Reinforcement Learning based Video Summarization with Combination of ResNet and Gated Recurrent Unit

Muhammad Sohail Afzal and Muhammad Atif Tahir
National University of Computer and Emerging Sciences, Karachi, Pakistan

Keywords: Video Summarization, Reinforcement Learning, Reward Function, ResNet, Gated Recurrent Unit.

Abstract: Video cameras are getting ubiquitous with passage of time. Huge amount of video data is generated daily in this world that needs to be handled efficiently in limited storage and processing power. Video summarization renders the best way to quickly review over lengthy videos along with controlling storage and processing power requirements. Deep reinforcement-deep summarization network (DR-DSN) is a popular method for video summarization but performance of this method is limited and can be enhanced with better representation of video data. Most recently, it has been observed that deep residual networks are quite successful in many computer vision applications including video retrieval and captioning. In this paper, we have investigated deep feature representation for video summarization using deep residual network where ResNet 152 is being used to extract deep videos features. To speed up the model, long short term memory is replaced with gated recurrent unit, thus gave us flexibility to add another RNN layer which resulted in significant improvement in performance. With this combination of ResNet-152 and two layered gated recurrent unit (GRU), we performed experiments on SumMe video dataset and got results not only better than DR-DSN but also better than several state of art video summarization methods.

1 INTRODUCTION

The deployment of cameras are getting proliferated on each successive day. As the number of cameras are increasing, we are gaining more and more video data. Almost 10 million GB of data related to videos is generated in a single week in this world (Lai et al., 2016). This needs to be efficiently stored and processed. But we always have some upper limit for available storage and processing power. It would be a wise decision if we reduce this data to summary of important frames and discard useless frames. Both the problems will be solved provided that the generated summary is good representation of complete video ensuring that important frames are not lost. To sum up long videos in just few frames is a challenging task and requires an approach to efficiently solve this problem with reasonable accuracy. Several supervised approaches have been proposed for video summarization but it is intimidate task to annotate whole video by just one label as there is no single ground truth for any single video. So DR-DSN (Zhou et al., 2018a) introduced unsupervised video summarization approach based on reinforcement learning reward function. Basic idea of how reinforcement learning interacts with video summarization is shown in Fig. 1. where agent generates summary and gives it to reward function for evaluation. Reward function gives evaluation feedback to agent through which agent learns to generate better summary next time.

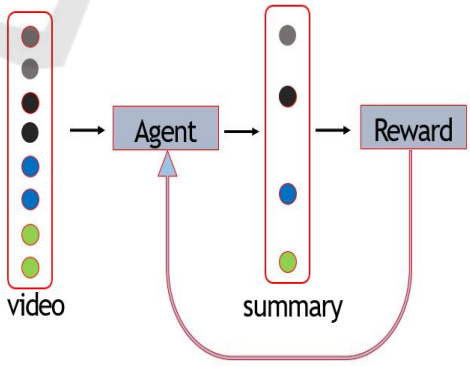

Figure 1: Basic idea of reinforcement learning in video summarization.

Residual network (ResNet) has always been of great importance in learning deep feature representation of data as these networks are being used in many computer vision challenges like ILSVRC and COCO competitions to get better results (He et al.,

Afzal, M. and Tahir, M.
Reinforcement Learning based Video Summarization with Combination of ResNet and Gated Recurrent Unit.
DOI: 10.5220/0010197402610268
In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - Volume 4: VISAPP*, pages 261-268
ISBN: 978-989-758-488-6

2016). Different ResNet models were employed for cancer prediction and malware detection on two different datasets (Khan et al., 2018) where ResNet-152 achieved top most accuracy. Deep ResNet has been used for classification of hyper spectral images and it was noticed that the more deeper the network is, the better features it can capture and hence the better result (Zhong et al., 2017). GoogleNet and ResNet were compared on malware detection problem (Khan et al., 2019) and it was concluded that ResNet gave more accurate results than GoogleNet but took more time than GoogleNet to solve problem at hand.

