IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 1, January 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 1, January 2021 ||

DOI:10.15680/IJIRSET.2021.1001032

Deep Learning Approach for Video to Text Summarization

Madhura Prakash M[1], Dr. Krishnamurthy G N[2]

Assistant Professor, Department of ISE, BNM Institute of Technology, Bangalore, Karnataka, India[1]

Principal, BNM Institute of Technology, Bangalore, Karnataka, India[2]

ABSTRACT: Advancement is multimedia technology; smart phones and the internet has incited the interest in everyone to capture videos. The videos captured might have many irrelevant parts that the user might not be interested in revisiting. However, there will be relevant and most interesting parts of the video that the end-user might want to revisit. Hence it is important to manage the Video Big Data and efficient techniques for extracting relevant information from the videos is required. The video analyst can be provided with the relevant information of the video in a textual format. Recent advances in Neural Networks have made it possible to annotate an image by identifying the objects in the image. Several Solutions to Video-to-Video Summarization techniques have been presented in the recent times. Video Summarization techniques fall into two categories namely Static Video summarization where relevant Key frames are extracted from the video and Dynamic Video Summarization where relevant video shots are extracted from the video. Video Summarization techniques may be based on the classification of the clustering-based approaches. To extract relevant part of the video several features of the frames like Colour Distribution, Contrast, sharpness, edges can be considered. In the recent time the state-of-the-art solution to feature extraction in Images is based on the Neural Networks. The Recurrent Neural Networks like LSTM (Long Short-Term Memory) have been used to annotate the images by giving the description of the image in the textual format. The same concept can be extended to annotate the extracted relevant parts of the summarized video. The textual description obtained can further be summarized using text summarization techniques to get the exact gist of the overall video.

KEYWORDS: Multimedia, Neural Networks, Classification, Clustering, LSTM

I. INTRODUCTION

The amount of multimedia data on the internet and other accessible sources is increasing each day because of factors like increased processing power, fast networks and low-cost storage devices with the introduction of smart capturing devices, it is a lot easier for users to watch and share videos personally. However, this huge amount of video data has caused a problem of efficient indexing and retrieval to save users from being overwhelmed in this pile of data. Browsing through these extensive videos to obtain the required video is time consuming. To extract the desired information from such a huge video repository in a minimal span of time is a challenging task. Hence there is need of technique to help this problem of video data management. A basic approach for managing video data is video summarization.

Video summarization is a technique that provide synoptic gist of larger video and significant events of any video to viewers in concise form and help them to watch summaries before deciding the worthiness of the video before downloading or purchasing the complete video. This enables a pleasant and efficient browsing experience to the users. Apart from smart browsing, the video summarizations also help in extracting significant or user defined events from the large videos [1]

Video summarization aims to reduce the amount of redundant data in order to extract the most salient information known as keyframe of the video. The resultant video summary represents the most significant video content. It enables viewers to quickly figure out the important contents of video and help them in searching and retrieval tasks. Video summarization plays an important role in handling a huge amount of data. But to generate a video summary, a full understanding of video is required, which is very difficult for contemporary machines. To summarize large scale open world web videos is difficult since they are unstructured and range over a wide variety of content. In video summarization, attaining the effective content of a video is important yet an unexplored aspect. Therefore, it is desired

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 1, January 2021 ||

DOI:10.15680/IJIRSET.2021.1001032

to develop a framework, which presents the effective elements in video data to the user by considering the meaningful information from the video [2].

Video summaries are generally performed by two different approaches: Static (key frame-based) and Dynamic (video skim-based) video summaries. Static video summary contains a collection of a small but meaningful number of silent frames known as key frames, while dynamic video summary contains a collection of important short video clips.Video summarization addresses the problem of subset selection of frames—which frames to be retained in a summary. Key frame extraction is one of the mechanisms to generate video summary. In general, the extracted key frames should both represent the entire video content and contain minimum redundancy
Two important properties of key frames are that

(1) Key frames should represent the entire content of a video and

(2) Each key frame should be as much different from each other as possible.

