Proceedings of 10th IOE Graduate Conference
Peer Reviewed
ISSN: 2350-8914 (Online), 2350-8906 (Print)
Year: 2021 Month: October Volume: 10

Video Summarization using Spatio-Temporal Features by Detecting Representative Content based on Supervised Deep Learning

Ramesh Kumar Sah [a], Sharad Kumar Ghimire [b]

[a, b] *Department of Electronics and Computer Engineering, Pulchowk Campus, IOE, Tribhuvan University, Nepal*
Corresponding Email: [a] 075msice016.ramesh@pcampus.edu.np, [b] skghimire@ioe.edu.np

Abstract
Video Summarization is the technique to generate the compact version of video keeping relevant content intact and eliminating redundancy that helps the user to browse and navigate through the video more efficiently and effectively. In this work, a framework has been proposed which makes use of the spatial and temporal features with self attention from the video sequences to identify the representative content by generating temporal proposals and supervised learning from the data manually created by humans or users. Existing Supervised methods don't deal with the temporal interest and its consistency. However the issue with these approaches is that for the same contextual segment, frame scores of the video alone cannot be sufficient enough to represent the semantic content. For that temporal uniformity is also necessary which can be addressed by predicting the temporal proposals of the video segment on the basis of action recognition task. These shortcomings have been addressed by this proposed work by treating it as temporal action detection which predicts importance score and location of the segments simultaneously by developing the anchor based method which generates anchors of varying lengths to identify interesting proposals. Moreover the extensive quantitative and qualitative analysis on SumMe and TVSum datasets justify that there is 15% and 5% improvements in F-score respectively compared to previous work with similar environment.

Keywords
Video Summarization, Self Attention, Deep Learning

1. Introduction

The exponential growth of the video consumption has brought up the new challenges for browsing and navigating through video more effectively and efficiently. Video has become the arguably the primary source for the data consumption with the emergence of the data centers and warehouses. Trend of streaming sites and distribution of the videos have become the mainstream in the world of the social media. Though, the consumer might not have enough time to watch the whole video and has to go through complete video to extract the information from it. In those cases, the consumer might just need to get overview of the video without watching the complete video which bestows more relevant event occurred in the video. The conventional news and media distribution methods are quickly replaced by video streaming sites such as YouTube, which are themselves compelled to accustom the rise of uploading videos rather than text and images.

One of the most fundamental measures in the field of video summarisation is the main frame video description. This method provides users with an accurate and portable representation of original video content. The basic concept of keyframe extraction converts the entire video frames to a lessser frames that represent most of the frames. Video synopsis greatly decreases details that must be reviewed for the Video Recovery Framework Based on Content (CBVRS). The majority of works extract keyframes after detecting videoshots in the sense of video summarization.

Even though unsupervised and weakly supervised methods are good performing they lack learning from human summaries which are manually created. This issue is addressed by supervised methods [1],[2]. Supervised Methods comprise summarization of video based on long short term memory, diverse

sequential subset selection, attention based encoder decoder networks. Existing Supervised methods don't deal with the temporal interest and its consistency. In this proposed work these research gaps have been addressed by adopting a new perspective to video summarization techniques. Our contributions are:

- Video summarization process has been treated as temporal action detection which predicts importance score and location of the segments simultaneously by developing the anchor based method which generates anchors of varying lengths to identify interesting proposals.
- Demonstrating that self attention mechanism along with visual features can be better approach for the video summarization.

2. Related Works

Video Summarization has allured a lot of consideration. Identifying and extracting the relevant information from the trivial content is the most daunting challenge in video summarization. There has been lot of researches in the domain of video summarization till date. These can be broadly classified into the following categories.

2.1 Unsupervised Video Summarization

K-Means clustering approaches has been prevailed in the video summarization in early works which utilizes the low level features and motion cues to leverage the summary[3]. These methods were able to achieve good performance although with the highly mobile camera and varying illumination condition causes degradation in the performance. Unsupervised approaches can further be divided into four different categories.1) Dictionary based learning[4],[5] takes the approach of formulating video by optimizing the loss function. Elhamifar et al. [6] is dictionary based approach. Similarly Roy et. al [7] forms representative method for summarization.2) Subset based selection methods selects the representative frames from the original videos. Elhamifar et al. [8] exploits this subset selection approach to determine the similarity between source and actual sets.3) Reinforcement learning has become one of emerging approach in the domain of video summarization which rewards and punish the agent based on action. It uses discrete sampling of action which gives the generated summaries. Zhou et al. [9] formulated deep network for the summarization based on diversity Representative reward. 4) Adversarial learning based methods uses the learning from the ground truth values and then discriminate the input and output accordingly. Mahasseni et al. [10] formed the network based on adversarial network i.e. on LSTM networks in which generated video is compared with the ground truth in the discriminator to get the summaries video. Rochan and Wang [11] developed the video summarization network using unpaired data. Yuan et al. [12] takes advantages of cycle consistent adversarial network to make summaries from corresponding videos.

