RED-SM: A reinforced encoder–decoder transformer for unsupervised video summarization

Venkatachalam ARULKUMAR[1]*, Rajendran LATHAMANJU[2], Rajamani THANGAM[2], Krishnamoorthy DURGA DEVI[3]

[1] Department of Computer Science and Business Systems, Easwari Engineering College, Ramapuram, Chennai, Tamil Nadu, India

arulkumar.v@eec.srmrmp.edu.in (*Corresponding author)

[2] Department of Electronics and Communication Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India

lathamar@srmist.edu.in, thangamr1@srmist.edu.in

[3] Department of Electronics and Communication Engineering, KGiSL Institute of Technology, Coimbatore, Tamil Nadu, India

durgadevi.k@kgkite.ac.in

Abstract: The rapid growth of the video content across online platforms has made it increasingly important to generate concise summaries that help users quickly understand and navigate long videos. However, creating high-quality video summaries typically requires large amounts of annotated data, which is costly and often unavailable. To address this challenge, the authors propose a fully unsupervised approach to video summarization built on Transformer architectures. The method introduces the Reinforced Encoder-Decoder Summarizer Model (RED-SM), which uses multi-head self-attention and feature extraction to identify informative video segments without human labels. RED-SM incorporates sparsity-promoting penalties and a reinforcement learning reward that balances diversity, representativeness, and temporal smoothness to guide frame selection. To further enhance the summarization quality, the RED-SM with a BERT-based text extractor is integrated, enabling multimodal fusion of visual and textual cues. The approach is evaluated on the SumMe and TVSum datasets, as well as a newly curated dataset of 30 categories of short videos. The experiments show that the method consistently produces concise and high-quality summaries across diverse domains. These results highlight the RED-SM as an effective and scalable solution for unsupervised video summarization in real-world applications.

Keywords: Stochastic Optimization, Reinforcement Learning, Video Summaries.

1. Introduction

Video summarization is increasingly important for structuring a massive amount of video resources from multiple platforms for better understanding. It involves compressing long video content to brief, yet informative summaries for efficient content searching, retrieving, and understanding. Supervised methods learn to summarize from a large corpus of human-annotated video summaries, whereas unsupervised methods aim to identify important segments automatically without such labels. The supervised methods learn to summarize from a large corpus of human-annotated video summaries, whereas unsupervised methods aim to identify important segments automatically without such labels. Classical methods often relied on hand-crafted features and clustering techniques. These approaches aim to identify important segments of a video computationally in the absence of any human description. But there are still open problems in conveying the relevance and semantic compactness of a video material.

Varieties of techniques for unsupervised video summarization have been proposed in the literature. For instance, Hong & Zhong (2021) integrates reinforcement learning, attention, and a bidirectional recurrent neural network (BRNN) for video summarization. It fuses BRNN for bi-directional sequential analysis and self-attention for generating an importance score with pre-trained CNN features. It has been shown that these LSTM-based approaches did not sufficiently account for context information and long-term dependencies between video frames. Several methods or approaches of unsupervised summarization of videos have been proposed in the literature. For instance, for video summarization, (Hong & Zhong (2021)) incorporates reinforcement learning,

https://doi.org/10.33436/v35i4y202506

attention mechanisms, and bidirectional recurrent neural networks (BRNN). It leverages the precomputed features of CNN and employs the bidirectional sequence processing in conjunction with the BRNN and self-attention to compute significance scores. As for effective contextual information capturing and long-range dependencies modelling across video clips, most of the LSTM-based methods have met challenges. Other works have explored GANs, reinforcement learning (RL), feature learning approaches, and self-supervised learning to unsupervised learning as well. However, there are still shortcomings that remain in those methods, such as failure to cover diverse content, challenges in training on long video inputs, and challenges to maintain the temporal coherence and representativeness in the summaries. Furthermore, numerous approaches are computationally expensive and rely on intricate architectures.

The authors address these problems with an original method for unsupervised video summarization in this study. The proposed framework integrates a modified Transformer architecture and feature extraction, which is demonstrated to be superior to LSTM in modelling the contextual information and long-range dependency. An effective summarization is achieved without annotated data by exploiting multi-head attention and encoder-decoder attention in the Transformer encoder and decoder layers. So as to encourage to generate the compact summaries, penalty terms are introduced into the present training time calculation of probabilities for attributes produced from frames. Frames are selected by maximizing rewards representing the three perspectives of diversity, representativeness, and temporal smoothness. Moreover, the authors attempted to mix their Red-Sum reinf-encoder decoder summarizer model with a model to provide a more comprehensive description of the video by leveraging both textual and visual cues. The latter focuses more on learning important clues from the source video's audio-based subtitles. The proposed approach has several advantages over the state-of-the-art techniques, including the efficiency for maintaining the temporal coherence and the representativeness in summaries, the simplicity for training, and the robustness for obtaining diverse information. Their methodology and experimental results are described in this paper, demonstrating that this approach indeed can produce clear and instructive video summaries on a broad range of datasets, including SumMe and TVSum. To enable transfer learning,a dataset that contains movies in 30 categories from the public domains is also generated.

The majority of videos cover a wide range of subjects and interests and last between three and eight minutes. Because of its diverse material, it is a valuable tool for research, analysis, and content production. It provides information on viewer trends and preferences across several categories using publicly accessible movies.

The remainder of this paper is structured as follows. Section 2 provides a comprehensive review of related work in unsupervised video summarization. Section 3 details the proposed RED-SM methodology. Section 4 presents the experimental setup, datasets, and results with a detailed discussion. Finally, Section 5 concludes the paper and suggests avenues for future research.

2. Literature survey

The following study provides a further account specific to recent advances in unsupervised video summarization and proposes a method that leverages GANs for summarizing videos.

GAN-based approaches

Mahasseni, Lam & Todorovic (2017) approach solves the problem of extracting a small number of video frames that well represent the original video content. The key idea is to reduce the gap between the original films and film summaries by training a summarizer network to identify the most essential frames and encode them. This is achieved by injecting a discriminator network within the architecture, which enables unsupervised learning. While the discriminator pulls apart the original and constructed movies, the summarizer selects, encodes, and decodes frames into the final video. It accomplishes an optimal frame selection process by making the discriminator unable to discriminate between them through adversarial training. The significant improvement brought by the method to video summarization comes at a cost of the enormous computational burden due to the complexity of training the discrimination and summarization networks jointly. The GAN-enabled methods may also struggle to handle a diverse range of video content with unequal saliencies.

