Video Summarization via Label Distributions Dual-Reward

Yongbiao Gao , **Ning Xu** and **Xin Geng** *
School of Computer Science and Engineering, Southeast University, Nanjing, China
{gaoyb, xning, xgeng}@seu.edu.cn

Abstract

Reinforcement learning maps from perceived state representation to actions, which is adopted to solve the video summarization problem. The reward is crucial for dealing with the video summarization task via reinforcement learning, since the reward signal defines the goal of video summarization. However, existing reward mechanism in reinforcement learning cannot handle the ambiguity which appears frequently in video summarization, i.e., the diverse consciousness by different people on the same video. To solve this problem, in this paper a label distribution is mapped from the CNN and LSTM-based state representation to capture the subjectiveness of video summaries. The dual-reward is designed by measuring the similarity between user score distribution and the generated label distribution. Not only the average score but also the the variance of the subjective opinions are considered in summary generation. Experimental results on several benchmark datasets show that our proposed method outperforms other approaches under various settings.

1 Introduction

Video summarization aims to produce a compact short video summary, which preserves the most representative sequence of frames/shots. Deep reinforcement learning has been introduced to interactive video summarization to capture the dynamic patterns of key-frames during the interactive with the video. However, reinforcement learning is limited to deal with the ambiguous applications due to the restricted scalar reward mechanism. Especially, in video summarization, the label ambiguity appears frequently since people have subjective consciousness about the importance of video frames. Inspired by label distribution learning (LDL) [Geng, 2016], a novel label distribution-based dual-reward is designed to guide the reinforcement learning agent to solve the video summarization task. Besides, the action form of reinforcement learning is redefined from the perspective of label distributed learning. The proposed method has great potential in various challenging sequential decision making scenarios that require both ambiguity modeling and long-term planning.

In recent years, video summarization has attracted a resurgence of interest [Li *et al.*, 2020; Zheng and Lu, 2020], especially with reinforcement learning. For example, DR-DSN [Zhou *et al.*, 2018a] designs a reward function that jointly accounts for the diversity and representativeness of the generated summaries. A weakly hierarchical reinforcement learning framework [Chen *et al.*, 2019] decomposes the whole task into several subtasks to enhance the summarization quality. The related works only consider the binary labels or the average user scores of the summary frames. However, video summarization is a highly typical problem with label ambiguity since people have subjective preferences over the summaries they would like to watch. Different users have various attentions on the same video. Through the analysis of video summarization, we draw two significant conclusions. As shown in Figure 1(a) and in Figure 1(b), 1) two frames may have the same average score, but the variance of the distribution and the image content are quite different. It reveals that a single scalar average score is insufficient to capture the true nature of the key-frame. 2) the larger the absolute importance score is, the smaller the variance of the user score distribution is. It indicates that the people's opinions on the key-frames are more consistent, and the opinions on the non-important frames are more scattered. Therefore, the score distribution maintains crucial information.

The previous works only use the average scores or the generated binary results to learn the policy to select the key-frames. As analysis above, the score distribution contains vital information to determine whether the video frame is a key-frame. Therefore, we consider not only the average scores of frames but also the score distributions. In addition, we develop a new label enhancement framework to transform the user scores or the binary summary results into label distributions by leveraging the relations among the label space.

In this paper, a specialized reinforcement learning algorithm with label distributions dual-reward is designed for the task of video summarization to solve the summary ambiguous problem. First, we introduce a new label enhancement framework to recover the label distributions from the annotated user scores or the binary results. Specially, the annotated user scores from multiply annotators can be transformed into label distributions according to the ratio of the sum of each