2.2 Weakly Supervised Video Summarization

These methods mainly focuses on the additional information which includes web priors[13],[14],video categories[15],[16], Video titles[17]. Khosla et al. [14] takes advantage of the web prior images for summarising the videos. Cai et al. [18] uses the variational autoencoder (VAE) [23] to train the web videos to get summaries of videos. Cai et al. [18] captured the key shots which has more visual contents based on the title of image search. s. Potapov et al. [16] developed a summarization method based on the categories of videos. Panda et al. [15] takes the derivative of classification loss to select the key segments in original videos.

2.3 Supervised Video Summarization

Recent advancement in deep learning and presence of abundant human created and annotated summaries supervised approaches have taken the huge step in performance. Gygli et al. [19] fomulated the video summarizing model which leverage the spatial and temporal information. . Gong et al. [20] as well as Sharghi et al. [21] developed the Detriminantal Point Process [22] as the video summarization model that is a non parametric approach to transfer the strucure from training videos to testing videos. Zhang et al. [23] takes advantage of deep network bidirectional LSTM that estimates importance score of each frames in video. Zhao et al. [24, 25] made use of fixed length hierarchical RNN to discover hierarchical structure of the videos. Video summarization is formed as sequence to sequence learning by Zhang et al. [26]. Hussain et al. [27] leverage the advantage of both CNN and Bi-LSTM to compute the multi-view approach for video summarization. [28] combined the encoder decoder architecture with attention model.

Further Fajtl et al. [29] uses self attention model which is extended version of attention based models.

2.4 Anchor-Based Models

The comprehensive progress in the computer vision specially in the object detection aid in the action localization tasks. Xu et al. [36] with the help predefined anchors it predicts the variable length proposals. Chao et al. [30] developed the model that is able to generate multi scale anchor segment for localizing the actions.

3. Methodology

In this section, our methodology is explained in details step by step. The complete methodolgy shown in figure 1 can be divided into four steps:

3.1 Extraction of Spatio-Temporal Features

In case of video sequences long range temporal information can be captured using CNN in order to recognize the characteristic frames and gives basic idea of video content. In addition to that long range representations are helpful for getting more context information. For that GoogleNet is used for feature extraction avoiding last three layers. Given the Video Sequences V of F Frames we will get the features vectors $v_j, j \in i,..,F$.. In case of videos sequences of the video frames are as important as the individual frames because they retains the flow of action in the video sequences and gives more contextual information about the event. Thus to capture long range features Temporal features are extracted using attention mechanism. Moreover other models like LSTM, Bi-LSTM , Graph convolution will also be investigated for their analysis. For the long term temporal features attention based mechanism will be used which will give the feature vector as w_j. Thus the final representations of the feature vector will be obtained as the concatenation of the two feature vectors as $x_j = w_j + v_j$.

3.2 Generation of Temporal Action Proposals

Video sequences has mostly the variable length duration that raise the concern for video summarizations when the temporal features are not taken into account that leads to problem of incomplete segmentation and irrelevant frames getting importance. In training process binary class labels will be assigned to the interest proposals. For that we will be calculating temporal Intersection over Union (tIoU) and compare with the threshold value to assign the binary labels either positive or negative. If greater that threshold positive value will be assigned and if less than threshold negative value will be assigned.

3.3 Proposal Classification and Regression

Temporal Features are average pooled and then fed next module. It bifurcates into two different smaller module i.e Classification and Regression each of these contain FC Layers. Classification gives the significance score and second output gives the center and length offset. This module consists of fully connected layers which is followed by tanh. It is followed by dropout(0.5) and layer normalization that divides into two outputs viz classification to predict importance score and regression to predicts segments. The Loss Function for training the network is defined as L and mathematically can be expressed below equation [31]:

$$L(p,p^*,t,t^*) = \frac{1}{N}\sum_i L_{cls}(p_i,p^*) + \frac{\lambda}{N_{pos}}\sum_i p_i^* L_{reg} \times (t_i,t_i^*) \tag{1}$$

where the λ is the hyper-parameter that balances the loss of classification and regression. N_{pos} denotes proposals with positive labels and N is the total labelled proposals.Similarly the p_i and p_i^* are importance score of predicted and GT respectively for the i^{th}proposals. Likewise L_{cls} is representation of cross entropy loss.
L_{reg} is the regression loss and it can be defined by the smooth absolute mean square loss function and mathematically can be written as:

$$L_{reg}(t_i,t_i^*) = \frac{1}{Q}\sum_{q=1}^{Q} L_{i_{smooth}}(t_{iq} - t_{iq}^*) \tag{2}$$

$$\begin{aligned} L_{i_{smooth}}(x) &= 0.5x^2 \quad if \quad |x| < 1 \\ &= |x| - 0.5 \quad otherwise \end{aligned} \tag{3}$$

These are the smooth absolute loss taken to consider the regression loss. In the equation 3.2 t_{iq} is the q^{th} loss for the element t_i. These losses are generated by comparing the predicted center offsets and length offset with that of ground truth offsets.