Yuan et al. (2019) summarize the key papers and introduce a method for unsupervised video summarization, which is called Cycle-SUM. Cycle-SUM adopts a particular architectural structure, which is referred to as a cycle-consistent adversarial LSTM. Such is its architecture that it emphasizes on retaining the vital aspects and keeping it as short as possible for the summary video. Cycle-SUM consists of two main components, i.e., a frame selector and a learning-based evaluator, which jointly search for the frames with the key information for the summary. In addition, Cycle-SUM employs two GANs for reconstructing the video in both the forward and backward directions, so that the summary retains the important information from the original content. It is operand to note that Cycle-SUM could be computationally expensive since it depends on complex topologies and adversarial training.

Reinforcement learning methods

Zhou, Qiao & Xiang (2018) summarize the key papers and introduce an unsupervised method using reinforcement learning with a diversity-representativeness Reward. It employs a Deep Summarization Network (DSN) to predict frame selection probabilities and a unique reward function to enhance diversity and representativeness in summaries. Their experiment results show superiority over any previous unsupervised methods and comparability to many supervised methods. The paper significantly contributes by introducing a reinforcement learning-based framework, extending it to supervised learning, and validating the approach through extensive experiments, marking the first application of reinforcement learning to unsupervised video summarization. This approach is evaluated on popular datasets such as SumMe and TVSum.

The problem of keyframe repetition during video summarization is explored in Hong & Zhong (2021), which is particularly common in user-generated videos with a variety of backgrounds and frequent shot changes. In response, the authors propose a spatial attention model designed to focus on large, moving objects in frames in a way that aligns with the human attentional patterns. Their model refines the keyframe selection probability by integrating features processed by a Bi-LSTM network, which consists of two LSTMs processing data in forward and backward directions to enhance contextual understanding. A notable aspect of this VSFB model is its use of an unsupervised diversity-representativeness reward for frame selection, unlike the approaches relying on supervised learning with human-annotated datasets.

Phaphuangwittayakul et al. (2021) integrates three key mechanisms: self-attention network, reinforcement learning, and bidirectional recurrent neural network (BRNN). The summarization network gathers frame details from the given video frames using a CNN network. Two different networks, the BRNN and the self-attention network, use these features as input. The self-attention network provides additional attention qualities for calculating the priority score as it enhances the summarization network. BRNN processes sequential data in both forward and backward directions. The reinforcement learning aspect of the model focuses on optimizing the actions in order to select the output frames. The model learns to describe the video in adversarial training without mode collapse. In addition, the attention map vector created by the self-attention network allows the summarization network to retain critical frames for the video's final summary. The account of shifting camera positions, focus changes, and slight frame composition changes during recording enhances the fairness and robustness of the summarization method. The shot-level semantics focus on the content of the scenes, the angles of the cameras, and the motions within the scenes, which curb user bias and provide a solid framework for the summarization algorithms. The objectivity and stability in this approach, the system used to create representative and diverse extracts, enhance the system's ability to create robust and diverse summaries. The goal of the shot-level semantics is to minimize user subjectivity in summarizing videos to construct objective and diverse summaries that capture the critical concepts of the input videos. Focusing exclusively on shot-level semantics may result in the overlooking of important contextual or visual clues in the video. In order to appreciate the nuances and details that contribute to the overarching narrative and thematic elements of the video, one must pay attention to these subtleties and clues.

Attention and transformer methods

Zhang et al. (2018) delves into the challenge of extracting significant segments from videos for unsupervised video summarization. It addresses two major obstacles: ineffective feature learning

and the complexity of training with lengthy video inputs. To tackle these issues, the authors introduce a variance loss to enhance feature learning effectiveness and unveil a unique two-stream network dubbed the Chunk and Stride Network. This network employs temporal perspectives at both local and global scales in video features. Additionally, an attention mechanism is incorporated to handle the dynamic video content. Through rigorous ablation studies, the effectiveness of these methods is thoroughly demonstrated. The final model achieves a remarkable performance, surpassing previous benchmarks on popular datasets such as SumMe and TVSum.

Jung et al. (2019) summarize the key papers and discuss an unsupervised object-level video summarization technique termed online motion auto-encoder (online motion-AE). It aims to capture detailed semantic and motion details in online videos without the need for manual labeling. The method works towards generating an automatic and prospective compact summary for the entire video by capturing important visual motion segments in it. Such a method makes the auto-encoder architecture to be well-suited for the super-segmented object motion clips. The effectiveness of the proposed approach is confirmed on a newly collected surveillance dataset named Orangeville, which provides spatial-temporal annotations, and also on other public datasets. One potential limitation of the strategy that is investigated in Jung et al. (2019) is that it might not be able to aggregate fine-grained semantic information, especially the content of complex videos featuring diverse content of different objects. Furthermore, this proposed method can forget about the larger context information that is distributed in the video. Lei et al. (2019) summarize the key papers Frame Rank and introduce a novel approach that discovers the reasonable and insightful parts in the video content by parsing the movie based on the information frame by frame. To do so, the first phase of this framework is to organize the frames of each video into temporally coherent segments. Then, it orders these segments with a graph-based approach to decide which ones are particularly important. Finally, Frame Rank selects the ranked segments for the video summarization. It focuses on temporal segments considering their frames' semantic depth by a graph based on Kullback–Leibler (KL) divergence. The author of this paper shows the effectiveness of Frame Rank on three datasets SumMe, TVSum, UGSum52 where it achieves state-of-the-art results. Lei et al. (2019) further introduces a new dataset for summarizing user-generated videos, i.e., UGSum52, to alleviate the absence of large-scale video datasets with human summaries.

Yuan et al. (2019) summarize the key papers and introduce a method for unsupervised video summarization, which is called Cycle-SUM. Cycle-SUM adopts a particular architectural structure, which is referred to as a cycle-consistent adversarial LSTM. Such is this architecture that it emphasizes on retaining the vital aspects and keeping it as short as possible for the summary video. Cycle-SUM consists of two main components, i.e., a frame selector and a learning-based evaluator, which jointly search for the frames with the key information for the summary. In addition, Cycle-SUM employs two GANs for reconstructing the video in both the forward and backward directions, so that the summary retains the important information from the original content. It is operand to note that Cycle-SUM could be computationally expensive since it depends on complex topologies and adversarial training.

Huang et al. (2021) summarize the key papers and propose a new unsupervised learning strategy that selects keyframes from videos without using any annotated labels by integrating visual and semantic information. The key to this approach is a Spatial Attention Module (SAM), which is designed to focus on large spatial, motion-driven features in movies known to capture human attention. A bi-directional LSTM network is applied to assist the SAM in guiding the keyframe selection process with semantic and saliency information.