It was noticed in a study related to recurrent neural network (Elsayed et al., 2018) that replacing long short term memory (LSTM) with gated recurrent unit (GRU) increased the performance of classification task. Furthermore, this GRU was combined with convolution neural network in the same classification task and it showed performance improvement which is what we also adopted in our proposed method of video summarization. To the best of our knowledge, this combination of ResNet and GRU has never been investigated before for video summarization methods. Our approach of video summarization is designed in a way to get better results in less amount of time. ResNet-152 is a deeper architecture with 152 layers but it takes more time to extract all important features so our recurrent neural network, which is basically a two layered gated recurrent unit, compensates for this time as it takes less time than long short term memory that was previously employed in DR-DSN. The proposed approach is evaluated on SumMe dataset which is widely used benchmark dataset for video summarization. The results indicate a significant increase in the performance when compared with the state-of-the-art video summarization methods.

The paper is organized as follows. Section 2 gives background of previously used approaches for video summarization. Section 3 elaborates our proposed methodology for video summarization with explanation of all components of our approach. Section 4 describes experimental settings and results that we got after comparison of our approach with DR-DSN and several other well known approaches. Section 5 finally concludes the paper.

2 RELATED WORK

Various studies have been conducted for solving video summarization through reinforcement learning (Masumitsu and Echigo, 2000)(Zhou et al., 2018a)(Zhou et al., 2018b)(Zhang et al., 2019)(Lei et al., 2018). Importance score of each frame was calculated (Masumitsu and Echigo, 2000) by projecting features in eigen space and then video was generated through reinforcement learning. Experiments were manipulated on soccer video game and higher precision values were achieved. In (Zhou et al., 2018a), a deep summarization network and an end to end reinforcement learning approach is proposed to generate high quality summary. Results were obtained on SumMe and TVSum dataset that were better than other state of art methods. Another approach based on action parsing is presented (Lei et al., 2018) in which video was first cut by action parsing and then summarized through reinforcement learning. Experiments were manipulated on SumMe and TVSum dataset achieving better F-score than several competitive methods. In (Zhou et al., 2018b), a summarization network is presented which was trained through deep Q-learning and tested on CoSum and TVSum dataset. Results were better than several advanced methods. MapNet is exploited (Zhang et al., 2019) to first map frames with respective queries and then summaries are generated through summNet. Experiments were manipulated on UT Egocentric dataset where state of art results were achieved.

Several significant researches related to video summarization have also been conducted being specific to surveillance cameras. A video summarization technique was proposed (Yang et al., 2011) for surveillance cameras based on key frame extraction. The most informative scenes were selected until required summarization rate was achieved. Experiments were manipulated on CAVIAR dataset and results were better than other famous techniques of optical flow and color spatial distribution. Target driven summarization of surveillance video for tracing suspects is proposed (Chen et al., 2013) that works in two steps. In first step, by using filtered summarized video of any camera, targets can be detected. This first step will filter for the categories of target along with the time information. After identification of targets, appearance signals are activated in other cameras. A perspective dependent model is then proposed based on grid that showed satisfactory results. Event based video summarization of surveillance cameras is proposed (Yun et al., 2014) that focuses on the appearance as well as patterns of movement. The problem of occlusion solved by separating local motion from global motion and hence greater accuracy was achieved. Another event driven approach was presented (Dimou et al., 2015) that focused on low level visual features. Score was assigned to each frame based on its importance. This work is more of a user centric one allowing users to select variable number of key frames for video summary.

A diversity based video summarization technique is introduced (Chen et al., 2015) based on dictionary learning while keeping in mind about relationships among different video samples which was the serious problem in dictionary learning methods. Geometrical distributions were drawn employing a strategy based on graph to address this problem and impressive results were obtained. Another diversity aware work (Panda et al., 2017) involves unsupervised framework to efficiently summarize videos using a minimization algorithm that minimizes overall loss function. A new Tour20 dataset was introduced to perform experiments and state of art results were achieved. In (Sharghi et al., 2016), a video summarization method is proposed that works by selecting key frames from a video with the help of user query. Datasets containing annotated videos were used in the experiment and state of art results were achieved. This method is capable of handling lengthy videos and also online streaming videos. Subjectiveness in video summarization is carefully analyzed (Sharghi et al., 2017) and some solutions were suggested. Determinant processes and memory networks were used in the summarizer to get better representative and diverse summary of original video. This method outperformed two popular methods of video summarization. One video was used for testing, one for validation and two videos were used for training making total of four experiment rounds.