*Corresponding Author

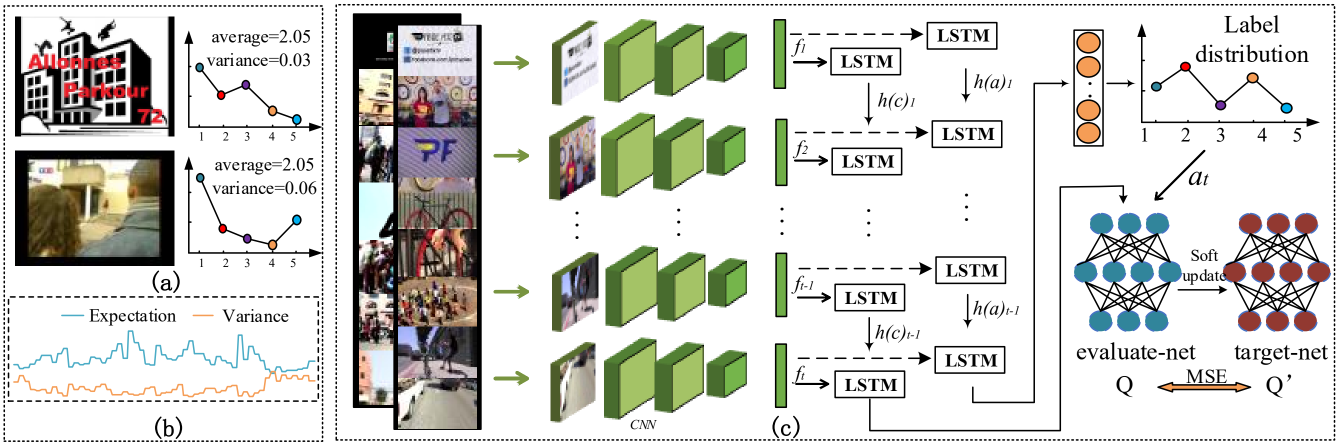

Figure 1: Even with the same average scores, the image content and the variance of distributions (frames 21 and 7212 from video 21) are different in TVSum dataset *(a)*. The relation between the expectation and the variance of the score distribution is illustrated in *(b)*. The framework of video summarization via label distributions dual-reward is shown in *(c)*.

annotated label to all annotators. The annotated binary results are converted into the normalized continuous scores via a proportionally scaling method to ensure the consistency of the relative distance among scores both in the original score space and the normalized score space. Then label distributions are recovered from the normalized scores in the training set by leveraging the topological information in the feature space and the correlation among the normalized scores by GLLE [Xu *et al.*, 2018]. The overall framework of video summarization is shown in Figure 1(c), CNN features of each frame are extracted by the pre-trained CNN model. Then the state representations are obtained by the LSTM architecture to exploit temporal dependency among video frames. Unlike the related works that output a scalar value to present the importance of the frame, our model predicts a label distribution, which includes the absolute importance information as well as the distribution of the importance score. The action of the reinforcement learning agent is mapped from the generated label distribution.

To better handle the ambiguous problem in video summarization, a novel dual-reward is designed to guide the agent to capture the true nature of the key-frames. The final reward consists of two parts, label distribution-based reward and the expectation-based reward. The dual-reward will encourage the agent to predict a distribution whose average and variance are both close to the groundtruth. The variance of the label distribution is also aggregated in the generation of video summaries. Therefore, the results are more consistent with human perceived summaries.

To demonstrate the effectiveness of our proposed method, we conduct various experiments on two widely used datasets, TVSum [Song *et al.*, 2015] and SumMe [Gygli *et al.*, 2014]. Experimental results indicate that our method achieves the better performance, especially under the much better rank correlation coefficient metrics setting. All the datasets, the code as well as the trained models have been be released [1]. The contributions of our paper are concluded as follows, 1) we propose a new label distribution-based dual-reward to cover the ambiguity in video summarization. To the best of our knowledge, we are the first to handle this problem with label distribution. 2) We demonstrate a new label enhancement framework to transform the user scores or the binary summary results into label distributions to capture the average scores as well as the distribution of human subjective opinions. 3) Our proposed method not only outperforms the reinforcement learning-based methods but also the supervised/unsupervised approaches.

2 Related Work

Label distribution learning and label enhancement. Label distribution learning (LDL) [Geng, 2016] has been explored a lot in recent years. LDL is successfully applied to multiple ambiguous applications such as partial multi-label learning [Lv *et al.*, 2020], head pose estimation [Geng *et al.*, 2020], and facial age estimation [Smith-Miles and Geng, 2020], etc. LDL-SCL [X *et al.*, 2018] is proposed to encode the influence of local samples by a local correlation vector for each instance. A label distribution learning forests algorithm based on differentiable decision trees is presented by [Shen *et al.*, 2017]. More recently, LDL-ALSG [Chen *et al.*, 2020] is proposed to leverage the topological information of the labels from related but more distinct tasks and PENCIL [Yi and Wu, 2019] is demonstrated to update both network parameters and label estimations as label distribution to solve the problem of noisy labels. Due to the difficulty of obtaining the label distributions directly, label enhancement [Xu *et al.*, 2018] is proposed to recover label distributions from the logical labels. RLLE [Gao *et al.*, 2020] algorithm formulates the label enhancement as a dynamic decision process to sequential adjust the label distribution via the prior knowledge. Inspired by LDL, we design a novel label distribution-based reward and propose a new label enhancement framework to generate the label distribution for video summarization.

