International journal of imaging science and engineering

ISSN: 1934--9955 www.ijise.net

Vol-20 Issue-01 April 2025

VIDSUM: VIDEO SUMMARIZATION USING DEEP LEARNING

Kurapati Jyotsna[1], Y. Vamshika[2] ,Pantham Lasya[3] , A. Sravanis[4]
[1,2,3] UG Scholar, Dept of IT, St. Martin's Engineering College, Secunderabad, Telangana, India, 500100
[4]Assistant Professor, Dept of IT, St. Martin's Engineering College, Secunderabad, Telangana, India, 500100
kurapatijyotsna@gmail.com

Abstract:

The aim of video summarization is to create a short summary video which captures the essence of the original video and contains all the important events of the original video. This is helpful because, now we don't have to go through the entire video and are able to get a gist of it from just a short summary video. Current Supervised learning video summarization methods, use Convolutional Neural Networks and some supervised learning techniques use Recurrent Neural Networks in addition to them. We propose VidSum, an architecture for Video Summarization using Deep Learning. We combine Long Short Term Memory (LSTM) Networks with Convolutional Neural Networks to solve the problem of Video Summarization. Our deep learning model is able to find the temporal importance of video frames and is able to generate video summaries which are temporally coherent and contain the important parts of a video clip. In our testing, our model outperforms other models on the famous TVSum and SumMe datasets for the task of Video Summarization.

Keywords: Video Summarization, Deep Learning, Recurrent Neural Networks, Convolutional Neural Networks, Long Short Term Memory.

1. INTRODUCTION

These days, due to the advent of social media and online video services like YouTube and Netflix, we consume and interact with a lot of video data. Every minute, hundreds of hours of video footage is uploaded to the internet. Other than that, due to cameras being present everywhere, ranging from our smartphones to cheap CCTV surveillance cameras, video data has grown multifold in the past decade and continues to grow at a tremendous rate. Therefore, it has become imperative to find suitable video summarization techniques, which capture the crucial segments of a video clip and thus help save human time and effort. Video summarization helps in efficient browsing of large amounts of video data by presenting us with a short summary video. It helps in making videos more informative, impressive and interesting by shrinking the original video to a few significant frames. Video Summarization has many useful real world applications. One very pertinent scenario is video surveillance. It is very time consuming and inefficient for humans to go through 24 hour or even 72 hour video surveillance footage. But using Video Summarization, investigating professionals only need to watch the summary video and don't need to watch hours of video footage to find who committed the crime or any important activity in the long surveillance footage as the Video Summarization model will find the important moments in the surveillance footage of, say, a parking lot for example, and extract the important scenes to include the in the summary video. This would be very helpful in legal matters and would expedite the investigation process, immensely helping the judicial system. Video Summarization can also be used to create automatically generated movie trailers. This would reduce the load on directors and movie editors. It can also be used to create beautiful video summaries of birthday parties and marriages from their original 6-7 hour long videos. Apart from that, video summarization can also be used in creating image and video carousels/ combined short 1-2 minute videos directly on our smartphone from our smartphone's photo gallery, similar to how they are made on our smartphone's photo gallery currently, but the output summary video would be much better. Thus, video summarization is very helpful in the real world and is an ongoing topic of research, with many researchers contributing to it and working on it.

2. LITERATURE SURVEY

T. -J. Fu, S. -H. Tai and H. -T. Chen. (2022) Attentive and Adversarial Learning for Video Summarization Attentive and Adversarial Learning for Video Summarization" focuses on improving the quality of video summarization by leveraging both attention mechanisms and adversarial learning frameworks. Attention mechanisms help in identifying the most relevant frames or segments of a video by focusing on critical features that are representative of the overall content. By dynamically assigning higher importance to key segments, attention models enable more informative summaries. Adversarial learning, on the other hand, introduces a discriminator that ensures the generated summaries.

This, leading to better summarization performance compared to traditional methods. Through the combination of these techniques, the system can adaptively capture significant video content while reducing redundancy.

