Enhanced Deep Video Summarization Network

Neeharika Gonuguntla[1]
gn12@iitbbs.ac.in
Bappaditya Mandal[2]
b.mandal@keele.ac.uk
Niladri B. Puhan[1]
nbpuhan@iitbbs.ac.in

[1] School of Electrical Sciences
Indian Institute of Technology
Bhubaneswar, Odisha, India

[2] School of Computing and Mathematics
Keele University
Staffordshire ST5 5BG, UK

Abstract

Video summarization is understanding video which aims to get an abstract view of the original video sequence by the concatenation of keyframes representing the highlights of the video. In this work, we propose an enhanced deep summarization network (EDSN) to summarize videos. We implement a reinforcement learning based framework to train our EDSN, where we design a novel reward function which considers the spatial and temporal features of the original video to be included in the summary. The reward function is formulated using the spatial and temporal scores obtained for each frame of the video using the temporal segment networks. During training, the reward function seeks to generate a summary by including the frames with high temporal and spatial scores, while the EDSN strives for earning higher rewards by learning to produce more diverse summaries. The method is completely unsupervised since no labels are required during training. Extensive experiments on two benchmark datasets show that the proposed approach achieves state-of-the-art performance.

1 Introduction

With the advent of technology, the number of videos and the amount of digital video data being generated has enormously increased. Video summarization is important in order to process this huge data. Video summarization aims to get concise and short summaries of large-scale videos that are representative of original videos. It makes the search way easier and useful than before. It is also a key tool where people can watch the important scenes without watching the full original video. This has lead to many recent competitions in life-log video summarization [3, 11]. Therefore, video summarization is important in this digital era to concise huge amount of visual information.

The importance of modeling temporal information has been differentiating the video and image models. Capturing the temporal information is necessary to understand the actions and sequences in a video [18]. Hence, extracting the temporal information is important to summarize the videos. In this paper, we aim to incorporate the long-range temporal structure, which plays an important role in understanding the dynamics in videos. This temporal information in a video is important in order to obtain an effective summary of the video. In

order to achieve this, we aim to design an effective video summarization technique using reinforcement learning and propose a reward function which is formulated using the spatial and temporal scores extracted from the video.

We propose an enhanced deep summarization network for video summarization. Our contribution is mainly towards developing a novel reward function incorporating the spatial and temporal information in a video. We formulate the reward function using the spatial and temporal scores extracted from the temporal segment networks in [22]. We conducted our experiments on two datasets, SumMe [9] and TVSum [20]. The SumMe dataset consists of 25 videos annotated with at least 15 human summaries whereas the TVSum dataset consists of 50 videos. Both the datasets consists of annotations, ground truth score for each frame of a video and the ground truth summary generated for each video. Experiments on the dataset showed that our approach achieves the state-of-the-art performance.

The rest of the paper is organized as follows: Section 2 covers the related work and the direction of the present work. Section 3 covers the proposed approach and our contribution to the new reward function which incorporates spatial and temporal information in a video. It also explains the reward function formulation from the scores and explains the method of extraction of the scores from the spatial and temporal ConvNets. Section 4 covers the experimental results and we conclude our work in Section 5.

2 Related Work

Video summarization has been of keen interest in recent years and lead to approaches of various techniques [1, 4]. For summarizing videos, important objects and people were identified in [7, 12]. Episodic visual memories are summarized and retrieved using semantic meaning from ego-centric videos in [2]. Large-scale egocentric visual data are summarized and retrieved using sparse graph representation in [16]. In [14] videos are summarized by modeling the viewer's attention. Inspired by works in text analysis, a story-driven video summarization is proposed in [13]. Key-frames are identified by clustering frames using conventional unsupervised approaches in [10]. On the other hand human annotated summaries are used to select informative and diverse subset of frames for summarizing the videos in [8]. In [9], scores of visual interestingness based on a set of low, mid and high-level features are obtained to create an interesting and informative summary. Long short-term memory is used in [24] to model the temporal dependency among the video-frames and derive compact summaries of the videos.

Key-frames are selected in [19], by employing reinforcement learning to train a summarization using the key-frame labels and category information. In [26], a deep summarization network with diversity- representativeness reward is proposed. This reward function is proposed in the way that labels or user interactions are not required at all during the learning process. A summarization network, which uses video-level category labels, using deep Q-learning is trained in [27], to identify key-frames.