Regardless, context issues routinely hinder unsupervised techniques, leading to shallow or overly simplistic summaries. In addition, the effectiveness of the SAM and Bi-directional LSTM is largely determined by the quality of the features that the SAM and Bi-directional LSTM extract. Problems of picking out the most salient features of videos may reduce the effectiveness of the summarizing process. In addition to this, the diversity of human perception poses a major problem towards the creation of a video summary that could be accepted universally because it is too difficult to be consistent with identifying the key frames or segments from almost all genres of video.

Clustering and feature-based methods

Jadon and Jasim (2020) sumarize the key papers and present an unsupervised framework for machine vision-centered video summarization. It includes various methods of keyframe extraction, including deep features from ResNet, histogram-based analysis, uniform sampling, as well as SIFT. Phaphuangwittayakul et al. (2021) where keyframes are compactly clustered with clustering algorithms such as Gaussian Mixture Models (GMMs) and K-means. Video skimming then works around the selected key-frames to obtain a complete summary. The efficacy of feature extraction via a deep-learning approach and associated with Gaussian clustering is empirically tested and proven on the SumMe dataset, throughout particular dynamic-viewing movies.

For video frames' features extraction, Majumdhar & Nayak (2021) summarize the key papers and suggest using more advanced texture descriptors, including LTP and LPQ, source. To improve the summarization effectiveness, it is necessary to detect both important and unnecessary frames. These methods improve the representation of content within the retrieved properties and contribute to better summarization. After feature extraction, similar feature frames can then be organized with clustering approaches, such as affinity propagation and BIRCH. By selecting the most remembered snaps to summarize, this ensures that the kaleidoscopic film does not sacrifice the original work's music and emotion. That is, the textural descriptors combined with the unsupervised clustering appear to provide an unambiguous technique to generate good summaries.

Graph neural network approaches

Automatically selecting salient snippets from videos has been studied in Gao et al. (2021) which summarizes the key papers without the use of pre-annotated data. It argues the necessity of understanding the relations of video clips by using nodes in a graph to represent clips and building a graph for learning in this case. GNNs, or Graph Neural Networks, are particularly useful for this purpose as they facilitate the communication and dependency capture between graph nodes. The vector representation of each node is determined by its neighbours, edge types, and attributes in the respective networks. Unlike the approaches that consider video clips as clips with soft importance scores, this approach guarantees the selection of clips that ensure that the summary is devoid of ambiguity. The reconstruction constraint to preserve the substance of the original video in the summary and the contrasting constraint to differentiate the summarized and non-summarized content are used to improve the framework using a multi-task loss optimization technique. However, issues like irregular graph topologies and sizes make it hard to use neural networks and conventional learning methods, which affects how the graph-structured data are processed. Furthermore, the classic Graph Neural Networks still have the issue of over-smoothing, which could make it harder to distinguish between different node representations. These difficulties demonstrate how complex it is to use GNNs for video summarization in an efficient manner.

While the existing methods show promise, they struggle with long-range dependencies and the multimodal integration. The proposed RED-SM model addresses these gaps by combining Transformer architectures with reinforced learning and flexible multimodal fusion, overcoming the limitations of the LSTM-based approaches and isolated modality processing.

3. Proposed methodology

3.1. Overview of RED-SM model architecture

Figure 1 shows the proposed methodology Reinforced Encoder Decoder Summarizer Model (RED-SM) employs an encoder-decoder structure, both of which are multi-layer, used to handle input features from the frames of a video. The encoder is composed of six Transformer encoder layers. Within the Transformer encoder layer, the input features are treated by a multi-head attention computation. The multi-head attention allows the model to effectively obtain dependency among different positions. Using eight independent attention heads, the model calculates attention weights in parallel and is therefore more flexible and consistent. Also, attempting to avoid overfitting and guaranteeing generalization, dropout regularization is used on the output of the attention mechanism. There is a significant change in the encoder layer, and there is an addition of a residual connection.

Once the attention output is calculated, it is added once more to the feedforward neural network output within the layer. The residual connection allows a smoother propagation of information and helps prevent the vanishing gradient problem during training, giving the model a better capacity to learn long-distance dependencies. The decoder uses the output of the encoder as its memory.

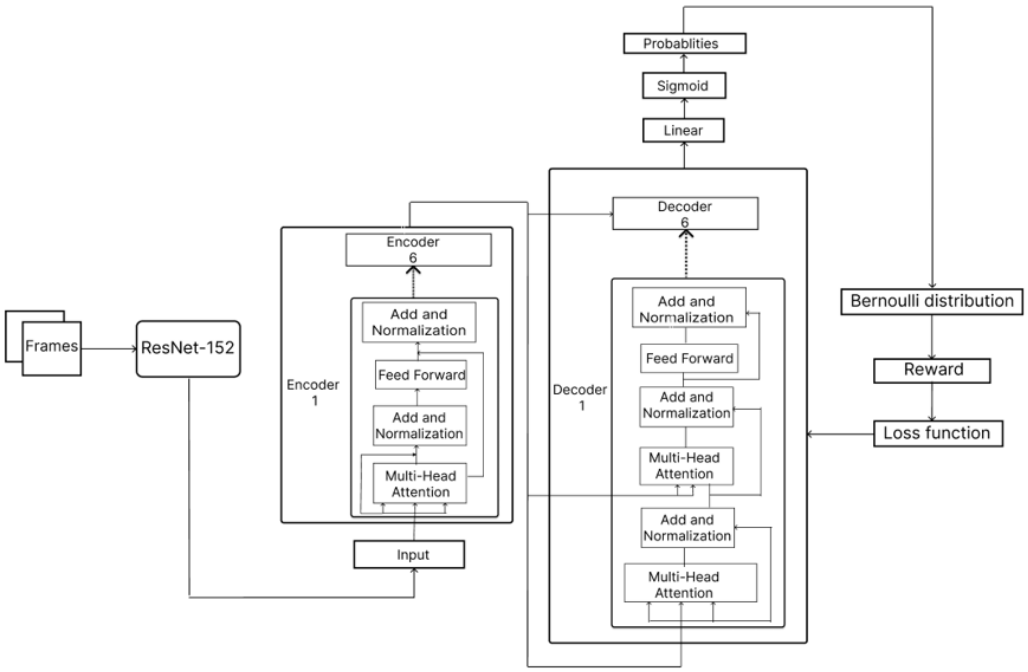

Figure 1. Reinforced encoder decoder summarizer model architecture (own research)

The decoder component consists of six Transformer decoder layers. Each Transformer decoder layer integrates two attention mechanisms: a multi-head attention one and an encoder-decoder attention one. The multi-head attention mechanism allows the decoder to attend to different positions within its own input sequence, capturing the dependencies within the sequence itself effectively. Meanwhile, the encoder-decoder attention mechanism enables the decoder to concentrate on relevant segments of the input sequence encoded by the encoder. This interaction between the encoder and decoder facilitates the generation of summaries that are coherent and contextually relevant.

