AUDIO VISUAL GENERATIVE ADVERSARIAL NETWORK BASED VIDEO SUMMARIZATION

A PROJECT REPORT

Submitted by

N. SATHWIK [CB.EN.U4CSE19037]
P. ROHIT REDDY [CB.EN.U4CSE19042]
M. VAISHNAV [CB.EN.U4CSE19228]
V. TEJASWI [CB.EN.U4CSE19259]

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

AMRITA SCHOOL OF COMPUTING

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE - 641 112

JUNE 2023

AMRITA VISHWA VIDYAPEETHAM

AMRITA SCHOOL OF COMPUTING, COIMBATORE – 641 112

BONAFIDE CERTIFICATE

This is to certify that the project report entitled **AUDIO VISUAL GENERATIVE ADVERSARIAL NETWORK BASED VIDEO SUMMARIZATION** submitted by N. Sathwik(CB.EN.U4CSE19037) , P. Rohit Reddy(CB.EN.U4CSE19037), M. Vaishnav(CB.EN.U4CSE19228), V. Tejaswi (CB.EN.U4CSE19259) in partial fulfillment of the requirements for the award of Degree **Bachelor of Technology** in Computer Science and Engineering is a bonafide record of the work carried out under our guidance and supervision at the Department of Computer Science and Engineering, Amrita School of Computing, Coimbatore.

Dr. Sikha OK
(Assistant Professor [SG])
Department of CSE

Dr.Vidhya Balasubramanian
Chairperson
Department of CSE

Evaluated on:

INTERNAL EXAMINER

EXTERNAL EXAMINER

DECLARATION

We, the undersigned solemnly declare that the project report **AUDIO VISUAL GENERATIVE ADVERSARIAL NETWORK BASED VIDEO SUMMARIZATION** is based on our own work carried out during the course of our study under the supervision of Dr. Sikha OK, (Assistant Professor [SG]), Computer Science and Engineering, and has not formed the basis for the award of any other degree or diploma, in this or any other Institution or University. In keeping with the ethical practice in reporting scientific information, due acknowledgement has been made wherever the findings of others has been cited.

N. SATHWIK[CB.EN.U4CSE19037]- signature

P. ROHIT REDDY [CB.EN.U4CSE19042]- signature

M. VAISHNAV[CB.EN.U4CSE19228]- signature

V. TEJASWI [CB.EN.U4CSE19259]- signature

ABSTRACT

The video summarization mainly addresses the problem of complexity of various kinds of videos like static, dynamic, noisy , speeches in crowds where temporal coherence between audio and visuals is preserved while summarizing . Here the audio and visual clips are classified and then converted into corresponding embeddings with the help of classifier where the precedence of visuals over audio or vise-versa is decided. here we are tackling this problem with Audio and Visual GAN based Video Summarization.
Abstract should be one page synopsis of the project report typed double line spacing. Just type in your abstract here.

ACKNOWLEDGEMENT

We would like to express our deep gratitude to our beloved Satguru **Sri Mata Amritanandamayi Devi** for providing the bright academic climate at this university, which has made this entire task appreciable. This acknowledgment is intended to thank all those people involved directly or indirectly with our project. We would like to thank our Pro Chancellor **Swami Abhayamritananda Puri**, Vice Chancellor **Dr.Venkat Rangan.P** and **Dr.Bharat Jayaraman**, Dean, Faculty of AI & Computing(Computing), Amrita Vishwa Vidyapeetham for providing us with the necessary infrastructure required for the completion of the project. We express our thanks to **Dr.Vidhya Balasubramanian**, Chairperson, Department of Computer Science Engineering and Principal, School of Computing, Amrita Vishwa Vidyapeetham, **Dr.C.Shunmuga Velayutham and Dr.N.Lalithamani,** Vice Chairpersons of the Department of Computer Science and Engineering for their valuable help and support during our study. We express our gratitude to our guide, **Dr. Sikha** for their guidance, support and supervision. We feel extremely grateful to **Dr. T.Senthil kumar**, **Dr. Mansi Sharma**, **Dr. Madhu Sudhanarao Nalluri** and **Mr. A Bhaskar** for their feedback and encouragement which helped us to complete the project. We would also like to thank the entire fraternity of the Department of Computer Science and Engineering. We would like to extend our sincere thanks to our family and friends for helping and motivating us during the course of the project. Finally, we would like to thank all those who have helped, guided and encouraged us directly or indirectly during the project work. Last but not the least, we thank God for his blessings which made our project a success.

N. SATHWIK [CB.EN.U4CSE19037]

P. ROHIT REDDY [CB.EN.U4CSE19042]

M. VAISHNAV [CB.EN.U4CSE19228]

V. TEJASWI [CB.EN.U4CSE19259]

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS

AD	Activity Detector
ASR	Automatic Speech Recognition
GAN	Generative Advesarial Network
LSTM	Long Short Term Memory
RNN	Recurrent Neural Network

Chapter 1

INTRODUCTION

Video summarization is a process that involves selecting and presenting the most informative or captivating content from a longer video, with the aim of providing a concise overview of its contents. The resulting summary typically consists of carefully chosen keyframes or video clips that have been extracted and edited from the source video. The primary objective of video summarization is to enable efficient browsing through large video collections, ensuring that users can quickly assess the video's content and decide whether they want to watch it in its entirety. The effectiveness and accessibility of the video material are key considerations in this process.

To evaluate the quality and informativeness of video summaries, usability studies are commonly conducted. These studies take into account the specific applications and target users of the video summarization system. By assessing user feedback and preferences, researchers can refine and optimize the summarization techniques to better meet the needs of users. This iterative process ensures that the generated summaries are not only informative but also relevant and engaging to the intended audience.

Video summarization is a significant task in the field of computer vision, specifically developed for video analysis. It leverages various techniques to effectively distill the relevant information from videos by extracting keyframes or key-shots. These extracted components serve as representative samples that encapsulate the essential content of the video. By condensing the video's content into a concise summary, video summarization enables efficient information retrieval and enhances the user experience in navigating and exploring video datasets.

In summary, video summarization plays a crucial role in facilitating video content browsing and accessibility. By creating succinct summaries, users can quickly assess the video's content and make informed decisions on whether to invest time in watching

the complete video. Usability studies and targeted research efforts ensure that video summarization techniques are refined and optimized for different applications and user preferences. As a fundamental computer vision task, video summarization effectively distills video information by extracting keyframes or key-shots, making video content more manageable and accessible for users .