Video summarization. Earlier works mainly focus on unsupervised basic visual features clustering. An unsupervised

[1] http://palm.seu.edu.cn/xgeng/

discriminator [Mahasseni *et al.*, 2017] is designed to compare the summaries generated by GANs and the original video. CSNet [Jung *et al.*, 2019] designs a variance loss to predict output scores for each frame with high discrepancy. k-SDPP [Zheng and Lu, 2020] partitions sampled frames of a video into segments as well as considering sequential nature of the frames. A supervised paradigm that predicts the importance scores of the frames/shots directly [Zhang *et al.*, 2016a; H *et al.*, 2018] is proposed in recent years. For example, stacked memory network [Wang *et al.*, 2019] is proposed to explicitly model the long dependency among videos summaries. More recently, reinforcement learning-based methods [Zhou *et al.*, 2018a; Zhou *et al.*, 2018b; Zhang *et al.*, 2019] have been proposed to obtain the policy for the key-frame selection. DR-DSN [Zhou *et al.*, 2018a] designs a reward function that jointly accounts for the diversity and representativeness of the generated summaries. The hierarchical reinforcement learning [Chen *et al.*, 2019] is used to decompose the whole task into several subtasks to enhance the summarization quality. However, related works seldom pay sufficient attention to the ambiguous problem in video summarization. Query-focused video summarization [Zhang *et al.*, 2019] only produces different summaries corresponding to different user queries. It cannot handle the video summarization with ambiguity. This fact encourages us to explore new methods for video summarization.

3 Approach

We formulate video summarization as a sequential-making process and study using label distributions dual-reward to guide the agent to learn the key-frame selection policy.

3.1 Label Enhancement

According to the methodology in [Geng, 2016], let $\mathcal{X} = \mathbb{R}^q$ denote that input space, $\mathcal{Y} = \{y_1, y_2, \ldots, y_c\}$ denote the complete set of labels, and $d_{\mathbf{x}}^{y}$ denote the description degree of the label $y \in \mathcal{Y}$ to the instance $\mathbf{x} \in \mathcal{X}$, where c indicates the number of labels. Without loss of generality, assume that $d_{\mathbf{x}}^{y} \in [0, 1]$. Further suppose that the label set is complete, *i.e.* all labels in the set can always fully describe the instance, $d_i = (d_{\mathbf{x}_i}^{y_1}, d_{\mathbf{x}_i}^{y_2}, \ldots, d_{\mathbf{x}_i}^{y_c})^T$. Then, $\sum_y d_{\mathbf{x}}^{y} = 1$. The benchmark dataset TVSum [Song *et al.*, 2015] provides average importance score for each video frame, and also releases the annotated information for every annotator. For TVSum dataset, inspired by [Geng and Hou, 2015], the annotated scores can be transformed into label distributions. The abscissa indicates the annotated scores y, and the ordinate represents the ratio of the corresponding score among all the annotators. In LDL formulation, it indicates the description degree $d_{\mathbf{x}}^{y}$. For SumMe [Gygli *et al.*, 2014] dataset, each annotator annotates a binary score, if the annotator considers it to be a key-frame/key-shot, it is annotated with 1, otherwise it is annotated with 0. The importance scores are obtained by the voting strategy. For example, for one frame/shot, 10 out of 15 annotators voted this frame/shot as the summary frame/shot. Then, the importance score is $10/15 \approx 0.67$. For those datasets without detailed score annotation, we propose a label enhancement framework to generated the label distribution. First, KTS [Potapov *et al.*, 2014] segmentation algorithm is used to partition the sequence of frames into shots. Second, inspired by [Zhang *et al.*, 2016a], the importance score of each shot is calculated by averaging frame-level scores within the same shot. Third, we propose a proportionally scaling function to transform the continuous score into a discrete label space while ensuring the relative distance among the discrete labels constant with that in the original score space. The scaling function converts the importance scores into a discrete label space of 1 to 5 as follows,

$$y = \lceil \frac{(y_{max} - y_{min}) * (score - score_{min})}{score_{max} - score_{min}} + y_{min} \rceil, \quad (1)$$

where y_{min} and y_{max} are the upper and lower limits of the label space. $score_{max}$ and $score_{min}$ are the maximum and minimum values of the importance score. The symbol $\lceil \cdot \rceil$ indicates the rounding up operation. Finally, the transformed discrete labels are used as the relevant labels to recover the label distribution by GLLE [Xu *et al.*, 2018]. The proposed label enhancement framework provides the solution that can be applied to any video even if there is no detailed user score annotation.

