TEMPORAL CONVOLUTIONAL NETWORK & CONTENT-BASED FRAME SAMPLING FUSION FOR SEMANTICALLY ENRICHED VIDEO SUMMARIZATION

S.Sumanth[1], T. Charan Durga[2], Ch. Yashwanth Sai[3], Suneetha Manne[4]

[1,2,3,4] Information Technology, VRSEC, Vijayawada, Andhra Pradesh, India.
sumanthsomireddy@gmail.com
thalapalacharandurga@gmail.com
yashwanthsai4266@gmail.com
suneethamanne74@gmail.com

Abstract:
Due to the exponential growth of video data, which makes it difficult to handle massive volumes of video content, video summarizing techniques have drawn a lot of interest recently. Although there are numerous video summary techniques, summarizing lengthy videos is still difficult since it takes a long time to process hundreds of frames. Modern video summarizing techniques that use Temporal Convolutional Networks to determine shot boundaries take a long time to process when dealing with large films. The Distributed Temporal Convolutional Networks method, which takes into account the shot length distribution's properties to effectively summarize lengthy films, is proposed in this paper as a novel solution to tackle this difficulty. By completely using the temporal coherence of lengthy films, the Content-Based Frame Sampling [3] technique is additionally suggested to improve the system's throughput. The fusion approach has a significant implication for improving video summarization techniques and enabling efficient video processing, paving the way for new applications in areas such as video surveillance, entertainment, and education.

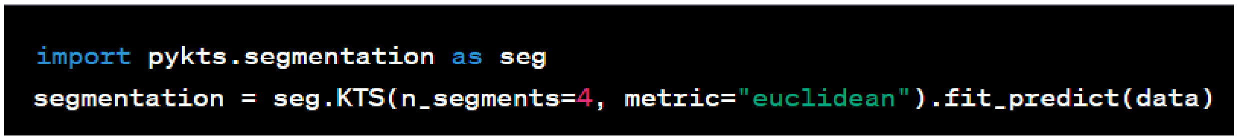
Keywords: Distributed-Temporal Convolution Neural Network, Content – Based Frame Sampling, Bidirectional and Auto-Regressive Transformer (BART).

1. Introduction:

By choosing and presenting the most significant and pertinent portions of the primordial film, video summarization is the process of producing compact version of a longer video. For effective browsing and content navigation, as well as for video analysis, retrieval, and organization, it is a crucial tool. The methods utilized in video summarizing range from automatic algorithms to manual procedures. Manual methods entail watching the full film and picking out the most significant and pertinent sections by hand. Although accurate and dependable, this method is time- and money-consuming [1]. Automatic techniques, on the other hand, rely on algorithms that assess the video footage and extract the most crucial parts based on different criteria, including motion, color.

The keyframe-based approach is one of the most widely used strategies for summarizing videos. This method picks out a group of keyframes from the original video that stand in for the most significant and instructive frames. The original video is then summarized by placing these keyframes in a chronological order. The shot-based technique is another method that chooses the most representative and instructive shots from the movie based on different characteristics, like motion and color. Object-based techniques identify important objects or regions in the video and extract segments that contain these objects.

Contribution:



```
import pykts.segmentation as seg
segmentation = seg.KTS(n_segments=4, metric="euclidean").fit_predict(data)
```

The above line of code snippet illustrates the direct execution of Temporal Convolutional Network [7] in Python. However, in this work, the default functionality of TCN is leveraged by the addition of Content-Based Frame sampling in order to attain a optimized space and time complexity for the task of extraction of key frames in summarization of long videos efficiently.

2. Keyframe-Based Techniques for Summarizing Video:

Among the most widely used methods for summarizing videos are keyframe-based techniques. Normally, characteristics like brightness, contrast, and color are taken into account while choosing the keyframes [5]. Keyframes can be chosen depending on a variety of factors, such as the maximum distance between adjacent keyframes or the permitted number of keyframes. The original video is then summarized by placing the keyframes in a temporal order. Although this method is quick and easy, it might not be able to catch all of the crucial details in reel. Keyframes are selected contingent on a vast set of variables, such the number of keyframes that are allowed or the maximum distance between neighboring keyframes [4]. The keyframes are then arranged in a chronological manner to summarize the original video. This method could fail to able to retrieve all of the important information from a video, despite how quick and simple it is.

