2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)

VIDEO SUMMARIZATION USING GLOBAL ATTENTION WITH MEMORY NETWORK AND LSTM

Dhruva Sahrawat[*†], *Mohit Agarwal*[*†], *Sanchit Sinha*[*†], *Aditya Adhikary*[*†], *Mansi Agarwal*[δ]
Rajiv Ratn Shah[†], *Roger Zimmermann*[γ]

[†]IIIT-Delhi, [δ]DTU-Delhi, [γ]NUS-Singapore

ABSTRACT

Videos are one of the most engaging and interesting mediums of effective information delivery and constitute the majority of the content generated online today. As human attention span shrinks, it is imperative to shorten videos while maintaining most of its information. The premier challenge is that summaries more intuitive to a human are difficult for machines to generalize. We present a simple approach to video summarization using Kernel Temporal Segmentation (KTS) for shot segmentation and a global attention based modified memory network module with LSTM for shot score learning. The modified memory network termed as Global Attention Memory Module (GAMM) increases the learning capability of the model and with the addition of LSTM, it is further able to learn better contextual features. Experiments on the benchmark datasets TVSum and SumMe show that our method outperforms the current state of the art by about 15%.

Index Terms— Video Summary, Shots, Frames, Deep Learning, Supervised Learning

1. INTRODUCTION AND RELATED WORK

Video content is highly personalizable, engages people across varying age groups and is more entertaining than textual media. With growth of mobile devices, the volume of videos created, viewed and shared online has grown exponentially [1]. The number of online videos on video sharing and social media websites such as Youtube, Dailymotion, Facebook, Twitter, Reddit, etc. have experienced an explosive growth in the past decade. Online users are also overwhelmed by the amount and variety of videos. Some studies have shown that the 'attention spans' of humans have been decreasing significantly in the past decade and some estimates have quantified it to about 9 seconds [2]. Hence, an effective technique which condenses the highlights or important points of a video in a short clip or a summary is required. Video summarization has huge potential in many use-cases such as highlights of sporting events or long term security monitoring and surveillance. Existing literature on methodologies for video summarization, in both unsupervised [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16] and supervised [17, 18, 19, 20, 21, 22, 23, 24, 25, 26] categories are well studied. Among the most naive approaches used for many years, clustering similar shots using hand crafted/deep features or graph-based hierarchical techniques have been used, such as in works [3, 4, 5, 6, 27]. These approaches have little or no holistic understanding as they did not consider any context and are also dissimilar to human understanding. Some methods like [7] used most frequent co-occurring shots from across videos in a dataset. LSTMs and RNNs, such as in works [9, 10, 11] model sequential attention and suggest a more holistic approach. However, using memory cells as is shows that they are not robust enough to hold information across large stretches of video. Recent GAN based approaches such as [13] used a Variational Autoencoder for selecting sparse frames (generator), and an RNN classifier for distinguishing original and summarized videos (discriminator). Another recent paper [14] used an adversarial training framework for *semi-supervised* video summarization, and achieved results comparable to the state-of-the-art. Thus, recent unsupervised approaches have so far been producing results comparable to the state-of-the-art.

Amongst supervised approaches, methods such as LSTM were combined with Determinantal Point Process (DPP), a kind of stochastic point process in works [21, 22] to model the variable-range temporal dependency among video frames, accounting for the sequential structure as well as long-term dependencies. Recent works such as that by Fajtl et. al [24] proposed a method for supervised, keyshot based video summarization by applying a self-attention mechanism, which performed the entire sequence-to-sequence transformation in a single feed forward and a single backward pass during training. The current state-of-the-art method by Feng at. al [17] uses a memory augmented neural network with an external memory, providing a more global understanding of the video frames while predicting the importance scores of the video shots. It attempts to understand the whole video, and the global attention mechanism captures information from all video frame.