Kanafani, Hussain, et al. (2021) Unsupervised Video Summarization via Multi-Source Features" explores the use of multiple feature sources to enhance the quality of video summaries without the need for labeled training data. In unsupervised learning, the system autonomously identifies keyframes or segments based on the inherent structure and patterns within the video. By incorporating multi-source features—such as visual, motion, and semantic cues—the model can better capture the diversity of video content. These features help in representing different aspects of the video, leading to a more comprehensive and balanced summary. This approach eliminates the dependency on labeled datasets while ensuring that the summarized content is diverse and informative, making it a highly efficient solution for video summarization.

E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris and I. Patras (2019) Connecting Actor-Critic and Generative Adversarial Networks for Unsupervised Video Summarization AC-SUM- GAN: Connecting Actor-Critic and Generative Adversarial Networks for Unsupervised Video Summarization" integrates reinforcement learning with adversarial learning to improve unsupervised video summarization. By combining the Actor- Critic framework with a Generative Adversarial Network (GAN), the model optimizes both the selection of keyframes and the quality of the generated summaries. The Actor-Critic mechanism helps in learning an optimal policy for selecting keyframes by balancing exploration and exploitation, while the GAN ensures that the generated summaries resemble human- like, coherent video summaries. This connection allows the system to learn robust summarization strategies without requiring labeled data, enhancing the quality and naturalness of the summaries in an unsupervised setting.

W. Zhu, J. LA Flexible Detect-to Summarize Network for Video Summarization. Trance imageu, J. Li, and J. Zhou. (2014) Does Explicit Information Security Policy Affect Employees A Flexible Detect to Summarize and Network for Video. Summarization" introduces a novel approach that integrates object detection with video summarization to enhance the adaptability and precision of summaries. The model uses a "detect-to- summarize" strategy, where it first identifies key objects or regions of interest within the video and then generates summaries based on the detected content. This flexible network architecture allows the system to focus on the most informative parts of the video, reducing redundancy while retaining important details. By combining detection with summarization, the approach adapts to various video domains, ensuring that the summaries are both contextually relevant and concise.

Cisco Systems. (2008b). Data leakage worldwide: The effectiveness of corporate security policies. Retrieved May 12, 2011, and also from the all over World Wide Web: http://www.cisco.com/en/US/solutions/collateral/ns170/ns896/ns 895/Cisco-STL- Data-Leakage-2008-.pdf Data leakage has become a critical global concern, with increasing incidents compromising sensitive information across various industries. The effectiveness of corporate security policies plays a crucial role in mitigating such risks. Research shows that well- implemented policies, including data encryption, access control, and employee training, can significantly reduce vulnerabilities.
However, the rapid evolution of cyber threats and the rise of remote work have exposed gaps in traditional security approaches. Studies highlight that while many organizations adopt robust policies, the lack of continuous monitoring, policy enforcement, and adaptation to new attack vectors often undermines their effectiveness, leading to persistent data leakage challenges worldwide.

3. PROPOSED METHODOLOGY

The proposed video summarization system represents a significant leap forward in automated content condensation technology. By leveraging state-of-the-art deep learning techniques, this system offers highly accurate and personalized video summaries with minimal user intervention. It employs a sophisticated multi-modal approach, analyzing visual, audio, and textual elements of videos to ensure comprehensive understanding of content. The system is designed with scalability in mind, capable of efficiently processing large volumes of data and handling high-definition videos without compromising on speed or quality. Its user-friendly interface makes it accessible to both technical and non-technical users, while its modular architecture allows for easy integration with various video platforms and tools. The system also incorporates adaptive learning capabilities, continuously improving its summarization quality based on user feedback and interactions.