Most of these methods processed video frames ignoring the inherent spatio-temporal patterns in the videos. To address this issue, we model video summarization via an enhanced deep summarization network (EDSN) to incorporate the spatio-temporal features in the summary. Our approach is based on training the deep network using reinforcement learning. Our work is mostly related to [26]. In [26], a deep summarization network with a diversity-representativeness reward is proposed. We significantly improve upon the deep summarization network by proposing a temporal-spatial reward which incorporates the spa-

tial and temporal features in the video.

3 Enhanced Deep Summarization Network

We develop an enhanced deep summarization network (EDSN) to predict probabilities for video frames and make decisions on which frames to select based on the predicted probability distributions. We design a temporal-spatial reward in order to ensure that the summary includes the frames with high temporal and spatial features. The enhanced deep summarization network (EDSN) is an enhanced version of the deep summarization network in [26]. Our contribution is mainly towards the design of the reward function which incorporates the temporal and spatial features within the video. The spatial and temporal scores for each frame of a video are extracted using the temporal segment networks (TSN). These scores are then used to obtain the reward function.

3.1 Extraction of Spatial and Temporal Cues

The temporal segment network framework in [22], which is composed of spatial stream ConvNets and temporal stream ConvNets, are used to obtain the spatial and temporal scores for each frame of the video. The temporal segment networks operate on a sequence of short snippets sparsely sampled from the entire video, instead of working on single frames or frame stacks. A preliminary prediction of the action classes is produced on its own by each snippet in this sequence. A consensus as the video-level prediction is then derived among the snippets. In the learning process, the loss values of video-level predictions, other than those of snippet-level predictions which were used in two-stream ConvNets, are optimized by iteratively updating the model parameters. The proposed framework is shown in Fig. 1.

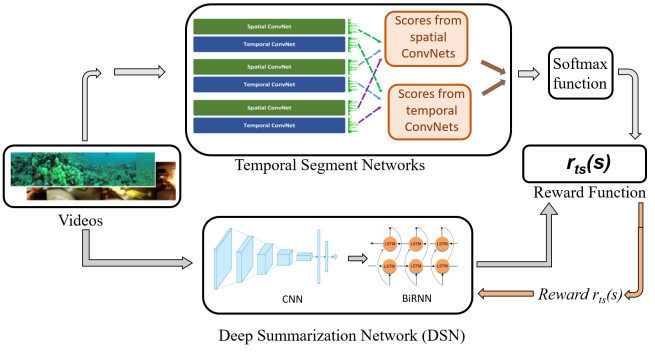

Figure 1: Enhanced Deep Summarization Network.

3.2 Temporal-Spatial Reward

We aim to incorporate the long-range temporal structure, which plays an important role in understanding the dynamics in videos. In general, we try to explore the temporal information within a video and propose the temporal-spatial reward which includes both the spatial and

temporal information within a video. The temporal-spatial reward is formulated as follows:

$$r_{\text{ts}}(s) = R_{\text{temp}} + R_{\text{spatial}}, \tag{1}$$

where R_{temp} and R_{spatial} are the temporal and spatial rewards, respectively. The temporal reward R_{temp} represents the temporal information whereas the spatial reward R_{spatial} represents the spatial information of each frame within the video. During training, the EDSN will receive a reward $r_{\text{ts}}(s)$ that evaluates the quality of generated summaries, and the objective of EDSN is to maximize the expected rewards over time by producing high-quality summaries.

3.2.1 Temporal Reward

This temporal information in a video is important in order to obtain an effective summary of the video. This information is included by formulating the temporal reward. The temporal reward R_{temp} is extracted from the scores of temporal relevance extracted for each frame of the video using the temporal ConvNets in [22]. The softmax function output for the scores of each frame x_t is assigned a probability $p(x_t)$ of being included in the video summary. The temporal reward R_{temp} is formulated from the scores as follows:

$$R_{\text{temp}} = \Sigma_{t=i}^{T} p(x_t), \tag{2}$$

where, $p(x_t)$ is the probability of a frame and T is the total number of frames in a video.

