yamasaki. Fully convolutional network with multi-step reinforcement learning for image processing. In *AAAI*, 2019.

[10] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool. Creating summaries from user videos. In *ECCV*, 2014.

[11] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv*, abs/1503.02531, 2015.

[12] Y. Jung, D. Cho, D. Kim, S. Woo, and I. S. Kweon. Discriminative feature learning for unsupervised video summarization. In *AAAI*, 2019.

[13] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *CVPR*, 2010.

[14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. Li. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[15] Z. Li and D. Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.

[16] R. Lin, J. Xiao, and J. Fan. Nextvlad: An efficient neural network to aggregate frame-level features for large-scale video classification. In *ECCV workshop*, 2018.

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

[18] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. In *NIPS Deep Learning Workshop*, 2013.

[19] M. Otani, Y. Nakashima, E. Rahtu, and J. Heikkila. Rethinking the evaluation of video summaries. In *CVPR*, 2019.

[20] M. Otani, Y. Nakashima, E. Rahtu, J. Heikkilä, and N. Yokoya. Video summarization using deep semantic features. In *ACCV*, 2017.

[21] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.

[22] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid. Category-specific video summarization. In *ECCV*, 2014.

[23] Z. Qiu, T. Yao, and T. Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *ICCV*, 2017.

[24] F. Sahba, H. R. Tizhoosh, and M. M. M. A. Salama. Application of opposition-based reinforcement learning in image segmentation. In *Proceedings of the IEEE Symposium on Computational Intelligence in Image and Signal Processing*, 2007.

[25] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *ICCV*, 2003.

[26] X. Song, K. Chen, J. Lei, L. Sun, Z. Wang, L. Xie, and M. Song. Category driven deep recurrent neural network for video summarization. In *ICME Workshop*, 2016.

[27] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes. Tvsum: Summarizing web videos using titles. In *CVPR*, 2015.

[28] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018.

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

[30] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[31] K. Zhang, W. Chao, F. Sha, and K. Grauman. Video summarization with long short-term memory. In *ECCV*, 2016.

[32] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu. Deep mutual learning. In *CVPR*, 2018.

[33] K. Zhou, Q. Yu, and X. Tao. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *AAAI*, 2017.