Video Summarization using Deep Reinforcement Learning

C Mithul
Department of Computer Science and engineering, Amrita School of Computing
Amrita Vishwa Vidyapeetham,
Bangalore, India
BL.EN.U4AIE21034@bl.students.amrita.edu

D Mohammad Abdulla
Department of Computer Science and engineering, Amrita School of Computing
Amrita Vishwa Vidyapeetham,
Bangalore, India
BL.EN.U4AIE21044@bl.students.amrita.edu

M Hari Virinchi
Department of Computer Science and engineering, Amrita School of Computing
Amrita Vishwa Vidyapeetham,
Bangalore, India
BL.EN.U4AIE21077@bl.students.amrita.edu

B Somasekhar
Department of Computer Science and engineering, Amrita School of Computing
Amrita Vishwa Vidyapeetham,
Bangalore, India
BL.EN.U4AIE21020@bl.students.amrita.edu

Dr. Amudha J
Department of Computer Science and Engineering, Amrita School of Computing
Amrita Vishwa Vidyapeetham,
Bangalore, India
j_amudha@blr.amrita.edu

***Abstract*—The exponential growth of video data, effective summarization techniques are essential to enhance content accessibility and reduce storage requirements. This paper presents a novel video summarization framework leveraging Deep Reinforcement Learning (DRL) to dynamically select key frames based on their informativeness, semantic significance, and temporal coherence. Using ResNet-50 for feature extraction, the DRL agents—Proximal Policy Optimization (PPO), Advantage Actor-Critic (A2C), and Deep Q-Networks (DQN)—are trained in a custom OpenAI Gym environment with a reward mechanism prioritizing relevance and diversity. Experimental results demonstrate the framework's ability to generate concise and high-quality summaries, outperforming traditional methods in scalability and adaptability. This approach holds significant potential for applications in surveillance, medical diagnostics, and multimedia content analysis.**

***Index Terms*—Video Summarization, Deep reinforcement learning, DQN, PPO, A2C**

I. Introduction

The exponential growth of digital video content on platforms such as YouTube, TikTok, surveillance systems, and professional multimedia repositories has created an urgent need for automated video summarization techniques. The process of video summarization involves generating a concise representation of a video by selecting key frames or segments that best encapsulate its core content while maintaining semantic and temporal coherence. This technology has become critical for various applications, including multimedia content management, surveillance analysis, entertainment, instructional content curation, and real-time decision-making.

Manual video summarization is increasingly impractical due to the sheer scale and diversity of video data. Social media platforms see thousands of hours of video uploaded every minute, while surveillance systems continuously generate streams of high-dimensional video footage. This abundance of data presents numerous challenges: storage and retrieval efficiency, content accessibility and navigation, and resource optimization for timely analysis. Furthermore, the computational burden of manually processing and summarizing such volumes of video data necessitates robust automated methods capable of adapting to diverse domains while preserving the narrative and semantic essence of the original content.

Developing effective video summarization systems involves addressing multiple technical challenges. Semantic understanding is a primary concern, requiring systems to accurately identify and preserve the most meaningful aspects of a video. Temporal coherence is another critical factor, ensuring that the relationships between frames reflect the logical progression of events in the original content. Additionally, achieving content diversity while minimizing redundancy is essential for maintaining representativeness and user satisfaction. Balancing these requirements with computational efficiency is particularly challenging when dealing with high-dimensional data from modern video formats.

In recent years, Deep Reinforcement Learning (DRL) has emerged as a powerful framework for tackling these challenges. DRL's ability to model sequential decision-making and learn adaptive strategies through interaction with the environment makes it an ideal candidate for video summarization

tasks. Unlike traditional methods that rely on handcrafted features or static selection criteria, DRL-based approaches integrate feature extraction, decision-making, and summary generation into a unified framework. This enables them to learn dynamically optimized strategies that account for both semantic importance and temporal dependencies.

This paper proposes a novel video summarization framework leveraging state-of-the-art DRL algorithms, including Proximal Policy Optimization (PPO), Advantage Actor-Critic (A2C), and Deep Q-Networks (DQN). High-dimensional feature representations are extracted using ResNet-50, a convolutional neural network pre-trained on ImageNet. These features are used within a custom OpenAI Gym environment to train DRL agents to maximize rewards based on informativeness, temporal coherence, and diversity. By incorporating these techniques, the proposed framework addresses the limitations of traditional methods, offering a scalable, adaptable, and context-aware solution for video summarization.