1.1 Problem Definition

We aim to develop a model that analyse both visual and audio content and decide what is prominent either audio or video or both to produce a summary and analyse that particular content to give out a summary in the form of a text

1.1.1 Major Contributions

Audio and visuals to text

Here the visual and audio content are extracted and converted into a textual information of the event happening or being discussed.

Generator and Discriminator

The generator is of encoder and decoder with attention mechanisms which generates summary and discriminator is of Long Short Term Memory (LSTM) layers which distinguishes between ground truth and the summary generated by generator.

General Domain

The model is aimed to work on general domain so that we could summarise various kinds of videos belonging to different sectors.

Web Application

We also integrated user friendly interface with the model to provide services.

Chapter 2

LITERATURE SURVEY

A fused representation of audio and visual with a bimodal attention mechanism is used to summarize video and Temporal coherence is preserved but computationally not optimal in the paper[1]. In the paper[2] Lstm and self-attention encoder are used to capture temporal dependencies and video is summarized but Lstms use up large space for video with multiple storylines. There are situations where Bi-Lstm is used to produce key shots and summarize like in [3] but Only visual content is captured. in [10] it mainly concentrates on two issues which are loss of shot term contextual information and inconsistency in summary produced and ground truth .these issues are dealt with by applying attention concept in encoder and decoder to capture short-term context and new loss function to decrease inconsistency of distribution and ground truth. Secondly, they introduced the multi-head self-attention model in order to get the long-range temporal dependencies.

in paper [11] they had taken text-based query as input and generated a video summary based on it. In this approach, there are 3 phases 1. Video summary controller In the beginning, we form a dictionary based on a bag of words that are collected from all the unique words of the training queries. Then, we encode an input query by exploiting the dictionary. After the encoding, we have a vector representation of the input query to represent the expected video summary content. 2. Video summary generator, The main idea of the video summary generator is to take a vector representation of an input text-based query and a video to generate a frame-based relevance score vector 3. Video summary output model. The main idea of this module is to output a video summary based on the relevance score prediction vector.

they initially divided the entire video into different images and passed through the difficulty predictor and each image will be awarded with difficultly score and then the images are sorted according to their difficulty scores and then passed through the dis-

criminator from least difficulty score to highest difficulty score which is real image.

Secondly, random noise vector is passed through a generator and the generator gives the fake images and this fake image given as an input to the discriminator and discriminator tries to differentiate and label and through these labels the video is summarized.

There are models like in [4] that generate video summaries by two uni-modal autoencoders, which embed the video frames and side information, It tries to evaluate based on side info and it's corresponding frames and give gate accuracy. Everything can be observed from different perspective based on the scenario just as in paper [5] where Based on user's query video summary is generated by frame based relevance score labels. Its produces Multiple Summaries are produced from various point of views of users.

A model that can leverage the spatiotemporal information Extractor and a transition effects detection (TED) method to segment the video streams into shots is used in paper[6]. It can capture Motion related information but complete semantic information is missing.

There was a need to capture both local and global inter-dependencies between video frames as in paper [7], so that we get summary closer to an actual event happening the video, even though it is unsupervised method , accuracy was on par with supervised and there are also reinforced algorithm based models like actor critic model with GAN which is also an unsupervised working well as in paper[9] in this paper they had used GAN'S in-order to build the architecture. So firstly they built their framework upon the GAN in an unsupervised manner. Specially the generator produces high level weighted frames features and predict the frame level importance scores, while discriminator tries to distinguish between weighted frame and raw frame features and further they introduced one conditional feature selector in-order to guide the GAN model to focus on the important temporal region from entire video. Ac model and GAN model together this is an unsupervised learning, Ac means actor critic model this follows reinforcement learning that means it rewards the accurate or correct values and penalizes the wrong values

this is what is meant by reinforcement learning. In this model each and every policy is passed into a function and based on the outputs which are getting more rewarded will be sent to a model so that we will get accurate results.

There were many models which majorly used supervised learning with Recurrent Neural Network (RNN) and encode-coders, like in the following paper[12] ,[13] The success of a recurrent neural network (RNN) in sequence processing tasks (e.g., machine translation and text classification), an RNN is introduced to the video summarization task which tackles the video data as a frame sequence and predicts the summary step by step and there are few issues with RNN so to address the issues a tension train hierarchical RNN is used (TTH-RNN).The video data are layered as frames and subshots intrinsically, where the subshot is formed by several frames, and the video is formed by several subshots.The standard RNN is extended from a feed forward network by adding a feedback connection. In this case, LSTM can deal with longer sequences. However, the favorable length is still limited to 80 frames for the video data. As a consequence, the standard LSTM cannot model the frame sequence in the video summarization task.So a tensor train embedding layer will also be present so that the above problem mentioned will be resolved.

There were also few problem like gradient decay and also interpreting and developing network models are difficult so to solve this issues the method in this research paper uses deep reinforcement learning together with independently recurrent neural networks for unsupervised video summarization.In this method a Leaky Rectified unit is used as an activation function to deal with decaying gradient and dying neuron problems. This model, which doesn't rely on any label or user interaction, is designed with a reward function that jointly accounts for uniformity, diversity and representativeness of generated summaries. In this way this model can create summaries as uniform as possible and also have more layers and can be trained with more steps out having any problems with gradients. these were rectified in paper [13].

2.1 Summary

From the literature survey it is evident that most of the papers are on domain specific models, which failed to capture semantic relationship between visual and audio on the general domain and they are quite complex .

2.2 Custom Made Data Set

It consists of sport videos of all categories and Has been made into csv file where its categories and ground truth are listed

DATASET DETAILS

	video_path	category(type)	summary
2	Sports\VolleyBall\sports_volleyball_v1		It is the volleyball game the spiker tried to smash the ball twice but failed and lost the point,
3	Sports\VolleyBall\sports_volleyball_v2		It is the volleyball game it was a super spike where the ball has directly smashed the opponets cort.
4	Sports\VolleyBall\sports_volleyball_v3		It is the volleyball game where game had lot of faints and exchange of rapid shots between two teams and finally spike
5	Sports\VolleyBall\sports_volleyball_v4		It is the volleyball game in this match the defender passed the ball to spiker and spiker smashed the ball
6	Sports\VolleyBall\sports_volleyball_v5		It is the volleyball game in this match the game started with the serve and opponent defended the ball and bump pass t
7	Sports\VolleyBall\sports_volleyball_v6		It is the volleyball game and in this match the defender passes the ball to the striker and striker failed to strike the ball a
8	Sports\VolleyBall\sports_volleyball_v7		It is the volleyball game in this video the team which serve the ball got the point by striking the ball to opponents court.
9	Sports\VolleyBall\sports_volleyball_v8		It is the volleyball game in this match the striker strikes the ball and defender tries to defend the ball but defender could
10	Sports\VolleyBall\sports_volleyball_v9		It is the volleyball game in this match there is a ball out challenge so challenger serves the ball and opponents striker str
11	Sports\VolleyBall\sports_volleyball_v10		It is the volleyball game in this match, the player servers and starts rally and other team through block they get the poin
12	Sports\Cricket\sports_cricket_v1		The batsman hit the ball towards the boundary and few defense shorts in cricket

