SELF-ATTENTION RECURRENT SUMMARIZATION NETWORK WITH REINFORCEMENT LEARNING FOR VIDEO SUMMARIZATION TASK

Aniwat Phaphuangwittayakul[1,*] *, Yi Guo*[1,2,3,†] *, Fangli Ying*[1,*] *, Wentian Xu*[1]*, and Zheng Zheng*[1]

[1]Department of Computer Science and Engineering, East China University of Science and Technology, China
[2]National Engineering Laboratory for Big Data Distribution and Exchange Technologies, China
[3]Shanghai Engineering Research Center of Big Data and Internet Audience, China
aniwat.pha@gmail.com, {guoyi, yfangli}@ecust.edu.cn, {2632883664, 423391432}@qq.com

2021 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-3864-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICME51207.2021.9428142

ABSTRACT

With the exponential growth of video data, video summarization techniques are urgently needed for reducing people's efforts in the videos' content exploration by generating succinct but informative summaries from original lengthy videos. Though supervised video summarization approaches have demonstrated the state-of-the-art performance, unsupervised methods are still highly demanded due to resourcefully expensive human annotations and the subjectiveness of video summarization tasks. In this paper, a novel unsupervised-based Deep Self-attention Recurrent summarization network with Reinforcement Learning (DSR-RL) for video summarization is proposed. The model can learn the input video sequence and suggest the key-shot summary without additional human annotations by integrating self-attention, BRNN, and reinforcement learning mechanisms. The DSR-RL improves not only importance score through the attention map vector of self-attention network but also the diversity of summaries via the reward function of reinforcement learning. Our method outperforms the state-of-the-art unsupervised video summarization methods on both SumMe and TVSum datasets. The source code is available at https://github.com/phaphuang/DSR-RL.

Index Terms— Video Summarization, Reinforcement Learning, Self-Attention, Video Representation

1. INTRODUCTION

The increasing popularity of online social network and the widespread availability cameras have led to an enormous and ever-growing collection of video data generated by users around the world. Video has rapidly become one of the most common forms of online information exchange. Consequently, there is an urgent demand to automatically extract representative parts from lengthy videos and effectively access video content [1]. Video summarization is one of the promising techniques to cater for this need, which shortens an original video to compact summary and offers a representative sequence of features. [2]. The video summarization approaches are mainly divided into two categories that are supervised and unsupervised learning methods [3]. Supervised learning methods can ensure that the model is straightforwardly learned as they include human annotations during the learning process. Nevertheless, the labels of existing real-world videos are insufficiently provided as well as the output summaries are often overfitting with the training set. To overcome these issues, the unsupervised learning methods are considered.

In this work, we investigate the unsupervised video summarization method on the basis of the attention network. The attention network was first proposed to support sequence modeling in various tasks such as language modeling neural machine translation task [4]. The attention networks can decrease computational complexity of models during learning process [5]. From the last few years, A number of attention-based supervised video summarization approaches are proposed. Fajtl *et al.* [6] designed a video summarization model by replacing the standard recurrent network with a self-attention network named VASNET. Even though this method showed the remarkable results for video summarization task, only a single attention map was not enough for the complicated scenes in the video. Liu *et al.* [7] performed H-MAN that was improved from the VASNET model by substituting the single self-attention network with multi-attention network. H-MAN comprises two models that are Bi-LSTM and multi-head attention model in the different stages. H-MAN performed marvelously against complicated scenes. Nonetheless, the structure of the hierarchical learning mechanism itself occurred loss of key-frames candidates from original da-

This research is financially supported by The National Key Research and Development Program of China (grant number 2018YFC0807105) and Science and Technology Committee of Shanghai Municipality (STCSM) (under grant numbers 17DZ1101003, 18511106602 and 18DZ2252300). Partially Supported by International College of Digital Innovation (ICDI), Chiang Mai University, Thailand.

† Corresponding Author

*Aniwat Phaphuangwittayakul and Fangli Ying contributed equally to this work.