3. Shot-Based Techniques for Summarizing Video:

Shot-based summary aims to choose a chosen few shots that best capture the video's content [2]. Segments a video into shots and selects the most representative shots that can capture the key content of the video. Shots are typically identified based on temporal or visual continuity. Typically, elements like motion, color, and texture are used to choose shots. The choice of shots can be made based on a variety of factors, including the maximum number of photos permitted or the similarity of adjacent shots. Because they capture more details about the video, shot-based techniques can be more productive than keyframe -based ones in summarizing length videos efficiently with excessive memory.

4. Object-Based Techniques for Summarizing Videos:

An emerging method of video summary is object-based methodology. These methods locate significant objects or areas in the movie using computer vision algorithms, then they extract the segments that contain those objects. Because they gather more semantic data about the video than other techniques, object-based techniques may be more efficient. They cost more to compute and are more complicated. These techniques use computer vision algorithms to find important objects or locations in the movie, then they extract the segments that contain those objects. Object-based strategies may be more effective since they collect more semantic information about the movie than other techniques [4]. They are more difficult to compute and more expensive These techniques use computer vision algorithms to identify pertinent items or regions in the video and then extract the portions of the movie that contain important information. Since they extract more semantic information from the video than other methods, object-based approaches can be more effective.

5. Architecture Diagram:

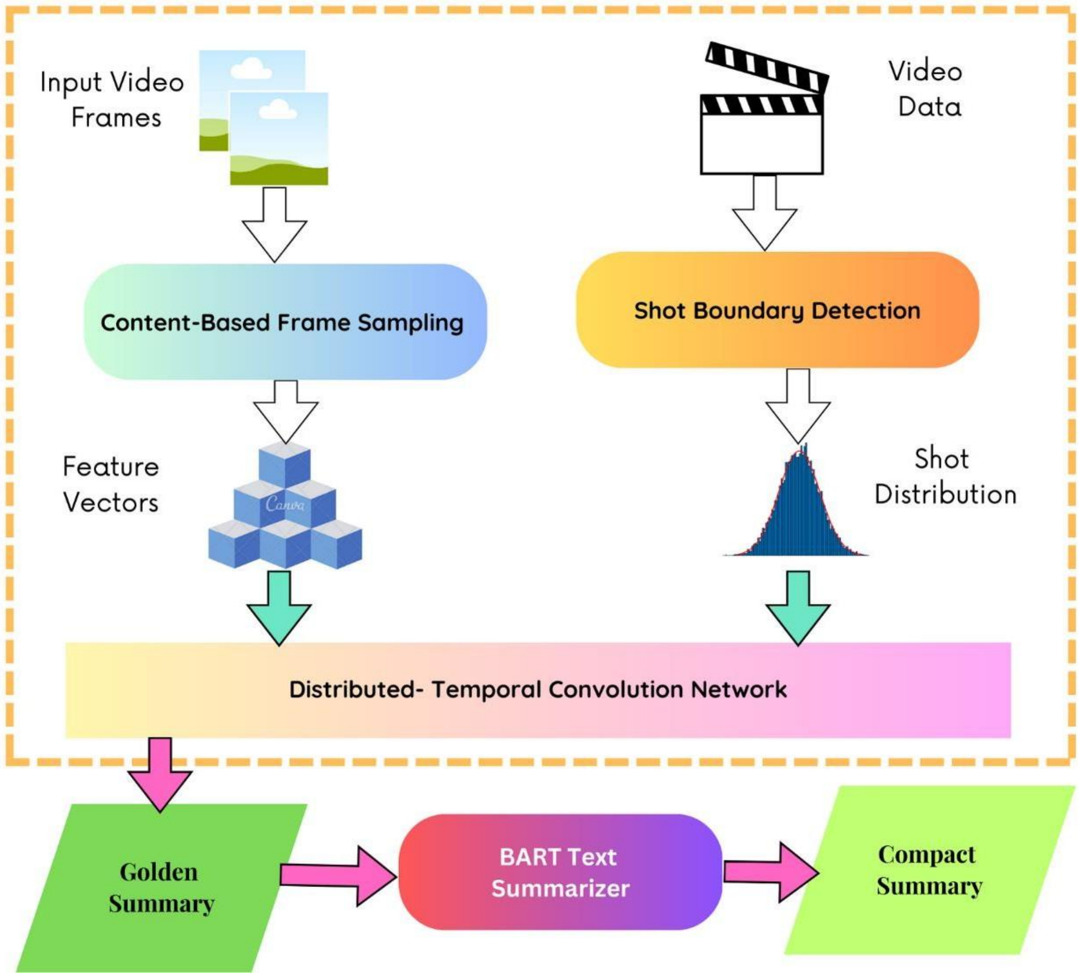


Fig-1: System Architecture Diagram

Fig-1:- The important phases are: (1) Frame Sampling depending on content; (2) Feature Engineering; (3) Temporal Convolution Network; (4) BART Summarization of Golden Summary (5) Compact Summary [8]. In phase of feature engineering, each frame's feature descriptor is stored in the form of vector , Content-Based Frame Sampling is used to accelerate that process. Then the shot distribution information is used to reduce the complexity of the Temporal Convolution Network (TCN). The rank of the frame and golden summary are determined, is generated using ViTAS by HuggingFace API in Python.

