Extractive Video Summarization

Mohit Agarwal
mohit15060@iiitd.ac.in
IIIT-Delhi

Sanchit Sinha
IIIT-Delhi
sanchit15083@iiitd.ac.in

Dhruva Sahrawat
IIIT-Delhi
dhuva15026@iiitd.ac.in

Aditya Adhikary
IIIT-Delhi
aditya15007@iiitd.ac.in

KEYWORDS

Video Summary, Shots, Frames, Deep Learning, Supervised Learning.

ACM Reference Format:
Mohit Agarwal, Sanchit Sinha, Dhruva Sahrawat, and Aditya Adhikary. 2019. Extractive Video Summarization. Multimedia Computing and Applications (MCA), IIIT, Delhi, India, 6 pages.

1 PROBLEM STATEMENT

Given a set of videos, what is the best possible way to extract a summary of each video from its frames which is both meaningful i.e it plays out without any disconnectedness, and stores all the crucial information pertinent to the video topic?

2 INTRODUCTION AND MOTIVATION

The number of online videos on video sharing websites such as Youtube, Dailymotion, etc. and social networking websites such as Facebook, Twitter, Reddit, etc. have experienced explosive growth in the past decade. As a statistic, 300 hours of video are uploaded to YouTube every minute - a figure which is increasing each year. As these trends show no signs of reversing, an efficient method of browsing videos is the need of the hour.

On the flip side, online users are bamboozled 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.
With low attention spans and abundance of videos, it becomes infeasible for an average person to browse through each and every video in its entirety. Hence, an effective technique which condenses the highlights or important points of a video in a short clip is the way forward. Keeping all of this in mind, video summarization is proposed.
Any such system has a potential to be an integral part of a variety of domains. A few of the many use-cases where video summarixzation might be helpful are as follows:

- Highlights of sporting events (example goals, red-cards, etc in football)
- Long term security monitoring and surveillance from Closed Circuit Television cameras
- Condensing political speeches or meetings for telecast over online or television mediums

3 LITERATURE SURVEY

There is plethora of literature on methods for video summarization. The state-of-the-art paper we are referring to uses a memory augmented neural network [4]. Their code is not publicly available.

3.1 Unsupervised approaches

- Among the most naive approaches used for many years, clustering similar shots using hand crafted features has been used, such as in [7]. These approaches have little or no holistic understanding as they do not consider any context.
- Some methods like [3] used most frequent co-occurring shots from across videos in a dataset. This is non-intuitive and not similar to a human understanding in any way, leading to unintelligible results.
- Making relationship graphs (like Facebook) and selecting those with high centrality. However, this was only used on movie datasets with no guarantee of results [13].
- LSTMs and RNNs, such as in [15] model sequential attention and suggest a more holistic idea. However, using memory cells as-is shows that they are not robust enough to hold information across large stretches of video and hence fall behind.
- Reinforcement Learning has been used to model the selection of a keyframe as a sequential process and allot a reward for selecting the correct frame [17]. However, video summarization is intuitively not a sequential process as there are no contextual cues while selecting the next keyframe, and a global attention is required.
- Recent GAN based approaches such as [10] use a Variational Autoencoder for selecting sparse frames (generator), and an RNN classifier for distinguishing original and summarized videos (discriminator). This has been used as one of the baselines. Another recent paper [5] uses an adversarial training framework for *semi-supervised* video summarization, and achieves results comparable to the state-of-the-art. Their discriminator is faced with the task of judging whether a summarized video is *from the summarization dataset* or is *generated* by the summarizer.

Thus, recent unsupervised approaches have so far been producing results comparable to the state-of-the-art.

3.2 Supervised approaches

- [11] used SVM for predicting the importance score of video shots, joining shots sequentially with higher scores. However, there was hardly any logic to this approach.

- Most early work such as [9] were dependent on hand-crafted features for video shot representation and were guided by manual labelling, which was time consuming and unscalable.
- CNN features were used in [8] to assign an importance score by comparing the most frequent (matching) shots, but most such methods was based on assumptions that summaries of similarly-structured videos would be similar.
- Methods such as LSTM were combined with DPP (Determinantal point process, a kind of stochastic point process) in [16] to model the variable-range temporal dependency among video frames, accounting for the sequential structure as well as long-term dependencies.