To evaluate the effectiveness of this approach, we use benchmark datasets such as SumMe and TVSum, which span a variety of domains including news, documentaries, instructional content, and personal videos. These datasets provide frame-level importance scores annotated by multiple human evaluators, offering a robust ground truth for evaluation. Through comprehensive experiments and comparative analysis with existing methods, our framework demonstrates significant improvements in summary quality, scalability, and adaptability.

The remainder of this paper is organized as follows: Section II provides a detailed review of related work in video summarization and reinforcement learning. Section III introduces the proposed methodology and system architecture. Section IV discusses experimental results and comparative performance evaluations. Finally, Section V concludes the paper with a summary of contributions and potential directions for future research.

II. Literature Review

Recent advancements in video summarization have been significantly influenced by reinforcement learning (RL) approaches, with various methodologies emerging to address the challenges of efficient content representation and frame selection. The integration of RL with deep learning architectures has shown promising results in both supervised and unsupervised contexts. Yuan and Zhang [1] proposed an unsupervised approach incorporating shot-level semantics with deep RL, utilizing an encoder-decoder framework comprising a CNN and bidirectional LSTM. Their method notably introduced objective factors derived from video shooting processes to reduce user subjectivity in summarization, achieving performance comparable to supervised methods on multiple benchmark datasets. Building on this semantic approach, Li and Yang [2] developed a weakly supervised framework featuring a dual-network architecture: a Video Classification Sub-Network guiding an unsupervised Summary Generation Sub-Network. Their innovation lay in introducing semantic similarity as a reward signal, eliminating the need for frame-level annotations while maintaining summary quality.

The challenge of capturing long-term dependencies in video sequences was addressed by Wang et al. [3], who proposed a lightweight DRL framework combining CNN encoding with one-way LSTM decoding. Their model, enhanced by an auxiliary summarization loss and novel dispersion reward function, demonstrated significant improvements in F-scores on standard benchmarks while maintaining mobile deployment capability. Liu et al. [4] further advanced the field by introducing the 3DST-UNet-RL framework, which uniquely processes 4D video features through a 3D spatio-temporal U-Net integrated with RL. This architecture's versatility was demonstrated through successful application across both general and specialized medical video summarization tasks.

Most recently, Wang et al. [5] introduced a Progressive Reinforcement Learning structure (PRLVS) that addresses the persistent challenges of limited training data and sparse rewards. Their hierarchical "T"-type thinking paradigm, implementing dual policies for coarse sampling and refinement, represents a significant advancement in unsupervised video summarization, achieving performance comparable to supervised techniques while providing continuous feedback throughout the summarization process.

Yoon et al. (2023) introduce a groundbreaking unsupervised video summarization method that leverages deep reinforcement learning combined with interpolation techniques, as detailed in their paper "Unsupervised Video Summarization Based on Deep Reinforcement Learning with Interpolation," published in Sensors [6]. The study outlines a novel framework that utilizes a temporal consistency reward function to enhance the uniform selection of keyframes, significantly addressing the challenges of representativeness and diversity in video summaries. Their lightweight architecture, which integrates transformer and convolutional neural networks, effectively captures both global and local context features, allowing for the accurate prediction of importance scores for video frames. Experimental results demonstrate that their approach achieves state-of-the-art performance on benchmark datasets, SumMe and TVSum, with improvements quantified as a 10(persent) increase in F1 scores compared to existing methods. This research not only advances the field of video summarization but also provides a robust solution for efficiently generating concise summaries from lengthy video content, thereby enhancing user experience on online platforms.

