Video Summarization Using Deep Learning Framework

Vaibhav Madavi[1], Dr. S.K. Shah[2] , Hrishikesh Shinde[3], Abhijeet Bhore[4]

Department of E&TC Engineering, SKNCOE, SPPU, Pune, India

[1]vaibhavmadavi77@gmail.com

[2]skshah@sinhgad.edu

[3]shinderushikesh30@gmail.com

[4]abhibhore708@gmail.com

Abstract

Video ssummary is a method of processing raw video into a short form without losing much detail. It covers the issue of choosing a short set of frames or shots for the most important content in the raw video. Most of these methods uses neural network to find difference between video frames without considering the specific details of extracted frames. Some methods use the opportunity to focus on exploring the features of certain frames while ignoring systematic information in video sequence. This paper addresses the matter of a supervised video summarization by using CNN.

Keywords*—CNN,Framwork,Deep Learning*

I. INTRODUCTION

Ideal video summarization is that, which may provide users the utmost information of the target video within the shortest time. it's also useful for several other practical applications, like video indexing, video retrieval, and event detection. Video summarization aims to present a meaningful abstract view of the whole video within a brief period of your time. the essential sort of video summarization is split into key frame extraction and video skim. Key-frames also are referred to as representative frames, still image summaries, or static storyboards. Key frame extraction refers to selecting a little number of image sequences from original videos, and these images are expected to be an approximate representation of the visual contents of the whole video. A video skims may be a video segment that has less duration than the first video. it's generally expected that the video segments can contain most parts of the first video, including visual images and sounds. Video summaries can greatly help people to quickly understand and master information.. Video summarization are often categorized into two forms:

- Static video summarization (key framing) and
- Dynamic video summarization (video skimming)

II. LITERATURE SURVEY

With the development of multimedia video technology, no matter in video browsing, video surveillance, or in artificial intelligence, there is explosive video data needed to be stored, analyzed or understood. The massive video information can be valuable resources if we can make full use of it to do many meaningful things, such as using artificial intelligence for security monitoring or intelligent analysis. But the large amount of data have a great impact on hardware processing devices, so the video summarization technology that can extract effective information from quantities of video data has attracted interests from researchers and becomes a popular research topic

TABLE 1

LITERATURE SURVEY

Sr. no.	**Paper name**	**Author name, publication, year**	**System proposed**	**Techniques/ tools**	**Remark**
1	Video Summarization with Attention-Based Encoder-Decoder Networks	Ji, Zhong; Xiong, Kailin; Pang, Yanwei; Li, Xuelong, IEEE, 2019	A deep attentive framework for supervised video summarization. Specifically, two attention-based deep models named A-AVS and M-AVS are developed, respectively.	BiLSTM, attention-based LSTM network	This model outperforms the competing methods on two benchmark datasets by 0.8%-3%.
2	Wide and Deep Learning for Video Summarization via Attention Mechanism and Independently Recurrent Neural Network	Juanping Zhou; Lu Lu, DCC, 2020	a novel video summarization methodology, Wide and Deep Summarization Network (WD-SN) is proposed. The encoder obtains a sequence of video features through CNN, and the decoder predicts frame-level importance scores from the wide and deep network	Independently Recurrent Neural Network	It captures both the wide independent characteristics and the deep interdependencies of a video sequence.
3	Novel Key-frames Selection Framework for Comprehensive Video Summarization	Huang, Cheng; Wang, Hongmei, IEEE, 2019	a novel framework for an efficient video content summarization as well as video motion summarization is proposed. Initially, Capsules Net is trained as a spatiotemporal information extractor and an inter-frames motion curve is generated based on those spatiotemporal features. Subsequently, transition effects detection (TED) method is proposed to automatically segment the video streams into shots. Finally, a self-attention model is introduced to select key-frames sequences inside shots, thus key static images are selected as video content summarization and optical flows can be calculated as video	transition effects detection (TED)	This method is competitive on VSUMM, TvSum, SumMe and RAI datasets about shot segmentation and video content summarization.

			motion summarization.		
4	Comprehensive Video Understanding: Video Summarization with Content-Based Video Recommender Design	Jiang, Yudong; Cui, Kaixu; Peng, Bo; Xu, Changliang, IEEE, 2019	Video summarization as a content-based recommender problem is formulated in this paper, which should distill the most useful content from a long video for users who suffer from information overload. A scalable deep neural network is proposed on predicting if one video segment is a useful segment for users by explicitly modelling both segment and video.	DNN	It shows that data augmentation and Multi-Task learning is helpful in solving the limitation of dataset problem. To better understand the video content, author also perform action and scene recognition in untrimmed video with state-of-the-art video classification algorithm.
5	Temporal U-Nets for Video Summarization with Scene and Action Recognition	Kwon, Heeseung; Shim, Woohyun; Cho, Minsu, IEEE, 2019	The proposed architecture is an encoder-decoder structure where the encoder captures long-term temporal dynamics from an entire video and the decoder predicts detailed temporal information of multiple contents of the video. Two-stream processing is adopted for obtaining feature representations, one for focusing on the spatial information and the other for the temporal information.	TUNet	TUNet captures long-term temporal dynamics of untrimmed videos through temporal convolutions and obtains detailed temporal information of multiple contents by segment-wise predictions.