Typically, key frame selection algorithms assume that a video has been segmented into shots, and then a small number of key frames are extracted from each shot [3]. Video summarization research has been largely driven by parallel advancements in video processing methods, intelligent selection of video frames, and start-of-the-art text summarization tools. The video data is a great asset for information extraction and knowledge discovery, due to its size and variability, it is extremely hard for users to monitor or find the occurrences in it.

Intelligent video summarization algorithms allow to quickly browse a lengthy video by capturing the essence and removing redundant information. Early video summarization methods were built mainly upon basic visual qualities (e.g., low-level appearance and motion features) while recently more abstract and higher-level cues are leveraged in the summarization frameworks. Automatic video captioning can generate semantic descriptions to a video segment, which can then be used for indexing and facilitating review [5]. Video Captioning has been taken as a challenging problem, as a description should not only capture the Semantic knowledge in the video but also express the spatio-temporal relationships in between and the dynamics in a natural sentence.

Video summarization is one of the promising approaches for effective comprehension of video content by selecting informative frames of the video. The aim is to produce a summary of the video which is interesting to the user and representing the whole video. Video summarization takes various features into account such as representativeness, uniformity, static attention, temporal attention and quality which includes colourfulness, brightness, contrast, hue count, edge distribution for selecting keyframes.

Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. Deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image. Description must capture not only the objects contained in an image, but it also must express how these objects relate to each other as well as their attributes and the activities they are involved in. Moreover, the above semantic knowledge has to be expressed in a natural language like English, which means that a language model is needed in addition to visual understanding.

For video captioning, earlier works used a template-based language model to generate natural language descriptions. Recently, with the success of image captioning using CNN (Convolution Neural Network) and RNN(Recurrent Neural Network) models, many researchers have adopted similar approaches to describe video content with natural language [5] .The large amount of video data makes it hard to review and navigate, particularly long videos such as surveillance videos and life-logging videos. Automatic video captioning can generate semantic descriptions to a video segment, which can then be used for indexing and facilitating review.The paper is organized as follows. Section II describes the related work. The Methodology to be followed is given in Section III. Section IV presents the evaluation process to be followed. Finally, Section V presents conclusion.

II. RELATED WORK

Bor-Chun Chen et al.[4] have proposed a work on Video to Text Summary (V2TS)to Summarize long video content. The aim is to avoid redundant or uninformative video captions resulting from applying segmentation and then captioning to long videos. Video summaries and captions are generated using a joint model. A supervised method with RNN is used for video summarization. Summarization weights are learned from supervised signals

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 1, January 2021 ||

DOI:10.15680/IJIRSET.2021.1001032

according to the labelers' understanding of highlights in the video content and are independent from the context of text generated during the decoding process. A variant of RNNs called Gated Recurrent Unit (GRU) and the CNN structure, VGGNet is used as basic building blocks. The length of the generated video summary is constrained to be less than 15% of the original video. The proposed system is built upon an encoder-decoder framework. By using multi-task learning, the shared RNN-based encoder is optimized over both the video summarization and video captioning tasks.

Each frame is input to a CNN that generates a frame-level feature representation. Frame-level CNN features are then fed into the RNN to sequentially generate an embedding at each time-step, followed by a multi-layer perceptron (MLP) to generate importance scores. Video is first segmented into small shots using the Kernel Temporal Segmentation (KTS) algorithm. Shots are selected as a video summary according to the knapsack problem MLP is used to obtain frame-wise importance scores that are then normalized by a softmax function. The output features of the encoder RNN are weighted by normalized importance scores before entering the decoder RNN. The proposed system is evaluated on several publicly available data set and the comparison is done with the following previous works. F-Score against the annotated summary is considered for evaluation of video to text summarization. The evaluation is against the ground-truth text descriptions using BLUE and METEOR.