Panagiota Alexoudi et al. (2023) explore innovative solutions to enhance unsupervised video summarization by addressing the challenge of local minima in deep reinforcement learning (DRL) frameworks [7]. Their paper, titled "Escaping local minima in deep reinforcement learning for video summarization," introduces a novel regularizer that effectively guides the training of DRL agents used for key-frame extraction. The proposed method involves a two-phase training strategy, where the first phase mirrors traditional training without the regularizer, followed by a second phase that incorporates an additive loss term designed to encourage exploration of the pa-

rameter space. This approach has been quantitatively validated on public datasets such as TVSum and SumMe, demonstrating significant performance improvements over existing state-of-the-art methods, with notable increases in summarization quality, thereby underscoring its potential impact on automated video summarization technologies.

Tianrui Liu et al. present a novel approach to video summarization tailored for medical applications, specifically focusing on fetal ultrasound imaging [8]. Their method employs deep reinforcement learning (RL) to efficiently condense lengthy ultrasound videos while preserving critical diagnostic information. By framing the summarization task as a decision-making process, the RL agents maximize rewards based on representativeness, diversity, and the likelihood of including standard diagnostic views. The authors demonstrate that their approach outperforms existing video summarization techniques, achieving superior F1-scores in both supervised and unsupervised settings. This work not only highlights the potential of deep RL in medical video analysis but also addresses the challenges of handling voluminous and often redundant video data in clinical practice.

Yujia Zhang et al. [9] propose a comprehensive framework for query-conditioned video summarization that leverages deep reinforcement learning to enhance personalization in video summaries. Their approach consists of two main components: a Mapping Network (MapNet), which establishes the relationship between video shots and user queries, and a Summarization Network (SummNet), which employs a reinforcement learning strategy to optimize the selection of video shots based on three key rewards: relatedness, representativeness, and diversity. By integrating these elements, the authors demonstrate that their method significantly outperforms existing state-of-the-art techniques on benchmark datasets, underscoring its effectiveness in generating tailored video summaries that cater to individual user interests.

Jie Lei et al. propose an innovative approach to video summarization that leverages reinforcement learning and action parsing to address the challenges of managing large volumes of video content [10]. Their model is divided into two main components: first, it employs a sequential multiple instance learning (SMIL) framework using weakly annotated data to effectively cut videos into segments containing single actions, thus mitigating the issues related to full annotation's time consumption and weak annotation's ambiguity. Second, it utilizes a deep recurrent neural network to summarize these segments by selecting the most distinguishable frames, with the quality of the extracted key frames evaluated through categorization accuracy. The experiments conducted demonstrate that their method outperforms existing state-of-the-art techniques, highlighting its potential for practical applications in video content management and indexing.

Yiyan Chen et al. (2019) introduce a weakly supervised hierarchical reinforcement learning framework for video summarization that addresses sparse rewards and reduces the need for extensive annotations. By using a Manager-Worker network, their approach effectively sets subgoals and predicts frame importance, achieving superior performance over traditional techniques [11]. Zhang et al. (2020) enhance key frame selection through a dual-stream architecture that integrates spatial and temporal features using an attention mechanism, leading to more coherent and contextually relevant summaries [12]. Li et al. (2021) propose an unsupervised method based on contrastive learning, which maximizes similarity within selected frames and improves the diversity of the summary without requiring labeled data [13].

Mehryar Abbasi et al. provide a comprehensive review of recent advancements in video summarization, emphasizing the shift towards unsupervised methods due to the high cost of manual annotation. Their review highlights frameworks using GANs and reinforcement learning, pointing to challenges in coherence and efficiency, and suggests future directions that include deep learning for greater summarization effectiveness [14]. Yunzuo Zhang et al. propose a joint reinforcement and contrastive learning framework, incorporating an Optimized Coding Module and a Dissimilarity-Guided Attention Graph for better feature aggregation and context modeling, showing superior results on benchmark datasets [15].

III. System Architecture and Methodology

The proposed video summarization framework represents a cutting-edge approach to automated content condensation, leveraging Deep Reinforcement Learning (DRL) techniques. As illustrated in Fig. 1 [1], the overall architecture provides a comprehensive strategy for transforming lengthy video content into concise, meaningful summaries.