3.2.2 Spatial Reward

While the temporal information is significant in understanding the dynamics of the videos, the spatial information is also equally important in creating meaningful summaries. Therefore, the spatial reward is proposed. The spatial reward R_{spatial} is extracted similar to the temporal reward from the scores obtained from the spatial ConvNets in [22]. The softmax function output for the scores of each frame x_t is assigned a probability $p(x_t)$ of being included in the video summary. The spatial reward R_{spatial} is formulated from the scores as follows:

$$R_{\text{spatial}} = \Sigma_{t=i}^{T} p(x_t), \tag{3}$$

where, $p(x_t)$ is the probability of a frame. The spatial and temporal scores are the class scores extracted for snippets of video frames from the spatial and temporal ConvNets, respectively. The softmax function output means that the score of each frame is used to assign the probability of the frame, *i.e.* frames with high scores are assigned higher probabilities.

3.3 Summary generation

For a test video, a trained EDSN predicts the frame-selection probabilities for each frame of a video. Then the shot-level scores are computed by averaging frame-level scores within the same shot and summaries are generated by maximizing the total scores while ensuring that the summary length does not exceed a limit, which is usually 15% of the video length.

4 Experiments

The reinforcement learning based approach EDSN for video summarization is implemented. Experiments are conducted on the widely used SumMe dataset [9] and the TVSum dataset

[20]. The SumMe dataset was annotated by multiple persons so there are multiple human summaries for each video. The videos are down-sampled by 2 fps. The importance scores are converted to shot-based summaries for evaluation following [26]. The model is implemented using PyTorch [17]. The GoogLeNet [21] pre-trained on ImageNet [5] is used to extract frame features from the dataset, which is the input to the decoder network shown as bidirectional recurrent neural network (BiRNN) in Fig. 1. We set the dimension of hidden state in the recurrent neural network (RNN) cell to 256 throughout this paper. We use the standard 5-fold cross validation, i.e., 80% of videos for training and the rest for testing to evaluate our method, same protocol as that used in [26]. The EDSN is trained with 60 epochs. For evaluation, the F-score is computed for each pair of machine summary and human summary and the average results for a single video are obtained.

For the TSN, the BNInception network architecture is implemented. The mini-batch stochastic gradient descent algorithm is used to learn the network parameters, where the batch size is set to 256 and momentum set to 0.9. The network weights are initialized with pre-trained models from ImageNet [5]. For spatial networks, the learning rate is initialized as 0.001. For temporal networks, we initialize the learning rate as 0.005. For the extraction of optical flow and warped optical flow, we choose the TVL1 [23] optical flow algorithm implemented in OpenCV with CUDA. The TSN is trained with 21 videos and tested using 4 videos for the SumMe dataset while 40 videos from TVSum are used for training the TSN and 10 videos for testing. The frame-level scores are obtained from both the spatial and temporal stream ConvNets, which are fused together to formulate the reward function.

4.1 Results

Table 1 shows the average F-scores obtained from 5 different random splits on the SumMe and TVSum datasets.

Table 1: Results showing the average F-scores obtained from different splits using the reward function $r_{\text{ts}}(s)$.

Split number	Average F-scores using $r_{\text{ts}}(s)$ on SumMe dataset	Average F-scores using $r_{\text{ts}}(s)$ on TVSum dataset
1	41.0	56.3
2	50.9	57.9
3	42.2	57.8
4	43.4	57.0
5	34.6	57.4
Average F-scores	42.6	57.3

Table 2 shows the experimental results of our EDSN approach compared with other unsupervised approaches on TVSum and SumMe datasets. It can be seen that our method outperforms the other unsupervised approaches by good margins. The results show that our EDSN framework can better capture the long-range temporal dependencies among the video frames. It can be seen that our method performs the best on the SumMe dataset which contains videos of different categories, while the results are close to the state-of-the-art performance on the TVSum dataset which contains 50 videos of 10 different categories (5 videos of each category). The results suggest that the EDSN was successful in capturing the dynamics of the video frames and thereby generating effective summaries.