Lianli Gao et al. [5] have proposed aLSTM, an attention-based LSTM model with semantic consistency, to transfer videos to natural sentences. An LSTM model is trained on video-sentence pairs and learns to associate a video to a sentence. LSTM is integrated with attention mechanism to capture salient structures of video, and explore the correlation between multimodal representations (i.e., words and visual content) for generating sentences with rich semantic content. Specifically, an attention mechanism that uses the dynamic weighted sum of local two-dimensional convolutional neural network representations is proposed. Then, an LSTM decoder takes these visual features at time t and the word-embedding feature at time t−1 to generate important words. Finally, multimodal embedding is used to map the visual and sentence features into a joint space to guarantee the semantic consistence of the sentence description and the video visual content. To extract meaningful spatial features, Inception-v3 neural network is adopted. To exploit temporal information, one-layer LSTM visual encoder is introduced to encode those spatial 2D CNN feature vectors. Then an attention mechanism which takes the dynamic weighted sum of local spatial 2D CNN feature vectors as the input for the LSTM decoder is proposed. Finally, multiword embedding and cross-view methodology is integrated to project the generated words and the visual features into a common space to bridge the semantic gap between videos and the corresponding sentences. An objective function is built by integrating two loss functions which simultaneously consider video translation and semantic consistency.

Shagan Sah et al [6] have proposed a novel technique for summarizing and annotating long videos. Interesting segments of long video are extracted based on image quality as well as cinematographic and consumer preference. Key frames from the most impactful segments are converted to textual annotations using sequential encoding and decoding deep learning models. Captions from interesting segments are fed into extractive methods to generate paragraph summaries from the entire video. Knobs to increase and/or decrease both the video and textual summary length to suit the application. Worthiness of a superframe to be included in the summary is determined by a non-linear combination of scores measuring a superframe cut's fitness regarding Boundary score (score is inversely proportional to the motion at each boundary neighborhood), Attention (based on temporal saliency, motion, m and variance, v), Contrast Score (standard deviation of luminance pixels.), Sharpness, Saturation, Facial impact. Each key frame is passed through the 152-layer ResNet CNN model pre-trained on ImageNet data, where the vector from the last pooling layer is used as a frame feature. These key frame feature vectors are passed sequentially into a Long Short-Term Memory (LSTM) network, a recurrent neural network. During training, a video sequence and corresponding text annotation pairs are input to the network. During testing, key frames around a superframe video segment are encoded into the trained neural network. Once all frames are processed, a begin of sentence keyword is fed into the network, triggering word generation until and end of sentence keyword is produced.

Rameswar Panda et al. [7] have proposed an unsupervised framework for summarizing a collection of videos. Each video in the collection may contain some information that other videos do not have, and thus the approach explores the underlying complementarity to create a diverse informative summary to efficiently solve the optimization problem, an alternating minimization algorithm that minimizes the overall objective function with

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 1, January 2021 ||

DOI:10.15680/IJIRSET.2021.1001032

respect to one video at a time while fixing the other videos is developed. A new benchmark data set, Tour20, that contains 140 videos with multiple manually created summaries is introduced.

Janya Sainui [8] has proposed the idea that Existing approaches heuristically select key frames; hence, the selected key frames may not be the most different frames and/or not cover the entire content of a video. Objective function is used for selecting key frames. In particular, a statistical dependency measure called quadratic mutual information is used as objective functions for maximizing the coverage of the entire video content as well as minimizing the redundancy among selected key frames. The proposed key frame extraction algorithm finds key frames as an optimization problem. Focus of the proposed work is on the static video summarization which is to extract a set of key frames from a video. Quadratic mutual information (QMI) is used to select m frames so that QMI over m key frames is minimized, while QMI between the set of selected key frames and the set of sampling frames is maximized. Objective functions for minimizing the redundancy among key frames as well as maximizing the coverage of the entire video content are presented.