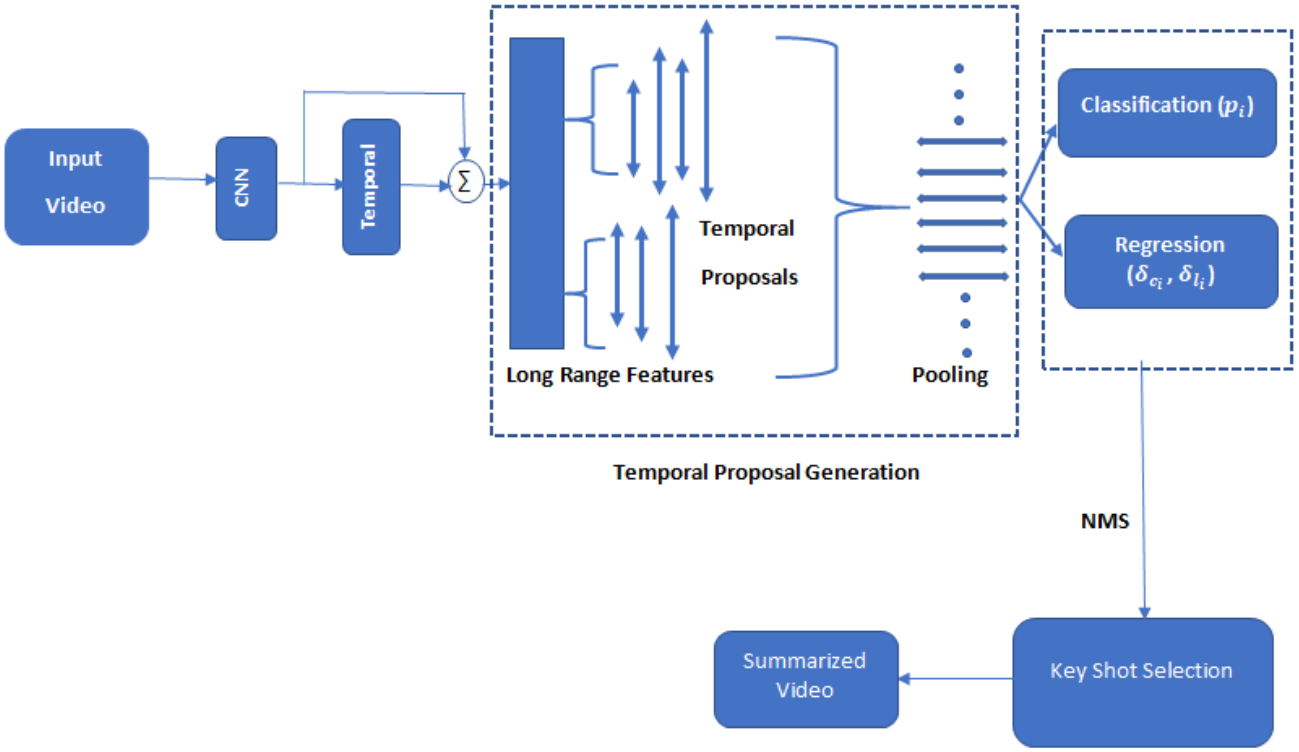

Figure 1: System flow diagram

3.4 Selection of the Keyshots

Finally after obtaining the importance score for each of the frames and offsets implementing classification and regression module refined segments are generated which is done as testing phase of the network. To eliminate the overlapping of low confidence segments Non-Maxima Suppression (NMS) will be performed which will mitigate the redundancy and the segment with low quality. Since the importance score are assigned as the frame level, shots need to be identified round the frames with high importance score. For the same Kernel Temporal Segmentation (KTS) algorithm will be implemented to get key shots which will take the consideration of both importance score of frames as well as the temporal features.

4. Implementations

4.1 Datasets

The algorithm has been experimented on the two standard Datasets i.e SumMe dataset and TVSum Dataset. TVSum and SumMe are currently the only datasets suitably labeled for keyshots video summarization and these data cannot be sufficient enough to train deep learning models to overcome this, OVP and YouTube datsets are used to augment the training dataset. These datasets are labelled using keyframes and thus they have to be converted to the frame-level scores and binary keyshot summaries.
The SumMe dataset is created by [19], the benchmark for evaluating the automatic summary for present and future approaches used for summarization of videos. It contains 25 videos with varying length ranging from one to five minutes.It includes summaries provided by various users, and the length of video is limited to 5% to 15%. The TVSum Dataset[17] has 50 videos sequences which are downloaded from YouTube. It contains videos like changing vehicle tyre,dog show etc. The ground truth segments are required for training Purpose.Two other datasets have been used for the experiments purpose in order to augment the training datasets. These are OVP[4] and YouTube[]. OVP contains 50 videos while YouTube contains 39 videos.
Videos in these datasets are at 30fps thus in order to handle computational complexities and reduce temporal redundancy these video sequences are downsampled to 2fps(Most of the previous works has been followed). Ground truths are required to train the model. SumMe and TVSum consists of frame level importance score which needed to be converted into shot level score for that we have followed KTS that segments videos into shots with their importance