6. Methodology:

6.1 Dataset Description:

Datasets from SumMe and TVSum are used for evaluation [11]. SumMe comprises of 25 films that range in length from one and a half to six and a half minutes and cover a variety of topics. Binary scores assigned at the frame level by at most 18 users are displayed underneath each video. On the other hand, TVSum has 50 videos, 5 each category, drawn from 10 categories. The TVSum video lengths range from one to ten minutes. Each film includes shot-level relevance scores that have been commented by 20 users, where every sequence lasts for two seconds.

6.2 Algorithms Usage:

6.2.1 Distributed- Temporal Convolutional Network:

Distributed-Temporal Convolution Network works by clustering video frames into representative keyframes using a kernel-based approach, and then selecting the most relevant keyframes to form the summary [8]

Table-1: Distributed- Temporal Convolution Network (TCN) Algorithm

Algorithm: Temporal Convolution Network

```
Input: X (batch_size, sequence_length, input_dim)
Output: Y (batch_size, sequence_length, output_dim)

Define TCN architecture:
- Number of layers: L
- Filter sizes: [$F_1$ , $F_2$ , ..., F_L]
- Number of filters: [$N_1$ , $N_2$ , ..., N_L]
- Activation function: activation_fn

Initialize TCN layers and parameters
Forward pass:
for each layer l in range(L):
    Apply 1D convolution: convolution_result = convolution(X, W_l) + b_l
    Apply activation function: activated_result = activation_fn(convolution_result)
    Apply temporal pooling: pooled_result = temporal_pooling(activated_result)
    X = pooled_result
```

```
Apply fully connected layers:
flatten_result = flatten(X)
for each fully connected layer m:
    fc_result = dot_product(flatten_result, W_fc_m) + b_fc_m
    flatten_result = activation_fn(fc_result)

Set final output: Y = flatten_result

Return output sequence Y
```

6.2.2 Content-based Frame Sampling:

Content-Based Frame Sampling is a video summarization method that selects keyframes based on their relevance and diversity. It uses a hash function to map the features of video frames into binary codes[9], and then selects representative keyframes based on their hamming distance.

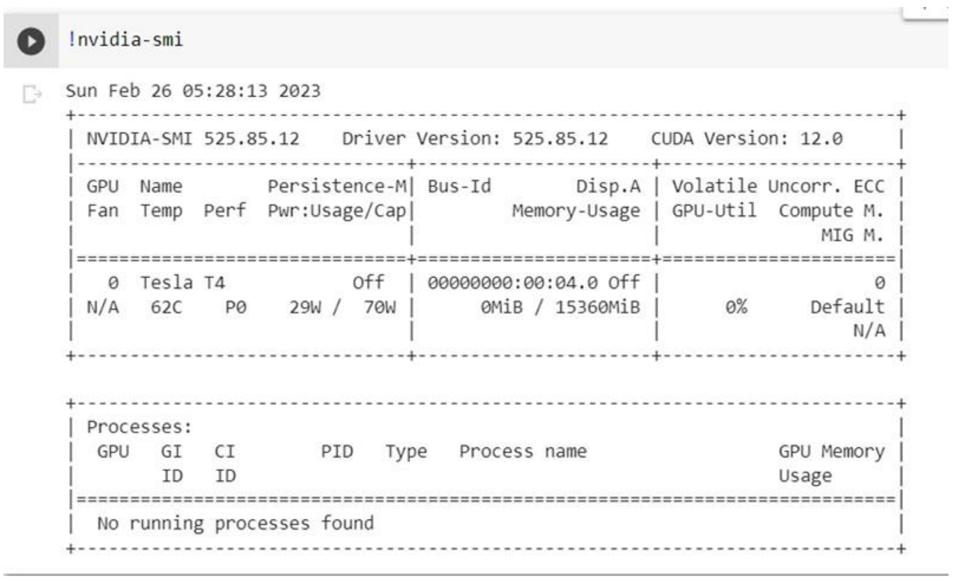
Table-2: Content-Based Frame Sampling Algorithm