978-1-6654-3864-3/21/$31.00 ©2021 IEEE

ta. There are very few an existing attention-based unsupervised video summarization methods. Apostolidis *et al.* [1] enhanced the SUM-GAN-sl [8] by combining the attention layer with variational auto-encoder (SUM-GAN-VAAE) and merging attention network with auto-encoder (SUM-GAN-AAE). The performance of SUM-GAN-AAE depended on the regularization factor. The optimal value of this factor number varied according to different datasets. Even though this method is unsupervised-based attention network, the proper factor selection and generative adversarial training are required. To avoid unstable learning process, reinforcement learning-based framework has been studied by optimizing the action for selecting output frames.

Zhou *et al.* [9] developed the reinforcement learning-based framework for video summarization called DR-DSN. By maximizing the reward function, DR-DSN produced more diverse and more representative summaries. Inspired by the previous works [6, 9], we propose a novel unsupervised video summarization framework called Deep Self-attention Recurrent summarization network with Reinforcement Learning (DSR-RL). With the reinforcement learning method, the model learns to summarize the video without mode collapse from adversarial training. Additionally, the attention map vector of self-attention allows the summarization network to monitor important frames for the output video summary. Our framework is based on three crucial mechanisms that are self-attention, bi-directional recurrent neural network, and reinforcement learning. Thus, the overall contributions of our work are:

- We present a new summarization network by concatenating the two features of the self-attention network and Bidirectional Recurrent Neural Network (BRNN). The self-attention network enhances the summarization network by providing additional attention features for calculating the importance score.
- We extend the objective of current reinforcement learning-based method for video summarization with two additional losses including regularization loss and reconstruction loss. The regularization loss prevents the attention weights out of a specific range. The reconstruction loss helps the model keep the key features of the original input for generate informative output summaries.

2. SELF-ATTENTION RECURRENT SUMMARIZATION NETWORK WITH REINFORCEMENT LEARNING

2.1. Summarization Network

The overall learning process of our summarization network is presented in Figure 1. First of all, the frame features $X = (x_1, \ldots, x_t, \ldots, x_T)$ of video length T are extracted from the input video frames $F = (f_1, \ldots, f_t, \ldots, f_T)$ by a pre-trained CNN network. The extracted features are then used as the input for two individual networks that are self-attention network and BRNN. The importance score is the probability distribution $P = (p_1, \ldots, p_t, \ldots, p_T)$ from fully connected layer with sigmoid activation. Finally, the policy function of the reinforcement algorithm is optimized by using stochastic gradient-based optimization. The detail of each component and objective function is described as the following section.

2.1.1. Self-Attention Network

We adopt the self-attention to extract the attention features. These features are useful for the summarization network to generate the important frames. The self-attention network takes the input features $X = (x_1, \ldots, x_t, \ldots, x_T)$ and generates attention features $Y = (y_1, \ldots, y_t, \ldots, y_T), y \in [0, 1)$, both features have length T where each x_t, y_t has 1024 dimensions. The self-attention network basically has three inputs that are query $Q = W_g x$, key $K = W_f x$, and value $V = W_h x$. Firstly, the multiplication between query and key are taken to produce the attention map. Self-attention vector of the attention map $e_{t,i}$ between the input feature at time t and every input features of entire sequence i is calculated by

$$e_{t,i} = \beta[(W_f x_i)^\intercal (W_g x_t)]; t = [1, T], i = [1, T] \tag{1}$$

With video length T, W_f and W_g are weight matrices of the linear transformation networks. The scale parameter β is set to 0.06. The scale parameter reduces the values of dot product between query Q and key K.

Secondly, the attention map vector e_t is passed through softmax layer followed by a dropout. The output of softmax layer is attention weights s_t at time t as

$$s_t = \frac{exp(e_{t,i})}{\sum_{k=1}^{T} exp(e_{t,k})} \tag{2}$$

Finally, output of the self-attention $y_{attn,t}$ is a linear transformation with weight matrix W_v of the dot product values between value V and attention weights s_t with dropout rate 0.5 as

$$y_{attn,t} = W_v(\sum_{i=1}^{T} \text{dropout}(s_{t,i}) W_h x_i) \tag{3}$$

2.1.2. Bi-directional Recurrent Neural Network (BRNN)

BRNN is a concatenation of two recurrent neural networks (RNNs) that are forward RNN and backward RNN. In this work, we evaluated two RNN units including Long-Shot Term Memory (LSTM) unit and Gated Recurrent Unit (GRU). The forward RNN firstly reads the input sequence X from x_1 to x_T and then calculates forward hidden states ($\overrightarrow{h_1}$ to $\overrightarrow{h_T}$).