In this paper, we build upon the global attention memory network for video summarization proposed by Feng et al. The objective behind using global memory is to emulate the

*Equal contribution

978-1-7281-5527-2/19/$31.00 ©2019 IEEE
DOI 10.1109/BigMM.2019.00042

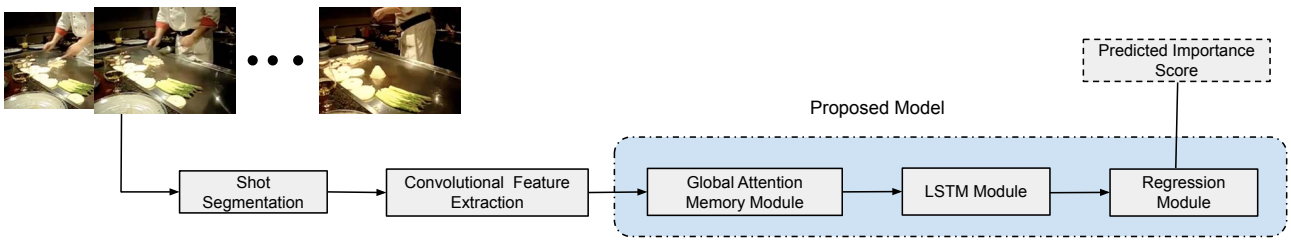

Fig. 1. Overall pipeline of our proposed approach

human thinking process for creating a video summary. As pointed out by the authors, global memory helps model the global dependency across frames - which is similar to the human behaviour of looking at not only the local sequences but also the context of the entire video to create a summary. The foremost step for a video summarization model is to implement shot detection and segmentation. We use Kernel Temporal Segmentation (KTS) [18] which produces rich, precise and visually coherent temporal segments. Our proposed model improves on the Memory Augmented Video Summarizer (MAVS) [17] by using a fully connected layer with ReLU in place of the embedding matrix, and also by adding an LSTM to the output of the memory module. After replacing with a fully connected layer, the model is able to learn better and more robustly due to the introduction of non-linearity. Our model adds on the LSTM unit to global attention network, thus incorporating advantages of using both the local memory which pays more attention to the local sequence and global memory units which attempts to mimic humans by looking at the whole video at once.

2. PROPOSED APPROACH

The goal of our model is Video Summarization, i.e., given a video, generate its summary. We model the problem as a subset selection problem. Each video is segmented into a sequence of video shots. The frames in each video shots are expected to be quite similar and represent a sub-event. The boundary frames between the shots represent scene changes i.e. frames where very significant change is detected, this can represent start of a new sub-event. Our proposed deep learning model takes in the sequence of video shots as input and predicts the importance score for each of the video shot. This importance score which is between 0 to 1 represents the likelihood that it will be included in the generated summary. This entire process is summarized in Figure 1.

2.1. Shot Detection and Segmentation

For this, we use Kernel Temporal Segmentation (KTS) algorithm, proposed by Potapov et al. [18]. Kernel change point analysis has been explored in signal processing and statistical studies before [28], but was exclusively used for video segmentation in the aforementioned works [18, 24]. This method's efficacy when taking high-dimensional descriptors (of each frame of the video in this case) as input is the result of strong theoretical verification as compared to other heuristic-based shot boundary detection techniques such as that used in the work [17]. The goal is to divide the video (or generally any noisy signal) into a set of non-overlapping segments, by differentiating between the cause of 'jumps' i.e whether they are resulting from noise or are caused by the signal itself. This is achieved by minimizing an objective function consisting of the within-segment variances and a penalty function for limiting the number of segments, using an iterative dynamic programming algorithm.

2.2. Shot Feature Representation

For each frame, the output of the second last (pool-5) layer of the GoogLeNet model [29] trained on ImageNet is utilized as the feature descriptor, which is of length 1024. Since the input to the model is a sequence of video shots, we represent each video shot with a single feature vector which is the average of the feature descriptors of each frame in it. We use shot-level features instead of frame level features to reduce the complexity of the model, since the number of videos is less. Also, using shot-level features reduces the memory footprint of our model which enables us to work on the entire video at once.