Figure 2.1: datset structure

2.3 Software/Tools Requirements

2.3.1 ReactJS- Frontend

Java script library based frame work whiich is used develop a vfast and resposive front pages.React.js has gained significant popularity due to its efficient rendering, modular component structure, and active community support. It is widely used for building single-page applications, progressive web applications, and mobile applications using frameworks like React Native. React.js is an open-source JavaScript library developed

by Facebook

2.3.2 Node.js

It has gained popularity for building scalable web applications, real-time applications, APIs, microservices, and more. and it is open source.NPM is the package manager for Node.js and the largest software registry in the world. it is Asynchronous and Event-driven and can be used for Server-side Development.

2.3.3 Pytorch

An open source ML/DL framework based on the Torch library, used for applications such as computer vision and natural language processing, this machine learning was framework developed by Facebook's AI Research lab.

2.3.4 Opencv

OpenCV is widely used in the fields of computer vision, robotics, augmented reality, and machine learning. Its extensive functionality, cross-platform support, and active development community make it a go-to library for a wide range of computer vision tasks like object detection and tracking, feature description, camera calibration .

2.3.5 Tensorflow (2.0)

A open source library used for machine learning applications developed by google and widely used for a variety of machine learning tasks, including image classification, object detection, natural language processing, and reinforcement learning. Its extensive ecosystem, along with its integration with other libraries and frameworks .

Chapter 3

PROPOSED SYSTEM

The proposed architecture operates as follows: It begins with a classifier that takes visual and audio clips as input and classifies them into three categories: visually prominent, prominent in audio, or both visually and audio prominent. Depending on the classification, different processing steps are applied to generate the video summary.

If the content is solely audio-based and the visual information doesn't contribute significantly to the summary, the audio clips are passed through an Automatic Speech Recognition (ASR) system. The ASR system transcribes the audio clips into text embeddings, which capture the textual representation of the audio content. These text embeddings are then used for further analysis and summary generation.

On the other hand, if the video is visually important and the audio content is not considered for the summary, the video frames are sent to an Activity Detection (AD) module. The AD module analyzes the activity being performed in the video by processing the frames and extracting relevant visual features. This analysis results in a string representation that captures the visually significant content of the video.

In cases where both the visual and audio aspects of the video are deemed important for the summary, the string generated from both the ASR and AD modules are concatenated. This combined string is then fed into a Generative Adversarial Network (GAN). The GAN, designed as a reinforcement learning agent, takes the concatenated string as input and generates an abstractive summary of the video. The goal is to produce a summary that effectively captures the key information from both the visual and audio components of the video.

By incorporating classification, ASR, AD, and GAN modules, the proposed architecture aims to leverage both visual and audio cues to generate comprehensive and informative

video summaries. This multi-step process enables the system to adaptively select the most relevant processing steps based on the content characteristics and optimize the summary generation process.

3.1 System Analysis

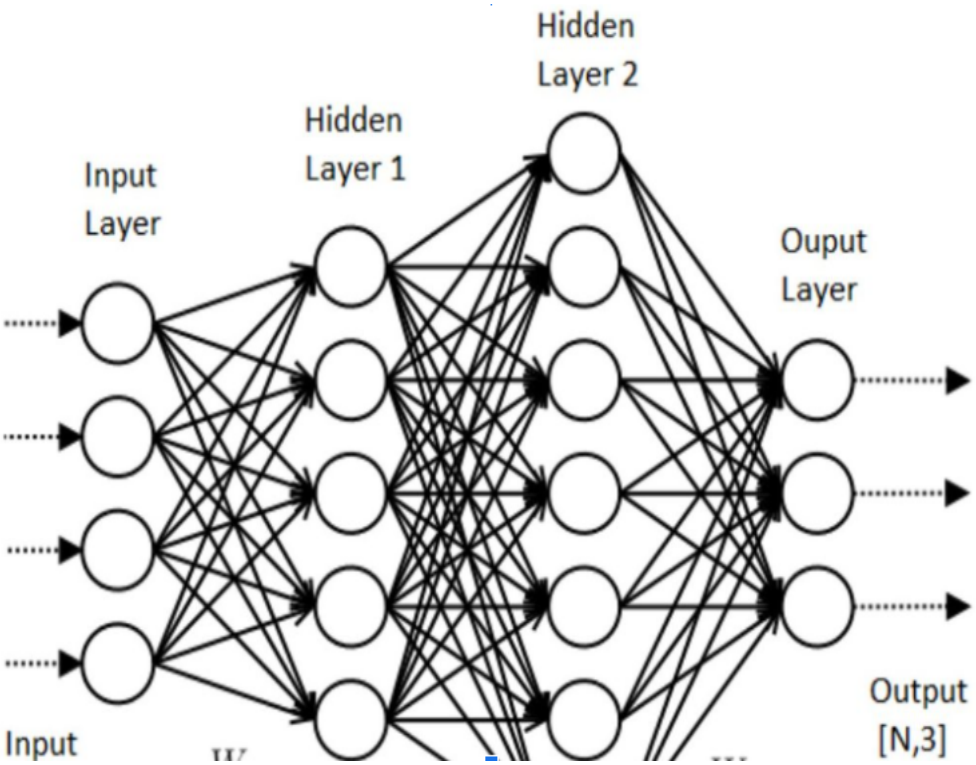

Figure 3.1: generator internal units

This model is used for classifying which path should be chosen to generate video summary so, here first the classifier gets two inputs one input is visual clips and the second input is the audio clip corresponding to that video input. Firstly, the classifier takes ten frames uniformly from the whole video.