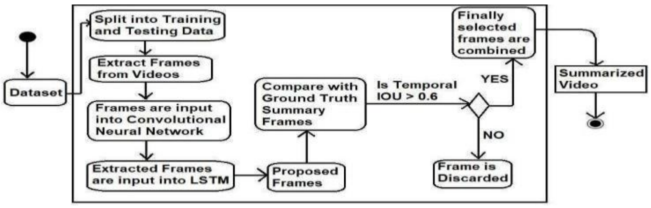
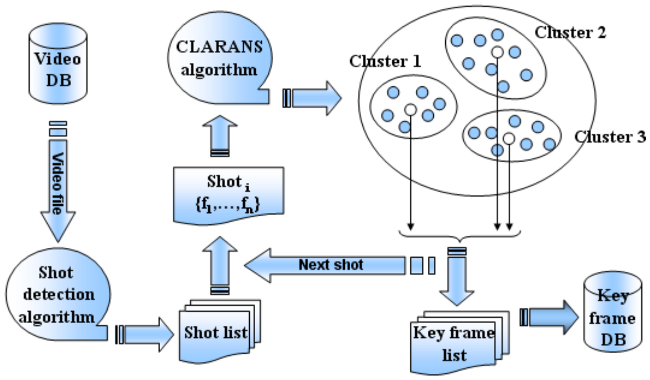

Dataset
Split into Training and Testing Data
Extract Frames from Videos
Frames are input into Convolutional Neural Network
Extracted Frames are input into LSTM
Proposed Frames
Compare with Ground Truth Summary Frames
Is Temporal IOU > 0.6
YES
NO
Frame is Discarded
Finally selected frames are combined
Summarized Video

Figure 1: Proposed System ApproachFlow Chart.

Our proposed solution is to first divide a video into its constituent video frames. These video frames are then entered into the Convolutional Neural Network, where Convolutional Layers extract spatial features from these frames. These extracted features are then fed to the LSTM network, in order to determine time based importance of these frames. After that, each of these images are classified into a yes or no, whether they will be included in the final summary or not by computing their Temporal Intersection over Union with the ground truth. In the last step, all the images which were classified as yes are compiled to create the summary. It has become imperative to find suitable video summarization techniques, which capture the crucial segments of a video clip and thus help save human time and effort. Video summarization helps in efficient browsing of large amounts of video data by presenting us with a short summary video. It helps in making videos more informative, impressive and interesting by shrinking the original video to a few significant frames. Video Summarization has many useful real world applications. This is helpful because, now we don't have to go through the entire video and are able to get a gist of it from just a short summary video.

Any given video consists of a number of video sequences and each video sequence consists of a number of images. So we first divide the video into sequences. Then we extract images from the video. Thus, we get an image representation of a video in the form of all the frames, which the video is made up of. For this we use the OpenCV library of Python. Most CCTV videos are 24fps (frames per second) which means that 1 second of video footage consists of 24 images. Thus 24 images are extracted from every second of video, and are thus put into the model. Our model supports other frame rates like 30fps and 60fps as well.

Architecture:



In order to determine whether a scene is important or not, we need to know what scene happened before that scene. Thus, we have used LSTM and so, our model has "memory" and can remember the previous scenes, so it can better determine which scenes in the video have a higher interest value in the summary. The model then proposes which frames are interesting and important in a time coherent manner with the help of LSTM**.**

Applications:

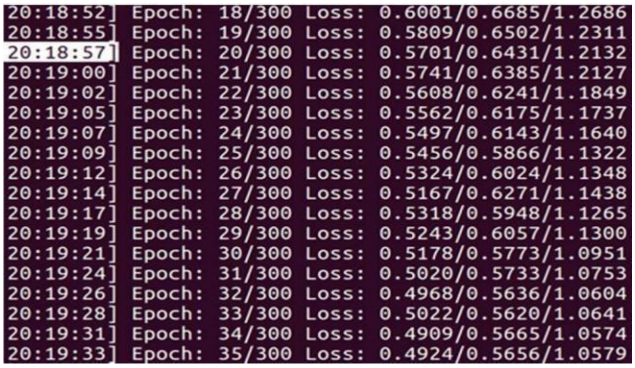
Surveillance Monitoring: Detects and highlights suspicious activities from lengthy CCTV footage**.**
Content Recommendation: Summarizes video content to suggest relevant clips to users.
Healthcare Monitoring: Analyzes and summarizes patient activities from video recordings for health tracking.
Legal Evidence Review: Summarizes hours of surveillance footage for faster legal analysis**.**
Social Media Content: Creates engaging short clips from longer videos for easy sharing**.**
Traffic Management: Summarizes traffic patterns from video feeds to identify congestion.