As the input features traverse through all Transformer decoder layers, they undergo further processing via a linear layer, yielding the final output tensor representing the probability associated with summarizing the input video frames.

3.2. Feature extraction

The authors utilize the ResNet-152 model to extract features from the frames of the video. The extracted features are then spatially aggregated using the help of an adaptive average pooling layer. Then, a fully connected layer is utilized for dimensionality reduction of the extracted features.

3.2. Model architecture overview

3.3. Model training

In the training phase, the model processes the input sequence to compute probabilities for each feature and subsequently calculates the cost, which includes a penalty term aimed at minimizing the summary length. This penalty term is derived from the average probability of the selected frames. Following this computation, the frames are chosen based on these probabilities using Bernoulli sampling, and rewards are then calculated.

The reward calculation involves three key functions: the temporal smoothness reward, the representativeness reward, and the diversity reward. The temporal smoothness reward is determined

by assessing the similarity of consecutive frame features using the cosine similarity. It accumulates dissimilarity by subtracting the cosine similarity from 1 to quantify the lack of smoothness between consecutive frames. The accumulated dissimilarity is then normalized by the number of transitions. Finally, the negative exponential of the normalized dissimilarity is returned as the temporal smoothness reward, penalizing the abrupt transitions.

Temporal Smoothness Reward:

$$S = \{s1, s2, \dots sn\} \tag{1}$$

be the selected frame features, where pickidxspickidxs contains indices of the selected frames. The dissimilarity DD is computed as:

$$D = \sum_{i=1}^{n-1}\left(1 - cosine\_similarity\left(S\left[pickidXs[i-1]\right], S\left[pickidXs[i]\right]\right)\right) \tag{2}$$

$$D_{trans} = \frac{1}{n_{trans}} \sum_{i=1}^{n-1}\left(1 - cosine_{similarity}\left(S\left[pickidXs[i-1]\right], S\left[pickidXs[i]\right]\right)\right) \tag{3}$$

$$R_{smooth} = e^{-D} \tag{4}$$

Representativeness Reward:

$$R_{repr} = \frac{1}{N} \sum_{i=1}^{N} \max_{j \in pickidxs} cosine_{similarity}(f_i, f_j) \tag{5}$$

We can use \operatorname{cosine_similarity} or simply define it as a function.

Alternatively, we can write: $R_{repr} = \frac{1}{N} \sum_{i=1}^{N} \max_{j \in pickidxs} sim(fi, fj)$

where f_i, f_j are all the frame features and where sim is the cosine similarity.

Subsequently, the representativeness reward evaluates the extent to which the selected frames accurately represent the overall content of the video. On the other hand, the diversity reward measures the diversity of the selected frames within the summary. To update the model, the expected reward is computed based on the log probabilities and actual rewards, and the cost is adjusted accordingly, factoring in the negative expected reward. The parameters of the model are then optimized using the Adam optimizer through backpropagation. This iterative process ensures that the model's parameters are fine-tuned to effectively summarize the input video sequences.

3.4. Video summary generation

The video input undergoes frame-by-frame processing, with ResNet-152, extracting features from every frame. These extracted features are then passed through the Reinforced Encoder Decoder Summarizer Model (RED-SM), which assigns importance scores to each frame. These importance scores indicate the significance of each frame in capturing the content of the video. The change points, or the significant changes in the content, are identified by calculating kernels between frames and identifying significant changes.

The change points also play the role of boundaries for dividing the video into meaningful segments. The segments of the video are ordered by the importance scores that have already been established by the Transformer model. The dynamic knapsack algorithm is subsequently employed to select segments for the end video summary. This algorithm chooses the entire importance score in the optimal manner by selecting the segments suitable for this task based on a given length constraint. The frames of these selected segments are then fused to produce the summarized video.

3.5. Frame selection using BERT-Based Extractive Summarizer

Audio extraction and transcription

The audio is pulled from the input video with the MoviePy library. It is then converted to text with the Whisper library. Whisper is a speech-to-text and transcription tool of high capacity that

offers reliable textual outputs of audio content.

Subtitle extraction

The text is scanned to extract subtitles and their timecodes. A subtitle is utilized to point to a part of the video and for how long it is.

Subtitle summarization

The bert-extractive-summarizer is utilized to summarize the subtitles with a specified ratio. An extractive summarizer is chosen to ensure that the generated summary retains the most important sentences from the subtitles without introducing new content.

Frame selection

Frames corresponding to the summarized subtitles are selected from the video. To accomplish this, the timestamps of the summarized subtitles are mapped to the corresponding video frames. If a frame falls within the timestamps, its corresponding index is set to true to indicate inclusion, otherwise, it is set to false to indicate exclusion.

3.6. Integration of models for frame selection

The results from both the Reinforced Encoder Decoder Summarizer Model (RED-SM) and the bert- extractive-summarizer model were integrated to select frames, adhering to specific criteria as shown in Figure 2. According to the standard evaluation process for the SumMe and TVSum datasets, the total frames selected for the summary should not surpass 15% of the total frames. To accommodate variations in audio quality, a weightage was assigned to the output of each model. When the signal-to-noise ratio is favorable, a 40% weightage of the total 15% is assigned to the frames selected by the bert-extractive- summarizer, with the remaining 60% allocated to the frames selected by the RED-SM. Conversely, when the signal-to-noise ratio is poor, a 80% weightage is given to the RED-SM and 20% to the bert-extractive-summarizer model. Finally, we combined the outputs from both models to generate the video summary.

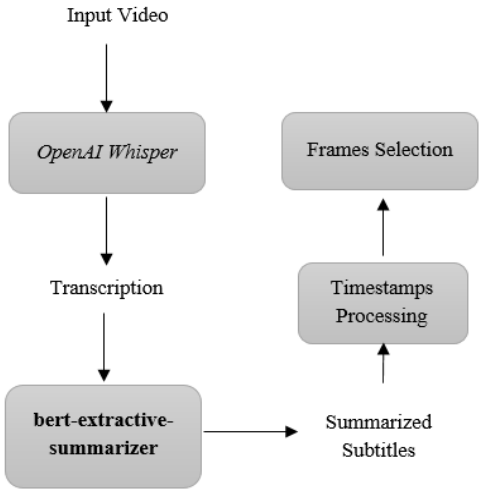

Figure 2. Frame selection using audio information (own research)

3.7 Architecture and tools used

Convolutional Neural Networks (CNNs): The ResNet-152 architecture is employed for feature extraction from video frames.