These video frames will be minimised to 100 by 100 and the pixel values will be normalised by dividing by 255 and thereby the mean of pixel values is calculated and we would get ten values from 10 different frames and finally form a 10 dimension vector and the same is done for audio where we collect all the wave lengths at 10 uniform interval of time and get ten dimensional vector out of it and concatenate the two vectors to get 2 by 10 dimension vector and finally pass this as input to dense layers to classy

them into three categories which are visually prominent class , prominent in audio and both are prominent so therefore the dense network's output layers has three neurons. the advantages of this method of classifying is it computationally efficient and less costly than using CNN's (Convolutional Neural Networks) and it would consider both visual and audio content indirectly for classification.

3.1.1 ASR

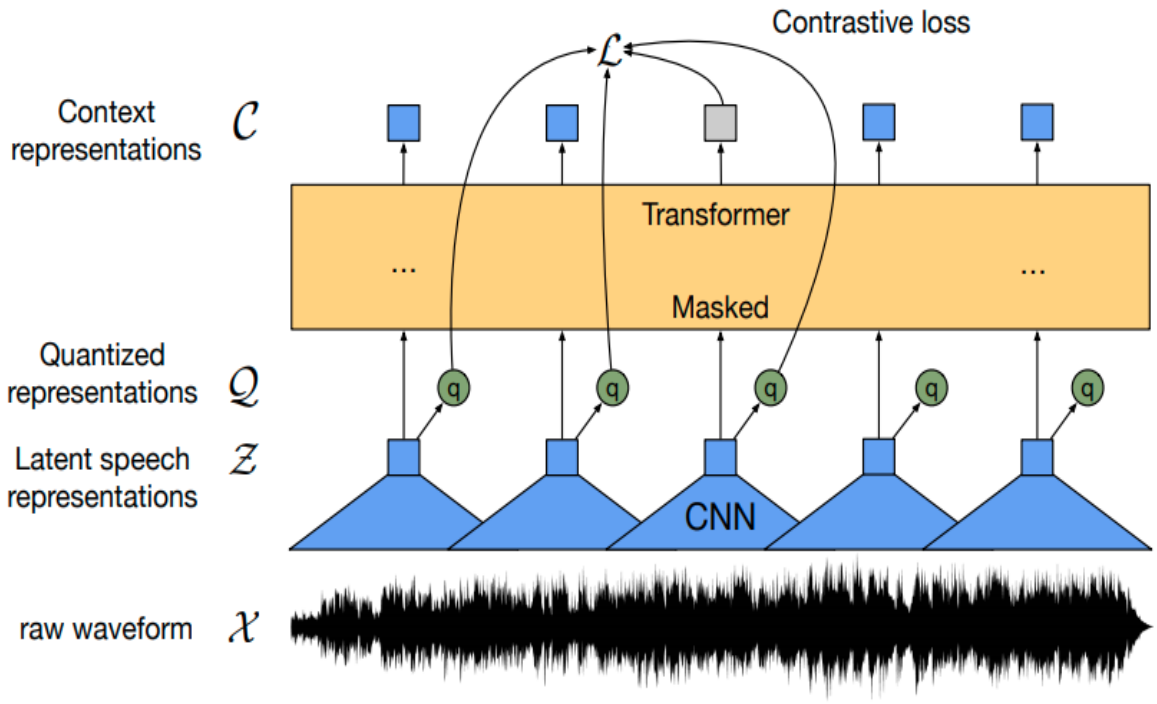

Figure 3.2: ASR , refered from https://arxiv.org/pdf/2006.11477.pdf

Wav2vec 2.0 model ("base" architecture with an extra linear module), is a pre-trained model on 960 hours of unlabeled audio from LibriSpeech dataset [Panayotov et al., 2015] , and fine-tuned for ASR on the same audio with the corresponding transcripts. Experiments using all labeled data of Librispeech achieve 1.8/3.3 WER on the clean/other test sets.it is supervised learning algorithm where the sound frequencies are are converted into spectrograms (i.e by applying log transformation to digital signal oof sound where plotting amplitude and time graph w.r.t sampling rate of a sound). these spectrogram are learned using CNN as feature extractor and are mapped to corresponding letters and finally all the letter are combined into word and words to sentence with the

help of encoders and decoders.

3.1.2 AD

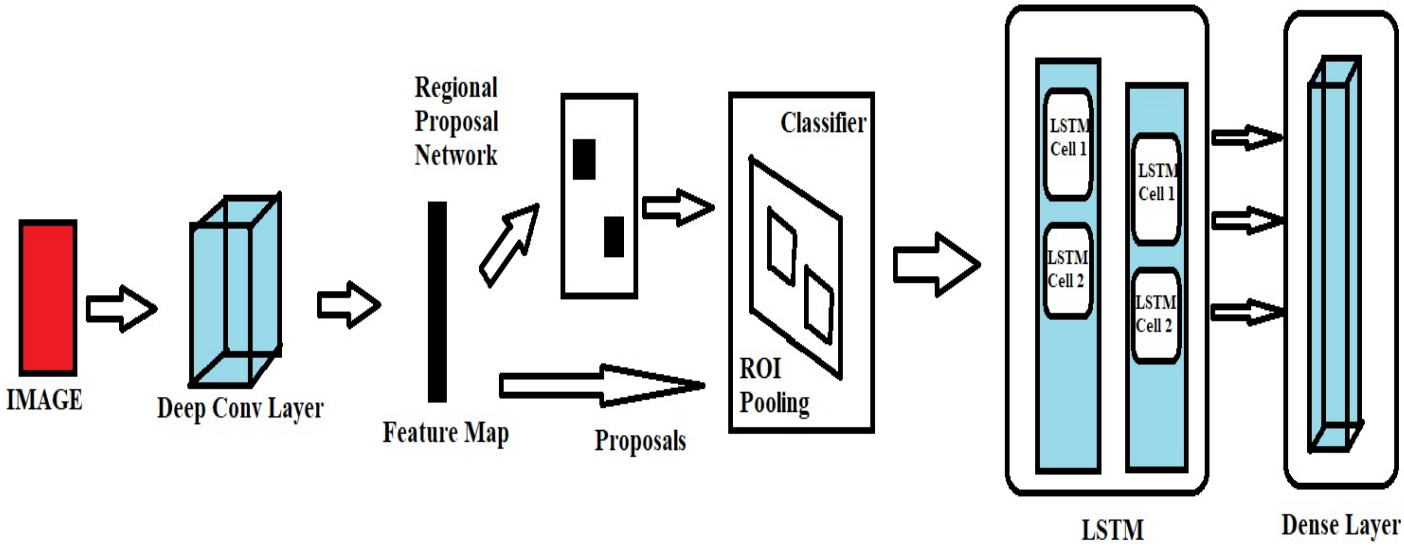

Figure 3.3: AD internal units

The AD comprises of three sub parts namely convolutional layers, Lstm layers and dense layers . It's aim is to detect the activity within the ten frames extracted during in the starting and finally produce a tentative summary sole based on visual content.