Datasets	No. of Videos	User Number	Contents	Annotation Type	Duration(avg)
SumMe	25	15-18	User generated Videos	Frame-Level Score	146s
TVSum	50	20	Web Videos	Frame-Level Score	235s
OVP	50	5	Various Genre Videos	KeyFrames	98s
YouTube	39	5	Web Videos	KeyFrames	196s

Table 1: Descriptions of datasets which will be used in this work

score by equation 4 [31]. Furthermore key shot based summaries are computed uisng knapsack algorithm. Similarly OVP and YouTube comprise of the keyframes so we did similar action to get shot level scores. It limits the summaries to be within 15% of the original videos. Table 1 shows the brief summary of the datasets used in this proposed work.

$$s_i = \frac{1}{l_i} \sum_{a=1}^{l_i} y_{i,a} \tag{4}$$

where l_i is the that particular i^{th} shot and s_i is the shot level score for the y_a^{th} frame.

4.2 Experimental Setups

There have been three experimental setups done with the datasets. They are Canonical, Augmented and Transfer settings. Datasets(TVSum and SumMe) are divided into 5 random splits. Model is trained using 80% of the data while remaining 20% of the data are for evaluation in canonical setting. In the Augmented settings again 5 random splits is taken for cross validation but in addition for training 80% of the data is augmented with other three dataset are used. For Example, to train SumMe in the augmented setting all samples from TVSum, OVP, and Youtube and 80% of the SumMe are taken as training sample while remaining 20% is used as evaluation set. Same goes for TVSum. While in transfer setup. Model is trained using three of the datasets and rest of one dataset is used for testing the modelIn this setting, Model is trained using three of the datasets and rest of one dataset is used for testing the model. For Example, If i take SumMe as evaluataion dataset then other three viz. TVSum, OVP and Youtube are used as training set. Similarly when SumMe , OVP, YouTube, are used as training set, TVSum is used as evaluation set.

4.3 Algorithm

- Step 1: Visual features are extracted from videos using GoogleNet trained on Imagenet as v_j
- Step 2: Temporal sequence is captured using self attention mechanism (w_j) and concatenated with v_j to get long range features, $x_j = w_j + v_j$
- Step 3: Temporal proposals are generated on each time stamp based on anchor mechanism of 4 anchor scales (4,8,16,32).
- Step 4: Each proposals are classified as positive and negative samples based on overlapping with ground truth by calculating Intersection over Union(IOU) value.
- Step 5: If IOU >0.6 it is positive sample and if IOU <0.3 it is negative sample and if 0<IOU<0.3 these samples are regarded as incomplete and unimportant proposals.
- Step 6: Classification module predicts the importance score and regression predicts the location offsets for each proposals.
- Step 7: Loss is calculated compared with ground truth using equation 1 during training of the model and minimizing loss.

4.4 Evaluation Metrics

For the evaluation of the result obtained from the experiments on the datasets, F- Measure will be used as the quantitative metrics. To asses the similarity between the machine summary and user summaries we use the harmonic mean of precision and recall expressed as the F-score in percentages. The F Measure is calculated by the following equations:

$$F - Measure = \frac{2P_i * R_i}{P_i + R_i} \tag{5}$$

Where $P_i$is the Precision and calculated by:

$$P_i = \frac{length(gs_i \cap gt_i))}{length(gs_i)} \quad (6)$$

Similarly $R_i$is recall and it is calculated by:

$$R_i = \frac{length(gs_i \cap gt_i))}{length(gt_i)} \quad (7)$$

gs_i is the generated summaries for i^{th} summary and gt_i is the annotated summary.

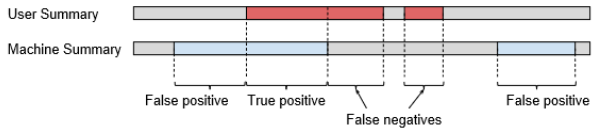

Figure 2: True Positive, False Positive and False Negative representation on per frame basis between ground truth and generated summary by the model

5. Results

Architecture used for visual features is GoogleNet from which 1024 dimensional features were extracted. Furthermore self attention mechanism helps to capture the temporal features which makes use of 8 heads. Dropout of 0.5 has been used with tanh as activation function for the fully connected layers.Drop out 0.5 is better for learning with lesser data like video data. Hyper parameter λ was set to 1 while threshold for the NMS is set to 0.5 as default. Model was trained for about 300 epochs with Adam optimizer and learning rate is defined as $5 * 10^{-5}$.
From the Table 2 it can be inferred that model performs well on Augmented Setting on TVsumm dataset while on SumMe dataset model performs well on canonical set up as compared with other set up. This can also be validated by looking at loss curve which shows that loss is minimum for the TVSum dataset in Augmented Setup.