Dynamic programming: The knapsack algorithm optimizes item selection within a weight limit to maximize the total value, employing dynamic programming to efficiently compute solutions. By iteratively filling a table with solutions for subproblems, it determines the maximum achievable value. It is employed for segment selection in the summarization process.

Computer vision Libraries: OpenCV is used for video processing tasks such as frame extraction, color space conversion, resizing etc.

Whisper: Whisper, crafted by OpenAI, is an ML library tailored for recognizing and transcribing spoken language. It is used to extract text from the audio of the video.

Machine learning framework: The authors used PyTorch, an acclaimed Deep Learning library with the reputation of being flexible and having wide library support. PyTorch has attributes like automatic differentiation and dynamic computation graphs, making it suitable for a range of machine learning applications. For the present work, the PyTorch library was used to create the Reinforced Encoder Decoder Summarizer model and the bert-extractive-summarizer model.

4. Experiments and results

4.1. Data set description

The present method was measured with the SumMe and TVSum datasets. SumMe comprises 25 user-uploaded videos on a variety of topics, such as holidays and sports, with the length of each video being from 1 to 6 minutes. TVSum comprises 50 videos on topics such as news and documentaries, which have lengths ranging between 2 and 10 minutes. TVSum also accompanies 20 annotators who provide importance scores frame by frame. Training and building video summarization algorithms are challenging tasks that depend on being able to access rich and diverse datasets that represent the real-world situations properly. Although there are available datasets, they always lack depth so as to train the summarization methods correctly over large sets of domains and content types. With the goal of improving the video summarization methods, the authors have embarked on creating a new dataset. They have obtained from different publicly available sources over 100 videos belonging to more than 30 different categories. Some of these categories are: gadget reviews, science experiments, stand-up comedy, music videos, cooking videos, workout tutorials, celebrity interviews, travel vlogs, and even tech news updates. This variety of content can be used to formulate new ideas that can be used to enhance the existing methods of video summarization. In order to guarantee that the domains and types of the content were diverse and well-spaced out, a video's duration was systematically collected along with the category, which created rich metadata for the dataset.

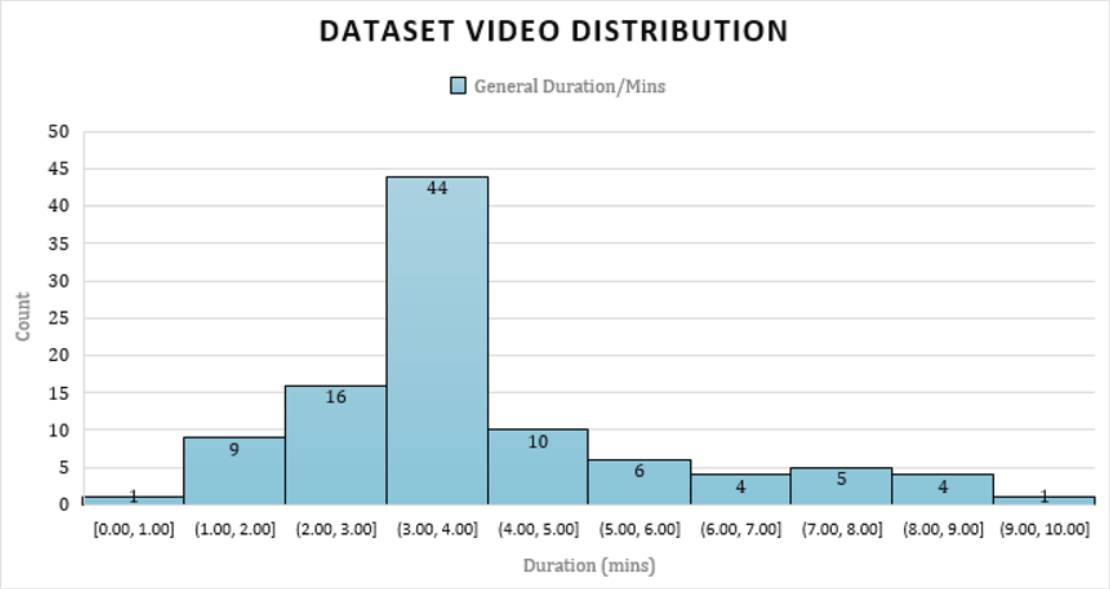

Figure 3. The dataset video distribution (own research)

Figure 3 shows the obtained dataset stands out because it encompasses a wide variety of categories, ensuring flexibility and potential for adaptation in numerous scenarios and types of content. This diversity significantly enhances its utility for studying summarization strategies in

different disciplines. This poses a fascinating opportunity to focus on improving the existing evaluation metrics or designing new ones tailored to specific content areas, summary objectives, or domain metrics on specialized content areas.

The authors would like to point out that, unlike the well-known benchmark datasets TVSum and SumMe, their dataset has the potential to create a domain shift. This points out the need to pay a special attention to the dataset characteristics and to any domain differences that influence the summarization methods.

4.2. Evaluation methods

The authors employ the two assessment settings proposed by Zhang in Zhang et al. (2016) their techniques. In the Canonical context, they first employ the conventional 5-fold cross-validation (5FCV) method, setting aside 80% of the videos for training and the remaining 20% for testing. Secondly, they train their model in the Transfer setting for a target dataset using our proprietary dataset, and then evaluate its performance on the SumMe and TVSum datasets

To ensure a balanced comparison with other methodologies, they adopt the commonly utilized protocol outlined in Zhang in Zhang et al. (2016) for computing the F-Score metric. The formula for F-Score calculation is given in Figure 4. This metric serves as a standard measure to evaluate the likeness between automatic summaries and ground truth summaries.

$$Precision\ (\boldsymbol{P}) = \frac{Overlapped\ Duration\ of\ Automatic\ Summary\ and\ Ground\ Truth\ Summary}{Duration\ of\ Automatic\ Summary}$$

$$Recall\ (\boldsymbol{R}) = \frac{Overlapped\ Duration\ of\ Automatic\ Summary\ and\ Ground\ Truth\ Summary}{Duration\ of\ Ground\ Truth\ Summary}$$

$$\boldsymbol{F - Score\ (F) = 2P \times R/(P + R) \times 100\%}$$

Figure 4. F-Score Calculation (own research)

4.3. Experimental setup

Ensuring an equitable comparison, the procedure outlined in Zhang et al. (2016) was followed, wherein video features were extracted from the output of the second-to-last layer (pool 5) of the GoogLeNet model, for videos in datasets including SumMe and TVSum. 80% of the videos in each dataset were utilized for training purposes, while the rest 20% were allocated for testing.