Algorithm: Content Based Frame Sampling

```
Input: Video frames (F), Hash function (H), Target number of frames (N)

1. Initialize an empty set to store selected frames: SelectedFrames = {}
2. Compute the hash value for each frame in the video:
   for each frame in F:
       hash_value = H(frame)

3. Sort the frames based on their hash values in ascending order:
   sorted_frames = sort(F, key=hash_value)

4. Compute the interval for frame selection:
   interval = length(F) / N

5. Select frames based on the computed interval:
   for i = 0 to N-1:
       index = floor(i * interval)
       selected_frame = sorted_frames[index]
       SelectedFrames.add(selected_frame)

6. Return the set of selected frames: SelectedFrames
```

7. Performance Evaluation and Results:

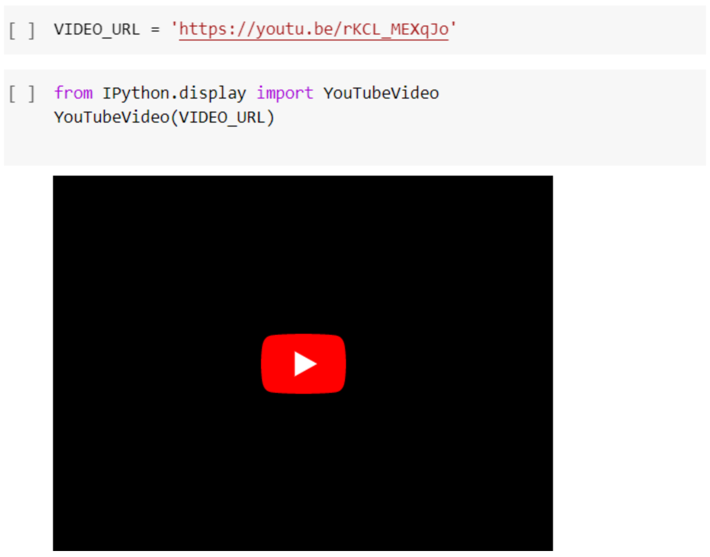
The model is trained on NVIDIA SMI GPU processor 525.85.12 with CUDA version 12.0. The video summarized result is integrated with BART text summarization to improve the compactness of the final summary.



```
!nvidia-smi

Sun Feb 26 05:28:13 2023
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 525.85.12    Driver Version: 525.85.12    CUDA Version: 12.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla T4            Off  | 00000000:00:04.0 Off |                    0 |
| N/A   62C    P0    29W /  70W |      0MiB / 15360MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

Fig-2: GPU Configuration Specifications

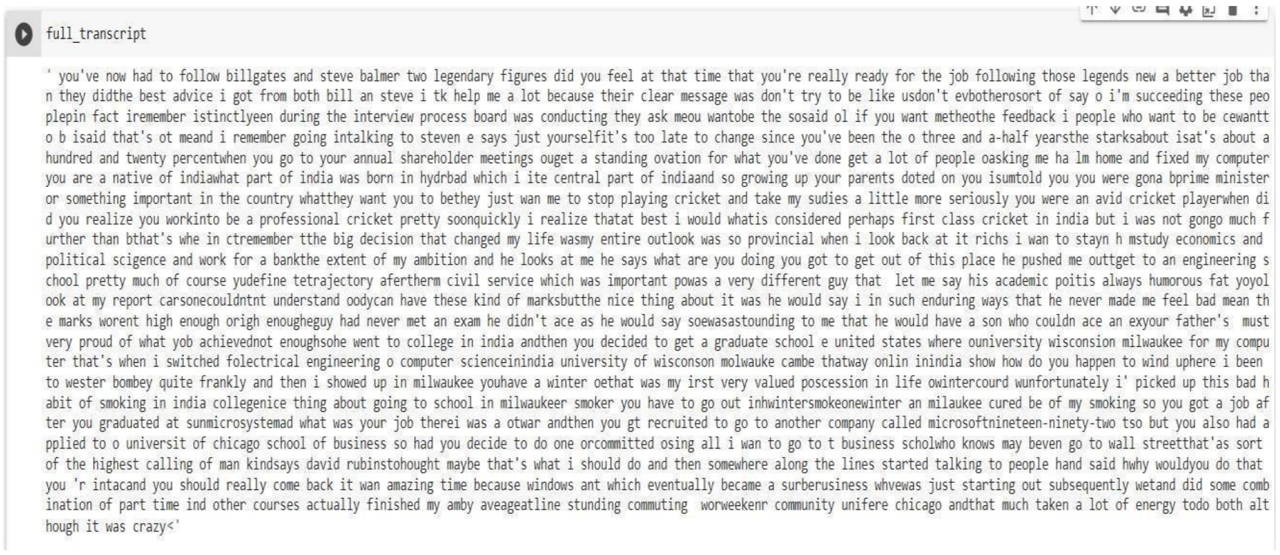
Fig-2 indicates the actual specifications of GPU used to execute the model. The GPU processor 525.85.12 is a powerful graphics processing unit designed for deep learning applications. It is optimized for running deep neural networks and can perform parallel computations at high speed.