ConVolutional layers help in feature extraction where given a frame it ties to give out ann embedding corresponding the object detected in the frames which is similar to RCNN families . and this model has been trained on MS Coco dataset which is most frequently or commonly seen objects in daily life . Later the embedding is sent to lstm layers.

The Backbone of this activity detection model is a faster rcnn architecture here after, The feature is extracted using convolution layers such a way that for every point in the output feature map the network will learn whether the object is in the input image and its corresponding location and size This is done with the help of the RPN which is Region Proposal Algorithm and there after the proposed regions . The ROl pooling layer then

pools features from the backbone feature map based on the bounding box proposals from the RPN. Based on the backbone feature map, ROI pooling basically consists of the following steps: a) Drawing the region associated with a proposal; b) Dividing the region into a fixed number of sub-windows; and c) Pooling the sub-windows to produce a fixed size.

Long-Short-Term-Memory receives the embedding and based on the embedding is trained in such a way that the output obtained will be a a text which comprises of activity being performed with help of the object detected during feature extraction with Convolutional layers.

3.1.3 Generator

```
generator.summary()

Model: "model_3"
_________________________________________________________________________________________________
 Layer (type)                Output Shape              Param #     Connected to
=================================================================================================
 input_7 (InputLayer)        [(None, 100, 100)]        0           []

 input_8 (InputLayer)        [(None, None, 1024)       0           []
                             ]

 encoder (LSTM)              [(None, 1024),            4608000     ['input_7[0][0]']
                              (None, 1024),
                              (None, 1024)]

 decoder (LSTM)              (None, 1024)              8392704     ['input_8[0][0]',
                                                                   'encoder[0][1]',
                                                                   'encoder[0][2]']

=================================================================================================
Total params: 13,000,704
Trainable params: 13,000,704
Non-trainable params: 0
_________________________________________________________________________________________________
```

Figure 3.4: generator internal units

In a Generative Adversarial Network (GAN), the discriminator plays a crucial role as a classifier. Its primary objective is to differentiate between real and generated data produced by the generator. The discriminator's architecture is tailored based on the specific data being classified. In the given context, the discriminator is utilized in the

task of text summarization, where its purpose is to assess the quality and authenticity of the generated summaries.

During the training phase of the GAN, the discriminator receives input from two distinct sources. Firstly, it is trained using real data instances, which refer to actual summaries of videos from the dataset. These real data instances act as positive examples for the discriminator. Secondly, the discriminator is exposed to fake data instances generated by the generator. These generated summaries, aiming to emulate human-generated summaries, serve as negative examples for the discriminator.

The process of training the discriminator with both real and generated data enables the GAN model to enhance the quality of the generated summaries. By accurately discerning between real and generated summaries, the discriminator provides valuable feedback to guide the generator in producing more realistic and higher-quality summaries. This adversarial training process establishes a competitive dynamic between the generator and discriminator, leading to the refinement and improvement of the text summarization model.

Overall, the discriminator's role within the GAN framework is to classify and distinguish between real and generated data. In the context of text summarization, the discriminator is trained using real summaries as positive examples and generated summaries as negative examples. This adversarial training approach drives the generator to create summaries that closely resemble human-generated ones, ultimately enhancing the performance and effectiveness of text summarization systems.

3.1.4 Discriminator

In a Generative Adversarial Network (GAN), the discriminator plays a crucial role as a classifier. Its primary objective is to distinguish between real and generated data produced by the generator. The discriminator is designed with an appropriate network architecture based on the type of data being classified. In the context of the given passage, the discriminator is employed in the text summarization task.

```
GAN.layers[1].summary()
```

```
Model: "sequential_4"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm_1 (LSTM)               (None, 30, 128)           117248

 dropout (Dropout)           (None, 30, 128)           0

 lstm_2 (LSTM)               (None, 128)               131584

 dropout_1 (Dropout)         (None, 128)               0

 dense (Dense)               (None, 32)                4128

 dropout_2 (Dropout)         (None, 32)                0

 dense_1 (Dense)             (None, 2)                 66

=================================================================
Total params: 253,026
Trainable params: 0
Non-trainable params: 253,026
```

Figure 3.5: discriminator internal units

During the training process of the GAN, the discriminator receives data from two sources. Firstly, it is trained using real data instances, which in this case refers to the actual summaries of videos from the dataset. These real data instances serve as positive examples for the discriminator. Secondly, the discriminator is exposed to fake data instances generated by the generator. These generated summaries, which aim to mimic human-generated summaries, are treated as negative examples for the discriminator.

By training the discriminator with both real and generated data, the GAN model learns to improve the quality of the generated summaries. The discriminator's ability to accurately distinguish between real and generated summaries helps guide the generator to produce more realistic and high-quality summaries. This adversarial training process fosters a competitive interplay between the generator and discriminator, leading to the refinement of the summarization model.

Overall, the discriminator in the GAN framework serves as a classifier that distinguishes between real and generated data. In the context of text summarization, the discrimi-

nator is trained using real summaries as positive examples and generated summaries as negative examples. This adversarial training process drives the generator to generate summaries that closely resemble human-generated ones, resulting in improved text summarization performance.