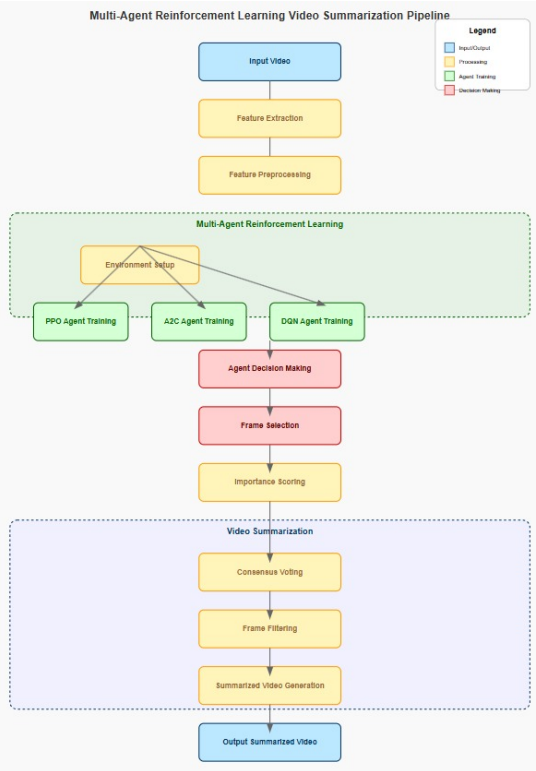

Fig. 1. Overall Architecture

The system architecture is meticulously designed to address the complex challenges of video summarization, integrating advanced machine learning methodologies with sophisticated decision-making algorithms. The primary objective is to develop an intelligent system capable of understanding and extracting the most significant moments from video content.

A. *Feature Extraction*

High-dimensional video features are extracted using ResNet-50, a pre-trained Convolutional Neural Network (CNN) optimized for feature representation. Each video frame is processed to generate a 2048-dimensional feature vector from the ResNet-50 network's penultimate layer. These feature vectors encapsulate the visual and semantic content of the frames, serving as the input state for the DRL agents. The key steps in this process include:

- **Frame Preprocessing:** Frames are resized, normalized, and transformed to align with the input requirements of ResNet-50.
- **Feature Encoding:** The processed frames are passed through ResNet-50 with its classification layer removed to extract feature vectors.
- **Temporal Representation:** The feature vectors maintain the temporal order of frames, providing contextual relationships for downstream decision-making.

B. *Reinforcement Learning Environment*

Fig .2. provides a detailed visualization of the module architecture, highlighting the intricate components of our reinforcement learning environment. A custom OpenAI Gym environment, tailored for video summarization, is developed to facilitate DRL training. The environment defines the state space, action space, and reward function, enabling agents to interact with video data.

- **State Space:** Each state corresponds to the feature vector of a video frame, representing its visual and semantic properties.
- **Action Space:** The action space is discrete, comprising two actions: select (include the frame in the summary) and skip (exclude the frame).
- **Reward Function:** A frame's reward is based on its importance score, which reflects its relevance to the overall video narrative. Rewards incentivize selecting frames that maximize informativeness and diversity while maintaining temporal coherence. Skipping frames yields no reward but contributes to avoiding redundancy.

C. *DRL Agent Design and Training*

The implementation harnesses three advanced DRL algorithms: Proximal Policy Optimization (PPO), Advantage Actor-Critic (A2C), and Deep Q-Networks (DQN). As depicted in Fig. 3, the training architecture represents a complex and dynamic learning process.

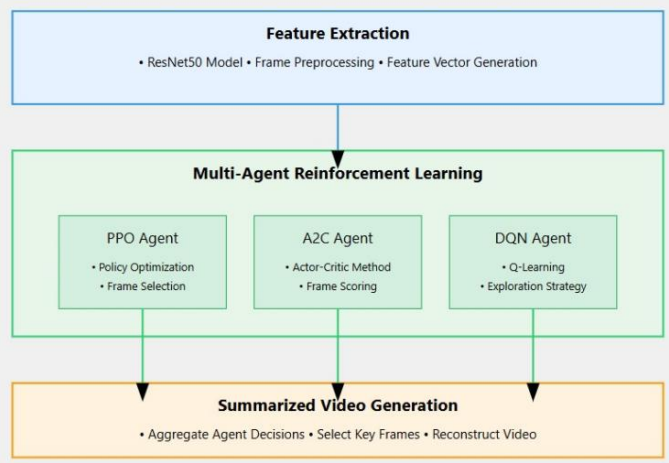

Fig. 2. Module Architecture

1) *Policy Optimization:*

- **PPO and A2C:** Use policy networks to map states to action probabilities. The stochastic nature of these policies ensures a balance between exploration and exploitation.
- **DQN:** Utilizes a Q-network to estimate action-value pairs, selecting actions with the highest Q-value. Epsilon-greedy exploration is used to encourage learning of diverse strategies.