Overall, supervised generally perform better than unsupervised but are non-scalable.

4 LIMITATIONS OF EXISTING WORK

The broad limitations in existing literature are mentioned here.

- As mentioned in the state-of-the-art reference, shot feature representation needs to be improved. Hand-crafted or CNN features have not so far been incorporated well in the model (it simply uses average pooled deep convolutional features using a pre-trained image classification model for the same). For memory efficiency purposes, taking the average is feasible, but loses motion information.
- The summarizer only uses visual information, as is the case with most other techniques mentioned in literature. The long-term goal would be to associate multiple modalities such as subtitles and sound for better summary.
- The reference uses cross (inter-dataset) training to prove their model has robust cross-dataset training ability, but applied it on only 2 datasets. To validate these findings, cross-training must be utilized on other datasets as well. Their model also did not mention what was considered in the difference of the properties of datasets while training, such as labeling of categories in one and not the other. For a more holistic approach, the model should be able to differentiate and utilize or discard such properties as per requirement.
- The paper limits the the length of the final summary paper to 15% of the actual video, which is set as the knapsack capacity. This is a simplifying assumption. Also, other evaluation metrics such as mutual information and joint entropy [1] have not been used for a well-rounded assessment.
- The shot segmentation approach used is based on an adaptive thresholding shot boundary detection algorithm [14], but it has not been mentioned whether the threshold has been varied for finding the best cuts, and could lead to a lot of false positives after training.

5 DATASETS

These are the two publicly available datasets used in the reference paper:

- **SumMe** [6] - This dataset was published and released in European Conference on Computer Vision, 2014 by researchers from ETH Zurich. It contains 25 videos and there are no specific categories. The video length typically varies from 1 to 6 minutes. Frame level importance scores have been provided by multiple human annotators. Each video contains at least 15 different human annotations and in total, there are 390 such annotations.
- **TVSum** [12] - This dataset was made available by researchers from Yahoo labs and was published in Computer Vision and Pattern Recognition Conference, 2015. It 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, so a total of 1000 annotations. It was a crowd-sourced annotation and was done by Amazon Mechanical Turk.

Additionally, we will attempt to use the following two datasets as they have been used across multiple other papers recently:

- **YouTube Dataset** [18] - This dataset has 50 videos from YouTube distributed across several genres like cartoons, news, commercials etc. Duration of these videos ranges from 1 to 10 minutes. It consists of 250 video summaries created manually by 50 users. Each user did 5 summaries and similarly each video has corresponding 5 summaries created by 5 different users.
- **LoL Dataset** [2] - This dataset was introduced in Conference on Empirical Methods in Natural Language Processing, 2017. It has 218 videos consisting of match highlights of League of Legends from North America League of Legends Championship Series (NALCS). Typical video duration is 30 to 50 minutes. Also, unlike the other datasets, LoL provides entire summarized videos instead of only providing the keyframes.

In addition to the above, we might randomly collect a few sample videos, produce video summaries using our proposed solution for them and then manually inspect to check the performance of our proposed approach.

6 STATE OF THE ART DESIGN

This has been derived from the work done by Feng *et al* [4].

6.1 Architecture

The model consists of an encoder and a global attention Module. CNN model pretrained on Imagenet is used as the encoder.

6.1.1 Encoder. Each frame from a shot is fed as the input to the encoder to generated the encoded features. Average of the encoded features from all the frames in a Video Shot is used as the final feature representation of the Video shot.

6.1.2 Global Attention Module. The global attention module is applied on the encoded feature of the video shot. It is explained in detail in the following points:

- The encoded feature of the kth shot is of dimension v and is represented as x_k.
- For every shot, its corresponding encoded feature x_k is converted to input memory feature a_k using embedding matrix A (of size $d * v$). Correspondingly output memory feature b_k is generated using embedding matrix B (of size $d * v$).
- The internal state u_i is generated from its corresponding encoded feature x_i using embedding matrix U.