3.2 State Representation

The state representation is based on CNN and LSTM networks. First, a convolutional neural network (CNN) is used to extract visual features $\{\mathbf{x}_t\}_{t=1}^{T}$ from the input video frames $\{v_t\}_{t=1}^{T}$ with the length T as shown in Figure 1(c). Then the sequential features are input into LSTM networks to capture the long-range temporal dependency of videos. The output of LSTM is defined as the state put into the summarization network to predict the label distribution. For fair comparison, we use the output from the penultimate layer of GoogLeNet [Szegedy and Liu, 2015], which is pre-trained on ImageNet [Deng *et al.*, 2009], as our CNN features (1024-dimensions). During the training process, the parameters of the feature extraction network remain unchanged, and only the LSTM and summarization networks are updated. Using the pre-trained model to extract CNN features can save the processing time of feature learning, thereby reducing the time required for policy learning.

The proposed video summarization architecture consists of two parts, one is the actor network that takes state representation s_t as input and maintains a parameterized actor function $(d_t, a_t) = u(s_t; \theta^u)$, which specifies the current policy by deterministically mapping state representation to a specific label distribution, where θ^u is the parameters of the actor function. Action a_t is mapped from the expectation of the label distribution. The details of the actor function $u(s_t; \theta^u)$ are as follows,

$$g_t = \sigma(s_t; \theta^u), \quad (2)$$

$$d_t = softmax(g_t), \quad (3)$$

$$a_t = \sum_{i=1}^{c} (i) * d_t^i, \quad (4)$$

where σ is the activate function, and g_t is the penultimate output of the actor network. According to the definition of label distribution learning, an instance is described by all labels, $\sum_y d_{\mathbf{x}}^y = 1$. Therefore, we use a softmax function to normalize the output to generate the label distribution. Another part is the critic network that takes state s_t as input and maintains a value network parameterized by θ^Q to approximate the state value, where Q indicates the critic network. More details of the learning algorithm are illustrated in Section 3.4. During training stage, at each time step t, the agent receives the state s_t and generates the label distribution d_t. Then action a_t is conducted based on the expectation of the generated label distribution.

3.3 Label Distributions Dual-Reward

The reward $R(s_t)$ is combined of label distribution-based reward and expectation-based reward. Labeling the importance of the frame/shot by the average score will lose the difference of image content and variance of the score distribution. We evaluate the generated label distribution by measuring the similarity with the ground truth.

$$R_{ld} = \begin{cases} exp(-dis(d_t, d_t^{'})), & \text{if } dis(d_t, d_t^{'}) > \delta, \\ 1, & \text{otherwise,} \end{cases} \tag{5}$$

where

$$dis(d_t, d_t^{'}) = \frac{1}{c}\sum_{i=1}^{c}(d_t^i - d_t^{i'})^2, \tag{6}$$

where d_t^i and $d_t^{i'}$ indicate the description degrees of the generated label distribution d_t and ground truth $d_t^{'}$ at time step t respectively. In practice, we give a larger reward when the distance is within a threshold δ to encourage the model to learn the policy with a quick convergence, which is proved effective in experiments. Intuitively, the label distribution-based reward maintains the generated label distribution to fit the corresponding ground truth. The closer the distance, the higher the reward that the agent can receive.

Predicting the importance score directly from the frame has been extensively studied [Zhang *et al.*, 2016a; H *et al.*, 2018; Gygli *et al.*, 2014; Zhou *et al.*, 2018b]. As mentioned above, the generated label distribution indicates the distribution of the score. According to the methodology in [Lillicrap *et al.*, 2016], the action should be a deterministic continuous value. We propose to use the expectation of the generated label distribution as both the agent's action and the part of the importance score, which can enhance the consistency with the ground truth label distribution. We propose an expectation-based reward, which evaluates the expectation of the distribution from the current video frame.

$$R_{ep} = exp(-\left\| E(d_t) - E(d_{t'}) \right\|_2), \tag{7}$$

where

$$E(d_t) = \sum_{i=1}^{c} i * d_t^{y_i}, \tag{8}$$

$$E(d_{t'}) = \sum_{i=1}^{c} i * d_{t'}^{y_i}, \tag{9}$$

where $E(\cdot)$ presents the expectation of the label distribution. The video summaries generation is based on both the expectation and the variance of the generated label distribution. Therefore, with this reward, the agent is encouraged to maintain the predicted importance score within a certain error range.