A query based video summarization method is introduced (Ji et al., 2017) that searches user preference content based on the query. This is based on sparse coding framework. Authors also introduced a public dataset containing total of 1000 videos that are annotated as well. Extensive experiments were manipulated on this public dataset and state of art results proved the effectiveness of this proposed method. Visual and textual embedding (Vasudevan et al., 2017) is employed so that better representative summary can be obtained through a user query. A new dataset related to selection of thumbnail was also used in this paper consisting of labeled videos. This approach proved that more advanced text model with better training goal and a better modelling quality gives highest performance gains. A multimodal approach for video summarization of cricket match (Bhalla et al., 2019) is introduced that notices important events in cricket match and then generates highlights that are good reflection of original video. In order to detect important things in ground like wickets and boundaries, techniques like optical character recognition were used. Events are then joined together to generate highlights for the entire cricket match. Accuracy of 89 percent was achieved by using this method. In (Ma et al., 2019), another video summarization method is proposed that is based on sparse dictionary selection that works by supposing relationship of linearity between frames. A non linear model is contructed and then video is mapped to high dimensional feature space with the help of kernel to convert non linearity into linearity. Furthermore, two greedy algorithms with strategy of back tracking are suggested to manipulate model. Experiments were conducted on two datasets that are SumMe and TVSum and it was concluded that summaries produced were better than summaries of other state of art video summarization algorithms.

3 PROPOSED APPROACH

The proposed approach is shown in Fig. 2. Videos are converted into frames and then these frames are fed to residual network. Residual network in our approach is ResNet-152 having total of 152 layers. As it is quite deep architecture so it will extract all deep and important features from videos that could be very important in generating better representative summaries. The extraction of these features is in 2048 dimensions for each video so that we do not loss important features that could be important for generation of better representative summaries. After extraction of these features, they will be further given to two layered gated recurrent unit which will generate hidden states from these features. These hidden states will be of two types i-e forward hidden states and backward hidden states. Sigmoid layer is mounted at the end of gated recurrent unit that will take these hidden states as input to finally generate probability scores for all the frames. Now each frame will be associated with a probability score. Frames with higher probability scores will be selected for further evaluation and will be fed to reinforcement learning reward function. Reward function is the sum of two other reward functions i-e diversity reward function and representative reward function. These two reward functions will evaluate these frames based on their diversity and representativeness. Diversity reward function will make sure that frames selected should be diverse enough to better represent all different parts of video. Representative reward function will check that frames selected are either better representing original video or not. Main goal is to maximize the sum of these two reward functions that will automatically generate better summary. In order to accomplish this task, a policy is learnt through adam's optimization till the reward function is maximized. The maximization of reward function will generate better summaries. This is

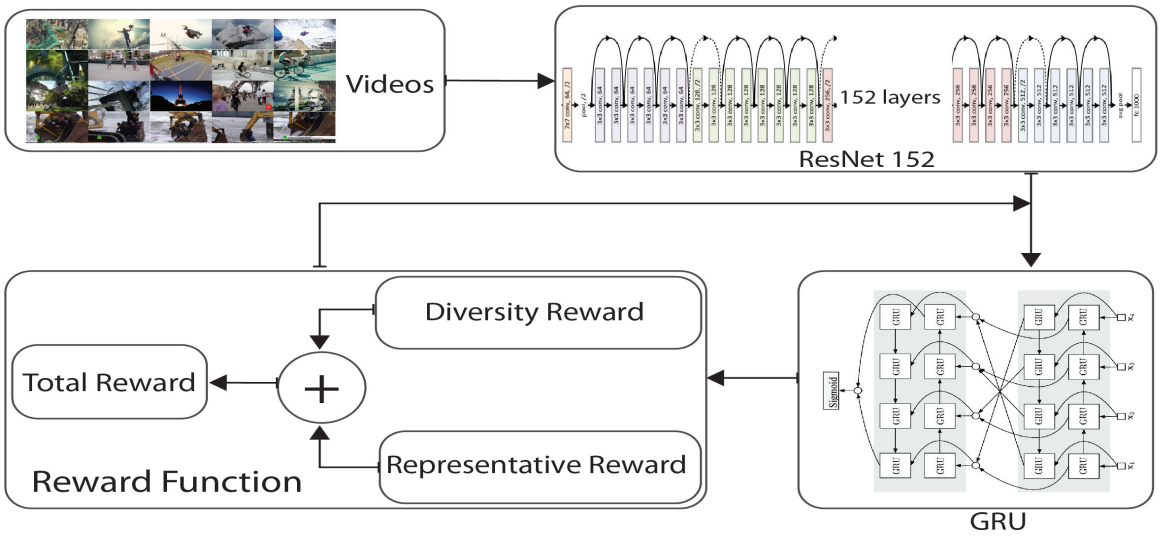

Figure 2: Our proposed approach.