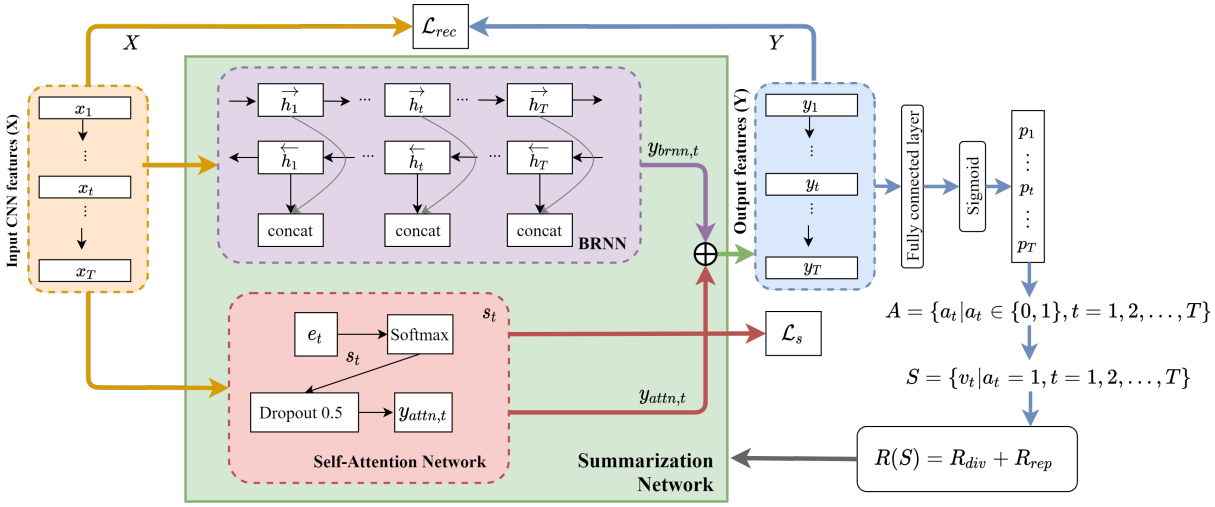


Fig. 1. An overview of our summarization network. The output is probability of selected frame at time t ($t = [1, T]$).

The dimensions of these hidden states are 512. The backward hidden states ($\overleftarrow{h_1}$ to $\overleftarrow{h_T}$) are retrieved by passing the sequence in reverse order through the backward RNN. The BRNN features of each frame at time t therefore have 1024 dimensions. Concatenating features from BRNN and self-attention network followed by a fully connected layer, we can obtain the importance score p_t as follows.

$$y_t = y_{brnn,t} \oplus y_{attn,t} \tag{4}$$

$$p_t = \sigma(W y_t) \tag{5}$$

where $y_{brnn,t}$ is features of BRNN network at time t. $y_{attn,t}$ is attention features of self-attention network at time t. σ is sigmoid activation function.

2.2. Reward Function

Working with reinforcement learning-based framework, a frame-selection action a_t is defined as

$$a_t \sim \text{Bernoulli}(p_t); a_t \text{ is 0 or 1} \tag{6}$$

Where $a_t = 1$ refers to the frame of video at time t (f_t) is selected into output video summary. Consequently, a video summary $S = \{f_t | a_t = 1, t = 1, 2, \ldots, T\}$ is composed of all selected frames.

Following DR-DSN [9], the reward function consists of diversity reward R_{div} and representativeness reward R_{rep}. Diversity reward R_{div} measures the dissimilarity of frames within the video summary. The dissimilarity function is computed as Eq. (8). The higher the diversity reward indicates that the more dissimilar of a particular selected frame to others. Representativeness reward R_{rep} measures the quality of the generated summary representing the content from the original video. The reward is calculated by the distance between each frame of the original video and frames in the generated summary. Given length of video summary as Υ and λ controls the degree of temporal distance, R_{div} and R_{rep} can be obtained by using Eq. (7) and Eq. (9), respectively.