A. Summary

As video summarization problem appeals to a lot of researchers, many state-of-the-art approaches have been presented in solving this problem [1-10].
Generally, they treat video summarization as sequence-to-sequence learning problem. Since RNN and its variants LSTM, GRU are very efficient in modelling long time dependencies under encoder-decoder architectures, many works in machine translation, image/video captioning, reading comprehension adopt these technologies.In this paper, encoder-decoder model and a key shot selection model is adopted in order to consider both the isolated segment, and its roles in the whole video.

III. PRAPOSED METHODOLOGY

Video summarization still remains a challenging task. Due to sufficient video data on the Internet, such task draws significant attention in the vision community and benefits a wide range of applications, e.g., video retrieval, search, etc. To effectively perform video summarization by deriving the keyframes which represent the given input video, we propose Neural Network based video summarization technique.

Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labeling or clustering raw input. The patterns they recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text or time series, must be translated.

METHODOLOGY

Proposed system is using software based tools like **python programming language and using several libraries integrated in python** which is necessary for running image processing as well as video processing ,etc. (And as we all know video is nothing but collection of images or frames). The Important python libraries we installed here **are Opencv, keras , tensorflow**, etc.

In this process, A video is provided as an input to our system which would be considered as original video, and the provided video undergoes pre-processing after which we get summarized video as a dersired output . The entire process is performed under the neural network libraries **(keras**) and several other libraries with trained models which pre-processes on the input video . The only reason behind using these three python libraries is that python is a programming language and individually can not do images processing or video processing. Using these three libraries and a certain code is created which we will use to run in Python IDE will helps to summarize the video as we required. (libraries are used in python IDE using **import** commands).

As it know there is a neural network model which does work like human brain . This neural network works to compares the frames which is extracted so that if the frame are same it will discard one of the frame and continuously checks for all the extracted frame and discard the unwanted frames and when this pre-processing is complete we will get the summarized video with no extra or useless frame .

IV. ALGORITHM

Neural networks help us cluster and classify. You can think of them as a clustering and classification layer on top of the data you store and manage. They help to group unlabeled data according to similarities among the example inputs, and they classify data when they have a labelled dataset to train on. (Neural networks can also extract features that are fed to other algorithms for clustering and classification; so you can think of deep neural networks as components of larger machine-learning applications involving algorithms for reinforcement learning, classification and regression. The agenda for this field is to enable machines to view the world as humans do, perceive it in a similar manner and even use the knowledge for a multitude of tasks such as Image & Video recognition, Image Analysis & Classification, Media Recreation, Recommendation Systems, Natural Language Processing, etc. The advancements in Computer Vision with Deep Learning has been constructed and perfected with time, primarily over one particular algorithm - **Convolutional Neural Network**.

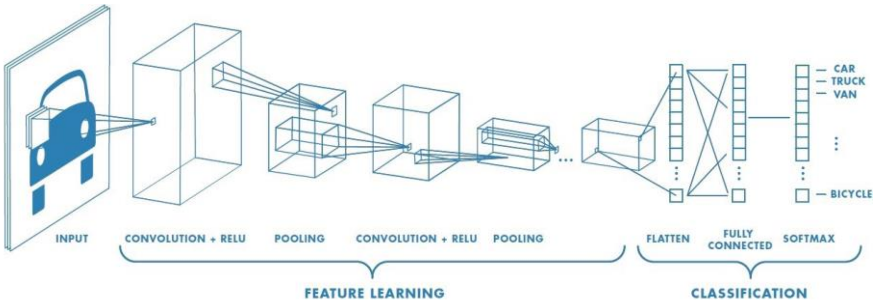
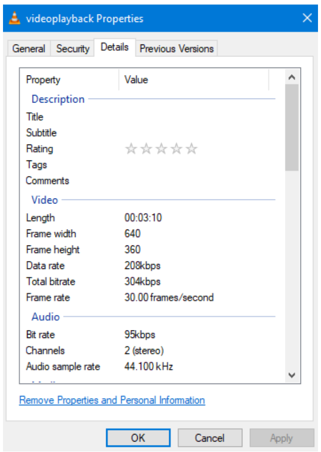


Fig1. Architechture CNN

A **Convolutional Neural Network (ConvNet/CNN)** is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other.