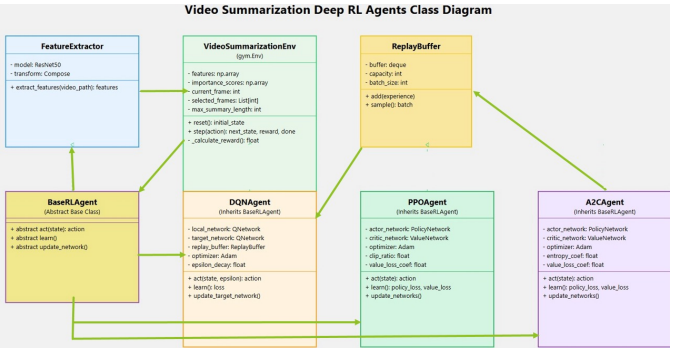

Fig. 3. Training Architecture

D. *Decision System*

The decision system drives the selection of frames and the generation of concise video summaries. It is composed of two main components: action selection and summary generation.

1) Action Selection:: The action selection mechanism helps the agent decide which frames to include in the summary. Key components include:

- **ϵ-greedy Strategy for Exploration:** The agent begins by exploring new actions (frame selections) with a high probability ϵ, gradually reducing ϵ over time to favor exploitation of learned knowledge. This approach helps the agent balance the exploration of new possibilities with the refinement of its policy.

- **Q-value based selection:** For DQN, the agent selects frames based on Q-values, which estimate the expected reward for including or skipping each frame. The frame with the highest Q-value is chosen, ensuring the selection of the most informative frames.
- **Policy network for action probability distribution:** For PPO and A2C, the agent uses a policy network to probabilistically choose actions. This network outputs a distribution over possible actions (select or skip), allowing for a balance between exploration and exploitation during decision-making.

E. Experience Management Strategy

The experience management approach is critical to the learning mechanism, utilizing two sophisticated components illustrated in Figures 4 and 5:

F. Experience Buffer

The Experience Buffer serves as a temporary storage mechanism, capturing recent episode interactions. This short-term memory allows for immediate learning and adaptation, providing a real-time mechanism for processing video content.

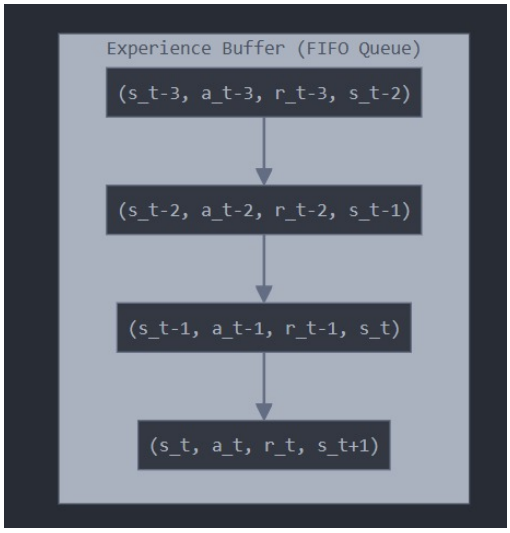

Fig. 4. Experience Buffer

G. Replay Buffer

The Replay Buffer maintains a comprehensive repository of experiences across multiple episodes. This component enables efficient off-policy learning, improving sample efficiency and allowing the agent to learn from a diverse range of historical interactions.

H. Reward Calculation Mechanism

The reward calculation strategy represents a sophisticated multi-component function designed to capture nuanced aspects of summary quality. By incorporating multiple evaluation criteria, including informativeness, diversity, relevance, and length considerations, the system creates a comprehensive scoring mechanism.