- The match between shot k and shot i is computed through finding probability vector p_i^k,
$$p_i^k = Softmax(u_i^T a_k)$$
.
- Memory output o_i corresponding to shot i is is sum of b_k weighted by the probability vector p_i^k.
$$o_i = \sum_k p_i^k b_k$$
- Modified internal state vector $u_i^{'}$ is generated through:
$$u_i^{'} = u_i \odot o_i$$
- the whole procedure is repeated for the required no of hops by taking modified internal state vector $u_i^{'}$ as the new state vector and repeating the above steps.

6.1.3 Importance Score prediction. The final internal state vector $u_i^{'}$ is then fed to a fully connected layer D to generate the importance score s_i for each video shot.

$$s_i = W_D.u_i^{'} + b_D$$

Linear L2 regression loss is applied on the predicted importance score and the ground truth importance score to compute the loss. The loss is then back-propagated for end-to-end learning of the entire network including all the three embedding matrix A, B and U.

The overall architecture can be seen in the Figure 1.
Shot segmentation is done using adaptive thresholding based shot boundary detection. Here, instead of the handcrafted features, they have used CNN features of SqueezeNet. Cosine similarity is used as the measure to get shot boundaries.
They have mainly compared against the following two baselines:

- Multi-Layer Perceptron (MLP) - The embedded video feature is directly used to predict importance scores which is then used to generate summaries.
- Long Short-Term Memory (LSTM) - LSTM has been known to perform well on sequential data. Since videos are also sequential, they have used LSTM as a baseline to generate importance scores to get the summary.

6.2 Experiments

They have performed three types of experiments:

- **Intra-Dataset** - This is independently done on the two datasets. Required train-test-validation splits are created and the architecture is tested upon them.
- **Inter-Dataset** - Usually, the video summarization algorithms are heavily dependent on the training dataset used and do not perform well on dataset with different variations. This particular experiment can be used to test the generalization ability of an algorithm. In this, the proposed approach is trained on one dataset and tested on the other. This cross of train-test helps us analyse the performance better.
- **Noisy Videos** - This experiment is performed on top of the same protocol used by the inter-dataset. In addition to that, they have randomly inserted 25% noisy shots in raw videos. Training is done using the noise-less video of one dataset and testing is performed on the noisy videos of the other dataset. This type of experiment is more real-life and further challenges the generalizability and robustness of an algorithm.

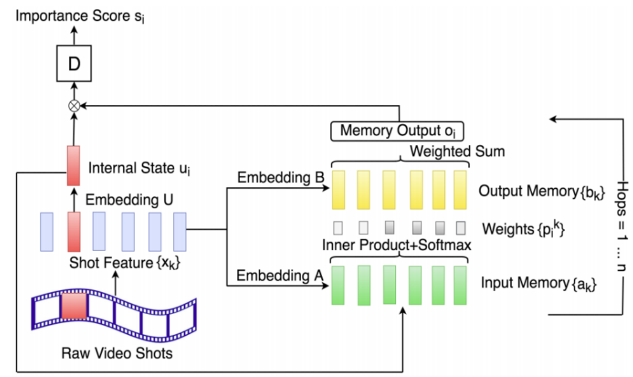

Figure 1: Architecture of the current state of the art design in the reference paper using a memory augmented network. (Image taken from the reference paper.)

The above experiments can be summarized and seen in the Figure 2.

6.3 Results of current SOTA

Results obtained using the current state of the art explained above on the experiments performed by them can be seen in the Figures 3, 4 and 5.