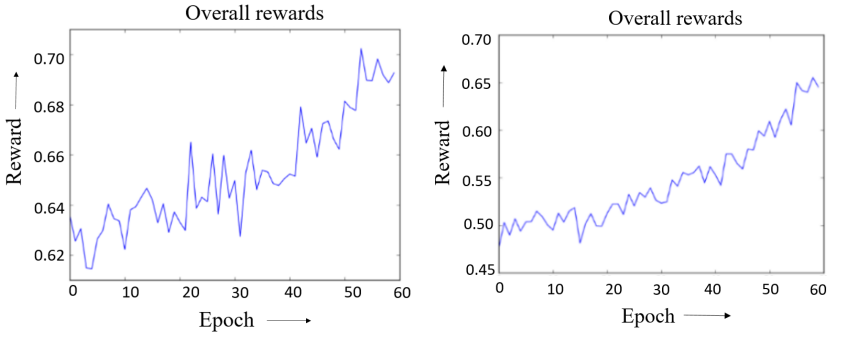

Figure 2: Plots showing the reward values obtained for each epoch. Left, plot of reward values vs epochs for split 1 in SumMe dataset. Right, plot of reward values vs epochs for split 1 in TVSum dataset.

The overall reward is the average of the rewards obtained for each video during training. The overall reward for each epoch during the training is obtained and the results are plotted in Fig. 2. We can observe that the reward generated increases as the number of epochs increases during training, i.e. the EDSN tries to maximize the reward and obtain meaningful summaries as the training progresses. The plots of overall reward obtained vs epochs for a random split selected from both the datasets are shown in Fig. 2.

The plots in Fig. 3 shows the ground truth scores (top in color red) and the predicted scores (bottom in color blue) for the test videos. The plots show that the distributions of output importance scores generated from the EDSN for each frame are almost uniform with low discrepancy, whereas the ground truth scores have a high discrepancy among the scores of each frame clearly showing that the scores are given by humans. The plots for a randomly chosen test video are shown in Fig. 3 left and right, from SumMe and TVSum datasets, respectively.

5 Conclusions and Future Work

In this work, we proposed a new reward function in order to include the temporal and spatial information for unsupervised video summarization. Extensive experiments on SumMe and TVSum datasets showed that using reinforcement learning with our unsupervised re-

Table 2: Results comparing F-scores on the TVSum and SumMe datasets using existing methodologies and our proposed EDSN approach.

Methods	F1-score on TVSum dataset	F1-score on SumMe dataset
Uniform Sampling	15.5	29.3
K-medoids [10]	28.8	33.4
Online sparse coding [25]	46.0	-
Dictionary selection [6]	42.0	37.8
GAN [15]	51.7	39.1
DR-DSN [26]	57.6	41.4
EDSN	57.3	42.6

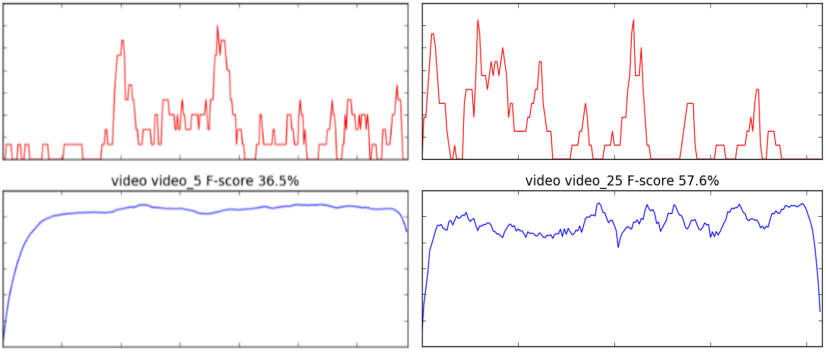

Figure 3: Left, plots showing the ground truth scores (left top) and the predicted scores from EDSN (left bottom) obtained for each frame of the video_5 from SumMe dataset. Right, plots showing the ground truth scores (right top) and the predicted scores from EDSN (right bottom) obtained for each frame of the video_25 from TVSum dataset. X-axis are the frames numbers and Y-axis are the score values.

ward function outperformed other state-of-the-art unsupervised alternatives and produced results comparable to most supervised methods [8]. In future, we aim to optimize the reward function by formulating the temporal and spatial information more rigorously by assigning weights to the scores obtained from the spatial and temporal stream ConvNets rather than directly obtaining their probabilities from the softmax function. Furthermore, issues like ineffective feature learning due to distributions of output importance scores for each frame are to be tackled. This can be tackled by proposing variance loss which allows a network to predict output scores for each frame with high discrepancy which enables effective feature learning and significantly improves the model performance.