2.3. Architecture

Our proposed approach takes in a video as input. Using the methods mentioned above a sequence of feature representation of each shot $\{x_i\}$ present in the input video is generated.

Our model is illustrated in Figure 1 and we now describe all of the modules which comprise the model in detail.

Our model uses Global Attention through a memory network module in a manner similar to how it is used in the works [17, 30]. We term this part of our proposed network as Global Attention Memory Module (GAMM). Each of the video shot's feature representation $\boldsymbol{x}_i$ is written into external memory as input and output memory vector represented by

$\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ respectively. We use separate fully connected layer with ReLU activation function to convert the input video shot feature vectors $\boldsymbol{x}_i$ to input $\boldsymbol{u}_i$ and output memory vectors $\boldsymbol{v}_i$.

Every input memory vector $\boldsymbol{u}_i$ encodes the information about its respective video shot. This information is later used to compare it with other video shots. This is done by comparing each of the feature representations $\boldsymbol{x}_i$ with every input memory vector $\boldsymbol{u}_j$. This generates global attention scores $\boldsymbol{a}_{i,j}$. When comparing instead of directly using feature representations $\boldsymbol{x}_i$ we first pass it through a fully connected layer with ReLU activation to generate projected feature representation $\boldsymbol{z}_i$. The normalized global attention vector $\boldsymbol{a}_{i,j}$ is then calculated as:

$$\boldsymbol{a}_{i,j} = \frac{\exp(\boldsymbol{z}_i^T \boldsymbol{u}_j)}{\sum_{j'} \exp(\boldsymbol{z}_i^T \boldsymbol{u}_{j'})} \tag{1}$$

The global attention vector $\boldsymbol{a}_{i,j}$ represent the relevance of the video shot j in the generation of importance score of video shot i. The context vector $\boldsymbol{c}_i$ is generated as the average of all the output memory vector $\boldsymbol{v}_j$ weighed by its corresponding global attention vector $\boldsymbol{a}_{i,j}$.

$$\boldsymbol{c}_i = \sum_j \boldsymbol{a}_{i,j} \boldsymbol{v}_j \tag{2}$$

The final output vector $\boldsymbol{o}_i$ of the memory module is calculated as element wise multiplication of the context vector $\boldsymbol{c}_i$ with its corresponding refined feature representation $\boldsymbol{z}_i$.

This whole process can be repeated again by applying the memory model again on the output of the GAMM where each application of the GAMM is a single hop. Since our dataset is small and also to reduce complexity, adjacency weight sharing is used similar to how it is used in works [17, 30].

The output generated after applying the memory module set number of hops is then fed into a Bidirectional-LSTM layer [31, 32] with number of cells equal to the number of shots in a video. This is the LSTM module of our model. Finally, a fully connected layer with sigmoid as activation function is used to performs regression on the output of the LSTM corresponding to each video shot to generate its importance score. This is the Regression module of our model. We use Mean Square Error (MSE) loss on the predicted and ground truth shot level importance score as our loss function.

Our model differs from the work [17] in many areas, instead of embedding matrix we use a fully connected layer with ReLU activation function which adds non-linearity and increases the learning power of the network while the number of parameters remains the same. This fully connected layer is used for refining the input feature vectors and allow for end-to-end training of the model.

We also added an LSTM to the output of the memory module. In the work [17], authors found that using multiple hops we can increase the performance of the model, but the computational complexity also increases rapidly. Instead, we propose to use a single hop and then use an LSTM on top of it. The LSTM helps to further refine the output of the memory model. The memory model applies global attention, the sequence of the video shots do not matter to the memory model, while in an LSTM the sequence is important. The output of a bidirectional-LSTM cell is affected more by the input to its nearer cell than by the distant cells. So by combining both the LSTM and Global attention using memory module we want to learn better contextual features for each video shots.