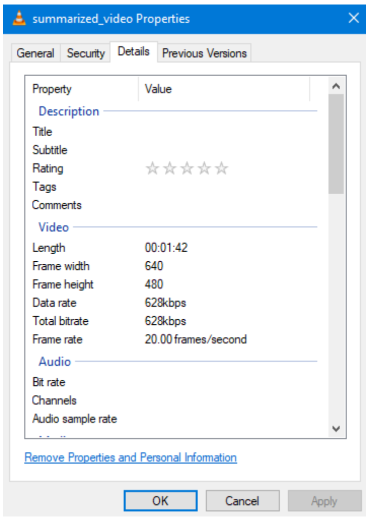
V. RESULT



Property	Value
Description	
Title	
Subtitle	
Rating	☆☆☆☆☆
Tags	
Comments	
Video	
Length	00:03:10
Frame width	640
Frame height	360
Data rate	208kbps
Total bitrate	304kbps
Frame rate	30.00 frames/second
Audio	
Bit rate	95kbps
Channels	2 (stereo)
Audio sample rate	44.100 kHz

Fig.2 Orignal video’s Properties screenshot image.

Above mentioned Image screen shot is of raw video file used as an input for our projects. This video has certain properties like length, size etc. This video is downloaded from Youtube.com. You can download or get any video file format for this project. From above image we can understand that our video length is 00:03:10 that is hr : min : sec.


summarized_video Properties
General
Security
Details
Previous Versions
Remove Properties and Personal Information
OK
Cancel
Apply

Property	Value
Description	
Title	
Subtitle	
Rating	☆☆☆☆☆
Tags	
Comments	
Video	
Length	00:01:42
Frame width	640
Frame height	480
Data rate	628kbps
Total bitrate	628kbps
Frame rate	20.00 frames/second
Audio	
Bit rate	
Channels	
Audio sample rate	

Fig. 3 Summarized video's Properties Screenshot image.

Above mentioned Image screen shot is of Summaried video file which we got as an desired out for our projects. This video has certain properties like length, size etc. This video is output of our project. We named output for every video as summarized _video. You can summarize any video for this project. From above image we can understand that our video length is 00:01:42 that is hr : min : sec.

VI. CONCLUSION

The aim of video summarization is to speed up browsing of a large collection of video data, and achieve efficient access and representation of the video content. By watching the summary, users can make quick decisions on the usefulness of the video. Dependent on applications and target users, the evaluation of summary often involves usability studies to measure the content informative-ness and quality of a summary. An Encoder decoder based video summarization system is proposed. The output summary of the proposed system is composed of a set of key frames extracted from the original video using Bidirectional LSTM Network encoder and attention based decoder and CNN network. To the best of our knowledge, our work is the first attempt to apply the attention mechanism in deep models for video summarization. This machine learning based approach can provide best results. We concluded our work. Here, we state how CNN is better than other algorithm and how we can improve the proposed system.

VII.REFERENCES

[1] Ji, Zhong; Xiong, Kailin; Pang, Yanwei; Li, Xuelong (2019). Video Summarization with Attention-Based Encoder-Decoder Networks. IEEE Transactions on Circuits and Systems for Video Technology, (), 1–1. doi:10.1109/tcsvt.2019.2904996

[2] Juanping Zhou; Lu Lu, "Wide and Deep Learning for Video Summarization via Attention Mechanism and Independently Recurrent Neural Network", 2020 Data Compression Conference (DCC) DOI: 10.1109/DCC47342.2020.00074

[3] Huang, Cheng; Wang, Hongmei (2019). Novel Key-frames Selection Framework for Comprehensive Video Summarization. IEEE Transactions on Circuits and Systems for Video Technology, (), 1–1. doi:10.1109/TCSVT.2019.2890899

[4] Jiang, Yudong; Cui, Kaixu; Peng, Bo; Xu, Changliang (2019). [IEEE 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW) - Seoul, Korea (South) (2019.10.27-2019.10.28)] 2019 IEEE/CVF International Conference on Computer Vision

Workshop (ICCVW) - Comprehensive Video Understanding: Video Summarization with Content-Based Video Recommender Design. , (), 1562–1569. doi:10.1109/ICCVW.2019.00195

[5] Kwon, Heeseung; Shim, Woohyun; Cho, Minsu (2019). [IEEE 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW) - Seoul, Korea (South) (2019.10.27-2019.10.28)] 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW) - Temporal U-Nets for Video Summarization with Scene and Action Recognition. , (), 1541–1544. doi:10.1109/iccvw.2019.00192

[6] Ji, Zhong; Xiong, Kailin; Pang, Yanwei; Li, Xuelong (2019). Video Summarization with Attention-Based Encoder-Decoder Networks. IEEE Transactions on Circuits and Systems for Video Technology, (), 1–1. doi:10.1109/tcsvt.2019.2904996

[7] Juanping Zhou; Lu Lu, “Wide and Deep Learning for Video Summarization via Attention Mechanism and Independently Recurrent Neural Network”, 2020 Data Compression Conference (DCC) DOI: 10.1109/DCC47342.2020.00074

[8] Jiang, Yudong; Cui, Kaixu; Peng, Bo; Xu, Changliang (2019). [IEEE 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW) - Seoul, Korea (South) (2019.10.27-2019.10.28)] 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW) - Comprehensive Video Understanding: Video Summarization with Content-Based Video Recommender Design. , (), 1562–1569. doi:10.1109/ICCVW.2019.00195