3.1.5 GAN

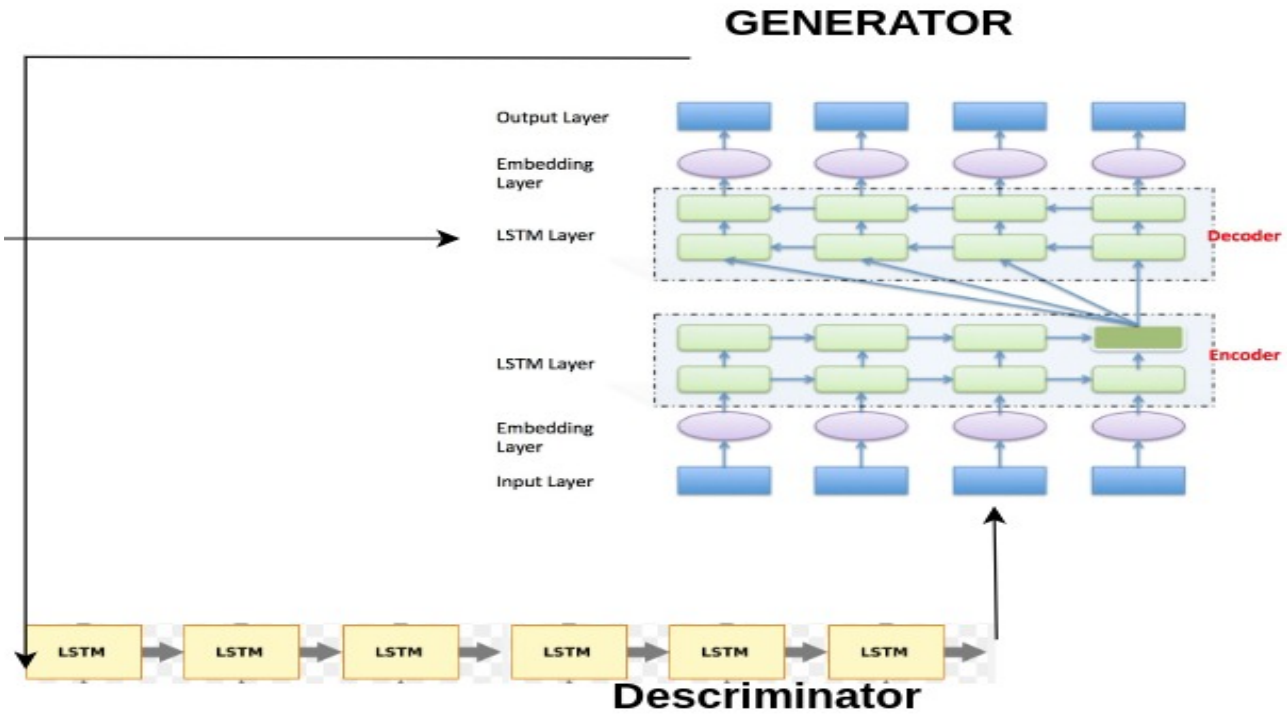

Figure 3.6: GAN

In the context of text summarization, an adversarial process is employed to train a generative model (G) and a discriminative model (D) simultaneously. The generative model, designed as a reinforcement learning agent, takes raw text as input and generates abstractive summaries. The goal of the discriminative model is to distinguish between the generated summaries and the ground truth summaries. By training these models together, the system aims to improve the quality and accuracy of the generated summaries.

The generative model (G) is trained using reinforcement learning techniques, allowing it to learn from rewards and penalties to generate more accurate and informative summaries. The discriminator (D) is trained to differentiate between the generated

summaries and the ground truth summaries. Through an adversarial training process, the generative model continuously improves its performance by attempting to fool the discriminator. This adversarial setup encourages the generative model to produce summaries that closely resemble human-generated summaries.

Experimental results on the CNN/Daily Mail dataset demonstrate the effectiveness of the GAN-based text summarization model. The model achieves competitive ROUGE scores, which are commonly used metrics for evaluating the quality of text summaries, when compared to state-of-the-art methods. This suggests that the adversarial training approach, with the generative and discriminative models working in tandem, can effectively generate high-quality abstractive summaries. Such advancements in text summarization techniques have the potential to enhance the efficiency and accuracy of information extraction from large volumes of text data.

Overall, the use of an adversarial training process, combining a generative model and a discriminative model, shows promise in improving the quality and accuracy of text summarization. The experimental results validate the effectiveness of this approach, and further research in this area may lead to more advanced and accurate text summarization models.

3.1.6 Website Development and workflow

We have developed a website which is very user friendly and we have used the latest technologies like ReactJS for frontend, and FastAPI for developing the API and by using a package and passing the video to the python file and then executing them internally and then finally the summary will be sent to NodeJS server and finally that will be sent to the frontend and the user can see the summary of the video. Total we have 2 parts 1)Frontend 2)FastAPI

Steps:-

- First the user will open the website and then upload video.
- Next that video will be sent to the API.

⇒ **Prototype of front end**

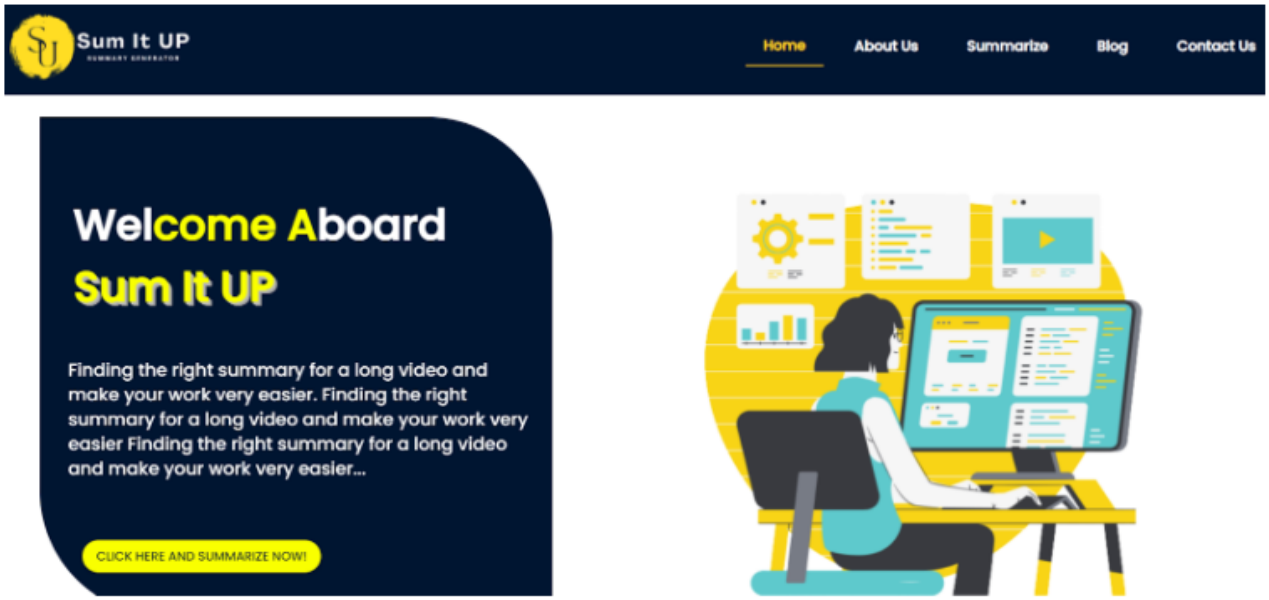

Figure 3.7: web page

- The API prepossesses the video and send it to the model to get the summary in the form of text and the text will returned to the front end .