Datasets	Canonical	Augmented	Transfer
TVSum	64.36	**64.87**	59.54
SumMe	**57.29**	56.59	47.90

Table 2: F-Score comparisons on TVSum and SumMe for different Settings

Datasets	LSTM	BiLSTM	Attention
TVSum	60.01	58.90	**64.36**
SumMe	51.21	52.56	**57.29**

Table 3: F-Score comparisons on TVSum and SumMe for different Temporal Models

Temporal Features have been extracted using self attention for the model while it has also been investigated with other temporal models like LSTM and BiLSTM to find out the effectiveness of the attention mechanism over others. Table 3 shows the F-Score of the each models based on the dataset while Figure 3 and 4 show the graphical representations of F-score on each split of TVSumm and SumMe dataset respectively. It signifies that attention mechanism is working better for TVSumm Dataset while for SumMe dataset all models have similar performance.

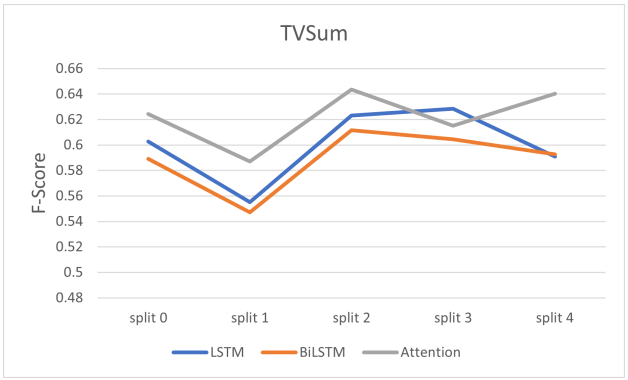

Figure 3: Comparison of Performance of LSTM,BiLSTM and Attention models on TVSumm Dataset

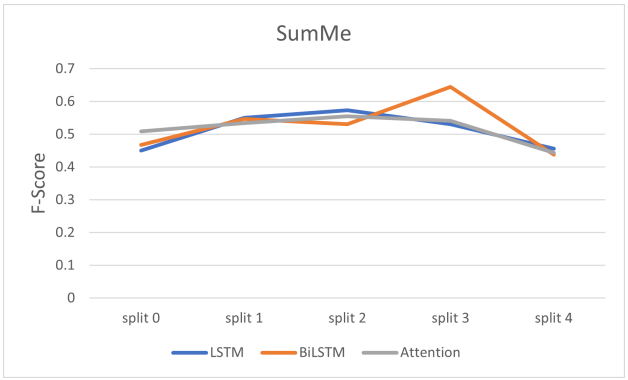

Figure 4: Comparison of Performance of LSTM,BiLSTM and Attention models on SumMe Dataset

5.1 Comparison with other video summarization methods

Results obtained from this work has been compared with other methods of the video summarization which has utilized deep learning methods and tested on the SumMe and TVSum dataset. It has been compared with those methods which has similar methodolgies and used deep learning approaches to get summarised results. Some of the methods are vsLSTM [23], dppLSTM [23], DR-DSN [9], SUM-GAN [10], VASNET [29], AVS [28]. Table 5.3 signifies that there is significant improvement in this works compared to other methods.

Methods	SumMe	TVSumm
Random Selection [17]	-	32
Uniform Sampling [17]	15	36
Clustering	17.5	39
vsLSTM [23]	37.6	54.2
dppLSTM [23]	38.6	54.7
SUM-GAN [10]	41.7	56.3
DR-DSN [9]	42.1	58.1
AVS [28]	43.9	59.4
VASNET [29]	49.71	61.42
Human [17]	64.2	63.7
Proposed Method	**57.29**	**64.36**

Table 4: F-Score comparison of other methods and this work

5.2 Diversity in the generated summaries

One of features of good summaries is that it should include diverse content which can be measured diversity score measurement.The degree of diversity of a generated summary is evaluated by measuring the dissimilarity among the selected frames in the feature space [9]. The diversity score is used to evaluate the diversity in the summaries generated by this algorithm on SumMe and TVSum dataset. Table 5 shows comparison of the diversity score of dppLSTM and DR-DSN. It can be observed that the summaries generated by this method has sufficiently able to get much diverse summaries as compared with those methods which have also calculated diversity score among the feature space.