R_{ld} and R_{ep} complement to each other, which not only guarantee the predicted label distribution to fit the ground truth label distribution as much as possible, but also maintain the expectation within a certain error range. The final dual-reward is jointly combined by the label distribution-based reward and the expectation-based reward.

$$R(s_t) = \alpha R_{ld}(s_t) + (1-\alpha)R_{ep}(s_t), \tag{10}$$

where α is a parameter to balance between the expectation-based reward and the label distribution-based reward. The dual-reward $R(s_t)$ guides the learning of the proposed model.

3.4 Training Procedures

We use the deep deterministic policy gradient (DDPG) algorithm [Lillicrap *et al.*, 2016] to train the model. It should be noted that the description degree of label distribution is not probability. It represents the degree to which each label describes the instance. The spaces of the label distribution and the action are continuous. Therefore, a deterministic policy is necessary to generate the label distribution.

DDPG is an actor-critic approach that maps state to a specific action. In our learning process, a replay buffer is used to disrupt the correlation among samples. In DDPG, at each timestep, the actor and critic networks are updated by sampling a minibatch uniformly from the replay buffer. Since DDPG is an off-policy algorithm, the replay buffer can be large, allowing the algorithm to benefit from learning across a set of uncorrelated samples. The actor policy is updated by using the sampled gradient,

$$\nabla_{\theta^u} u|_{s_i} \approx \frac{1}{N}\sum_i \nabla_a Q(s,a|\theta^Q)|_{s=s,a=u(s_i)} \nabla_{\theta^u} u(s|\theta^u)|_{s_i}, \tag{11}$$

where

$$u(s_i) = E(d_i), \tag{12}$$

Action a is mapped from the expectation of the label distribution. The critic network is updated by minimizing the square loss,

$$L = \frac{1}{N}\sum_t (z_i - Q(s_i, a_i|\theta^Q)^2), \tag{13}$$

where

$$z_i = r_i + \gamma Q'(s_{i+1}, u^{'}(s_{i+1}|\theta^{u^{'}})|\theta^{Q^{'}}), \tag{14}$$

Since the network $Q(s,a|\theta^Q)$ being updated is used in calculating the target value in Eq.13, the Q update is prone to divergence. The target network using "soft" updates, rather than directly copying the weights.

$$\theta^{Q^{'}} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q^{'}}, \tag{15}$$

$$\theta^{u^{'}} \leftarrow \tau\theta^u + (1-\tau)\theta^{u^{'}}, \tag{16}$$

This soft update method can make the value changing of the target network slowly, greatly improving the stability of learning. Exploration-exploitation dilemma is a major challenge of reinforcement learning. Similar to previous work, we construct an exploration policy $u^{'}$ by adding a random noise process $\mathcal{N}$ to our actor policy.

$$u^{'}(s_t) = u(s_t|\theta_t^u) + \mathcal{N}, \tag{17}$$

In practice, we add a Gaussian random process at every step. The mean of the Gaussian distribution is the expectation of the generated label distribution. And the variance is a variable, which gradually decreases from a constant value $\mathcal{N}$ to zero.

3.5 Video Summarization Inference

During testing, similar to [Zhou *et al.*, 2018a; Zhou *et al.*, 2018b], we use KTS [Potapov *et al.*, 2014] for the temporal segmentation. Shot-level scores are computed by averaging frames scores within the same shots. We apply the trained model to predict the frame-selection importance score. To maximize the use of the predicted label distribution information. The predicted importance score is obtained by calculating both the expectation and the variance of the generated label distribution.

$$score_{pre} = E(d_t) + (1 - D(d_t)), \tag{18}$$

where $D(d_t)$ indicates the variance of the generated label distribution. Video summaries are generated by maximizing the total scores while ensuring that the length of summary does not exceed a limit ϕ. During testing, the reward is not required.

4 Experiments

We carry out experiments for video summarization under both canonical, augmented, and transfer settings to validate the effectiveness and robustness.

4.1 Datasets

We evaluate our approach on two widely used benchmark datasets, SumMe [Song *et al.*, 2015] and TVSum [Gygli *et al.*, 2014]. Both datasets are annotated by multiple persons. SumMe contains 25 personal videos downloaded from YouTube. Each video ranges from 1 to 6 minutes. The dataset provides shot-level annotated scores annotated by 15 to 18 persons. We use the proposed label enhancement framework illustrated in Section 3.1 to generate the label distributions. TVSum contains 50 videos with 10 categories. The duration of each video varies from 2 to 10 minutes. Each frame is annotated by 20 annotators. The annotated scores range from 1 to 5 in TVSum. Therefore, the abscissa of the label distribution is from 1 to 5, the ordinate presents the ratio of each annotated score to the total number of annotations. Following [Zhang *et al.*, 2016a; Zhou *et al.*, 2018a], we use other two datasets from YouTube [Avila *et al.*, 2011] and Open Video Project (OVP) [2] as auxiliary datasets to conduct augmented and transfer experiments.

