Spatial attention model-modulated bi-directional long short-term memory for unsupervised video summarisation

Rui Zhong,1,✉ Diyang Xiao,1 Shi Dong,1 and Min Hu2
1School of Computer, Central China Normal University, Wuhan, China
2NERCMS, Wuhan University, China
✉Email: zhongr@mail.ccnu.edu.cn

Compared with surveillance video, user-created videos contain more frequent shot changes, which lead to diversified backgrounds and a wide variety of content. The high redundancy among keyframes is a critical issue for the existing summarising methods in dealing with user-created videos. To address the critical issue, we designed a salient-area-size-based spatial attention model (SAM) on the observation that humans tend to focus on sizable and moving objects in videos. Moreover, the SAM is taken as guidance to refine frame-wise soft selected probability for the bi-directional long short-term memory model. The reinforcement learning framework, trained by the deep deterministic policy gradient algorithm, is adopted to do unsupervised training. Extensive experiments on the SumMe and TVSum datasets demonstrate that our method outperforms the state-of-the-art methods [7, 6] by 21.48%–22.95% and 1.56% F-score improvements, respectively.

Introduction: Video summarisation aims to represent an original video through a brief sketch, which can generally be applied in the video pre-view, storage, retrieval, and management. Ji et al. [1] first used the classical attention-based encoder-decoder structure to emphasise the salient information of images for the supervised video summarisation method. The attention mechanism can efficiently alleviate the limitation of the encoder-decoder framework in which the scant output length of the encoder leads to the abandonment of salient features [3]. Additionally, Elhamifar proposed the classic dictionary selection-based unsupervised video summarisation approach [4], and then Mahasseni presented a variational autoencoder-based generative adversarial networks (GAN) to summarise videos [5]. Given the excellent performance of reinforcement learning in video representation, Zhou [6] subsequently presented an end-to-end long short-term memory (LSTM)-based unsupervised learning method for video summarisation, deep reinforcement-deep summarisation network (DR-DSN), which comprises a novel feedback reward with the aspects of diversity and representativeness. Afterwards, Li [7] proposed an unsupervised cycle-consistent adversarial LSTM network for video summarisation (Cycle-sum) by integrating a frame selector and a cycle-consistent learning-based evaluator.

Even with reinforcement learning, the above LSTM- and GAN-based unsupervised methods could not achieve high accuracy of selecting keyframes. To enhance the accuracy of selecting keyframes, we propose a spatial attention model (SAM)-modulated bi-directional long short-term memory (Bi-LSTM), within which the SAM is established based on salient-area-size-guided attention measuring according to the fact that humans often concentrate on sizable and moving objects in a video.

Specifically, the SAM adopts the attention mechanism [8] to rank the frame-wise saliency level and then select keyframes with a saliency level higher than a threshold. Furthermore, the scant output length of the gate structure in Bi-LSTM causes the problem of ignoring salient features, which leads to low accuracy in the probability of being selected as keyframes in video summarisation named soft selected probability (SSP) in this letter. Thus, the SAM is stacked up with the Bi-LSTM as a modulation to improve the accuracy of predicting the SSP. We first adopt the reinforcement learning framework, deep deterministic policy gradient (DDPG) algorithm [9], to backpropagate loss, which eventually alleviates the convergence problem that occurred in DR-DSN [6].

1. To overcome the issue of high redundancy among keyframes in user-created video summarisation, we propose a SAM-modulated Bi-LSTM model, which further distinguishes images’ differences to promote the accuracy of SSP prediction.
2. We propose the salient-area-size-based SAM on the observation that sizable and moving objects attract more attention.

Fig 1. Our proposed spatial attention model (SAM) where SAM is stacked up behind the bi-directional long short-term memory network to further filter redundant keyframes.

3. We first adopt DDPG to extend our model into unsupervised fashion under the reinforcement learning framework.
4. The detailed experiments, conducted on two general datasets with user-created videos, SumMe and TVSum, demonstrate that our method outperforms the state-of-the-art methods [7, 6] by 21.48%–22.95% and 1.56% F-score improvements, respectively.

Comparing with the video summarisation model with a single feature [6], the combination of semantic and saliency features within the proposed SAM-modulated Bi-LSTM method can improve the capability of filtering redundant frames in selecting the keyframes.