$$R_{div} = \frac{1}{\Upsilon(\Upsilon - 1)} \sum_{t \in \Upsilon} \sum_{\substack{t' \in \Upsilon \\ t' \neq t}} d(x_t, x_{t'}) \tag{7}$$

$$d(x_t, x_{t'}) = \begin{cases} 1, & \text{if } |t - t'| > \lambda \\ 1 - \frac{x_t^\intercal x_{t'}}{\|x_t\|_2 \|x_{t'}\|_2}, & \text{otherwise} \end{cases} \tag{8}$$

$$R_{rep} = \exp(-\frac{1}{T} \sum_{t=1}^{T} \min_{t' \in \Upsilon} \|x_t - x_{t'}\|_2) \tag{9}$$

The total reward $R(S)$ is calculated by

$$R(S) = R_{div} + R_{rep} \tag{10}$$

2.3. Policy Gradient

The objective of optimizing reward function is to train the model with policy gradient based on reinforcement algorithm [10]. The algorithm generally performs reliable estimation; however, it remains the high variance problem affecting the convergence of network. The original gradient equation is thus subtracted with a constant b from the reward. The objective function of this optimization algorithm is

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T} (R_n - b) \nabla_\theta \log \pi_\theta(a_t | h_t) \tag{11}$$

where N is number of episodes. R_n is reward at episode n. b is the moving average reward of each video.

2.4. Regularization and Reconstruction Loss

Our framework can be viewed as an encoder-decoder network, where pre-trained CNN network extracted input video sequence as the encoder and summarization network predicted the output importance score as the decoder. In order to constrain the attention weights and assess the error of output features in learning encoder-decoder network, the reconstruction and regularization loss are included during training summarization networks inspired by SUM-GAN-sl and SUM-GAN-AAE [3, 8]. These two terms are defined as follows.

Summary-Length Regularization loss $\mathcal{L}_s$: Calculating the sparsity between attention weights and selected regularization factor.

$$\mathcal{L}_s = |\frac{1}{M}\sum_{t=1}^{M} s_t - \delta| \tag{12}$$

where M is the total number of video frames and δ is an a tunable hyper-parameter representing a regularization factor. In this work, we set δ to 0.15.

Reconstruction loss $\mathcal{L}_{rec}$: is simply Euclidean distance between the input sequence and output features. The equation of reconstruction loss is defined as Eq. (13).

$$\mathcal{L}_{rec} = \|X - Y\|_2 \tag{13}$$

2.5. Loss Optimization

To achieve the highest reward and the lowest loss, the parameters θ of policy function was optimized through Adam optimizer [11] which is one of stochastic gradient-based methods with learning rate $\alpha = 10^{-5}$.

$$\theta = \theta - \alpha\nabla_\theta(-J(\theta) + \mathcal{L}_s + \mathcal{L}_{rec}) \tag{14}$$

3. EXPERIMENTAL RESULTS

3.1. Datasets

The proposed framework was evaluated on SumMe [12] and TVSum [13] datasets. These two datasets have ground-truth annotations and meta-data for computing evaluation metrics. Thus, they were commonly used as benchmark datasets in most of the existing works. The duration of videos is ranging from 1.5 minutes to 10 minutes. The videos are extracted from multiple events such as news, sports, holidays, and cooking. Each video in these datasets is attached with various ground-truth summaries from users with the number from 15 to 20.

3.2. Implementation Details

The same setting environments is adopted for comparing with most of the previous works [1, 14, 6, 9]. The input features of each video are extracted from pool5 layer of GoogLeNet [15] trained on ImageNet [16] dataset. The input features therefore have 1024 dimensions following the pool5 layer of GoogLeNet. We used two kinds of network as frame selector. The former is bidirectional recurrent neural network (BRNN) cell is combined two layers of 512 hidden units. Our work used two kinds of BRNN cells that are LSTM [17] and GRU [18]. The latter is a self-attention network. Following DR-DSN work [9], we set the value of temporal distance $\lambda = 20$ in Eq. (8), and the number of episodes $N = 5$.

This summarization network predicts the importance score p_t in form of selected frame probability at time t. The final summary S can be generated by selecting key-shots, which is a collection of key-frame. To produce key-shots for summarization with provided key-frame scores, the videos are temporally segmented using Kernel Temporal Segmentation (KTS) algorithm [19]. The output is ranked based on their fragment-level importance scores. These importance scores are the average score of each key-shot. The subsets of key-shots are selected by using Knapsack algorithm. The generated video summary is formed with not exceed 15% of the full video length. The training process is stopped when the model reaches 100 epochs. In every iteration of 10 epochs, the model with the highest F-measure is selected to evaluate testing results. This work is implemented with PyTorch machine learning library on GPU NVIDIA GeForce GTX 1650.