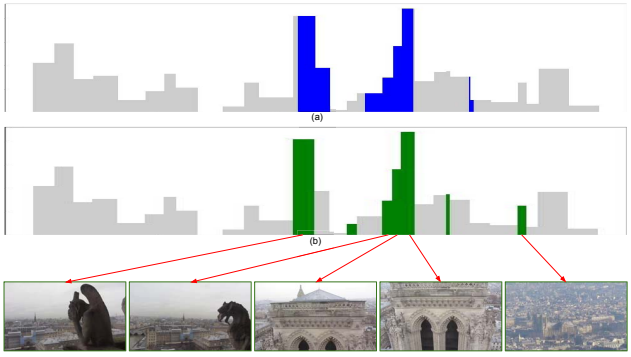

Fig. 2. Visualization[2] of ground truth shot scores (grey), (a)ground truth summary (blue), (b) model generated summary (green) for the video "Notre_Dame" in SumMe dataset.

3. EXPERIMENTS

We first describe the datasets we use for evaluation. Then we define the various experimental settings used for training the model. Finally, we present and discuss the results.

3.1. Dataset

We evaluate the performance of our model on the following two popular publicly available Video Summarization datasets:

SumMe [34]: This dataset contains 25 videos and there are no specific categories. The video length typically varies from 1 to 6 minutes. Each video contains at least 15 different human annotations of frame level importance scores.

TVSum [35]: This dataset comprises of 50 videos with the typical video length ranging from 2 to 10 minutes. It has 10 categories with 5 videos each. Each video has exactly 20 human annotations of frame level importance scores.

3.2. Experimental Configurations

We use Adam optimizer [36] to train our model. The learning rate is taken as 10^{-3} and the batch size which the number of input videos to the model is set to 1. The size of the feature representation d is taken as 1024, the number of hidden units in the fully connected layers (except the last one) f and the embedding size of the LSTM is taken as 512. The

Table 1. Performance comparison of our proposed model with other recent Video Summarization approaches on SumMe and TVSum datasets.

	Methods	hops	SumMe (F1-score in %)	TVSum (F1-score in %)
Baselines	DPP-LSTM[9]	-	38.6	54.7
	Summary-transfer [33]	-	40.9	-
	GAN-based[13]	-	41.7	56.3
	RL-based[12]	-	42.1	58.1
	Temporal-tessellation[20]	-	41.4	64.1
	MAVS [17]	1	39.8	67.0
	MAVS[17]	4	43.1	67.5
	MAVS[17]	6	40.3	66.8
Proposed	KTS + GAMM	1	49.3	81.5
	KTS + GAMM	4	50.0	79.1
	KTS + GAMM	6	50.7	78.6
	KTS + GAMM	10	**53.8**	78.4
	KTS + GAMM + LSTM	1	41.2	**83.1**

maximum length of the video in terms of number of shots is taken as 500. If the number of video shots is less than 500, the rest of the video shots are assumed to be zero-vector. We use identical training and testing ground truth data used by works [13, 21, 37]. The predicted summaries from predicted shot level importance score and ground truth shot-level importance are obtained in a manner similar to work [17].

3.3. Result

We compare our proposed model with prior state-of-art works on Video Summarization. We use F1-score for comparison. The results are summarised in Table 1. The best performing results on each dataset are in bold. An example of a visually interpretable result can be seen via Figure 2.

Note that KTS+GAMM represents the model where only both KTS (Kernel Temporal Segmentation) and GAMM (Global Attention Memory Module) parts of the network are used. KTS+GAMM+LSTM represents our proposed model.

Our proposed model (KTS+GAMM+LSTM) gives state of the art result, it shows an improvement of about 16% on the TVSum dataset compared to the previous state-of the art result of 67.3% while its result is comparable to the previous state-of-art on the SumMe dataset. The model KTS+GAMM also shows a significant improvement and beats the the previous state-of-art models on both TVSum and SumMe.