Fig. 5. Replay Buffer

I. Training Process Dynamics

The training process follows a meticulously structured approach:

- Extracting high-dimensional frame features using ResNet-50
- Facilitating DRL agent interactions within the custom Gym environment
- Collecting and storing diverse experiences
- Optimizing policies through iterative learning
- Continuously refining frame selection strategies

J. Hyperparameter Configuration

Extensive hyperparameter tuning was conducted to optimize learning dynamics:

- Learning rate: 0.0001
- Discount factor: 0.95-0.99
- Exploration strategy: ϵ-greedy
- Batch size: 64-128

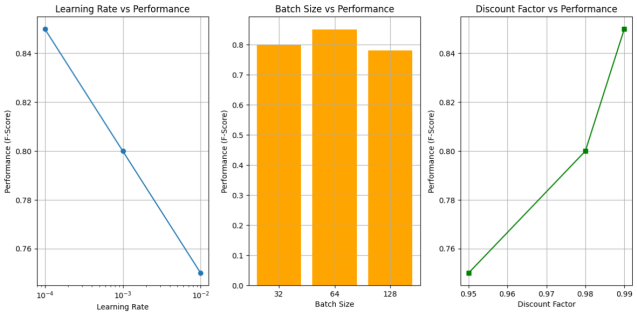

Fig. 6. Shows how performance changes with different parameters

K. Summary Generation Mechanism

The summary generation process unfolds through a systematic approach:

- Scoring individual frames using learned policy networks
- Applying intelligent thresholding to select key frames
- Reconstructing a sequential summary using OpenCV

L. *Evaluation Metrics*

Performance assessment leverages standard metrics such as F-score and Regret. These metrics provide a comprehensive evaluation by comparing selected frames against ground truth annotations, emphasizing crucial aspects like representativeness, diversity, and narrative coherence.

M. *Technological Innovations*

The implemented framework introduces several groundbreaking innovations:

- Unsupervised learning capabilities
- Dynamic and adaptive frame selection
- Advanced temporal context preservation
- Robust performance across diverse video domains

The proposed implementation demonstrates a revolutionary approach to video summarization, seamlessly integrating advanced deep reinforcement learning techniques to transform complex video content into concise, informative summaries.

IV. Results

A. *Performance Evaluation*

The performance of the proposed DRL-based video summarization methods (PPO, A2C, and DQN) was evaluated using the SumMe and TVSum datasets, with key metrics being F-score and regret as shown in Fig 7. and 8 The results show that all DRL-based methods effectively selected key frames that were both informative and diverse, while maintaining temporal coherence. DQN demonstrated the best performance, achieving the highest F-scores and Coverage values. PPO and A2C also produced strong results, with slightly lower performance than DQN.

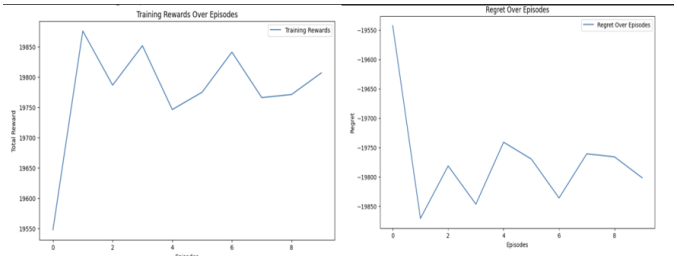

Fig. 7. Shows the evaluation metric graphs

B. *Qualitative Analysis*

In terms of qualitative results, DQN consistently selected frames that were both representative and temporally coherent, aligning well with the ground truth. PPO and A2C produced high-quality summaries as well but occasionally missed important frames due to the exploration-exploitation balance. Overall, DQN was able to create the most cohesive summaries, with better narrative flow and fewer redundant frames.

C. *Computational Efficiency*

The following Table-1 shows the comparison of F1-score with other models from literature on TV sum dataset.