working of our approach. Now we will separately discuss all important components of this approach which are ResNet, Gated recurrent unit and reinforcement learning reward function.

3.1 ResNet-152

Residual network (Nguyen et al., 2018), which is also known as ResNet, belong to deep neural network family having same structure but with different depths. It avoids degradation in neural networks through a unit known as residual learning unit. It is well known architecture to extract deep features from images. The structure of residual learning unit in ResNet is a feed forward network which is capable of adding new inputs in the network and thus generating new outputs through a shortcut connection. The main advantage of ResNet is it produces better results with higher accuracy with out increasing complexity. Fig. 3. shows basic architecture of ResNet-152.

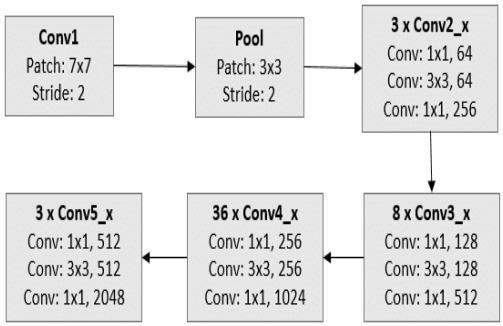

Figure 3: The basic architecture of ResNet-152 (Nguyen et al., 2018).

So ResNet will take all the frames for every single video as an input and generate feature matrices for each video in 2048 dimensions. As it has network depth of 152 so it will make sure that all deep and important features have been extracted from each video. Extraction of important features from each video is very necessary because if we miss important features from videos then we can not generate better hidden sates through gated recurrent unit as important features were not fed to it. So all further work will be useless. That is why ResNet-152 serves the purpose here as it is quite deeper architecture.

3.2 Gated Recurrent Unit (GRU)

Gated recurrent unit is a sequential type of encoder. The advantage of gated recurrent unit over long short term memory is it has less parameters than long short term memory and this results in less training time with a very low risk of overfitting (Chung et al., 2014). Its performance has also bypassed performance of LSTM in several cases (Elsayed et al., 2018) including our case too. In our approach, first we checked performance of LSTM on all videos and then we replaced it with GRU. We discovered that though LSTM gave good results on fewer videos where GRU did not but overall GRU was performing much better on most of the videos with less training time.

So GRU gets all the important features that were previously extracted by ResNet-152 and generates hidden states that are forward hidden states and backward hidden states. GRU has a sigmoid layer at the end where all these hidden states are transformed into different probability scores. Actions will be performed to select frames with higher probability scores

and these selected frames will further be given to reinforcement learning reward function.

3.3 Reward Function

The purpose of reinforcement learning agent is to maximize its reward function after learning from the environment. Interaction of reward function with video summarization system is also shown in Fig. 1. Reward function here is basically combination of two other reward functions which are diversity reward function and representative reward function as can be seen in Fig. 2.

3.3.1 Diversity Reward Function

It is responsible for finding key frames from the selected ones. It keeps on comparing every frame with every other to find the most dissimilar frames. Suppose $\mathcal{K}$ is the set of selected frames which are given to diversity reward function and 'b' denotes each single frame then this reward function is given by the equation :

$$R_{\text{div}} = \frac{1}{|\mathcal{K}|(|\mathcal{K}|-1)} \sum_{t\in\mathcal{K}} \sum_{\substack{t'\in\mathcal{K} \\ t'\neq t}} (1 - \frac{b_t^T b_{t'}}{\|b_t\|_2 \|b_{t'}\|_2}) \quad (1)$$