The difference in performance of KTS+GAMM+LSTM and KTS+GAMM on SumMe dataset can be explained by the fact that the SumMe dataset is smaller, therefore the parameters in the LSTM layer is not able to train well. In the TVSum the opposite happens, since it is larger than SumMe, the parameters of LSTM are able to train well, therefore KTS+GAMM+LSTM perform better than KTS+GAMM. One interesting fact is that KTS+GAMM beats MAVS [17] for both the datasets. This shows the benefit of using KTS as shot detection algorithm well as replacement of the embedding matrices with the fully connected layers. As expected increasing the number of hops leads to increase in performance as can be seen in the case of KTS+GAMM for both the SumMe and TVSum dataset. One of the issues that occurs with increasing the number of hops is the rapid increase in computation time. This is one of the reasons why we use only one hop in case of KTS+GAMM+LSTM. The LSTM layer is able to model the relationship between the memory network module output corresponding to each shots. This helps in generating better contextual features and makes the number of hops redundant.

[1] Original, ground truth and model generated videos can be seen at `http://tiny.cc/363gaz`

4. CONCLUSION

Owing to the increase in the number of videos and reduction in the attention span of people, we present a general approach to create meaningful video summaries. Our method utilizes both, the Global Attention Memory Module (GAMM) comprising of fully connected layer with ReLU, and LSTM for incorporating local attention. It leads to an increase in the learning capability of the model and better feature learning for video shots. Results on the two benchmark datasets *SumMe* and *TVSum* show that our method performs better than the previously existing state-of-the-art to the best of our knowledge and as a result can be used to create relevant and valid video summaries.

5. ACKNOWLEDGEMENT

Rajiv Ratn Shah is partly supported by the Infosys Center for AI, IIIT Delhi and ECRA Grant by SERB, Government of India. This research is supported by Singapore Ministry of Education Academic Research Fund Tier 1 under MOE's official grant number T1 251RES1820.

6. REFERENCES

[1] "55 video marketing statistics for 2019," https://biteable.com/blog/tips/video-marketing-statistics/, 2019.

[2] "As attention spans get shorter, content gets even shorter," https://www.huffpost.com/entry/as-attention-spans-get-shorter/-content-gets-shorter_b_5a57ae42e4b00a8c909f7f1e/, 2018.

[3] 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*, New York, NY, USA, 2006, SAC '06, pp. 1400–1401, ACM.

[4] Luciana dos Santos Belo, Carlos Antônio Caetano, Zenilton Kleber Gonçalves do Patrocínio, and Silvio Jamil Ferzoli Guimarães, "Summarizing video sequence using a graph-based hierarchical approach," *Neurocomput.*, vol. 173, no. P3, pp. 1001–1016, Jan. 2016.

[5] Qing-Ge Ji, Zhi-Dang Fang, Zhen-Hua Xie, and Zhe-Ming Lu, "Video abstraction based on the visual attention model and online clustering," *Image Commun.*, vol. 28, no. 3, pp. 241–253, Mar. 2013.

[6] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya, "Video summarization using deep semantic features," *CoRR*, vol. abs/1609.08758, 2016.

[7] Wen-Sheng Chu, Yale Song, and Alejandro Jaimes, "Video co-summarization: Video summarization by visual co-occurrence," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[8] C. Tsai, L. Kang, C. Lin, and W. Lin, "Scene-based movie summarization via role-community networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 11, pp. 1927–1940, Nov 2013.

[9] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman, "Video summarization with long short-term memory," *CoRR*, vol. abs/1605.08110, 2016.

[10] Bin Zhao, Xuelong Li, and Xiaoqiang Lu, "Hierarchical recurrent neural network for video summarization," in *ACM Multimedia*, 2017.

[11] Bin Zhao, Xuelong Li, and Xiaoqiang Lu, "Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[12] Kaiyang Zhou and Yu Qiao, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," *CoRR*, vol. abs/1801.00054, 2018.