4.2 Evaluation Metrics

We follow the F-score protocol to evaluate our framework as in most of the previous works [Zhou *et al.*, 2018a; Zhou *et al.*, 2018b; Otani *et al.*, 2017; Otani *et al.*, 2019]. Besides, as suggested by [Otani *et al.*, 2019], two rank order correlation metrics, Kendall's τ [KENDALL and G., 1945] and Spearman's ρ [Kokoska and Zwillinger, 1999], are also adopted to validate the results comparing with three video summarization methods. We calculate the Kendall's τ and Spearman's ρ correlation coefficients between the generated scores with respect to each human annotated reference scores. The final correlation coefficient is then obtained by averaging over the individual results.

4.3 Implementation Details

For fair comparison, we use the output from the penultimate layer of GoogLeNet pre-trained on ImageNet as our features (1024-dimensions), which is same as the previous work [Zhou *et al.*, 2018a]. We downsample videos by 2 fps. The LSTM layer includes 128 units. The time step of LSTM is 10. The actor network has one label distribution layer. The label distribution layer has 5 units. The critic has two fully connected layers including 32 and 64 units for SumMe dataset, 300 and 600 units for TVSum dataset respectively. The output layer of the critic network has 1 unit. The parameters y_{min} and y_{max} are 1 and 5 in Eq.1. The hyperparameters δ in Eq.5, α in Eq.10, τ in Eq.15 and 16, $\mathcal{N}$ in Eq.17 are 0.2, 0.3, 0.001 and 4, respectively. The learning rate is $1e-04$ for actor and $1e-03$ for critic. The batch size is 32. The discount factor γ is 0.99. And the size of the memory capacity is 10000 for TVSum and 5000 for SumMe. The limited length of video summaries ϕ is 15% of the whole video length.

We use three settings to evaluate our method. (1) Canonical: we use the standard 5-fold cross validation (5FCV). (2) Augmented: we still use the 5FCV with more training data of OVP and YouTube. (3) Transfer: for a target dataser, SumMe or TVSum, the other three datasets are used as training data to test the transfer ability of our model.

In our evaluation, we select reinforcement learning/unsupervised/supervised video summarization approaches to compare with our method. DR-DSN [Zhou *et al.*, 2018a], Hier-PG [Chen *et al.*, 2019] and DQSN [Zhou *et al.*, 2018b] are reinforcement learning-based methods. $\mathrm{GAN_{dpp}}$ [Mahasseni *et al.*, 2017] and Co-archetypal [Song *et al.*, 2015] are unsupervised baselines. vs-LSTM and dpp-LSTM [Zhang *et al.*, 2016a] are supervised LSTM-based methods. Summary transfer [Zhang *et al.*, 2016b] and SASUM [H *et al.*, 2018] are also supervised approaches by using the semantic information to select the keyframes. $\mathrm{DR-DSN_{sup}}$ and $\mathrm{GAN_{sup}}$ are two augmented supervised methods extended from DR-DSN [Zhou *et al.*, 2018a] and $\mathrm{GAN_{dpp}}$ [Mahasseni *et al.*, 2017]. As analysis by [Otani *et al.*, 2019], video segmentation has significant impact on the performance. For fair comparison, all the comparison methods use the KTS [Potapov *et al.*, 2014] segmentation algorithm for evaluation.

4.4 Results

Table 1 shows the compared results with three reinforcement learning-based methods on SumMe and TVSum datasets. As

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

Methods	SumMe			TVSum		
	canonical	augmented	transfer	canonical	augmented	transfer
DR-DSN	41.4	42.8	42.4	57.6	58.4	57.8
DQSN	-	-	-	58.6	-	-
Hier-PG	43.6	44.5	42.4	58.4	58.5	58.3
$\mathrm{GAN}_{\mathrm{dpp}}$	39.1	43.4	-	51.7	59.5	-
Co-archetypal	-	-	-	50.0	-	-
Random	41.0	-	-	57.0	-	-
vs-LSTM	37.6	41.6	40.7	54.2	57.9	56.9
SASUM	40.6	-	-	53.9	-	-
dpp-LSTM	38.6	42.9	41.8	54.7	59.6	58.7
$\mathrm{GAN}_{\mathrm{sup}}$	41.7	43.6	-	56.3	**61.2**	-
Summary Transfer	40.9	-	-	-	-	-
$\mathrm{DR}-\mathrm{DSN}_{\mathrm{sup}}$	42.1	43.9	42.6	58.1	59.8	**58.9**
our method	**44.7**	**46.1**	**44.0**	**60.7**	60.9	58.6