Datasets	dppLSTM [23]	DR-DSN [9]	Proposed Method
SumMe	0.591	0.594	**0.6549**
TVSum	0.463	0.464	**0.4748**

Table 5: Diversity Score comparison

6. Parameter Analysis and Ablation Study

6.1 Influence of the average pooling layer(temporal)

In this model average pooling layer has been implemented in order to handle the variable length of the proposals. This layer has significant influence in the classification and regression module so in order to investigate the importance of this layer. Table 6 shows the F-Score variation because of the pooling layer presence and absence on two different datasets. It can be inferred from the table that with pooling layer performance of the model is better as compared that of without pooling layers.

6.2 Analysis of NMS Threshold

NMS thresholds has been used for the removal of redundant and low quality proposals from the output of classification and regression section which signifies that it directly affects the performance. High qualities proposals can be refined when higher threshold is chosen retaining low qualities proposals. Thus it is necessary to analyze the influence of the NMS threshold value in the model. Figure 5 and Figure 6 show the value of F-Score corresponding of the different NMS threshold values on SumMe and TVSum datasets respectively. It can be observed that as threshold increases corresponding f-score increases and after nms 0.5 value it starts decreasing this shows that changes in the value of threshold directly influence the value of F-score thus menthod performance. The default NMS threshold chosen is 0.5.

6.3 Significance of Temporal Sequence and continuity

In order to get video summary temporal sequence and the continuity is major concern. Thus to investigate the significance of the temporal continuity some experiments were performed. To esnure the continuity

Pooling Layer	SumMe			TVSum		
	Canonical	Augmented	Transfer	Canonical	Augmented	Transfer
×	51.4	52.3	44.9	61.2	61.9	56.7
✓	57.2	56.59	47.90	64.36	64.87	59.54

Table 6: Effect of with and without average pooling layer by showing F-Score(%) comparison on TVSumm and SumMe Datasets

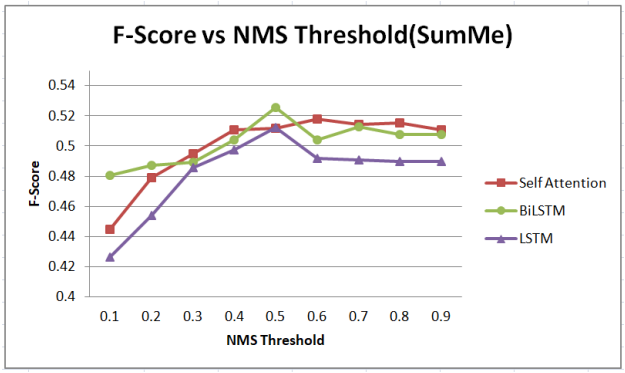

Figure 5: NMS Threshold analysis on SumMe dataset (Default is set at 0.5)

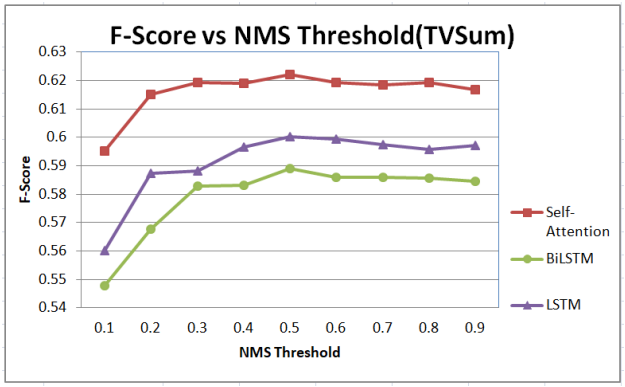

Figure 6: NMS Threshold analysis on TVSum dataset (Default is set at 0.5)

in summary relevant proposals are being selected which are refined by using the regression module. The result has to be compared with the reference values. The reference values are calculated by without generating proposals and importance scores are directly predicted using self attention mechanism. Table 7 shows the results obtained from the experiment done using self attention.$\lambda = 0$ shows parameter λ is zero in the loss function in equation 1 and proposals are only generated but not refined using regression. It can be observed that the result has been degraded in such case as compared when refining of the proposals are performed. With Refining of proposals the permanence is superior in both the datasets.

Methods	Relevant Proposal	Refined Proposal	SumMe	TVSumm
Reference	×	×	48.8	59.6
$\lambda = 0$	✓	×	47.4	60.7
$\lambda = 1$	✓	✓	57.29	64.36

Table 7: F-Score comparison in terms of temporal continuity and refined proposals

Runtime	SumMe	TVSumm
Average Frames(number)	293	470
Average time(ms)	17.25	31.18

Table 8: Runtime Analysis(average)

6.4 Runtime Analysis

The inference time has been calculated of the model to demonstrate the effectiveness of the method. To calculate the inference time runtime is calculated after the extraction of features in GoogleNet. Table 8 shows the inference time averaged on SumMe and TVSum datasets. It is computed per video basis on the average of 15 frames per second. The runtime has been expressed in the ms and frames is in number.