3.3.2 Representative Reward Function

It will make sure that selected frames are good representation of original video and this is done by selecting those frames that are nearer to clusters in the whole feature space. This reward function is given by the equation:

$$R_{\text{rep}} = \exp\left(-\frac{1}{T}\sum_{t=1}^{T} \min_{t'\in\mathcal{K}} \|b_t - b_{t'}\|_2\right) \quad (2)$$

Total reward function is sum of both of these reward functions :

$$R(total) = R_{\text{div}} + R_{\text{rep}} \quad (3)$$

3.4 Learning Descend Policy Gradient

Agent should learn a policy for maximizing rewards. If *'p'* is the probability score and θ is the policy function parameter then objective function can be given by:

$$L(\theta) = \mathbb{E}_{p(a|\pi)}\left[R_{div} + R_{rep}\right] \quad (4)$$

In order to find gradient, we need to find derivative of objective function. So derivative will be:

$$\nabla_\theta L(\theta) \approx \frac{1}{N}\sum_{k=1}^{N}\sum_{t=1}^{T}\left(R_{div}^k + R_{rep}^k - m\right)\nabla_\theta \log\pi(a_t|h_t;\theta) \quad (5)$$

where π is the policy function, *'m'* is the rewards moving average, h_t are the hidden states generated by GRU and a_t are the actions taken by model.

Now optimization is performed through Adam's algorithm (Kingma and Ba, 2014). It will aid agent to take those steps more and more that can lead to better reward and avoid those steps that can lead to low reward value.

4 EXPERIMENTS AND RESULTS

4.1 Dataset

SumMe dataset is used for experiments that contains total 25 videos from various topics such as sports, holidays and various other events. Each video has length of approximately 1 to 6 minutes and each is annotated by 15 to 18 persons which means there are multiple ground truth summaries available for each video.

4.2 Evaluation Metric

In order to assess automatic summaries that are generated by our summarization network, F- measure is used as evaluation metric to find the difference between ground truth summaries and automatic summaries. F-measure is harmonic mean of precision and recall so here it is used to evaluate on testing dataset.

4.3 Evaluation Settings

Leave one out cross validation is used for evaluation of our system. Model is trained with $N-1$ videos and evaluated on remaining one video. This process is repeated N times with different training sets of size $N-1$.

4.4 Implementation Details

We implemented our approach using pytorch on GTX 1050 Ti GPU. Features were extracted through ResNet in 2048 dimensions and were given to model with 256 hidden units having two RNN layers. We ran training for 70 epochs. Total 30 steps were set to decay the learning rate.

4.5 Comparison with DR-DSN

We replicated DR-DSN (Zhou et al., 2018a) and ran it on each video of SumMe dataset. We got same results for DR-DSN as given in paper (Zhou et al., 2018a).

Table 1: Comparison with DR-DSN (Zhou et al., 2018a) on SumMe dataset.

Video	Video Name	Camera Type	Video Length	Frames	F-measure	
					DR-DSN	Our method
Video 1	Air Force One	static	2 min 60 sec	4494	60	60
Video 2	Base Jumping	egocentric	2 min 39 sec	4729	22.3	28.7
Video 3	Bearpark Climbing	moving	2 min 14 sec	3341	35.2	54.1
Video 4	Bike Polo	egocentric	1 min 43 sec	3064	54.4	54.9
Video 5	Bus in Rock Tunnel	moving	2 min 51 sec	5131	29.7	36.2
Video 6	Car RailCrossing	moving	2 min 49 sec	5075	21	21
Video 7	Cockpit Landing	moving	5 min 2 sec	9046	29.2	29.2
Video 8	Cooking	moving	1 min 27 sec	1286	49.4	50.1
Video 9	Eiffel Tower	moving	3 min 20 sec	4971	30.9	32.6
Video 10	Excavator River Cross	moving	6 min 29 sec	9721	26.5	31.4
Video 11	Fire Domino	static	0 min 55 sec	1612	60.2	60.2
Video 12	Jumps	moving	0 min 39 sec	950	0	40.4
Video 13	Kids Playing in Leaves	moving	1 min 46 sec	3187	50.2	22.9
Video 14	Notre Dam	moving	3 min 12 sec	4608	44.6	37.4
Video 15	Paintball	static	4 min 16 sec	6096	55.1	55.2
Video 16	Playing on Water Slide	moving	1 min 42 sec	3065	34.9	34.9
Video 17	Saving Dolphines	moving	3 min 43 sec	6683	32.7	43
Video 18	Scuba	egocentric	1 min 14 sec	2221	67.5	67.5
Video 19	St. Marten Landing	moving	1 min 10 sec	1751	60.8	41.7
Video 20	Statue of Liberty	moving	2 min 36 sec	3863	61.4	43.4
Video 21	Uncut Evening Flight	moving	5 min 23 sec	9672	17.4	27.8
Video 22	Valparaiso Downhill	egocentric	2 min 53 sec	5178	39	44.1
Video 23	Car Over Camera	static	2 min 26 sec	4382	66.7	66.7
Video 24	Paluma Jump	moving	1 min 26 sec	2574	29.5	30.6
Video 25	Playing Ball	moving	1 min 44 sec	3120	55.3	77.7
RESULT (Average F-measure)					41.4	43.7