Problem formulation: Given a long video \( X = \{x_t\}_{t=1}^T \), where \( x_t \in \mathbb{R}^{w \times h \times c} \), \( w, h, c \) is the size of the width, height, and channel number for each frame, \( t \) is the index of the frames. \( Y = \{y_t\}_{t=1}^T, y_t \in \{0, 1\} \) is the \( t \) binary label that represents whether the \( t \) frame is selected as a keyframe. The collection of the selected frames composes the final video summarisation represented by \( X_{sms} = \{x_t | t \in E\} \), where \( E = \{t \in [1, T]|y_t = 1\} \).

Generating video summarisation: The binary possibility set \( Y = \{y_t\}_{t=1}^T \) is computed on the Bernoulli distribution [10] of the frame-wise SSP [8] \( B = \{\beta_t\}_{t=1}^T, \beta_t \in [0, 1] \) as shown in Equation (1):

\[
y_t = \text{Bernoulli}(\beta_t), y_t \in \{0, 1\},
\]

where \( y_t = 1 \) alternatively represents that the \( t \) frame with the SSP \( \beta_t \) is selected, and vice versa.

Main components of our model: As illustrated in Figure 1, we propose a SAM-modulated Bi-LSTM model that performs better in discriminating the difference among images with the integration of semantic and saliency features. First, a saliency map histogram is rendered from the saliency map generated by the saliency detector [2]. Furthermore, the SAM is modelled by measuring the frame-wise spatial importance scores with the input of saliency features. Last, the semantic feature is learned via a pre-trained GoogLeNet on ImageNet [9] and then processed by the Bi-LSTM network for predicting the initial SSP.

SAM-modulated Bi-LSTM: The SSP \( \beta_t \) is modelled as the initial probability of the Bi-LSTM, modulated by the spatial attention \( l_s \), computed as Equation (2):

\[
\beta_t = W_p \cdot \text{N}(p_t) + W_l \cdot l_s,
\]

where \( N(\cdot) = \exp (\cdot) / \sum \exp (\cdot) \) is the normalisation operator, and \( W_p \) and \( W_l \) represent the weight corresponding to the probability \( p_t \) and \( l_s \), respectively.

Salient-area-size-based SAM: Inspired by the idea that human beings frequently pay more attention to the sizable moving objects in a video, we propose a saliency feature-based model SAM to formulate spatial
attention on the statistics of the saliency map of moving objects. Specifically, the spatial attention \( l_i = [l_i(r)]_{t=1}^T \) is calculated as follows:

\[
l_i = N(\lambda_i), \quad i \in [0, 1],
\]

where \( \lambda_i \) is the visual importance score for the \( l_i \) frame.

In Figure 1, for the frame-wise saliency map \( z_s \), we rank the area of salient regions, and then set the salient regions whose area \( r_z \) is less than the threshold \( r_t \) as non-salient regions. To calculate the optimal value of \( r_z \) for determining the core salient regions, we conduct experiments by ranging \( r_z \) from 20 to 25 and randomly sample 150 videos for testing. As Figure 2 shows, when \( r_z = 21 \), SAM arrives the best result in terms of \( F \)-score. Thus, the visual importance score \( \lambda_i \) is computed as the sum of the saliency histogram of the core salient regions by Equation (4):

\[
\lambda_i = \sum_{r_z \in r_z} \text{bins}(r_z),
\]

We define a saliency histogram for the frame-wise salient map \( z_s \), whose horizontal coordinate is saliency level \( r_z \) while the vertical coordinate is frequency \( \text{bins}(\cdot) \). The range of horizontal coordinate is \( 26 \) saliency levels presented as \( \{r_0, r_1, \ldots, r_25\} \), \( \partial \in [0, 25] \), where the class interval of each level is 10 (except that the special case corresponds to \( r_z = 5 \)). In the saliency histogram, the frequency of vertical coordinate is computed by the discrete function \( \text{bins}(\cdot) \), written as Equation (5):

\[
\text{bins}(r_z) = \frac{q_{r_z}}{w \times h},
\]

where the \( q_{r_z} \) is the number of pixels at the salient regions with the level \( r_z \). Specifically, a saliency detector [2] is adopted to detect the salient regions of moving objects and generate a sequence of the saliency map \( Z = [z_s]_{t=1}^T, z_s \in \mathbb{R}^w \times h \) denoted in Equation (6):

\[
z_s = [z_s(i, j)] \in [0, 255], [i \in [0, w], j \in [0, h]],
\]

where \((i, j)\) represents the pixel-wise coordinate of \( z_s(i, j) \).