Advantages:

Simple Shot Boundary Detection: Quickly identifies transitions between scenes, reducing redundant frames for efficient summarization.
Basic Frame Differencing Techniques: Detects frame changes by comparing consecutive frames, allowing extraction of relevant content.
Rule-based Keyframe Extraction: Selects keyframes based on predefined rules, ensuring that important scenes are highlighted.
Unsupervised Clustering for Scene Segmentation: Groups similar frames into clusters without labeled data, enabling automatic scene segmentation and reducing manual effort.

4. EXPERIMENTAL ANALYSIS

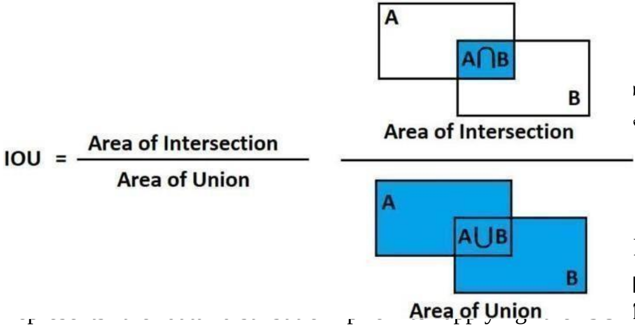
Figure 1 shows custom Machine learning Model made up of Convolutional layers and Maxpool2D layers in between them. The Convolutional Neural Network extracts spatial features from the images and trains on them, so that it can better predict the right frame to be selected for the summary video.



```
20:18:52] Epoch: 18/300 Loss: 0.6001/0.6685/1.2686
20:18:55] Epoch: 19/300 Loss: 0.5809/0.6502/1.2311
20:18:57] Epoch: 20/300 Loss: 0.5701/0.6431/1.2132
20:19:00] Epoch: 21/300 Loss: 0.5741/0.6385/1.2127
20:19:02] Epoch: 22/300 Loss: 0.5608/0.6241/1.1849
20:19:05] Epoch: 23/300 Loss: 0.5562/0.6175/1.1737
20:19:07] Epoch: 24/300 Loss: 0.5497/0.6143/1.1640
20:19:09] Epoch: 25/300 Loss: 0.5456/0.5866/1.1322
20:19:12] Epoch: 26/300 Loss: 0.5324/0.6024/1.1348
20:19:14] Epoch: 27/300 Loss: 0.5167/0.6271/1.1438
20:19:17] Epoch: 28/300 Loss: 0.5318/0.5948/1.1265
20:19:19] Epoch: 29/300 Loss: 0.5243/0.6057/1.1300
20:19:21] Epoch: 30/300 Loss: 0.5178/0.5773/1.0951
20:19:24] Epoch: 31/300 Loss: 0.5020/0.5733/1.0753
20:19:26] Epoch: 32/300 Loss: 0.4968/0.5636/1.0604
20:19:28] Epoch: 33/300 Loss: 0.5022/0.5620/1.0641
20:19:31] Epoch: 34/300 Loss: 0.4909/0.5665/1.0574
20:19:33] Epoch: 35/300 Loss: 0.4924/0.5656/1.0579
```

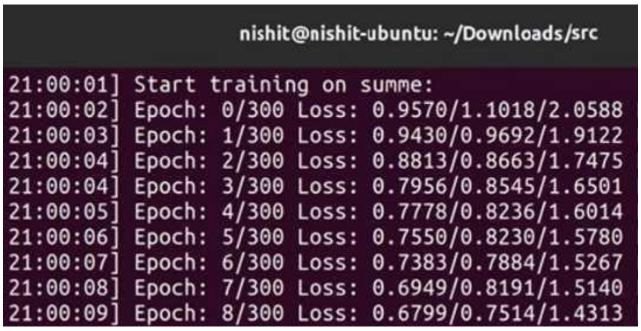
Figure 1: Convolutional Neural Networktraining and extracting