7 PROPOSED SOLUTIONS

- Using better encoded video shot representations. In [4], average of all the deep feature representation of the images in a video shot is taken as the encoded video shot representation. A lot of information including motion information etc. is lost. To tackle this instead of using average pooling we can use weighted averaging of each frame or some other form of pooling.
- We can also use local attention. One way local attention can be incorporated is that we can use the attention scores to weigh each frame for weighted average pooling to generate the encoded video shot feature representation.
- We can also make use of Motion information which is lost in each shot of the video. We can use motion features which are typically used for tracking, action recognition etc. They can be concatenated with other features to generate the encoded Video Shot feature representation.
- We can also extract audio features from the video and use that in addition to the original features for prediction of importance score. In a video, both audio and image features are important. Audio can give extremely important cues for scene and shot importance prediction. Thus, we can try to work using both the features as well.
- Using pretrained encoders which are trained on other tasks like activity recognition or Video captioning for example

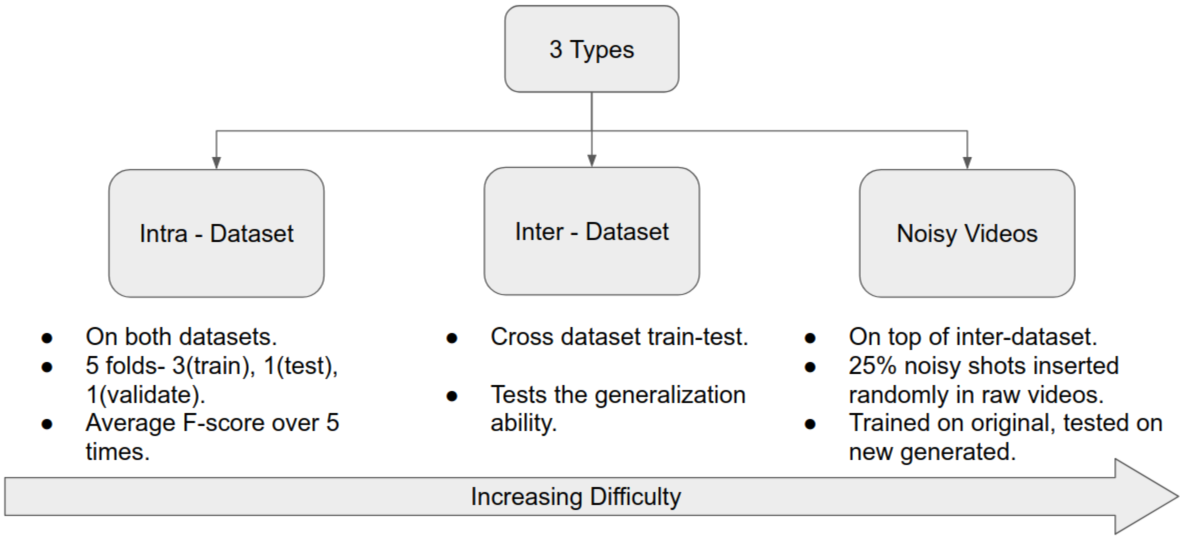

Figure 2: Types of experiments performed in the reference paper.

Methods	hops	SumMe	TVSum
DPP-LSTM (canonical)	-	38.6	54.7
Summary-transfer	-	40.9	-
GAN-based (SUM-GAN$_{sup}$)	-	41.7	56.3
RL-based (DR-DSN$_{sup}$)	-	42.1	58.1
Temporal-tesselation (Unsupervised)	-	41.4	64.1
MAVS	1	39.8	67.0
MAVS	2	42.5	67.2
MAVS	3	42.6	67.3
MAVS	4	43.1	67.5
MAVS	5	42.3	67.3
MAVS	6	40.3	66.8

Figure 3: Results of intra-dataset experiment from the reference paper.

Methods	hops	SumMe2TVSum	TVSum2SumMe
MLP1	-	64.5	36.5
MLP2	-	62.1	35.5
MLP3	-	61.7	36.4
MLP4	-	61.4	35.8
LSTM1	-	62.8	36.6
LSTM2	-	61.5	36.9
LSTM3	-	60.1	34.6
MAVS	1	63.2	37.3
MAVS	2	65.5	38.8
MAVS	3	66.2	39.8
MAVS	4	66.4	41.7
MAVS	5	66.3	41.2
MAVS	6	65.8	40.2

Figure 4: Results of inter-dataset experiment from the reference paper.