Manasa Srinivas et al. [9] have taken various features into account such as representativeness, uniformity, static attention, temporal attention and quality which includes colorfulness, brightness, contrast, hue count, edge distribution for selecting keyframes. Keyframe selection is mainly divided into 3 stages. The scores for each frame are computed in the first stage and keyframes are selected based on the combined scores in the second stage. Finally, the elimination of duplicate frames is performed in the third stage.

Subhashini Venugopalan et al [10] have proposed a model that take a short video as input and output a natural language description of the main activity in the video. An end-to-end deep model for video-to-text generation that simultaneously learns a latent "meaning" state, and a fluent grammatical model of the associated language is developed. LSTM is used to "decode" a visual feature vector representing the video to generate textual output. The first step in this process is to generate a fixed-length visual input that effectively summarizes a short video. For this a CNN, specifically the publicly available Caffe reference model, a minor variant of AlexNet is used. The net is pre-trained on the 1.2M image of the ImageNet dataset and hence provides a robust initialization for recognizing objects and thereby expedites training. Frames are sampled in the video (1 in every 10 frames) and the output of the fc7 layer is . A mean pooling over the frames is performed to generate a single 4,096-dimension vector for each video. The resulting visual feature vector forms the input to the first LSTM layer. Another LSTM layer is stacked on top, and the hidden state of the LSTM in the first layer is the input to the LSTM unit in the second layer. A word from the sentence forms the target of the output LSTM unit. Words are represented using "one-hot" vectors (i.e 1-of-N coding, where is N is the vocabulary size).The two-layer LSTM model is trained to predict the next word Swt in the sentence given the visual features and the previous t − 1 words, p(Swt |V, Sw1 , . . . , Swt−1). During training the visual feature, sentence pair (V, S) is provided to the model, which then optimizes the log-likelihood over the entire training dataset using stochastic gradient descent.

Jeff Donahue et al. [11] have proposed a novel recurrent convolutional architecture suitable for large-scale visual learning which is end-to-end trainable, and have demonstrated the value of the model on benchmark video recognition tasks, image description and retrieval problems, and video narration challenges.Long-term RNN models are appealing in that they directly can map variable-length inputs (e.g., video frames) to variable length outputs (e.g., natural language text) and can model complex temporal dynamics; yet they can be optimized with backpropagation. The proposed recurrent long-term models are directly connected to modern visual convnet models and can be jointly trained to simultaneously learn temporal dynamics and convolutional perceptual representations.

Aidean Sharghi et al. [12] have studied the subjectiveness in video summarization. Dense per video-shot concept annotations is collected. The semantic information in each video shot is represented by a binary semantic vector, in which the 1's indicates the presence of corresponding concepts in the shot. A lexicon of concepts is constructed by overlapping nouns in the video shot captions, with those in the SentiBank. Those nouns serve as a great starting point since they are mostly entry-level words. The concepts that are weakly related to visual content are pruned out. And, the redundant concepts are merged. Some new concepts are added in order to construct an expressive and comprehensive dictionary.

Ana Garcia del Molino et al. [13] have done an extensive analysis of FPV (First Person View) video summarization approaches. The introduction of wearable video cameras (e.g., GoPro) in the consumer market has

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 1, January 2021 ||

DOI:10.15680/IJIRSET.2021.1001032

promoted video life-logging, motivating users to generate large amounts of video data. This increasing flow of first-person video has led to a growing need for automatic video summarization adapted to the characteristics and applications of egocentric video. The main challenge yet to be solved is the personalization of the summary, either from the perspective of the wearer or a third party viewing the summary. The summary quality perception is subjective, and therefore, a good summary should strive to that specific user's preferences.