Then we ran our approach on each video and compared results of both approaches. Results are shown in Table 1. It can be clearly seen that our approach performed 2.3 percent better than DR-DSN approach. Even there is one video in the dataset i-e Video 12 where DR-DSN failed completely as given in Table 1 while our approach gave reasonable F-measure of 40.4 on this particular video. Our approach performed better on most of the videos and hence overall, our proposed method is better than DR-DSN.

Table 2: Comparison with other approaches.

Method	Dataset	F-measure
CSUV	SumMe	23
Uniform Sampling	SumMe	29.3
Vsumm	SumMe	33.7
Dictionary Selection	SumMe	37.8
GAN dpp	SumMe	39.1
DR-DSN	SumMe	41.4
Ours	SumMe	43.7

4.6 Comparison with Other State of Art Approaches

Several unsupervised video summarization approaches have been proposed till now as can be seen in Table 2. We replicated CSUV approach (Gygli et al., 2014) on SumMe dataset that basically works through superframe segmentation. Results achieved from our model were far better than this approach. Uniform sampling (Jadon and Jasim, 2019) is also popular unsupervised technique for video summarization that works using key frame extraction. It tries to extract important parts of video but results showed that it is not effective summarization approach. Clustering has always been very common in video summarization and Vsumm approach does the same through k means. It makes clusters of similar type of frames. Frames will be part of those clusters where their distance from centroid is minimal.So this way k-means algorithm works for summarizing videos but it is not good enough as our proposed approach.

Dictionary selection video summarization approach (Ma et al., 2019) works by assuming relationship of linearity between frames. A non linear model is contructed and then video is mapped to high dimensional feature space with the help of kernel to convert non linearity into linearity. Furthermore, two greedy algorithms with strategy of back tracking were suggested to manipulate model in dicitonary selection but the final results were not better than our proposed approach. GAN dpp (Mahasseni et al., 2017) summarization approach works through discriminator and a summarizer. Long short term memory acts as summarizer as well as discriminator. As a discriminator, it distinguishes between summary generated by the system and original summary. This approach is based on generative adversarial network but it is not as effective as our approach. Comparison of all of these approaches with our approach on SumMe dataset can be seen in Table 2 where it can be clearly seen that our approach is leading all other approaches.

5 CONCLUSIONS

In this paper, we proposed improved video summarization method that outperformed several state of art methods. Residual network ResNet-152 was employed with gated recurrent unit having two RNN layers. We performed detailed comparison of our approach with DR-DSN by providing results on each and every video in SumMe dataset. Furthermore, we compared overall average F-score of our approach with average F-score of several other state of art video summarization methods and concluded the fact that our method is best in terms of generating better representative summaries of original videos.

ACKOWLEDGEMENT

We thank Kaiyang Zhou for detailed discussion about his paper (Zhou et al., 2018a) . This work is supported by Video Surveillance Lab, Karachi, Pakistan affiliated from National Center of Big data and Cloud Computing, Pakistan.

REFERENCES

Bhalla, A., Ahuja, A., Pant, P., and Mittal, A. (2019). A multimodal approach for automatic cricket video summarization. In *2019 6th International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 146–150. IEEE.

Chen, S.-C., Lin, K., Lin, S.-Y., Chen, K.-W., Lin, C.-W., Chen, C.-S., and Hung, Y.-P. (2013). Target-driven video summarization in a camera network. In *2013 IEEE International Conference on Image Processing*, pages 3577–3581. IEEE.