4.4. Results discussion

Table 1. Comparison of different methods with the canonical setting on SumMe and TVSum datasets

Methods	Datasets	
	SumMe	TVSum
SUM - GAN (dpp)	39.1	52.7
DR - DSN	41.4	57.6
Online Motion - AE	37.7	51.5
FrameRank	45.3	60.1
Cycle - SUM	41.9	57.6
RCL	48.6	58.4
RED - SM (theirs)	50.4	60.3

Table 2. Results of transfer learning: The model was trained using their own dataset and evaluated on SumMe and TVSum datasets

Method	Datasets	
	SumMe	TVSum
RED - SM (theirs)	42.0	57.8

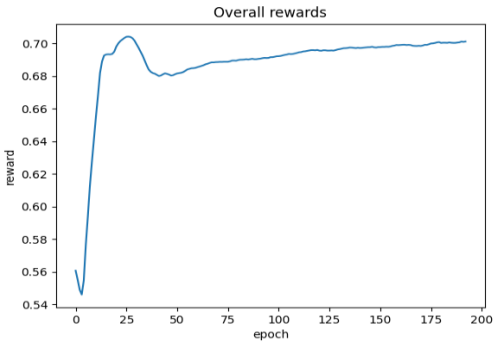

Figure 5a. Rewards obtained when training SumMe dataset across different data splits

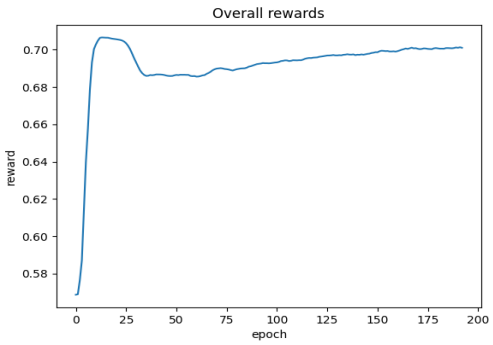

Figure 5b. Rewards obtained when training SumMe dataset across different data splits

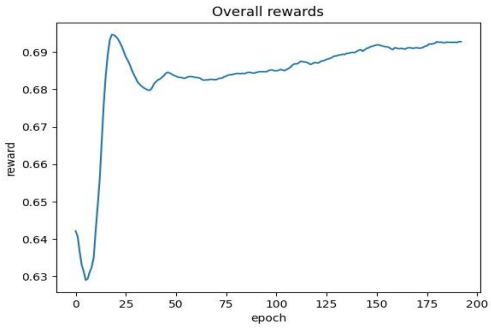

Figure 5c. Rewards obtained when training SumMe dataset across different data splits

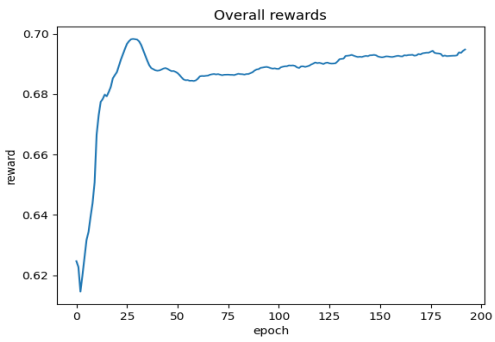

Figure 5d. Rewards obtained when training SumMe dataset across different data splits

(Source: Authors own research)

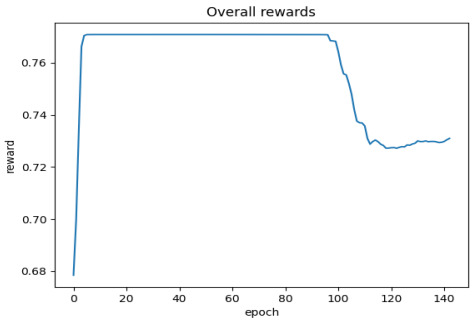

Figure 6a. Rewards obtained when training TVSum dataset across different data splits

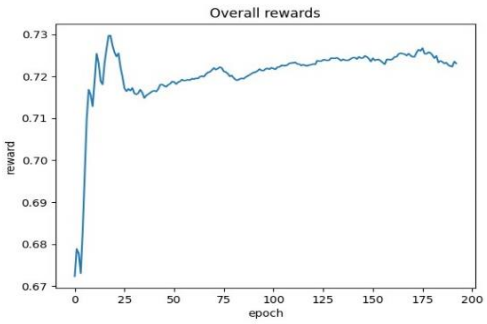

Figure 6b. Rewards obtained when training TVSum dataset across different data splits

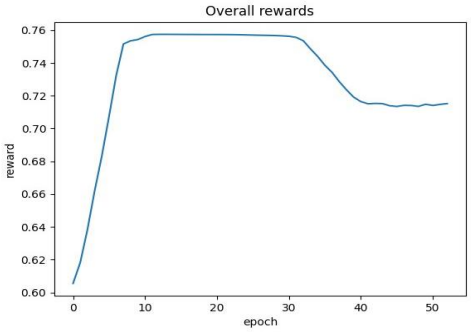

Figure 6c. Rewards obtained when training TVSum dataset across different data splits

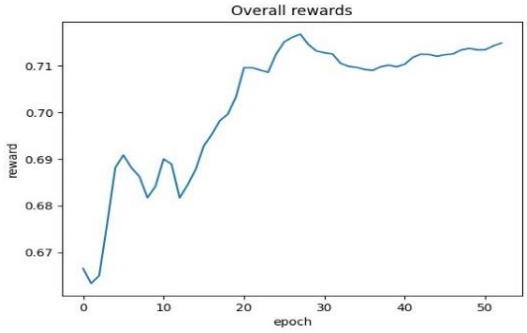

Figure 6d. Rewards obtained when training TVSum dataset across different data splits

(Source: Authors own research)

Table 1 shows the comparison of different methods with the canonical setting on SumMe and TVSum datasets and Table 2 shows the results of transfer learning: The model was trained using their own dataset and evaluated on SumMe and TVSum datasets. Figure 5a, 5b, 5c and 5d shows the **r**ewards obtained during training using the SumMe dataset across different data splits and Figure 6a,6b,6c and 6d shows the rewards obtained during training using TVSum dataset across different data splits. When the authors combined the output of the Reinforced Encoder Decoder Summarizer Model (RED-SM) with the output generated using the bert-extractive-summarizer model for the SumMe and TVSum datasets, there was no significant improvement in the average F-score. This lack of improvement can be attributed to the poor signal-to-noise ratio (SNR) of the audio and videos in the SumMe dataset. Conversely, some videos in the TVSum dataset had good SNR, resulting in higher F-scores, while others had poor SNR, leading to lower F-scores. These variations in SNR levels resulted in limited information extraction from the generated subtitles.