References

[1] Alejandro Betancourt, Pietro Morerio, Carlo S. Regazzoni, and Matthias Rauterberg. The evolution of first person vision methods: A survey. *IEEE Trans. Circuits Syst. Video Techn.*, 25(5):744–760, 2015.

[2] Ana Garcia del Molino, Bappaditya Mandal, Liyuan Li, and Joo-Hwee Lim. Organizing and retrieving episodic memories from first person view. In *2015 IEEE International Conference on Multimedia & Expo Workshops, ICME Workshops 2015, Turin, Italy, June 29 - July 3, 2015*, pages 1–6, 2015.

[3] Ana Garcia del Molino, Bappaditya Mandal, Jie Lin, Joo-Hwee Lim, Vigneshwaran Subbaraju, and Vijay Chandrasekhar. Vc-i2r@imageclef2017: Ensemble of deep learned features for lifelog video summarization. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.*, 2017.

[4] Ana Garcia del Molino, Cheston Tan, Joo-Hwee Lim, and Ah-Hwee Tan. Summarization of egocentric videos: A comprehensive survey. *IEEE Transactions on Human-Machine Systems*, 47(1):65–76, 2017.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255, 2009.

[6] Ehsan Elhamifar, Guillermo Sapiro, and René Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 1600–1607, 2012.

[7] Tian Gan, Yongkang Wong, Bappaditya Mandal, Vijay Chandrasekhar, and Mohan S. Kankanhalli. Multi-sensor self-quantification of presentations. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 601–610, 2015.

[8] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems, Montreal, Quebec, Canada*, pages 2069–2077, 2014.

[9] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc J. Van Gool. Creating summaries from user videos. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII*, pages 505–520, 2014.

[10] Youssef Hadi, Fedwa Essannouni, and Rachid Oulad Haj Thami. Video summarization by k-medoid clustering. In *Proceedings of the 2006 ACM Symposium on Applied Computing (SAC), Dijon, France, April 23-27, 2006*, pages 1400–1401, 2006.

[11] ImageCLEFlifelog. Subtask 2: Lifelog summarization (lst). https://www.imageclef.org/2017/lifelog, 2019.

[12] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 1346–1353, 2012.

[13] Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 2714–2721, 2013.

[14] Yu-Fei Ma, Lie Lu, HongJiang Zhang, and Mingjing Li. A user attention model for video summarization. In *Proceedings of the 10th ACM International Conference on Multimedia 2002, Juan les Pins, France, December 1-6, 2002.*, pages 533–542, 2002.

[15] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. Unsupervised video summarization with adversarial LSTM networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2982–2991, 2017.

[16] Wu Min, Xiao Li, Cheston Tan, Bappaditya Mandal, Liyuan Li, and Joo-Hwee Lim. Efficient retrieval from large-scale egocentric visual data using a sparse graph representation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2014, Columbus, OH, USA, Jun*, pages 541–548, 2014.

[17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[18] Sibo Song, Ngai-Man Cheung, Vijay Chandrasekhar, and Bappaditya Mandal. Deep adaptive temporal pooling for activity recognition. In *ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, Oct*, pages 1829–1837, 2018.

[19] Xinhui Song, Ke Chen, Jie Lei, Li Sun, Zhiyuan Wang, Lei Xie, and Mingli Song. Category driven deep recurrent neural network for video summarization. In *2016 IEEE International Conference on Multimedia & Expo Workshops, ICME Workshops 2016, Seattle, WA, USA, July 11-15, 2016*, pages 1–6, 2016.

[20] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. TVSum: Summarizing web videos using titles. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 5179–5187, 2015.

[21] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1–9, 2015.

[22] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, Oct*, pages 20–36, 2016.

[23] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-L^1 optical flow. In *Pattern Recognition, 29th DAGM Symposium, Heidelberg, Germany, September 12-14, 2007, Proceedings*, pages 214–223, 2007.

[24] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, Oct*, pages 766–782, 2016.

[25] Bin Zhao and Eric P. Xing. Quasi real-time summarization for consumer videos. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 2513–2520, 2014.

[26] Kaiyang Zhou, Yu Qiao, and Tao Xiang. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pages 7582–7589, 2018.

[27] Kaiyang Zhou, Tao Xiang, and Andrea Cavallaro. Video summarisation by classification with deep reinforcement learning. In *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*, page 298, 2018.