3.2 System Design

3.2.1 Flow diagram of the system

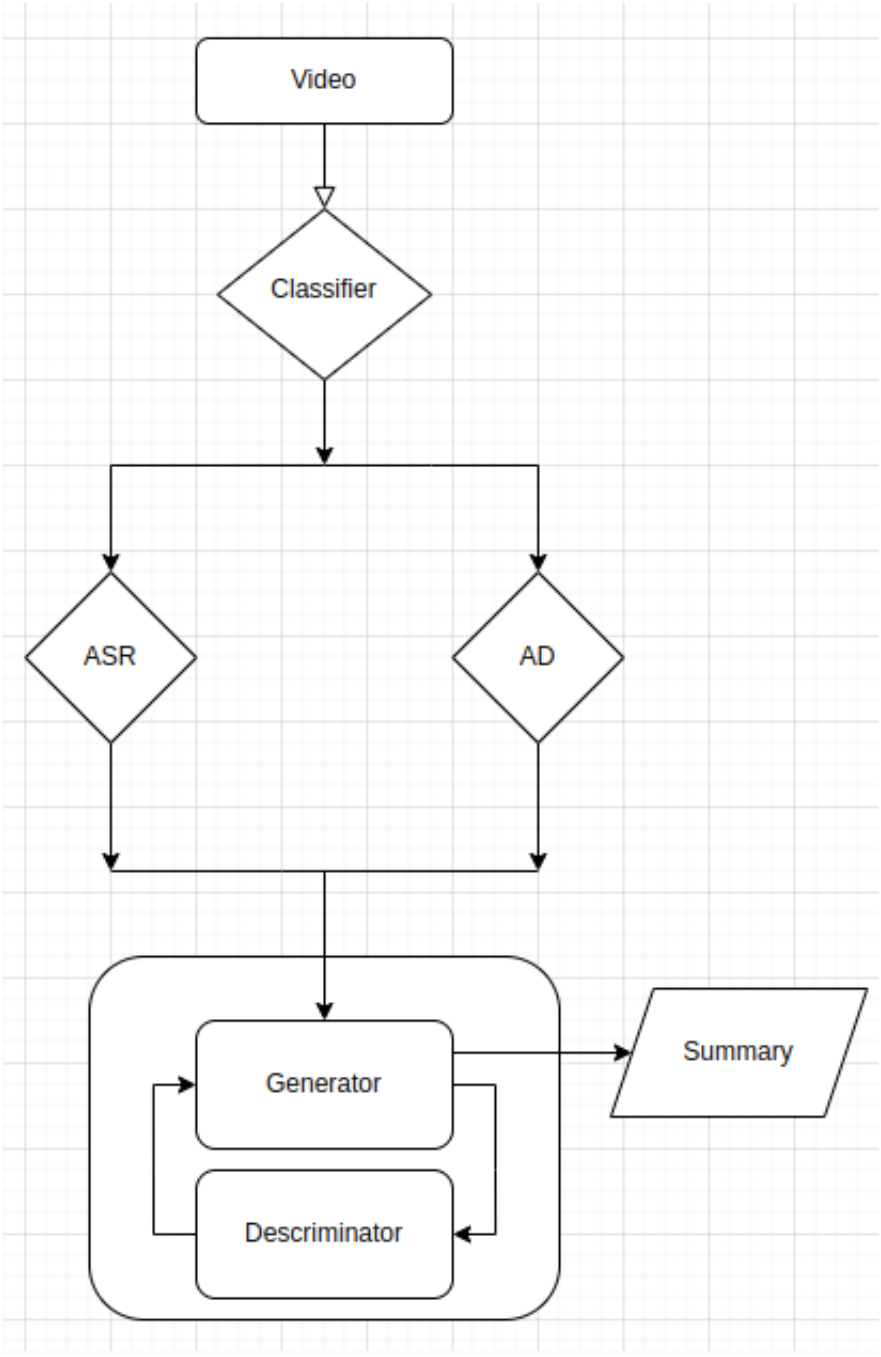

Figure 3.8: Flow Diagram

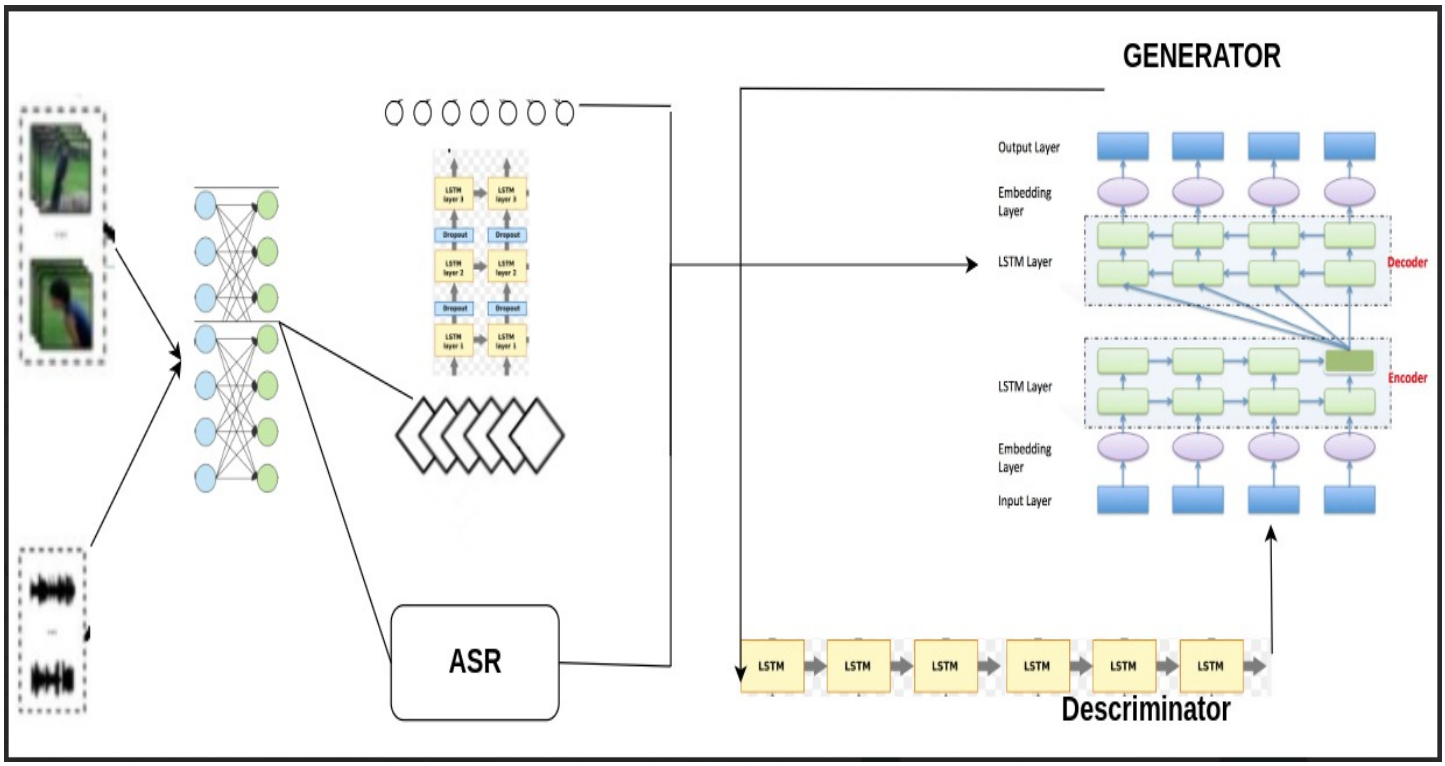

Figure 3.9: Complete Architechture

Chapter 4

IMPLEMENTATION AND TESTING

Here we have test the model on customade dataset and evaluated it with Blue score evaluation matrics and abtained a score of 0.53

The mathematical details

Mathematically, the BLEU score is defined as

BLEU = min(1,exp(1 - reference-length/output-length($\prod_1^4 precision_i)^1/4$.

precision$_i$ = $\frac{\sum snt^i \in Cand-Corpus \sum (i' \in snt)(m^i_{cand}, m^i_{ref})}{w^i_t = \sum_{snt^i \in Cand-Corpus} \sum_{i' \in snt'} m i' cand //}$

where :-

- mi cand is the count of i-gram in the candidate matching the reference translation.
- mi ref is the count of i-gram in the reference translation.
- wi t is the total number of i-grams in candidate translation.

Chapter 5

RESULTS AND DISCUSSION

Sub Module Accuracy

1. classifier accuracy 91.2
2. Activity Dection Accuracy 82.5
3. GAN accuracy 53.8

Overall Model Performance

BLUE Score = 0.53

Output for test data

Figure 5.1: A day of remembrance held in China to honour those who died of coronavirus.The Qingming festival is usually a time when people visit the graves of friends and family.

Figure 5.2: A special vault in Arctic to store thousands of seeds. Scientists fear the impact of climate change, devastating consequences on food crops around the world.

Figure 5.3: Everest's 'worst disaster' in 60 seconds - BBC News":"Everest was affected by the earthquake. The earthquake in Nepal caused Everest's worst ever disaster.The quake caused multiple avalanches across the Himalayas.

Chapter 6

CONCLUSION

Current state of art in visual and audio based video highlight detection uses bimodal attention mechanism which is effective but complex and was not for general domain where as the architecture which we have come up with is light weight and takes in both visuals and audio Therefore we have built an unsupervised AVG(Audio visual GAN) based video summarization model which is lightweight and fast and also works on general domain of videos of time interval 3 minutes to 30 minutes .

Chapter 7

FUTURE ENHANCEMENT

In future we want implement the GAN with inbuilt AD and make it more efficient by introducing attention mechanism in between the encoders and decoder layers and try to increase accuracy and decrease time . we want further extent our scope to the dataset consisting of even you tube shots also .