7. Qualitative Results

For the intuitive interpretation of the result we have done some qualitative analysis which shows the effectiveness of the framework. Figure 7 show the comparison of the frames or segments using different methods and ground truth of playing ball video from SumMe dataset. It shows that our method has successfully selected the segments similar to that of the ground truth meanwhile other methods like VASNet[29] and dppLSTM[23] have also selected those frames which are less relevant to the summaries.

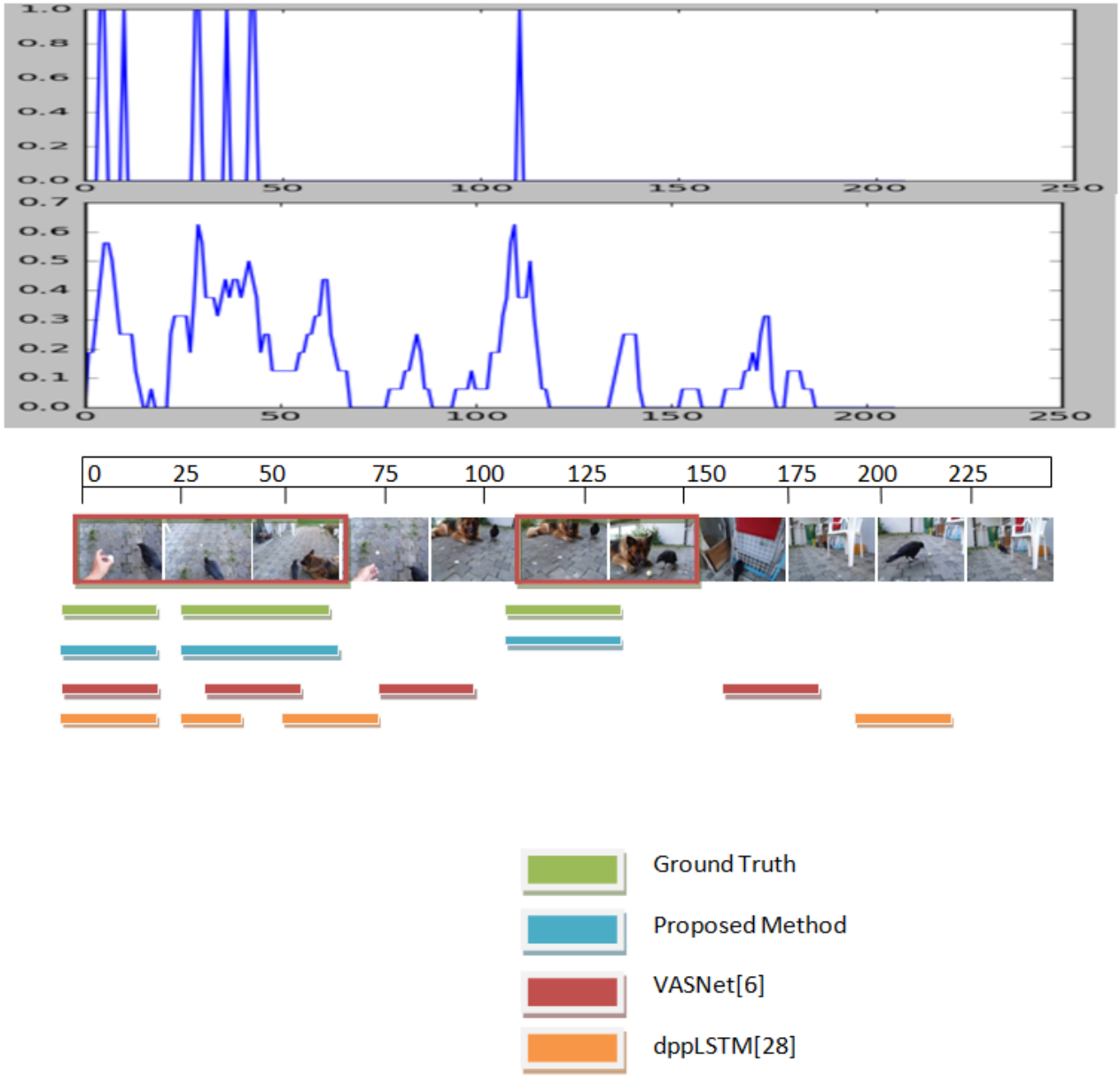

Figure 7: Comparison of frames/segments selected using different methods and ground truth of Playing ball video from SumMe dataset

Moreover figure 8 reveals the selected frames from the video of plane landing of TVSum dataset which shows that most of the representative frames are being selected from which we can describe the content of the videos.It can be observed that the segments selected for summaries of the video are more consistent to the original video ground truth which clarifies the efficiency of the method.