Table 1: Results (%) on SumMe and TVSum datasets under canonical, augmented, and transfer settings.

can be observed, our method outperforms the other three reinforcement learning-based video summarization methods. On the canonical setting, our method achieves 44.7% (SumMe) and 60.7% (TVSum), outperforming 43.6% (SumMe) and 58.6% (TVSum) obtained by Hier-PG and DQSN. There are mainly two reasons for the good performance of the proposed framework. First, our method uses the annotated scores to design a novel reward function. DQSN uses the categories of videos, which is a high-level semantic information. There is a deviation between the category information and the frame-level score annotations. Second, the ambiguity of the subjectivity can be captured by the label distribution.

Table 1 also reports the results compared with unsupervised and supervised approaches. From the results, we can notice that our approach performs marginally better than the supervised baseline dpp-LSTM by 6.1% and 6.0% and unsupervised baseline $\mathrm{GAN}_{\mathrm{dpp}}$ by 5.6% and 9.0% on the canonical setting. Our method beats $\mathrm{GAN}_{\mathrm{sup}}$ method (44.7% vs. 41.7% on SumMe and 60.7% vs. 56.3% on TVSum) as well as the Summary Transfer and the random results. The performances of our method are also better than the most comparable methods $\mathrm{DR}-\mathrm{DSN}_{\mathrm{sup}}$ (44.7% vs. 42.1% on SumMe and 60.7% vs. 58.1% on TVSum). In addition to the augmented and transfer settings on TVSum dataset, our method gets the best performance. Because the label distribution of the auxiliary datasets is generated from label enhancement algorithm. There is a certain difference between the enhanced label distribution and the distribution annotated by the annotators. Therefore, Our method does not achieve the state-of-the-art results on the TVSum dataset under augmented and transfer settings. In general, the results verify the effectiveness of our proposed approach.

Rank correlation coefficient. As analysis in [Otani *et al.*, 2019], F1-measure evaluation approach is relevant to the preprocessing stage. Random results also achieve comparable results. Therefore, Kendall's τ and Spearman's ρ correlation coefficients are recommended to measure the performance of video summaries. We compare our method with dpp-LSTM, DR-DSN and Hier-PG on TVSum dataset. The results are illustrated in Table 2. From the results, we can conclude that the margin from the dpp-LSTM, DR-DSN and Hier-PG is very significant. It illustrates that the results from our method are more consistent with the groundtruth from human annotated scores.

Methods	Kendall's τ	Spearman's ρ
Random	0.000	0.000
DR-DSN	0.020	0.026
dpp-LSTM	0.042	0.055
Hier-PG	0.078	0.116
our method	**0.1778**	**0.2316**

Table 2: Results of Kendall's τ and Spearman's ρ correlation coefficients on TVSum dataset.

5 Conclusion and Future Work

In this paper, we propose a label distributions-based dual-reward to capture the ambiguity problem in reinforcement learning. The dual-reward is designed by applying the label distribution-based and the expectation-based rewards to guide the agent to learn the policy. In addition, we propose a new label enhancement framework to transform the annotated scores or the binary results into label distributions. We redefined the action form by mapping from the label distribution. DDPG algorithm is used to train the proposed model for video summarization. The experimental results show that our method achieves good performance, especially under the rank correlation coefficient metrics. The text description of the video always implies the summary information. Therefore, we will explore the semantic information to improve the performance in the future.

Acknowledgments

This research was supported by the National Key Research and Development Plan of China (No. 2017YFB1002801), the National Science Foundation of China (62076063), the Collaborative Innovation Center of Novel Software Technology and Industrialization, and the Collaborative Innovation Center of Wireless Communications Technology.

References

[Avila *et al.*, 2011] Sandra Avila, Ana Lopes, and Antonio da Luz. Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. *PRL*, 32:56–68, 01 2011.

[Chen *et al.*, 2019] Y Chen, L Tao, X Wan, and Toshihiko Yamasaki. Weakly supervised video summarization by hierarchical reinforcement learning. In *ACM Multimedia Asia*, 2019.

[Chen *et al.*, 2020] S Chen, J Wang, Y Chen, Z Shi, X Geng, and Y Rui. Label distribution learning on auxiliary label space graphs for facial expression recognition. In *CVPR*, 2020.