5. Conclusion and future directions

In this paper, the authors propose a Reinforced Encoder Decoder Summarizer Model (RED-SM) for unsupervised video summarization. The key to their model is a hybrid use of the state-of-the-art methods, including feature extraction, self-attention mechanisms, and reward-based training to generate fluent and precise video summaries. They utilized the ResNet-152 model to perform feature extraction, allowing the important visual cues to be effectively summarized via spatial pooling and dimension reduction.

Figure 5 shows the RED-SM architecture, which has an encoder encoder-decoder structure and consists of several layers that process input features from the frames of video. They use the same encoder in the two models: a modified version of the Transformer architecture, adapted to the video summarization task. It is comprised of six modified Transformer encoder layers, tailored for the needs of video summarization, and utilizes multi-head attention mechanisms that are essential for modelling dependencies across different positions in the sequence.

Additionally, the decoder of their architecture consists of six sequential Transformer decoder layers, enabling a full coverage and the ability to generate appropriate summaries. They also have multi-head attention and encoder-decoder attention within the layers to help improve the summarization.

In order to improve generalization and avoid over-fitting, strategies such as dropout regularization, residual connections, and normalization within the model's architecture were employed.

When training the model, they employ a reward-based learning strategy where the model gets rewarded when generating high-quality summaries to foster better selection of video content frames. Some of the different rewards used include temporal smoothness, representativeness, and diversity. In detecting important cuts for the summarization, they analyse frames for content feature change points. The segments are ranked based on their self-relevance and a knapsack algorithm is applied to select the frames for the summary. This holistic approach enables their summaries to maintain coherence and compactness while providing an extensive coverage of the video content.

Despite the positive implications of the results obtained from the RED-SM model, there remains ample room for refinement and innovation. One of the approaches for improvement is the integration of semantic understanding, which enhances the model's understanding of the video frames' underlying semantics. This can be aided with the integration of natural language processing techniques or the use of language models to foster the understanding of the model in the visual context. Also, the exploration of adaptive attention techniques could be beneficial in the improvement of the summarization skills. The model could enhance its summarization performance by flexibly adjusting the attention to frame or region significance and dynamically concentrating on critical information while disregarding less relevant features.

Furthermore, tackling the long-term dependencies in the context of movies remains extremely challenging. Working on the use and capture of temporal information for an extended timeframe in the generation of summaries has the potential to significantly improve the temporal coherence and context understanding of the generated summaries. Moreover, the model's comprehension of the temporal structure and the narrative flow of a video can be augmented by the application of techniques that deal with the integration of distant frames or segments, resulting in more comprehensive and contextually enriched summaries.

Author contributions

Conceptualization: A.V., L.R., D.K. and T.R.; Data Curation: A.V., L.R., D.K. and T.R.; Project administration: A.V., L.R., D.K. and T.R.; Supervision: A.V., L.R.; Validation: D.K. and T.R.; Writing—original draft: A.V., L.R. and E.M.; Writing—review and editing: A.V., L.R., D.K. and T.R. All authors have read and agreed to the published version of the manuscript.

Submission received: 23 July 2025; Revised: 20 November 2025; Accepted: 28 November; Published: 12 December 2025.

REFERENCES

Abbasi, M. & Saeedi, P. (2023) Adopting Self-Supervised Learning into Unsupervised Video Summarization through Restorative Score. In *2023 IEEE International Conference on Image Processing (ICIP)* (pp. 425-429). IEEE. https://doi.org/10.1109/ICIP49359.2023.10222350.

Apostolidis, E., Adamantidou, E., Metsai, A.I., Mezaris, V. & Patras, I. (2020) AC-SUM-GAN: Connecting actor-critic and generative adversarial networks for unsupervised video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(8), pp.3278-3292. https://doi.org/10.1109/TCSVT.2020.3037883.

Gao, J., Yang, X., Zhang, Y. & Xu, C. (2021) Unsupervised Video Summarization via Relation-Aware Assignment Learning. *IEEE Transactions on Multimedia*. 23, 3203-3214. https://doi.org/10.1109/TMM.2020.3021980.

Gygli, M., Grabner, H., Riemenschneider, H. & Van Gool, L. (2014) Creating summaries from user videos. In *European conference on computer vision* (pp. 505-520). Cham: Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-319-10584-0_33.

Hong, Z. & Zhong, R. (2021) Visual and Semantic Feature Coordinated Bi-Lstm Model for Unsupervised Video Summarization. In *2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 05-09 July 2021*. IEEE. pp.1-6. https://doi.org/10.1109/ICME51207.2021.9428250.

Hu, J., Sui, C., Wang, H., Hong, D., Gong, Q., Zhou, S. & Wang, A. (2023) Context-guided Unsupervised Video Summarization Network. In *2023 3rd International Conference on Electronic Information Engineering and Computer Science (EIECS)* (pp. 1069-1072). IEEE. https://doi.org/10.1109/EIECS59936.2023.10435485.

Huang, Y., Zhong, R.Yao, W. & Wang, R. (2021) Unsupervised Learning of Visual and Semantic Features for Video Summarization. In *2021 IEEE International Symposium on Circuits and Systems (ISCAS), Daegu, Korea, 22-28May, 2021. IEEE*, pp.1-5. https://doi.org/10.1109/ ISCAS51556.2021.9401310.

Jadon, S. & Jasim, M. (2020) Unsupervised video summarization framework using keyframe extraction and video skimming. In Proceedings of IEEE 5th International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 30-31 October 2020. IEEE, pp.140-145. https://doi.org/10.1109/ICCCA49541.2020.9250764.

Jung, Y., Cho, D., Kim, D. et al. (2019) Discriminative Feature Learning for Unsupervised Video Summarization. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence. 33(1), 8537-8544. https://doi.org/10.1609/aaai.v33i01.33018537.

Kaseris, M., Mademlis, I. & Pitas, I., (2021) Adversarial unsupervised video summarization augmented with dictionary loss. In *2021 IEEE International Conference on Image Processing (ICIP)* (pp. 2683-2687). IEEE. https://doi.org/10.1109/ICIP42928.2021.9506088

Lei, Z., Zhang, C., Zhang, Q. & Qiu, G. (2019) Frame Rank: A Text Processing Approach to Video Summarization. To be published in *Computation and Language*. [Preprint] https://doi.org/10.48550/arXiv.1904.05544 [Accessed:12 April 2019].