[13] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic, "Unsupervised video summarization with adversarial lstm networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2982–2991, 2017.

[14] Tsu-Jui Fu, Shao-Heng Tai, and Hwann-Tzong Chen, "Attentive and adversarial learning for video summarization," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[15] Rajiv Ratn Shah, Yi Yu, Akshay Verma, Suhua Tang, Anwar Dilawar Shaikh, and Roger Zimmermann, "Leveraging multimodal information for event summarization and concept-level sentiment analysis," *Knowledge-Based Systems*, vol. 108, pp. 102 – 109, 2016, New Avenues in Knowledge Bases for Natural Language Processing.

[16] Rajiv Ratn Shah, Yi Yu, and Roger Zimmermann, "Advisor: Personalized video soundtrack recommendation by late fusion with heuristic rankings," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, New York, NY, USA, 2014, MM '14, pp. 607–616, ACM.

[17] Litong Feng, Ziyin Li, Zhanghui Kuang, and Wei Zhang, "Extractive video summarizer with memory augmented neural networks," in *Proceedings of the 26th ACM International Conference on Multimedia*, New York, NY, USA, 2018, MM '18, pp. 976–983, ACM.

[18] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid, "Category-specific video summarization," in *ECCV - European Conference on Computer Vision*, David Flee, Tomas Pajdla, ernt Schiele, and Tinne Tuytelaars, Eds., Zurich, Switzerland, Sept. 2014, vol. 8694 of *Lecture Notes in Computer Science*, pp. 540–555, Springer.

[19] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov 2004.

[20] Dotan Kaufman, Gil Levi, Tal Hassner, and Lior Wolf, "Temporal tessellation for video annotation and summarization," *CoRR*, vol. abs/1612.06950, 2016.

[21] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman, "Video summarization with long short-term memory," in *ECCV*, 2016.

[22] Aidean Sharghi, Jacob S. Laurel, and Boqing Gong, "Query-focused video summarization: Dataset, evaluation, and A memory network based approach," *CoRR*, vol. abs/1707.04960, 2017.

[23] Y. Yuan, T. Mei, P. Cui, and W. Zhu, "Video summarization by learning deep side semantic embedding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 1, pp. 226–237, Jan 2019.

[24] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino, "Summarizing videos with attention," *CoRR*, vol. abs/1812.01969, 2018.

[25] Y. Yu, S. Tang, K. Aizawa, and A. Aizawa, "Category-based deep cca for fine-grained venue discovery from multimodal data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 4, pp. 1250–1258, April 2019.

[26] Yi Yu, Suhua Tang, Francisco Raposo, and Lei Chen, "Deep cross-modal correlation learning for audio and lyrics in music retrieval," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 15, no. 1, pp. 20:1–20:16, Feb. 2019.

[27] Rajiv Ratn Shah, "Multimodal analysis of user-generated content in support of social media applications," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, New York, NY, USA, 2016, ICMR '16, pp. 423–426, ACM.

[28] Zaïd Harchaoui, Eric Moulines, and Francis R. Bach, "Kernel change-point analysis," in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., pp. 609–616. Curran Associates, Inc., 2009.

[29] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[30] Minh-Thang Luong, Hieu Pham, and Christopher D Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

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

[32] M. Schuster and K.K. Paliwal, "Bidirectional recurrent neural networks," *Trans. Sig. Proc.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[33] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman, "Summary transfer: Exemplar-based subset selection for video summarization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1059–1067.

[34] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool, "Creating summaries from user videos," in *ECCV*, 2014.

[35] Yale Song, J. Vallmitjana, A. Stent, and A. Jaimes, "Tvsum: Summarizing web videos using titles," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, vol. 00, pp. 5179–5187.

[36] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[37] Kaiyang Zhou, Yu Qiao, and Tao Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," *arXiv:1801.00054*, 2017.