Training and optimisation: In this letter, we adopt the DDPG [8] to train the SAM-modulated Bi-LSTM.

Experimental results: For a comparison with the other methods, we adopt the commonly used evaluation metric, the \( F \)-score [1]. The experiments on the SumMe and TVSum user-created video-based datasets show that our SAM-modulated Bi-LSTM outperforms the state-of-the-art methods.

Comparison with unsupervised approaches: As Table 1 shows, the \( F \)-score of the proposed SAM-modulated Bi-LSTM is 22.95% and 1.56% higher than DR-DSN [6] on the SumMe and TVSum datasets, respectively. Further, our method also performs better than the state-of-the-art Cycle-SUM [7], like 21.48% and 5.76% \( F \)-Score against

| Mode | Baseline | Deep deterministic policy gradient | Spatial attention model | SumMe | TVSum |
|------|----------|----------------------------------|------------------------|-------|-------|
| 1    | ✓        | ✓                                | ✓                      | 39.5  | 56.2  |
| 2    | ✓        | ✓                                | ✓                      | 45.9  | 57.8  |
| 3    | ✓        | ✓                                | ✓                      | 47.8  | 58    |
| 4    | ✓        | ✓                                | ✓                      | 50.9  | 58    |

Table 1. Results (\( F \)-score %) of comparisons

16.2% and 2.85% \( F \)-score gain, respectively, compared with the baseline (mode 1: the Bi-LSTM model). Moreover, the combination of the SAM and the DDPG (mode 4) rises by 28.86% and 4.09% \( F \)-Score against mode 1.

Conclusion: To accomplish the goal of high efficiency for user-created video summarisation, we propose a SAM-modulated Bi-LSTM model to discriminate the difference among images with the integration of semantic and saliency features. The high redundancy among keyframes has been proven from the low accuracy of SPP of Bi-LSTM, which leads to the low efficiency of unsupervised user-created video summarisation. According to the fact that humans tend to concentrate on sizeable moving objects, the SAM is established via a salient-area-size-guided saliency attention model. Moreover, we innovatively adopt DDPG to fulfill the unsupervised learning under the reinforcement learning framework. The extensive experiments demonstrate that the proposed SAM-modulated Bi-LSTM model outperforms the state-of-the-art methods.

Acknowledgements: This work is supported by the National Natural Science Foundation of China (Grant No. 62002130, 61702472) and the Hubei Province Technological Innovation Major Project (No. 2019AAA0409).

© 2021 The Authors. Electronics Letters published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited. Received: 25 October 2020 Accepted: 24 January 2021 doi: 10.1049/ell.121111

References
1 JI, Z., et al.: Video summarization with attention-based encoder-decoder networks. IEEE Trans. Circuits Syst. Video Technol. 30(6), 1709–1717 (2020)
2 Wang, W., Shen, J.: Video salient object detection via fully convolutional networks. IEEE Trans. on Image Processing 27(1), 38–49 (2017)
3 Zhang, K., et al.: Video summarization with long short-term memory. In: European Conference on Computer Vision - ECCV 2016, PT VII, Amsterdam, The Netherlands, vol. 9911, pp. 766–782 (2016)
4 Elharrar, E., Sapuro, G., Vidalt, R.: Spiece modeling for finding representative objects. In: CVPR, Providence, RI, USA, pp. 1600–1607 (2012)
5 Mahasseni, B., Lam, M., Todorovic, S.: Unsupervised video summarization with adversarial LSTM networks. In: CVPR, Hawaii, USA, pp. 2982–2991 (2017)

6 Zhou, K., Qiao, Y., Xiang, T.: Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In: AAAI 2018, Louisiana, USA, pp. 7582–7589 (2018)

7 Li, Y., et al.: Unsupervised video summarization with cycle-consistent adversarial LSTM networks. *IEEE Trans. Multimedia* 22(10), 2711–2722 (2020)

8 Treisman, A., Riley, J.: Is selective attention selective perception or selective response? A further test. *J. Exp. Psychol.* 79(1), 27–34 (1969)

9 Lillicrap, T., et al.: Continuous control with deep reinforcement learning. In: 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico (2015)

10 El-Ghoroury, H., Gupta, S.: Additive Bernoulli noise linear sequential circuits. *IEEE Trans. Comput.* 21(10), 1119–1124 (1972)