E-learning Platforms: Generates concise summaries of educational videos for quick learning.
Sports Analytics: Extracts key highlights and critical moments from sports events.
News Broadcasting: Summarizes lengthy news reports for quick public consumption**.**
Video Archiving: Compresses lengthy video archives by



unt
kely
the
the
the
hen
plot
ITE
technique, likely used for data analysis or modeling.

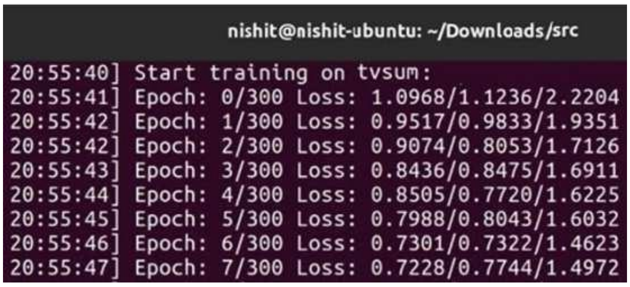
Figure 2: Calculating Temporal Intersection over Union



```
nishit@nishit-ubuntu: ~/Downloads/src
21:00:01] Start training on summe:
21:00:02] Epoch: 0/300 Loss: 0.9570/1.1018/2.0588
21:00:03] Epoch: 1/300 Loss: 0.9430/0.9692/1.9122
21:00:04] Epoch: 2/300 Loss: 0.8813/0.8663/1.7475
21:00:04] Epoch: 3/300 Loss: 0.7956/0.8545/1.6501
21:00:05] Epoch: 4/300 Loss: 0.7778/0.8236/1.6014
21:00:06] Epoch: 5/300 Loss: 0.7550/0.8230/1.5780
21:00:07] Epoch: 6/300 Loss: 0.7383/0.7884/1.5267
21:00:08] Epoch: 7/300 Loss: 0.6949/0.8191/1.5140
21:00:09] Epoch: 8/300 Loss: 0.6799/0.7514/1.4313
```

Figure 3: Training our model on SumMe Dataset

Figure 3 shows that the We trained the model for 10,000 iterations for 30 hours. We made changes to the model multiple times and trained the model again after making changes to increase its accuracy.



```
nishit@nishit-ubuntu: ~/Downloads/src
20:55:40] Start training on tvsum:
20:55:41] Epoch: 0/300 Loss: 1.0968/1.1236/2.2204
20:55:42] Epoch: 1/300 Loss: 0.9517/0.9833/1.9351
20:55:42] Epoch: 2/300 Loss: 0.9074/0.8053/1.7126
20:55:43] Epoch: 3/300 Loss: 0.8436/0.8475/1.6911
20:55:44] Epoch: 4/300 Loss: 0.8505/0.7720/1.6225
20:55:45] Epoch: 5/300 Loss: 0.7988/0.8043/1.6032
20:55:46] Epoch: 6/300 Loss: 0.7301/0.7322/1.4623
20:55:47] Epoch: 7/300 Loss: 0.7228/0.7744/1.4972
```

Figure 4: Classification report of RFC

Figure 4 shows the We trained the model for 10,000 iterations for 30 hours. We made changes to the model multiple times and trained the model again after making changes to increase its accuracy. We added a few dropout layers to stop overfitting. This was done in order to make the model more generalized and so that it is better able to learn the important segments in a video clip.