[Deng *et al.*, 2009] J Deng, W Dong, Richard Socher, L Jia L, and F-F Li. Imagenet: a large-scale hierarchical image database. In *CVPR*, 2009.

[Gao *et al.*, 2020] Y Gao, Y Zhang, and X Geng. Label enhancement for label distribution learning via prior knowledge. In *IJCAI*, 2020.

[Geng and Hou, 2015] X Geng and P Hou. Pre-release prediction of crowd opinion on movies by label distribution learning. In *AAAI*, 2015.

[Geng *et al.*, 2020] X Geng, X Qian, Z Huo, and Y Zhang. Head pose estimation based on multivariate label distribution. In *TPAMI*, 2020.

[Geng, 2016] X Geng. Label distribution learning. *TKDE*, 28(7):1734–1748, 2016.

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

[H *et al.*, 2018] Wei H, Ni B, Yan Y, Yu H, and Yang X. Video summarization via semantic attended networks. In *AAAI*, 2018.

[Jung *et al.*, 2019] Y Jung, D Cho, Dahun Kim, Sanghyun Woo, and In So Kweon. Discriminative feature learning for unsupervised video summarization. *AAAI*, 33:8537–8544, 2019.

[KENDALL and G., 1945] KENDALL and M. G. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945.

[Kokoska and Zwillinger, 1999] S. Kokoska and D. Zwillinger. Crc standard probability and statistics tables and formulae. *Technometrics*, 43(2):249–250, 1999.

[Li *et al.*, 2020] Y Li, W Lin, T Wang, Q Guo, and S Xu. Video summarization via cluster-based object tracking and type-based synopsis. In *MIPR*, 2020.

[Lillicrap *et al.*, 2016] Timothy P. Lillicrap, Jonathan J. Hunt, and Alexander Pritzel. Continuous control with deep reinforcement learning. In *ICLR*, 2016.

[Lv *et al.*, 2020] J Lv, M Xu, L Feng, G Niu, and X Geng. Progressive identification of true labels for partial-label learning. In *ICML*, 2020.

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

[Otani *et al.*, 2017] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkil, and Naokazu Yokoya. Video summarization using deep semantic features. In *ACCV*, 2017.

[Otani *et al.*, 2019] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne. Rethinking the evaluation of video summaries. In *CVPR*, 2019.

[Potapov *et al.*, 2014] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. Category-specific video summarization. In *ECCV*, 2014.

[Shen *et al.*, 2017] W Shen, K Zhao, Y Guo, and Alan L Yuille. Label distribution learning forests. In *NeurIPSWS*, pages 834–843, 2017.

[Smith-Miles and Geng, 2020] K Smith-Miles and X Geng. Revisiting facial age estimation with new insights from instance space analysis. In *TPAMI*, 2020.

[Song *et al.*, 2015] Y Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *CVPR*, 2015.

[Szegedy and Liu, 2015] Christian Szegedy and Wei. Liu. Going deeper with convolutions. In *CVPR*, 2015.

[Wang *et al.*, 2019] J Wang, W Wang, Z Wang, L Wang, D Feng, and T Tan. Stacked memory network for video summarization. In *ACM Multimedia*, 2019.

[X *et al.*, 2018] Zheng X, Jia X, and Li W. Label distribution learning by exploiting sample correlations locally. In *AAAI*, pages 4556–4563, 2018.

[Xu *et al.*, 2018] N Xu, A Tao, and X Geng. Label enhancement for label distribution learning. In *IJCAI*, 2018.

[Yi and Wu, 2019] K Yi and J Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *CVPR*, 2019.

[Zhang *et al.*, 2016a] K Zhang, W Lun Chao, and Kristen Grauman. Video summarization with long short-term memory. In *ECCV*, 2016.

[Zhang *et al.*, 2016b] K Zhang, W Lun Chao, F Sha, and Kristen Grauman. Summary transfer: Exemplar-based subset selection for video summarization. In *CVPR*, 2016.

[Zhang *et al.*, 2019] Y Zhang, Michael K, X Zhao, and M Tan. Deep reinforcement learning for query-conditioned video summarization. *Applied Sciences*, 9(4), 2019.

[Zheng and Lu, 2020] J Zheng and G Lu. k-sdpp: Fixed-size video summarization via sequential determinantal point processes. In *IJCAI-PRICAI*, 2020.

[Zhou *et al.*, 2018a] K Zhou, Yu Q, and Tao X. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *AAAI*, 2018.

[Zhou *et al.*, 2018b] K Zhou, Tao X, and Andrea C. Video summarisation by classification with deep reinforcement learning. In *BMVC*, 2018.