Sinnu Susan Thomas et al. [14] have proposed a summarization technique that incorporates the peculiar features of human perception. The features of the human visual system within the summarization framework itself are used to allow for the emphasis of perceptually significant events while simultaneously eliminating perceptual redundancy from the summaries. The human visual system (HVS) has peculiar properties, degradation. For example, viewers do not focus on the whole image but only around a small so-called foveation region, an important region of the scene that requires higher weightage in the video summary than the peripheral regions. This fact allows for an intelligent perceptual redundancy reduction. The perceptual feature or any measurable attribute of an object, such as motion, size, shape, speed, contrast, and color, is taken into consideration while selecting the frames for summarizing a video. Predominantly, a moving object in an image catches the attention of the viewer. Human vision tends to concentrate on moving objects rather than on the static object. Thus, motion is one important cue as a perceptual feature, and includes both global motion and local motion. The motion is taken as a cue for drawing attention in the top–down model. In the case of the bottom–up attention model, the cues taken are size, shape, speed, contrast, and color. The proposed framework is a combination of the top–down and bottom–up attention model.

Yunan Ye et al. [15] have studied the problem of video question answering by modeling temporal dynamics with frame-level attention mechanism. The video content often contains the evolving complex interactions and the simple extension of image question answering is thus ineffective to provide the satisfactory answers. This is because the relevant video information is usually scattered among the entire frames. Furthermore, a number of frames in video are redundant and irrelevant to the question. The attribute augmented attention network learning framework that enables the joint frame-level attribute detection and unified video representation learning for video question answering is proposed. Then the multi-step reasoning process for the proposed attention network to further improve the performance is incorporated. A largescale video question answering dataset is created. Experiments are conducted on both multiple-choice and open-ended video question answering tasks to show the effectiveness of the proposed method.

Kuo-Hao Zeng et al. [16] have proposed a scalable approach to learn video-based question answering (QA): to answer a free-form natural language question about the contents of a video. The approach automatically harvests a large number of videos and descriptions freely available online. Then, a large number of candidate QA pairs are automatically generated from descriptions rather than manually annotated. Next, these candidate QA pairs are used to train a number of video-based QA methods. In order to handle non-perfect candidate QA pairs, a self-paced learning procedure is proposed to iteratively identify them and mitigate their effects in training. Finally, the performance is evaluated on manually generated video-based QA pairs. The results show that the self-paced learning procedure is effective

Rameswar Panda et al. [17] have proposed a Novel unsupervised framework via joint embedding and sparse representative selection. The objective function is two-fold. The first is to capture the multiview correlations via an embedding, which helps in extracting a diverse set of representatives. The second is to use a norm to model the sparsity while selecting representative shots for the summary. The work is to jointly optimize both of the objectives, such that embedding can not only characterize the correlations, but also indicate the requirements of sparse representative selection. An efficient alternating algorithm based on half-quadratic minimization to solve the proposed non-smooth and non-convex objective with convergence analysis is proposed. A key advantage of the proposed approach with respect to the state-of-the-art is that it can summarize multi-view videos without assuming any prior correspondences/alignment between them, e.g., uncalibrated camera networks.

Prashant M.Baluragi et al. [18] have extracted the key-frames with higher motion intensity and higher audio energy in MFCC(Mel-frequency cepstral coefficients) domain to detect those events that draws maximum audience's attention. The input video stream is read and the audio track and video is separated manually using external tools. The separated audio track and video are given to the matlab module. The frames are converted from RGB to Gray scale for separated video track. Those frames with the higher motion intensity and higher audio

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 1, January 2021 ||

DOI:10.15680/IJIRSET.2021.1001032

energy in MFCC domain values are found and combined with the help of Index match to create the summary video.

Irfan Mehmood et al. [19] have proposed a visual saliency driven framework for keyframe extraction that provides concise versions of video by extracting semantically relevant frames. The proposed visual saliency model helps to bridge the gap between low-level features and high-level information. The visual saliency model is build using static and dynamic saliency maps. The static saliency is derived from colour opponent component space using center surround measure. The dynamic saliency is determined using motion intensity and its phase coherence. Then a two-dimensional visual saliency curve is estimated by fusing static and dynamic saliency maps. Finally, peak points are calculated in the visual saliency curve that leads to the extraction of the keyframes.