REFERENCES

[1] T. Badamdorj, M. Rochan, Y. Wang and L. Cheng, "Joint Visual and Audio Learning for Video Highlight Detection," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8107-8117, doi:10.1109/ICCV48922.2021.00802.

[2] B. Zhao, M. Gong and X. Li, "AudioVisual Video Summarization," in IEEE Transactions on Neural Networks and Learning Systems,doi: 10.1109/TNNLS.2021.3119969.

[3] Zhong Ji1 , Kailin Xiong1 , Yanwei Pang1 , and Xuelong Li2 "Video Summarization with Attention-Based Encoder-Decoder Networks"

[4] Y. Yuan, T. Mei, P. Cui and W. Zhu, "Video Summarization by Learning Deep Side Semantic Embedding," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 1, pp. 226-237, Jan. 2019, doi: 10.1109/TCSVT.2017.2771247

[5] Bin Zhao , Member, IEEE, Maoguo Gong , Senior Member, IEEE, and Xuelong Li , Fellow, IEEE “AudioVisual Video Summarization”

[6] Z. Ji, Y. Zhao, Y. Pang, X. Li and J. Han, "Deep Attentive Video Summarization With Distribution Consistency Learning," in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 4, pp. 1765-1775, April 2021, doi: 10.1109/TNNLS.2020.2991083.

[7] . Lei, Q. Luan, X. Song, X. Liu, D. Tao and M. Song, "Action Parsing-Driven Video Summarization Based on Reinforcement Learning," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 7, pp. 2126-2137, July 2019, doi: 10.1109/TCSVT.2018.2860797.

[8] arXiv:2101.06072 [cs.CV]

[9] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris and I. Patras, "AC-SUM-GAN: Connecting Actor-Critic and Generative Adversarial Networks in IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 8, pp. 3278-3292, Aug. 2021, doi: 10.1109/TCSVT.2020.3037883. .

[10] Ji, Y. Zhao, Y. Pang, X. Li and J. Han, "Deep Attentive Video Summarization

With Distribution Consistency Learning," in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 4, pp. 1765-1775, April 2021, doi: 10.1109/TNNLS.2020.2991083.
[11] https://doi.org/10.48550/arXiv.1910.08967
[12] B. Zhao, X. Li and X. Lu, "TTH-RNN: Tensor-Train Hierarchical Recurrent Neural Network for Video Summarization," in IEEE Transactions on Industrial Electronics, vol. 68, no. 4, pp. 3629-3637, April 2021, doi: 10.1109/TIE.2020.2979573.
[13] Yaliniz, G., Ikizler-Cinbis, N. Using independently recurrent networks for reinforcement learning based unsupervised video summarization. Multimed Tools Appl 80, 17827–17847 (2021). https://doi.org/10.1007/s11042-020-10293-x.