3.3. Quantitative Results

Both SumMe and TVSum datasets are divided into 80% of videos for training and 20% of videos for testing. The standard five-fold cross-validation is utilized for evaluating our method. The F-measure [20] was adopted for evaluation. The F-measure is a common metric for determining the similarity between the ground-truth summary and the generated summary of each original video.

We evaluated our method and state-of-the-art approaches on SumMe and TVSum datasets by using F-Measure evaluation. Table 1 illustrates the comparison in terms of F-measure between our model and state-of-the-art supervised/unsupervised video summarization approaches. Our DSR-RL model outperforms the existing unsupervised learning methods in Canonical (C), Augmented (A), and Transfer (T) settings. Moreover, the DSR-RL achieves comparable results with state-of-the-art supervised-based video summarization framework.

Based on the experiments, SumMe dataset includes a short sequence of videos that requires less memory consumption. As a result, the model with Gated Recurrent Unit (GRU) reveals higher F-measure than standard Long-Shot Term Memory (LSTM) on SumMe dataset. Nevertheless, the

Table 1. Quantitative comparison F-Measure (%) between our unsupervised method and state-of-the-art supervised/unsupervised video summarization approaches on SumMe [12] and TVSum [13].

Method	SumMe			TVSum		
	C	A	T	C	A	T
Supervised Method						
SUM-FCN [21]	47.5	51.1	44.1	56.8	59.2	58.2
SUM-DeepLab [21]	48.8	50.2	45.0	58.4	59.1	57.4
VASNET [6]	49.7	51.1	-	**61.4**	**62.4**	-
CSNetsup [22]	48.6	48.7	44.1	58.5	57.1	57.4
H-MAN [7]	**51.8**	**52.5**	48.1	60.4	61.0	59.5
Unsupervised Method						
DR-DSN [9]	41.4	42.8	42.4	57.6	58.4	57.8
SUM-FCNunsup [21]	41.5	-	-	52.7	-	-
UnpairedVSN [23]	47.5	-	-	55.6	-	-
SUM-GAN-sl [8]	47.3	-	-	58.0	-	-
FrameRank [24]	45.3	-	-	60.1	-	-
SUM-GAN-VAAE [1]	45.7	-	-	57.6	-	-
SUM-GAN-AAE [1]	48.9	-	-	58.3	-	-
DSR-RL-LSTM (ours)	43.8	43.3	44.2	**61.4**	60.2	**59.9**
DSR-RL-GRU (ours)	50.3	48.5	**48.7**	60.2	59.2	58.9

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

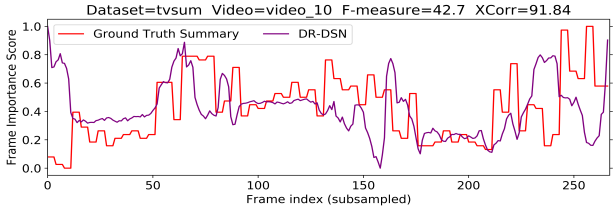
Method	τ	ρ
dppLSTM	0.042	0.055
DR-DSN [9]	0.020	0.026
Hierarchical RL [25]	0.078	**0.116**
DSR-RL-LSTM (ours)	**0.0862**	0.113
DSR-RL-GRU (ours)	**0.0863**	0.114
Human Annotation	0.177	0.204

model with LSTM unit shows better performance on TVSum dataset as the memory in LSTM unit learns more complex patterns in the long-length video.

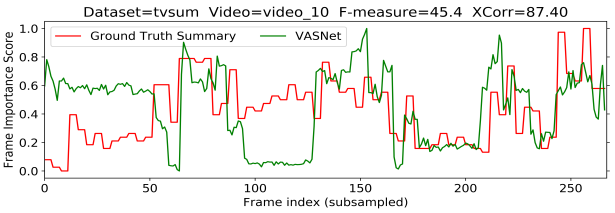
Besides F-Measure, Kendall's τ and Spearman's ρ correlation coefficients [26], metrics for evaluating the summarization performance, are calculated. From Table 2, our method achieves the highest Kendall's and comparable Spearman's to the current approaches.