8. Conclusion and Future Enhancements

In this paper we have proposed a method for summarizing video which takes advantage of the anchor based mechanism to generate interest proposals of varying length to show the most representative content of the video. In the contrast of the other supervised methods this proposed predicts the important segments from the video assigning the importance score as well as the regress through the segment simultaneously. It deals with the varying length of the segments with the help of varying length

Figure 8: Some of the selected frames of the Landing plane video of TVSum dataset

proposals generated that handles the incomplete and incorrect segments. The result shows the effectiveness of the proposed work. Although the video has been summarised using visual features audio has not been included because of lack of sufficient labelled dataset. There are lots of research going in the domain of video summarization to get summaries which can represent original video more appropriately. This work can also be further researched to include the audio as well for training and evaluation. Development of sufficient labelled datasets are needed to make the model more efficient.

References

[1] Tsu-Jui Fu, Shao-Heng Tai, and Hwann-Tzong Chen. Attentive and adversarial learning for video summarization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1579–1587. IEEE, 2019.

[2] Siyu Huang, Xi Li, Zhongfei Zhang, Fei Wu, and Junwei Han. User-ranking video summarization with multi-stage spatio–temporal representation. *IEEE Transactions on Image Processing*, 28(6):2654–2664, 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*, pages 1400–1401, 2006.

[4] Shiyang Lu, Zhiyong Wang, Tao Mei, Genliang Guan, and David Dagan Feng. A bag-of-importance model with locality-constrained coding based feature learning for video summarization. *IEEE Transactions on Multimedia*, 16(6):1497–1509, 2014.

[5] Qiao Luan, Mingli Song, Chu Yee Liau, Jiajun Bu, Zicheng Liu, and Ming-Ting Sun. Video summarization based on nonnegative linear reconstruction. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2014.

[6] Ehsan Elhamifar, Guillermo Sapiro, and Rene Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1600–1607. IEEE, 2012.

[7] Rameswar Panda and Amit K Roy-Chowdhury. Multi-view surveillance video summarization via joint embedding and sparse optimization. *IEEE Transactions on Multimedia*, 19(9):2010–2021, 2017.

[8] Ehsan Elhamifar, Guillermo Sapiro, and S Shankar Sastry. Dissimilarity-based sparse subset selection. *IEEE transactions on pattern analysis and machine intelligence*, 38(11):2182–2197, 2015.

[9] Kaiyang Zhou, Yu Qiao, and Tao Xiang. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[10] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 202–211, 2017.

[11] Mrigank Rochan and Yang Wang. Video summarization by learning from unpaired data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7902–7911, 2019.

[12] Li Yuan, Francis EH Tay, Ping Li, Li Zhou, and Jiashi Feng. Cycle-sum: cycle-consistent adversarial lstm networks for unsupervised video summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9143–9150, 2019.

[13] Wen-Sheng Chu, Yale Song, and Alejandro Jaimes. Video co-summarization: Video summarization by visual co-occurrence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3584–3592, 2015.

[14] Aditya Khosla, Raffay Hamid, Chih-Jen Lin, and Neel Sundaresan. Large-scale video summarization using web-image priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2698–2705, 2013.

[15] Rameswar Panda, Abir Das, Ziyan Wu, Jan Ernst, and Amit K Roy-Chowdhury. Weakly supervised summarization of web videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3657–3666, 2017.

[16] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[17] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5179–5187, 2015.

[18] Sijia Cai, Wangmeng Zuo, Larry S Davis, and Lei Zhang. Weakly-supervised video summarization using variational encoder-decoder and web prior. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 184–200, 2018.

[19] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *European conference on computer vision*, pages 505–520. Springer, 2014.

[20] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. *Advances in neural information processing systems*, 27:2069–2077, 2014.

[21] Aidean Sharghi, Boqing Gong, and Mubarak Shah. Query-focused extractive video summarization. In *European Conference on Computer Vision*, pages 3–19. Springer, 2016.

[22] Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*, 2012.

[23] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *European conference on computer vision*, pages 766–782. Springer, 2016.

[24] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. Hierarchical recurrent neural network for video summarization. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 863–871, 2017.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[26] Ke Zhang, Kristen Grauman, and Fei Sha. Retrospective encoders for video summarization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 383–399, 2018.

[27] Tanveer Hussain, Khan Muhammad, Amin Ullah, Zehong Cao, Sung Wook Baik, and Victor Hugo C de Albuquerque. Cloud-assisted multiview video summarization using cnn and bidirectional lstm. *IEEE Transactions on Industrial Informatics*, 16(1):77–86, 2019.

[28] Zhong Ji, Kailin Xiong, Yanwei Pang, and Xuelong Li. Video summarization with attention-based encoder–decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(6):1709–1717, 2019.

[29] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Summarizing videos with attention. In *Asian Conference on Computer Vision*, pages 39–54. Springer, 2018.

[30] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1130–1139, 2018.

[31] Le Yang, Houwen Peng, Dingwen Zhang, Jianlong Fu, and Junwei Han. Revisiting anchor mechanisms for temporal action localization. *IEEE Transactions on Image Processing*, 29:8535–8548, 2020.