5. CONCLUSION

this paper presents our framework for video summarization. Contrary to current video summarization models that learn every video frame's importance score only, our video summarization model thinks of video summarization as a temporal-interest detection problem. This helps it to learn temporal coherence between the frames of a video and thus it performs much better than current supervised models at the task of video summarization. In the end, our model became robust and was able to successfully determine which segments of a video footage are important to be included in the summary video and the model was able to separate them out from the rest of the video clip. We were able to improve accuracy for the problem of video summarization as our model was able to create a good summary video containing all the important parts in a video clip. In the future, we will attempt to incorporate more interest1based coherence into our unified framework. We could also attempt to increase the temporal coherence between scenes in a video, which could improve the performance of our model.

The VIDSUM: deep learning approach to video summarization presents our framework for video summarization. Contrary to current video summarization models that learn every video frame's importance score only, our video summarization model thinks of video summarization as a temporal-interest detection problem. This helps it to learn temporal coherence between the frames of a video and thus it performs much better than current supervised models at the task of video summarization. In the end, our model became robust and was able to successfully determine which segments of a video footage are important to be included in the summary video and the model was able to separate them out from the rest of the video clip. We were able to improve accuracy for the problem of video summarization as our model was able to create a good summary video containing all the important parts in a video clip. In the future, we will attempt to incorporate more interest1based coherence into our unified framework. We could also attempt to increase the temporal coherence between scenes in a video, which could improve the performance of our model.

REFERENCES

[1] T. -J. Fu, S. -H. Tai and H. -T. Chen, "Attentive and Adversarial Learning for Video Summarization," 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2019, pp. 1579- 1587, doi: 10.1109/WACV.2019.00173.

[2] Kanafani, Hussain, et al. Unsupervised Video Summarization via Multi-Source Features. arXiv, 26 May2021.arXiv.org, https://doi.org/10.48550/arXiv.2105.12532.

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

[4] W. Zhu, J. Lu, J. Li, and J. Zhou. 2021. DSNet: A Flexible Detect-to1Summarize Network for Video Summarization. Trans. Img. Proc. 30 (2021), 948–962.

[5] S. Lan, R. Panda, Q. Zhu, A.K. Roy-Chowdhury. (2018). "FFNet: Video Fast-Forwarding via Reinforcement Learning". 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6771- 6780.

[6] K. Zhou, Y. Qiao, and T. Xiang. 2018. "Deep reinforcement learning for unsupervised video summarization with diversity1representativeness reward". In Proceedings of the Thirty-Second AAAI

Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'18/IAAI'18/EAAI'18). AAAI Press, Article 929, 7582–7589.

[7] P. Papalampidi, F. Keller, and M. Lapata, "Movie Summarization via Sparse Graph Construction", AAAI 2021, 2021 [8] Fred Hohman, Sandeep Soni, Ian Stewart, and John Stasko, 2017, A Viz of Ice and Fire: Exploring Entertainment Video Using Color and Dialogue,
VIS4DH Workshop 2017

[8] Arun Balajee Vasudevan, Michael Gygli, Anna Volokitin, and Luc Van Gool, 2017, Query-adaptive Video Summarization via Quality-aware Relevance Estimation, MM '17: Proceedings of the 25th ACM International Conference on Multimedia, pp. 582–590, https://doi.org/10.1145/3123266.31232972017

[9] Mrigank Rochan, Linwei Ye and Yang Wang, 2018, Video Summarization Using Fully Convolutional Sequence Networks, ECCV 2018, 2018 [11] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino, 2019, Summarizing Videos with Attention

[10] Shruti Jadon, and Mahmood Jasim, 2020, Unsupervised video summarization framework using keyframe extraction and video skimming, 2020 IEEE 5th International and the more Conference on Computing Communication and Automation (ICCCA), 2020, doi: 10.1109/ICCCA49541.2020.9250764

[11] Yair Shemer, Daniel Rotmanand, and Nahum Shimkin, 2019, the ILS1SUMM: Iterated local search for unsupervised for the video summarization, 2020 25th International Conference on Pattern Recognition (ICPR), 2020, doi:10.1109/ICPR48806.2021.9412068

[12] Jia-Hong Huang, and Marcel Worring, 2019, Query controllable on video summarization, ICMR '20: Proceedings of the 2020 International Conference on Multimedia