III. METHODOLOGY TO BE FOLLOWED

To generate visual summaries of lengthy videos, and to annotate and generate textual summaries of the videos, neural networks can be used. Interesting segments of long video can be extracted based on image quality as well as cinematographic and consumer preference. Key frames from the most impactful segments can be converted to textual annotations using sequential encoding and decoding deep learning models.

Key frames can be extracted from interesting segments whereby deep visual-captioning techniques generate visual and textual summaries. Captions from interesting segments can be fed into extractive methods to generate paragraph summaries from the entire video. The paragraph summary is suitable for search and organization of videos, and the individual segment captions are suitable for efficient seeking to proper temporal offset in long videos

Key frame extraction from the video can be based on

- Temporal motion to define a visual attention score.
- Using spatial saliency at the frame level.
- Cinematographic rules which pull segment boundaries to locations with minimum motion.
- Favouring frames with higher contrast and sharpness
- Favouring more colourful frames,
- people and object content,
- The role facial content plays in image preference.
- Tracking objects across a long video to discover story content.

Given descriptive captions at key frame locations, text extraction methods can be used for summarization.

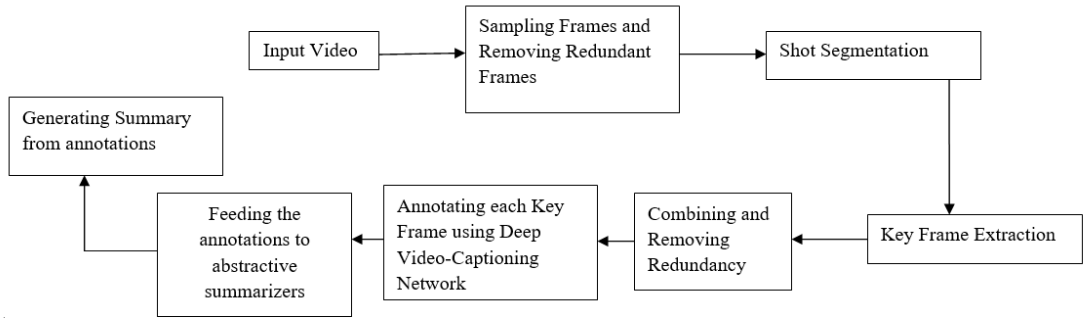

Fig 1: General Framework of Video to Text Summarization

The general framework of video to text summarization is as shown in Figure 1. Abstractive Summarization technique can be applied to the captions generated from Key Frames. Abstractive based summarization is a Natural Language Generation technique. The following are the stages in Abstractive text summarization: -

1. Content Determination
2. Document Structuring
3. Aggregation (Merging of Similar Sentences)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 1, January 2021 ||

DOI:10.15680/IJIRSET.2021.1001032

4. Lexical Choice (Choice of words)
5. Referring expression generation (Pronouns and Anaphora)
6. Realization (Creating text according to the rules of syntax, morphology and orthography)

Abstractive summarization is classified into two categories: structured base and semantic based methods [20]. Structure based Abstractive Summarization Methods includes several methods such as rule based method, tree-basedmethod, Ontology method, lead and body phrase method, graph-based method. Semantic based Abstractive Summarization Methods includes the several methods such as multimodal semantic model, information item-based method, semantic graph-based method, semantic text representation model etc. The basic form subject-verb-object of the sentences is considered for abstractive summarization method. Some steps of this method are given below [21]:

Preprocessing: It is considered for each sentence to create a semantic graph. The actions performed in this step are sentence position, word position, stop word removal and stemming.

Sentence position: It is concerned with counting the words in the given sentences.