Mahasseni, B., Lam, M. & Todorovic, S. (2017) Unsupervised Video Summarization with Adversarial LSTM Network. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21-26 July 2017*. IEEE, pp.2982-2991. https://doi.org/10.1109/CVPR.2017.318.

Majumdhar, J. & Nayak, S.K. (2021) A Novel Method on Summarization of Video Using Local Ternary Pattern and Local Phase Quantization. In *2021 2nd International Conference on Range Technology (ICORT), Chandipur, Balasore, India, 5-6 August 2021*. IEEE. pp. 1-6. https://doi.org/10.1109/ICORT52730.2021.9581941.

Phaphuangwittayakul, A., Yi Guo, Fangli Ying, et al. (2021) Self-Attention Recurrent Summarization Network with reinforcement learning for video summarization task. In *Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5-9 July 2021*. IEEE. pp.1-6. https://doi.org/10.1109/ICME51207.2021,9428142.

Song, Y., Vallmitjana, J., Stent, A. & Jaimes, A. (2015) Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5179-5187).

Sun, H., Zhu, X. and Zhou, C. (2022) Deep reinforcement learning for video summarization with semantic reward. In *2022 IEEE 22nd International Conference on Software Quality, Reliability, and Security Companion (QRS-C)* (pp. 754-755). IEEE. https://doi.org/10.1109/QRS-C57518.2022.00119

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I. (2017) Attention is all you need. NeurIPS 30. *Neural Information Processing Systems Foundation*, pp.5998-6008.

Yuan, L., Tay, F.E., Li, P. et al. (2019) Cycle-SUM: Cycle-Consistent Adversarial LSTM Networks for Unsupervised Video Summarization. In *Thirty-Third Association for the Advancement of Artificial Intelligence Conference (AAAI-19) Conference on Artificial Intelligence*. 33(01), 9143-9150. https://doi.org/10.1609/aaai.v33i01.33019143.

Yuan, Y. & Zhang, J. (2022) Unsupervised video summarization via deep reinforcement learning with shot-level semantics. *IEEE Transactions on Circuits and Systems for Video Technology*. 33(1), 445-456. https://doi.org/10.1109/TCSVT.2022.3197819

Zhang, K., Chao, W-L., Sha, F. & Grauman, K. (2016) Video Summarization with Long Short-Term Memory. In Leibe, B., Matas, J., Sebe, N. & Welling, M. (eds.) *Computer Vision – European Conference on Computer Vision (ECCV).* ECCV 2016. Lecture Notes in Computer Science (9911) Springer, Cham. https://doi.org/10.1007/978-3-319-46478-7_47.

Zhang, Y., Liang, X., Zhang, D. et al. (2018) Unsupervised Object- Level Video Summarization with Online Motion Auto-Encoder. To be published in *Computer Vision and Pattern Recognition.* [Preprint] https://doi.org/10.48550/arXiv.1801.00543 [Accessed: 11 August 2018].

Zhang, Y., Liu, Y., Kang, W. & Zheng, Y., (2023). Mar-net: motion-assisted reconstruction network for unsupervised video summarization. *IEEE Signal Processing Letters*. 30, 1282-1286. https://doi.org/10.1109/LSP.2023.3313091.

Zhang, Y., Liu, Y., Zhu, P. & Kang, W. (2022) Joint reinforcement and contrastive learning for unsupervised video summarization. *IEEE Signal Processing Letters*. 29, 2587-2591. https://doi.org/10.1109/LSP.2022.3227525.

Zhou, K., Qiao, Y. & Xiang, T. (2018) Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 32(1), 7582-7589. https://doi.org/10.1609/aaai.v32i1.12255.

Venkatachalam ARULKUMAR[*] is an Associate Professor in the Department of Computer Science and Business Systems Easwari Engineering College, Ramapuram, Chennai, Tamil Nadu, India. He earned his B.E degree in Information technology in the year 2003 from Periyar University, Salem and his masters in computer science and engineering from Anna University, Chennai in the year 2008. He completed his Ph.D. from the faculty of Information and Communication Engineering under Anna University, Chennai in the area of Cloud Computing by 2020. His teaching and research interests include Cloud computing, Scheduling and distributed algorithms. He has 18 years of teaching experience from reputed Engineering Colleges. He is an active member in the professional bodies of IEEE.

Rajendran LATHAMANJU is an Associate Professor in the Department of Electronics and Communication Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India. She completed B.E degree in ECE by 2004 from Government College of Engineering, Bargur and Post graduate in Communication Systems by 2007 from Anna University, Chennai. She completed her Ph.D. degree under Anna University, Chennai in the area of wireless sensor Networks. Her teaching and research interests includes Wireless sensor networks, Internet of Things, Cloud computing and having around 18 years of teaching experience from various reputed Engineering Colleges. She is an active member in professional bodies of ISTE, Filed and published 3 patents and 2 patents grant, at present guiding two Ph.D. scholar and also acted as a reviewer in conferences and peer reviewed journals.

Krishnamoorthy DURGA DEVI is an Assistant Professor in the Department of Electronics and Communication Engineering, KGiSL Institute of Technology, Coimbatore,641035 Tamil Nadu, India. She completed B.E degree in Electronics and Communication Engineering by 2011 and Post graduate in Communication Systems by 2013. She completed her Ph.D. degree in College of Engineering Guindy Campus, Anna University, Chennai in the field of Image Processing in 2019. Her teaching and research interests includes Networks, Internet of Things, Image Processing, Machine learning. She is an active member in professional bodies of ISTE, Filed and published 3 patents in 2021 and 2022, at present guiding one Ph.D. Scholar and also acted as a reviewer in conferences and peer reviewed journals.

Rajamani THANGAM working as an Assistant Professor in the Department of Electronics and Communication Engineering at SRMIST, Ramapuram. graduated in Electrical and Electronics Engineering from Madras University, Chennai, she earned a Master of Engineering in Applied Electronics from Anna University, Chennai. Her academic journey culminated in a Ph.D. from MIT, Anna University, where she specialized in digital pulse skipping modulation strategies for DC-DC converters. With over 14 years of teaching experience, Dr. Thangam has been instrumental in shaping the next generation of engineers. She is passionate about integrating innovative teaching methods to enhance student learning and engagement. Her research interests include power electronics and modulation techniques, contributing to advancements in efficient energy conversion. Dr. Thangam actively participates in academic conferences and workshops, sharing her knowledge with peers and students alike. Her dedication to education and research has made her a respected in the academic community, inspiring students to excel in their studies and pursue their interests in engineering.