Chen, X., Li, X., and Lu, X. (2015). Representative and diverse video summarization. In *2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, pages 142–146. IEEE.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Dimou, A., Matsiki, D., Axenopoulos, A., and Daras, P. (2015). A user-centric approach for event-driven summarization of surveillance videos.

Elsayed, N., Maida, A. S., and Bayoumi, M. (2018). Deep gated recurrent and convolutional network hybrid model for univariate time series classification. *arXiv preprint arXiv:1812.07683*.

Gygli, M., Grabner, H., Riemenschneider, H., and Van Gool, L. (2014). Creating summaries from user videos. In *European conference on computer vision*, pages 505–520. Springer.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jadon, S. and Jasim, M. (2019). Video summarization using keyframe extraction and video skimming. *arXiv preprint arXiv:1910.04792*.

Ji, Z., Ma, Y., Pang, Y., and Li, X. (2017). Query-aware sparse coding for multi-video summarization. *arXiv preprint arXiv:1707.04021*.

Khan, R. U., Zhang, X., and Kumar, R. (2019). Analysis of resnet and googlenet models for malware detection. *Journal of Computer Virology and Hacking Techniques*, 15(1):29–37.

Khan, R. U., Zhang, X., Kumar, R., and Aboagye, E. O. (2018). Evaluating the performance of resnet model based on image recognition. In *Proceedings of the 2018 International Conference on Computing and Artificial Intelligence*, pages 86–90.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lai, P. K., Décombas, M., Moutet, K., and Laganière, R. (2016). Video summarization of surveillance cameras. In *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 286–294. IEEE.

Lei, J., Luan, Q., Song, X., Liu, X., Tao, D., and Song, M. (2018). Action parsing-driven video summarization based on reinforcement learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(7):2126–2137.

Ma, M., Mei, S., Wan, S., Wang, Z., and Feng, D. (2019).

Video summarization via nonlinear sparse dictionary selection. *IEEE Access*, 7:11763–11774.

Mahasseni, B., Lam, M., and Todorovic, S. (2017). Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 202–211.

Masumitsu, K. and Echigo, T. (2000). Video summarization using reinforcement learning in eigenspace. In *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*, volume 2, pages 267–270. IEEE.

Nguyen, L. D., Lin, D., Lin, Z., and Cao, J. (2018). Deep cnns for microscopic image classification by exploiting transfer learning and feature concatenation. In *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE.

Panda, R., Mithun, N. C., and Roy-Chowdhury, A. K. (2017). Diversity-aware multi-video summarization. *IEEE Transactions on Image Processing*, 26(10):4712–4724.

Sharghi, A., Gong, B., and Shah, M. (2016). Query-focused extractive video summarization. In *European Conference on Computer Vision*, pages 3–19. Springer.

Sharghi, A., Laurel, J. S., and Gong, B. (2017). Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4788–4797.

Vasudevan, A. B., Gygli, M., Volokitin, A., and Van Gool, L. (2017). Query-adaptive video summarization via quality-aware relevance estimation. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 582–590.

Yang, Y., Dadgostar, F., Sanderson, C., and Lovell, B. C. (2011). Summarisation of surveillance videos by key-frame selection. In *2011 Fifth ACM/IEEE International Conference on Distributed Smart Cameras*, pages 1–6. IEEE.

Yun, S., Yun, K., Kim, S. W., Yoo, Y., and Jeong, J. (2014). Visual surveillance briefing system: Event-based video retrieval and summarization. In *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 204–209. IEEE.

Zhang, Y., Kampffmeyer, M., Zhao, X., and Tan, M. (2019). Deep reinforcement learning for query-conditioned video summarization. *Applied Sciences*, 9(4):750.

Zhong, Z., Li, J., Ma, L., Jiang, H., and Zhao, H. (2017). Deep residual networks for hyperspectral image classification. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 1824–1827. IEEE.

Zhou, K., Qiao, Y., and Xiang, T. (2018a). Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhou, K., Xiang, T., and Cavallaro, A. (2018b). Video summarisation by classification with deep reinforcement learning. *arXiv preprint arXiv:1807.03089*.