Word position: Split the sentence into words

The stop words (a, the, of and an) Stemming – A word can be found in different forms in the same document. These words need to be converted to their original form for simplicity. The stemming algorithm is used to transform words to their canonical forms

Lemmatization: Replacing words with their base forms

A semantic graph can be generated from the text. After generating semantic graph from the original document, a reduced semantic graph can generate accessing Domain ontology & Wordnet. Domain ontology contains the information needed in the same domain of semantic graph to generate the final texts and Wordnet is accessed to generate multiple texts according to the synonym of the word. [22]. Abstractive rewriting techniques like sentence compression, sentence fusion and sentence splitting can be followed.

IV. EVALUATION PROCESS

A good summary is one that has the following properties:

• Representativeness. The set of videos should be reconstructed with high accuracy using the extracted summary.
• Interestingness. The summary should contain most interesting parts of the input videos
• Sparsity. Although the summary should be representative and interesting, the length should be as small as possible.
• Diversity. The summary should be diverse as much as possible capturing different aspects of the input video collection. In other words, the amount of content redundancy should be small in the final set of extracted summaries.

BLEU and METEOR are two metrics are widely used in the machine translation as well as video captioning works and have been shown to have good correlation with human judgments.ROUGE is a set of objective metrics of summarization quality that can be calculated automatically, making them ideal for development and comparison of summaries generated by multiple summarization models. These metrics rely on methods of measuring word overlap between the output of a summarization system and one or more human generated reference summaries. Although simple, the ROUGE metrics correlate very highly with human evaluations. ROUGE-2, which measures the number of bigrams (i.e., two-word sequences) appearing in the summarization output that also appear in the reference summaries.

V. CONCLUSION

Integrating visual content with language learning to generate descriptions for images, especially for videos, has been regarded as a challenging task and is a critical step towards machine intelligence and many applications in daily scenarios such as image/video retrieval, video understanding, blind navigation and automatic video subtitling. Compared with image captioning, video captioning is more difficult for the diverse sets of objects, scenes, actions, attributes and salient contents.One of the main obstacles to the research on video summarization is the user subjectivity — users have various preferences over the summaries. The subjectiveness causes at least two problems. First, no single video summarizer fits all users unless it interacts with and adapts to the individual users. Second, it is very challenging to evaluate the performance of a video summarizer. Visual information retrieval (VIR) is the information delivery mechanism that enables users to post their queries and then obtain the answers from visual contents

Despite the great progress, video captioning suffers from similar issues as image captioning: (1) it is fairly easy to generate a relevant, but non-specific, natural language description (2) it is hard to evaluate the quality of the generated open-ended natural language description. The context of the video summarization task can improve captioning because it helps the captioning model focus on the features in important frames. The context of the video captioning task can

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 1, January 2021 ||

DOI:10.15680/IJIRSET.2021.1001032

also enhance the video representation of the summarization model because the understanding of semantic video content such as objects, persons, and scenes, can help video summarization.Finally, as compared with image content, video has both spatial and temporal structure. In order to efficiently translate video to language, approaches should take both temporal and spatial information into account.