```
[ ] VIDEO_URL = 'https://youtu.be/rKCL_MEXqJo'

[ ] from IPython.display import YouTubeVideo
    YouTubeVideo(VIDEO_URL)
```

Fig-3: Video Input

Fig-3 shows that the model takes in a YouTube URL directly to carry out the summarization process. However, there's also a feature of manual upload of any video whose summarization must be generated.

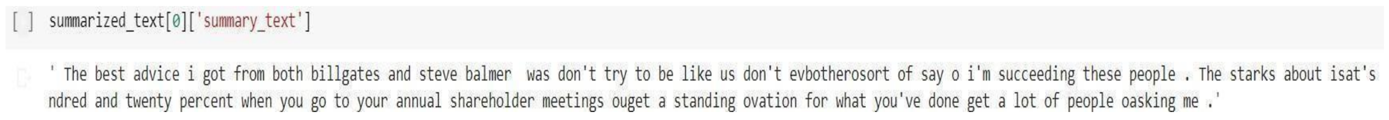


```
full_transcript
```

```
' you've now had to follow billgates and steve balmer two legendary figures did you feel at that time that you're really ready for the job following those legends new a better job tha
n they didthe best advice i got from both bill an steve i tk help me a lot because their clear message was don't try to be like usdon't evbotherosort of say o i'm succeeding these peo
plepin fact iremember istinctlyeen during the interview process board was conducting they ask meou wantobe the sosaid ol if you want metheothe feedback i people who want to be cewantt
o b isaid that's ot meand i remember going intalking to steven e says just yourselfit's too late to change since you've been the o three and a-half yearsthe starksabout isat's about a
hundred and twenty percentwhen you go to your annual shareholder meetings ouget a standing ovation for what you've done get a lot of people oasking me ha lm home and fixed my computer
you are a native of indiawhat part of india was born in hydrbad which i ite central part of indiaand so growing up your parents doted on you isumtold you you were gona bprime minister
or something important in the country whatthey want you to bethey just wan me to stop playing cricket and take my sudies a little more seriously you were an avid cricket playerwhen di
d you realize you workinto be a professional cricket pretty soonquickly i realize thatat best i would whatis considered perhaps first class cricket in india but i was not gongo much f
urther than bthat's whe in ctremember tthe big decision that changed my life wasmy entire outlook was so provincial when i look back at it richs i wan to stayn h mstudy economics and
political scigence and work for a bankthe extent of my ambition and he looks at me he says what are you doing you got to get out of this place he pushed me outtget to an engineering s
chool pretty much of course yudefine tetrajectory afertherm civil service which was important powas a very different guy that  let me say his academic poitis always humorous fat yoyol
ook at my report carsonecouldntnt understand oodycan have these kind of marksbutthe nice thing about it was he would say i in such enduring ways that he never made me feel bad mean th
e marks worent high enough origh enougheguy had never met an exam he didn't ace as he would say soewasastounding to me that he would have a son who couldn ace an exyour father's  must
very proud of what yob achievednot enoughsohe went to college in india andthen you decided to get a graduate school e united states where ouniversity wisconsion milwaukee for my compu
ter that's when i switched folectrical engineering o computer scienceinindia university of wisconson molwauke cambe thatway onlin inindia show how do you happen to wind uphere i been
to wester bombey quite frankly and then i showed up in milwaukee youhave a winter oethat was my irst very valued poscession in life owintercourd wunfortunately i' picked up this bad h
abit of smoking in india collegenice thing about going to school in milwaukeer smoker you have to go out inhwintersmokeonewinter an milaukee cured be of my smoking so you got a job af
ter you graduated at sunmicrosystemad what was your job therei was a otwar andthen you gt recruited to go to another company called microsoftnineteen-ninety-two tso but you also had a
pplied to o universit of chicago school of business so had you decide to do one orcommitted osing all i wan to go to t business scholwho knows may beven go to wall streetthat'as sort
of the highest calling of man kindsays david rubinstohought maybe that's what i should do and then somewhere along the lines started talking to people hand said hwhy wouldyou do that
you 'r intacand you should really come back it wan amazing time because windows ant which eventually became a surberusiness whvewas just starting out subsequently wetand did some comb
ination of part time ind other courses actually finished my amby aveageatline stunding commuting  worweekenr community unifere chicago andthat much taken a lot of energy todo both alt
hough it was crazy<'
```

Fig 4: Golden Summary extracted by ViTAS

Fig-4 shows the golden summary extracted by the model. The key frames are selected using Content Based Frame Sampling and they are segmented using D-Temporal Convolutional Network. The segmented frame samples act as input to ViTAS model of HuggingFace API [10] which results in the generation of golden summary.