3.4. Qualitative Results

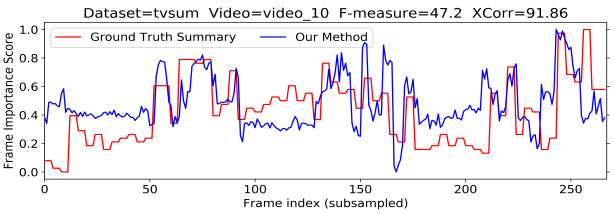
The importance scores of two baselines (DR-DSN and VASNet) and our method compared with ground truth on video 10 of TVSum dataset are illustrated as Figure 2. Our method shows the correlation higher than baselines especially during frame index from 50 to 100. The correlation determines the similarity between the predicted importance score of each model and the ground truth.



(a) DR-DSN



(b) VASNet



(c) Our Method

Fig. 2. F-measure and Correlation score (XCorr) between ground truth (red) and predicted summary by various methods for video 10 of TVSum dataset.

4. CONCLUSION

In this paper, a novel unsupervised summarization framework with self-attention and deep reinforcement learning is proposed. Our summarization network is mainly based on the combination of two mechanisms that are self-attention network and Bi-directional recurrent network (BRNN). With attention map vector of self-attention network and policy gradient in reinforcement learning algorithm, the model outperforms the existing state-of-the-art unsupervised learning video summarization approaches. Future work direction will focus on the real-time applications of video summarization and the development of more stable deep reinforcement learning method.

References

[1] E. Apostolidis, E. Adamantidou, A. I Metsai, V. Mezaris, and I. Patras, "Unsupervised video summarization via attention-driven adversarial learning," in *International Conference on Multimedia Modeling*. Springer, 2020, pp. 492–504.

[2] Y. Shemer, D. Rotman, and N. Shimkin, "Ils-summ: Iterated local search for unsupervised video summarization," *CoRR*, vol. abs/1912.03650, 2019.

[3] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic, "Unsupervised video summarization with adversarial lstm networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 202–211.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[5] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015, pp. 2048–2057.

[6] J. Fajtl, H. S. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino, "Summarizing videos with attention," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 39–54.

[7] Y.-T. Liu, Y.-J. Li, F.-E. Yang, S.-F. Chen, and Y.-C. F. Wang, "Learning hierarchical self-attention for video summarization," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 3377–3381.

[8] E. Apostolidis, A. I Metsai, El. Adamantidou, V. Mezaris, and I. Patras, "A stepwise, label-based approach for improving the adversarial training in unsupervised video summarization," in *Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery*, 2019, pp. 17–25.

[9] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[10] R. J Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[11] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR*, Yoshua Bengio and Yann LeCun, Eds., 2015.

[12] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *European Conference on Computer Vision*. Springer, 2014, pp. 505–520.

[13] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "Tvsum: Summarizing web videos using titles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5179–5187.

[14] T.-J. Fu, S.-H. Tai, and H.-T. Chen, "Attentive and adversarial learning for video summarization," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1579–1587.

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *CoRR*, vol. abs/1409.1259, 2014.

[19] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *European Conference on Computer Vision*. Springer, 2014, pp. 540–555.

[20] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *European Conference on Computer Vision*. Springer, 2016, pp. 766–782.

[21] M. Rochan, L. Ye, and Y. Wang, "Video summarization using fully convolutional sequence networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 347–363.

[22] Yunjae Jung, Donghyeon Cho, Dahun Kim, Sanghyun Woo, and In So Kweon, "Discriminative feature learning for unsupervised video summarization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 8537–8544.

[23] M. Rochan and Y. Wang, "Video summarization by learning from unpaired data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7902–7911.

[24] Zhuo Lei, Chao Zhang, Qian Zhang, and Guoping Qiu, "Framerank: a text processing approach to video summarization," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 368–373.

[25] Y. Chen, L. Tao, X. Wang, and T. Yamasaki, "Weakly supervised video summarization by hierarchical reinforcement learning," in *Proceedings of the ACM Multimedia Asia*, pp. 1–6. 2019.

[26] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkila, "Rethinking the evaluation of video summaries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7596–7604.