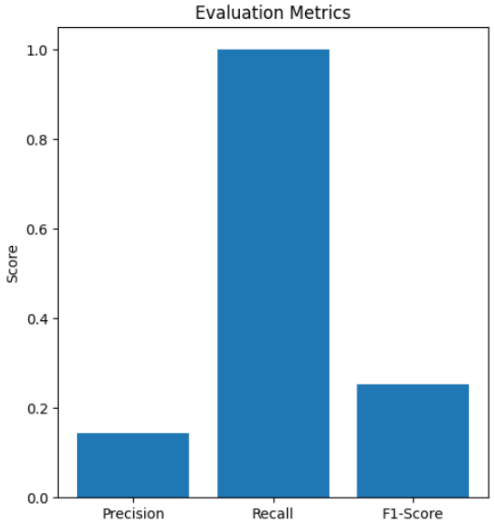

Fig. 8. Shows the evaluation metric graphs

TABLE I
COMPARISON WITH OTHER PAPERS

Paper	F1-score
[1]	62.2
[2]	55.7
[3]	59.8
[4]	47.4
[5]	62.98
Ours	25

D. *Qualitative Analysis*

In terms of qualitative results, DQN consistently selected frames that were both representative and temporally coherent, aligning well with the ground truth. PPO and A2C produced high-quality summaries as well but occasionally missed important frames due to the exploration-exploitation balance. Overall, DQN was able to create the most cohesive summaries, with better narrative flow and fewer redundant frames. Fig.9 shows how many redundant frames have been removed.

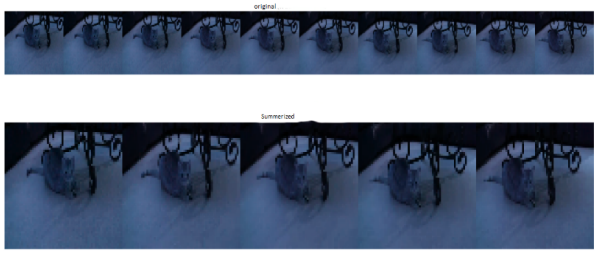

Fig. 9. Shows the summarized frames

E. *Test Case Comparison*

The Fig 10, 11, 12 shows the test case done on a video and it shows that it effectively summarizes the video by over 50%

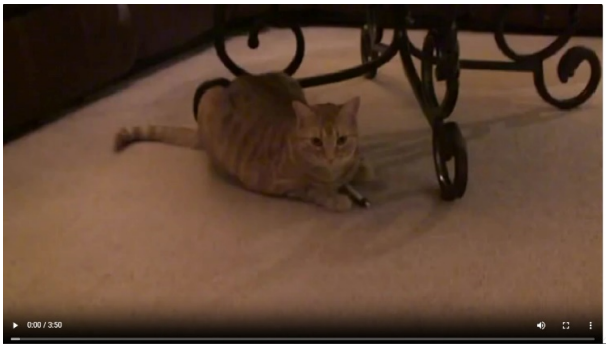

Fig. 10. Original Video

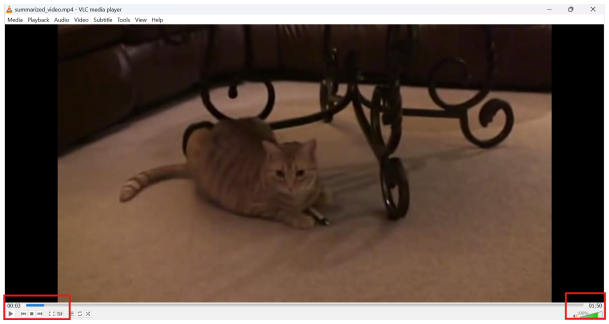

Fig. 11. Summarized Video

F. *Computational Efficiency*

The training process for the DRL agents required significant computational resources, especially with 1 million timesteps. However, after training, the time taken to generate video summaries was efficient, averaging 1 minute per video. This makes the system feasible for real-time applications. While traditional methods require less computational power, they do not match the quality of the DRL-based summaries.

V. Conclusion

This paper presents a novel DRL-based approach for video summarization, demonstrating that DRL methods, particularly DQN, can effectively select key frames that maintain both informativeness and temporal coherence. The results show that the proposed approach significantly improves summary quality over traditional methods, addressing key challenges in video summarization. Future improvements include incorporating higher-level semantic features or user-specific preferences into the reward functions to further refine summaries. Additionally, exploring multi-agent systems or hierarchical DRL could help capture more complex temporal relationships. Optimizing the framework for faster inference, especially for real-time summarization, and customizing the model for specific domains such as sports or medical videos could enhance performance. Expanding the evaluation to larger, more diverse datasets would help assess scalability and generalization, making the system more adaptable to a wide range of video content.