```
[ ]  summarized_text[0]['summary_text']
```

```
' The best advice i got from both billgates and steve balmer  was don't try to be like us don't evbotherosort of say o i'm succeeding these people . The starks about isat's
ndred and twenty percent when you go to your annual shareholder meetings ouget a standing ovation for what you've done get a lot of people oasking me .'
```

Fig-5: Compact Summary extracted by BART

Fig-5 shows the final compact summary of the video input which is a result of processing of golden summary by BART Text Summarization model [12,13]. In general, the extractive summary of the video is generated by BART. A pre-trained model of BART is taken, which can be used as a starting point for fine-tuning on specific domains. The pre-trained model is trained on vast amount of training data namely DailyMail and Arxiv [6].

8. Analysis of Performance:

Due to the typical TCN's performance, which takes into account every conceivable interval in the video, it is revealed that large videos cannot be handled professionally or effectively. After examining the shot length distribution, Distributed-TCN is suggested to address the problem brought on by the traditional TCN [6]. The Content-Based Frame Sampling method, here acts as a catalyst to skip some frames while maintaining performance in order to substantially reduce the feature engineering cost.

Table-3: Time and Memory Consumption of Traditional-TCN

Video Length	60	180	300	420
Processing Time	41.25	1057.81	4800.53	130577.74
Memory (MB)	178.25	756.75	1988.13	3902.19

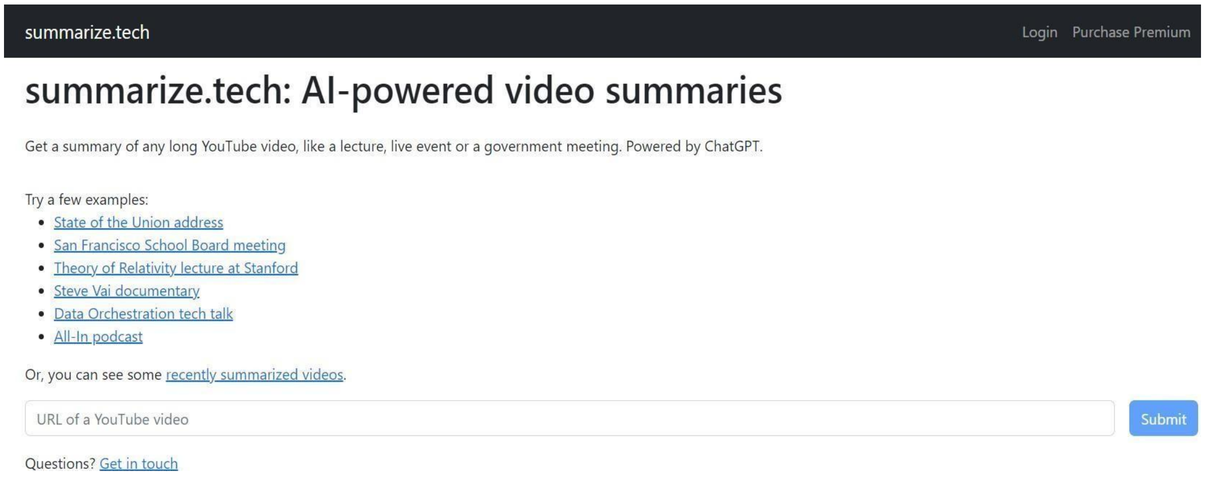
Table-4: Time and Memory Consumption of Distributed-TCN

Video Length	60	180	300	420
Processing Time	21.37	193.77	548.12	1052.04
Memory (MB)	120.41	196.26	281.88	350.61

Table 3 &4 depicts the time and memory consumed for different lengths of videos and compares the parameters between Traditional-TCN and Distributed- TCN. It is clearly evident that for the same lengths of input videos, Distributed-TCN has optimized in terms of time and space complexity of video processing.

9. Result Analysis:

Using the same data, comparison between the compact summaries created by BART and summarize.tech using GPT-4 summarization is caried out to evaluate the suggested methods. To determine the Cosine Similarity Index, Jaccard Similarity Index, and Levenshtein Similarity Index, the model's output is compared to summarize.tech's output. 100 videos are taken in a pipeline for evaluation, and they are then compared using similarity indexes. In summary, the choice of similarity measure depends on the nature of the data being comparedand the type of analysis being performed. Euclidean similarity is suitable for continuous numerical data, Jaccard similarity is suitable for binary or categorical data, and Levenshtein similarity index is suitable for text data with spelling variations or typos.