REFERENCES

[1] Shagan Sah et al. “Semantic Text Summarization of Long Videos” 2017 IEEE Winter Conference on Applications of Computer Vision (WACV)
[2] Manasa Srinivas et al., “An Improved Algorithm for Video Summarization – A Rank Based Approach” Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016) Elsevier-Science Direct
[3] Wu, Zuxuan & Yao, Ting & Fu, Yanwei & Jiang, Yu-Gang. (2016). Deep Learning for Video Classification and Captioning.
[4] Chen, Bor-Chun & Chen, Yan-Ying & Chen, Francine. (2017). Video to Text Summary: Joint Video Summarization and Captioning with Recurrent Neural Networks. 10.5244/C.31.118.
[5] L. Gao, Z. Guo, H. Zhang, X. Xu and H. T. Shen, "Video Captioning With Attention-Based LSTM and Semantic Consistency," in IEEE Transactions on Multimedia, vol. 19, no. 9, pp. 2045-2055, Sept. 2017, doi: 10.1109/TMM.2017.2729019.
[6] S. Sah, S. Kulhare, A. Gray, S. Venugopalan, E. Prud'Hommeaux and R. Ptucha, "Semantic Text Summarization of Long Videos," 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, 2017, pp. 989-997, doi: 10.1109/WACV.2017.115.
[7] R. Panda, N. C. Mithun and A. K. Roy-Chowdhury, "Diversity-Aware Multi-Video Summarization," in IEEE Transactions on Image Processing, vol. 26, no. 10, pp. 4712-4724, Oct. 2017, doi: 10.1109/TIP.2017.2708902.
[8] Sainui, J.. “Key Frame Based Video Summarization via Dependency Optimization.” World Academy of Science, Engineering and Technology, International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering 12 (2018): 111-117.
[9] Srinivas, Manasa & M M, Manohara & Pai, Radhika. (2016). An Improved Algorithm for Video Summarization – A Rank Based Approach. Procedia Computer Science. 89. 812-819. 10.1016/j.procs.2016.06.065.
[10] Venugopalan, Subhashini & Xu, Huijuan & Donahue, Jeff & Rohrbach, Marcus & Mooney, Raymond & Saenko, Kate. (2014). “Translating Videos to Natural Language Using Deep Recurrent Neural Networks”. 10.3115/v1/N15-1173.
[11] J. Donahue et al., "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 677-691, 1 April 2017, doi: 10.1109/TPAMI.2016.2599174.
[12] Sharghi, Aidean & Laurel, Jacob & Gong, Boqing. (2017). Query-Focused Video Summarization: Dataset, Evaluation, and a Memory Network Based Approach. 2127-2136. 10.1109/CVPR.2017.229.
[13] A. G. del Molino, C. Tan, J. Lim and A. Tan, "Summarization of Egocentric Videos: A Comprehensive Survey," in IEEE Transactions on Human-Machine Systems, vol. 47, no. 1, pp. 65-76, Feb. 2017, doi: 10.1109/THMS.2016.2623480.
[14] S. S. Thomas, S. Gupta and V. K. Subramanian, "Perceptual Video Summarization—A New Framework for Video Summarization," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 27, no. 8, pp. 1790-1802, Aug. 2017, doi: 10.1109/TCSVT.2016.2556558.
[15] Ye, Yunan & Zhao, Zhou & Li, Yimeng & Chen, Long & Xiao, Jun & Zhuang, Yueting. (2017). Video Question Answering via Attribute-Augmented Attention Network Learning. 10.1145/3077136.3080655.
[16] Zeng, Kuo-Hao & Chen, Tseng-Hung & Chuang, Ching-Yao & Liao, Yuan-Hong & Niebles, Juan Carlos & Sun, Min. (2016). Leveraging Video Descriptions to Learn Video Question Answering.
[17] R. Panda and A. K. Roy-Chowdhury, "Multi-View Surveillance Video Summarization via Joint Embedding and Sparse Optimization," in IEEE Transactions on Multimedia, vol. 19, no. 9, pp. 2010-2021, Sept. 2017, doi: 10.1109/TMM.2017.2708981.
[18] Y. S. Kumar, S. K. Gupta, B. R. Kiran, K. R. Ramakrishnan and C. Bhattacharyya, "Automatic summarization of broadcast cricket videos," 2011 IEEE 15th International Symposium on Consumer Electronics (ISCE), Singapore, 2011, pp. 222-225, doi: 10.1109/ISCE.2011.5973819.
[19] Mehmood, Irfan, Muhammad Sajjad, and Sung Wook Baik. "Visual attention based extraction of semantic keyframes." Advances in Information Science and Applications 1 (2014).
[20] Rupal Bhargava et al “ATSSI: Abstractive Text Summarization using Sentiment Infusion” Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016), Elsevier, Science Direct
[21] Sabina Yeasmin et al, “Study of Abstractive Text Summarization Techniques” American Journal of Engineering Research (AJER) 2017
[22] N. Moratanch et al “A Survey on Abstractive Text Summarization” 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT]

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details