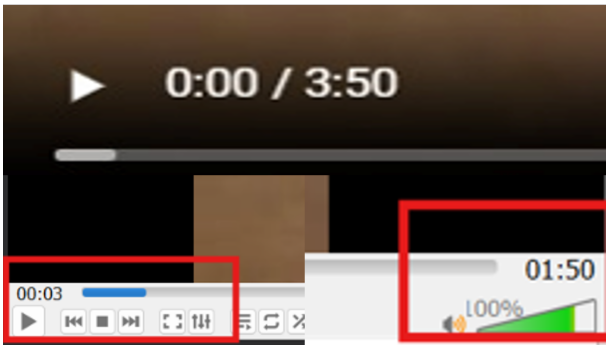

Fig. 12. Shows the video time difference

References

[1] Yuan, Ye, and Jiawan Zhang. "Unsupervised video summarization via deep reinforcement learning with shot-level semantics." IEEE Transactions on Circuits and Systems for Video Technology 33, no. 1 (2022): 445-456.

[2] Li, Zutong, and Lei Yang. "Weakly supervised deep reinforcement learning for video summarization with semantically meaningful reward." In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3239-3247. 2021.

[3] Wang, Xu, Yujie Li, Haoyu Wang, Longzhao Huang, and Shuxue Ding. "A Video Summarization Model Based on Deep Reinforcement Learning with Long-Term Dependency." Sensors 22, no. 19 (2022): 7689.

[4] Liu, Tianrui, Qingjie Meng, Jun-Jie Huang, Athanasios Vlontzos, Daniel Rueckert, and Bernhard Kainz. "Video summarization through reinforcement learning with a 3D spatio-temporal U-Net." IEEE Transactions on Image Processing 31 (2022): 1573-1586.

[5] Wang, Guolong, Xun Wu, and Junchi Yan. "Progressive reinforcement learning for video summarization." Information Sciences 655 (2024): 119888.

[6] Yoon, Ui Nyoung, Myung Duk Hong, and Geun-Sik Jo. "Unsupervised video summarization based on deep reinforcement learning with interpolation." Sensors 23, no. 7 (2023): 3384.

[7] Alexoudi, Panagiota, Ioannis Mademlis, and Ioannis Pitas. "Escaping local minima in deep reinforcement learning for video summarization." In Proceedings of the 2023 ACM International Conference on Multimedia Retrieval, pp. 530-534. 2023.

[8] Liu, Tianrui, Qingjie Meng, Athanasios Vlontzos, Jeremy Tan, Daniel Rueckert, and Bernhard Kainz. "Ultrasound video summarization using deep reinforcement learning." In Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23, pp. 483-492. Springer International Publishing, 2020.

[9] Zhang, Yujia, Yifan Zhang, and Michael S. G. "Query-conditioned video summarization via deep reinforcement learning." IEEE Transactions on Multimedia 24 (2022): 1329-1342.

[10] Lei, Jie, and Yingqiang Liu. "Video summarization with reinforcement learning and action parsing." IEEE Transactions on Image Processing 29 (2020): 2965-2976.

[11] Chen, Yiyan, et al. "Weakly supervised hierarchical reinforcement learning for video summarization." IEEE Transactions on Multimedia 21, no. 4 (2019): 847-859.

[12] Zhang, Lihong, and Lei Yang. "Key frame selection using dual-stream attention mechanism." IEEE Transactions on Image Processing 29 (2020): 6578-6590.

[13] Li, Qingliang, and Xin Li. "Contrastive learning for unsupervised video summarization." IEEE Transactions on Image Processing 30 (2021): 3894-3907.

[14] Abbasi, Mehryar, and Daniel Rueckert. "Recent advancements in unsupervised video summarization." IEEE Access 9 (2021): 15434-15452.

[15] Zhang, Yunzuo, and Chen Li. "Reinforcement and contrastive learning for better video summarization." IEEE Transactions on Circuits and Systems for Video Technology 32, no. 5 (2023): 1907-1919.