CNN+LSTM architecture to generate better encoded feature representations of video shots and then apply the global Attention Mechanism on it to generate the final Importance Score.

- Instead of using Pretrained encoders we can apply the global attention mechanism to the encoded representation generated using GANs.
- Better analysis of global attention mechanism on video summarizer, including ablation study, application of it on other tasks and checking its performance on random videos for better understanding of it.
- The threshold used in the adaptive thresholding algorithm for shot segmentation [14] could be fine-tuned to avoid the possibility of false positives after training.
- In the reference paper [4] experiments have been performed only on SumMe and TVSum datasets, we plan to perform the experiments on Youtube Dataset and LoL Dataset also which is explained in more details in the Dataset Section.
- As explained in the Experiments Section, more experiments on Inter-dataset, Intra-Dataset, Cross-Dataset and Noisy Videos will be performed to better analyse the performance of the model.
- The final summary on which the evaluation is carried out is generated using the confidence score predicted by the model. It is only a fraction of the whole video (in this case 15%). None of this is incoporated in the learning objective, there is only a regression loss on the importance score of a single shot. We can probably address this and improve the

Methods	Hops	SumMe2TVSum	TVSum2SumMe
MLP1	-	45.1	18.8
MLP2	-	47.6	19.7
MLP3	-	47.9	18.4
MLP4	-	46.1	18.8
LSTM1	-	40.7	24.0
LSTM2	-	41.7	24.6
LSTM3	-	41.1	24.2
MAVS	1	50.8	19.1
MAVS	2	57.2	20.7
MAVS	3	57.9	22.6
MAVS	4	59.4	25.9
MAVS	5	58.3	25.3
MAVS	6	58.5	25.4

Figure 5: Results of noisy-dataset experiment from the reference paper.

performance of the model by using some sort of sequence loss.

8 PLANS AND TIMELINE

Timeline of the project is split into three parts based on the multifold evaluation stages spreading across the whole semester.

- Intermediate Report 1 (Scheduled submission on *22 February 2019*) - Dataset review and preprocessing, Baseline (state of the art) reproduction.
- Intermediate Report 2 (Scheduled submission on *22 March 2019*) - Implementation and reproduction of other baselines (MLP, LSTM etc.), experiments with our proposed solutions.
- Final Report (Scheduled submission on *12 April 2019*) - Finalization and implementation of our final proposed model, experiments and analysis on different datasets using different protocols. The plan is to make our study and analysis as exhaustive as possible.

In general, the whole work will be divided equally among all members. Rough distribution of work can be as followed:

- Aditya Adhikary - Data preprocessing, Baseline 1 (MLP) reproduction, modifications to current state of the art, combining analysis from all to proposed a final proposed model, intra-dataset analysis.
- Dhruva Sahrawat - Shot segmentation, Baseline 2 (LSTM) reproduction, modifications to shot segmentation, combining analysis from all to proposed a final proposed model, noisy video analysis.
- Mohit Agarwal - Features production, Baseline 3 (GAN) reproduction, experimenting with new proposed, combining analysis from all to proposed a final proposed model, inter-video analysis.
- Sanchit Sinha - State of the art reproduction, experimenting with other newly proposed, combining analysis from all to proposed a final proposed model, metric computations.

9 CONCLUSION

Video summarization as explained has a lot of applications. It can be used in diverse types of videos like sports, entertainment, educational. This has a potentital to aid both the users and creators. Users can directly view the summaries and creators wouldn't have to edit and make separate summary videos. The reference paper [4] tries to follow the holistic understanding of a video to generate its summary, and proposes an external memory-aided neural network for predicting the weightage of shots, which lays emphasis on the global attention of the video. This is different from the models previously used which looked at only the local span in videos. This idea of global attention is derived from the fact that humans also look at the whole context of the video to generate summaries. It can be used and exploited further to make a better state of the art model. We propose to implement the same and attempt the improvements mentioned above. We might also need to look into an altogether different model if some of the above mentioned improvements don't work out well with this.