summarize.tech

summarize.tech: AI-powered video summaries

Get a summary of any long YouTube video, like a lecture, live event or a government meeting. Powered by ChatGPT.

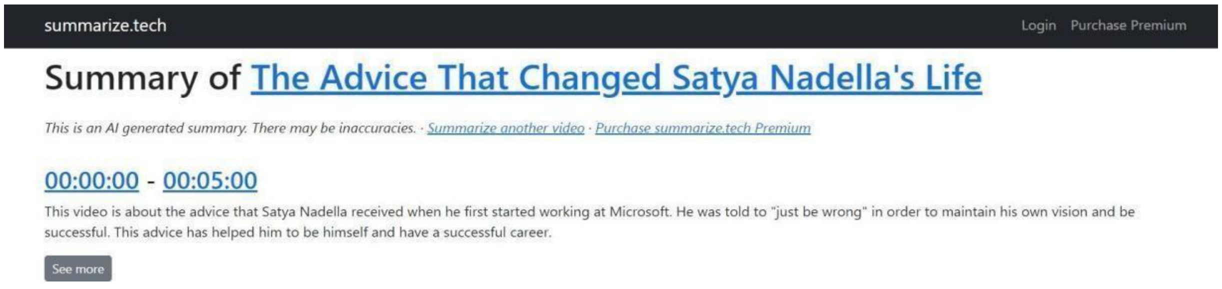
Try a few examples:
- State of the Union address
- San Francisco School Board meeting
- Theory of Relativity lecture at Stanford
- Steve Vai documentary
- Data Orchestration tech talk
- All-In podcast

Or, you can see some recently summarized videos.

Questions? Get in touch

Fig 6: Home page of summarize.tech

Fig 6 shows the overview of summarize.tech portal which is available on the internet. Utilising the GPT-4 algorithm, the online tool Summarize.tech automatically reduces any text, webpage, or video while retaining the key ideas.


summarize.tech

Summary of The Advice That Changed Satya Nadella's Life

This is an AI generated summary. There may be inaccuracies. · Summarize another video · Purchase summarize.tech Premium

00:00:00 - 00:05:00

This video is about the advice that Satya Nadella received when he first started working at Microsoft. He was told to "just be wrong" in order to maintain his own vision and be successful. This advice has helped him to be himself and have a successful career.

Fig 7: Summary generated by summarize.tech

Fig 7 shows the summary generated by summarize.tech for the same video input used in the proposed model.

Table 5: Similarity Index Measures

Model output vs summarize.tech output	**Similaity Index (%)**
Cosine	72.7098
Levenshtein	51
Jaccard	62.7657

An insightful comparison between the compact summary of the proposed model and the summary generated by summarize.tech is done using three similarity indices in Table 5: Cosine Similarity Index, Jaccard Similarity Index and Levenshtein Similarity Index respectively. Considering all the factors namely numerical data, categorical data, spelling mistakes, grammar, typos in terms of the above used similarity indices, we can observe that the model’s output has extracted relevant content in the video input

10. Conclusion and Future Study:

According to the experiment's findings, Distributed-TCN and Content-Based Frame Sampling surpassed the traditional method measured in variables of processing times as well as memory use while compromising just a slight degree of accuracy. These adjustments addressed the problem of processing lengthy videos and were helpful in lowering the resource consumption of video summarizing techniques. The performance bottleneck of cutting-edge video summarizing techniques [14] for processing lengthy movies was addressed by the authors' two innovative optimizations, Distributed-TCN and Content-Based Frame Sampling, which were offered in their conclusion. Using a commonly used video summarizing technique, these optimizations were tested on well-known datasets, and the findings demonstrated that they can reduce memory usage and increase processing speed while compromising a tiny degree of accuracy.