10 ACKNOWLEDGMENTS

We would like to thank Dr Rajiv Ratn Shah, IIIT-Delhi for providing us the opportunity and resources to work on this project. We would also like to extend our gratitude towards the TAs Vani Agarwal and Saurabh Gupta.

REFERENCES

[1] Z. ˇCerneková, C. Nikou, and I. Pitas. 2002. Entropy Metrics Used for Video Summarization. In *Proceedings of the 18th Spring Conference on Computer Graphics (SCCG '02)*. ACM, New York, NY, USA, 73–82. https://doi.org/10.1145/584458.584471

[2] Mohit Bansal Cheng-Yang Fu, Joon Lee and Alexander C. Berg. 2017. Video Highlight Prediction Using Audience Chat Reactions. In *EMNLP*.

[3] Wen-Sheng Chu, Yale Song, and Alejandro Jaimes. 2015. Video Co-summarization: Video Summarization by Visual Co-occurrence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[4] Litong Feng, Ziyin Li, Zhanghui Kuang, and Wei Zhang. 2018. Extractive Video Summarizer with Memory Augmented Neural Networks. In *Proceedings of the 26th ACM International Conference on Multimedia (MM '18)*. ACM, New York, NY, USA, 976–983. https://doi.org/10.1145/3240508.3240651

[5] Tsu-Jui Fu, Shao-Heng Tai, and Hwann-Tzong Chen. 2019. Attentive and Adversarial Learning for Video Summarization. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

[6] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating Summaries from User Videos. In *ECCV*.

[7] Youssef Hadi, Fedwa Essannouni, and Rachid Oulad Haj Thami. 2006. Video Summarization by K-medoid Clustering. In *Proceedings of the 2006 ACM Symposium on Applied Computing (SAC '06)*. ACM, New York, NY, USA, 1400–1401. https://doi.org/10.1145/1141277.1141601

[8] Dotan Kaufman, Gil Levi, Tal Hassner, and Lior Wolf. 2016. Temporal Tessellation for Video Annotation and Summarization. *CoRR* abs/1612.06950 (2016). arXiv:1612.06950 http://arxiv.org/abs/1612.06950

[9] David G. Lowe. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60, 2 (01 Nov 2004), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

[10] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. 2017. Unsupervised Video Summarization with Adversarial LSTM Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 2982–2991.

[11] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. 2014. Category-specific video summarization. In *ECCV - European Conference on Computer Vision (Lecture Notes in Computer Science)*, David Flee, Tomas Pajdla, ernt Schiele, and Tinne Tuytelaars (Eds.), Vol. 8694. Springer, Zurich, Switzerland, 540–555. https://doi.org/10.1007/978-3-319-10599-4_35

[12] Yale Song, J. Vallmitjana, A. Stent, and A. Jaimes. 2015. TVSum: Summarizing web videos using titles. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 00. 5179–5187. https://doi.org/10.1109/CVPR.2015.7299154

[13] Chia-Ming Tsai, Li-Wei Kang, Chia-Wen Lin, and Weisi Lin. 2013. Scene-Based Movie Summarization Via Role-Community Networks. *Circuits and Systems for Video Technology, IEEE Transactions on* 23 (11 2013), 1927–1940. https://doi.org/10.1109/TCSVT.2013.2269186

[14] Yusseri Yusoff, William J. Christmas, and Josef Kittler. 2000. Video Shot Cut Detection using Adaptive Thresholding. In *BMVC*.

[15] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Video Summarization with Long Short-term Memory. *CoRR* abs/1605.08110 (2016). arXiv:1605.08110 http://arxiv.org/abs/1605.08110

[16] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Video Summarization with Long Short-term Memory. In *ECCV*.

[17] Kaiyang Zhou and Yu Qiao. 2018. Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward. *CoRR* abs/1801.00054 (2018). arXiv:1801.00054 http://arxiv.org/abs/1801.00054

[18] Kaiyang Zhou, Yu Qiao, and Tao Xiang. 2017. Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward. *arXiv:1801.00054* (2017).