A larger and more balanced training set should help us achieve better results in this area. This model necessitates a large amount of training data, and computing resources. The model must be updated to work on journals as well as generate abstract summary apar from extractive summary. A unified web UI is to be designed for end-users to directly generate the required format of summary. This can be achieved using top cloud computing systems like Microsoft Azure and Amazon Web Services.

References:

[1] X. Ke, B. Chang, H. Wu, F. Xu and S. Zhong, "Towards Practical and Efficient Long Video Summary," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022.

[2] Pang, Zongshang & Nakashima, Yuta & Otani, Mayu & Nagahara, Hajime. (2023). Contrastive Losses Are Natural Criteria for Unsupervised Video Summarization. 2009-2018.

[3] Kaiyang Zhou, Yu Qiao, Tao Xiang, "Deep Reinforcement Learning for Unsupervised Video Summarization With Diversity-Representativeness Reward", AAAI, p., 2018.

[4] W. Zhu, J. Lu, J. Li and J. Zhou, "DSNet: A Flexible Detect-to-Summarize Network for Video Summarization," in IEEE Transactions on Image Processing.

[5] Zhang, Y., Kampffmeyer, M., Liang, X. et al. Dilated temporal relational adversarial network for generic video summarization, (2019).

[6] M. Elfeki and A. Borji, "Video Summarization Via Actionness Ranking," 2019 IEEE Winter Conference on Applications of Computer Vision (WACV).

[7] M. Ma, S. Mei, S. Wan, J. Hou, Z. Wang and D. Feng, "Video Summarization via Simultaneous Block Sparse Representation," 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA).

[8] L. Yuan, F. E. H. Tay, P. Li and J. Feng, "Unsupervised Video Summarization With Cycle-Consistent Adversarial LSTM Networks," in IEEE Transactions on Multimedia, Oct. 2020.

[9] Sheng-hua Zhong, Jiaxin Wu, Jianmin Jiang, Video summarization via spatio-temporal deep architecture, Neurocomputing, 2019.

[10] E. Apostolidis, G. Balaouras, V. Mezaris and I. Patras, "Combining Global and Local Attention with Positional Encoding for Video Summarization," 2021 IEEE International Symposium on Multimedia (ISM), 2021.

[11] 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, April 2021.

[12] Sijia Cai, Wangmeng Zuo, Larry S. Davis, and Lei Zhang. 2018. Weakly-Supervised Video Summarization Using Variational Encoder-Decoder and Web Prior. In Computer Vision – ECCV 2018: September 8–14, 2018.

[13] Sheng-Hua Zhong, Jingxu Lin, Jianglin Lu, Ahmed Fares, and Tongwei Ren. 2022. Deep Semantic and Attentive Network for Unsupervised Video Summarization. ACM Trans. Multimedia Comput. Communication May 2022.

[14] Guoqiang Liang, Yanbing Lv, Shucheng Li, Shizhou Zhang, Yanning Zhang, Video summarization with a convolutional attentive adversarial network, Pattern Recognition, 2022.