IEEE IEACon 2023
INDUSTRIAL ELECTRONICS AND APPLICATIONS CONFERENCE

An Attention based Video Summarization Technique for Wireless Capsule Endoscopy Data

[1]Department of Electronics and Communication Engineering, National Institute of Technology Calicut, India
[2]Department of Computer Science, Norwegian University of Science and Technology Gjøvik, Norway

Authors

Nitish Kumar[1]
Sudhish N George[1]
Kiran Raja[2]

Organizer: IEEE

Technical Co-Sponsor: IAS IEEE INDUSTRY APPLICATIONS SOCIETY Linking Research to Practice

Table of content

Introduction

Wireless Capsule Endoscopy (WCE)

- WCE is a non-invasive medical diagnosis tool which is used to record a video of the patient's gastrointestinal (GI) tract.
- A small pill sized camera capsule is swallowed by the patient and the capsule captures and sends its images to an external receiver worn by the patient.
- It is suitable to capture images of the small intestine which is not accessible with conventional endoscopic methods.

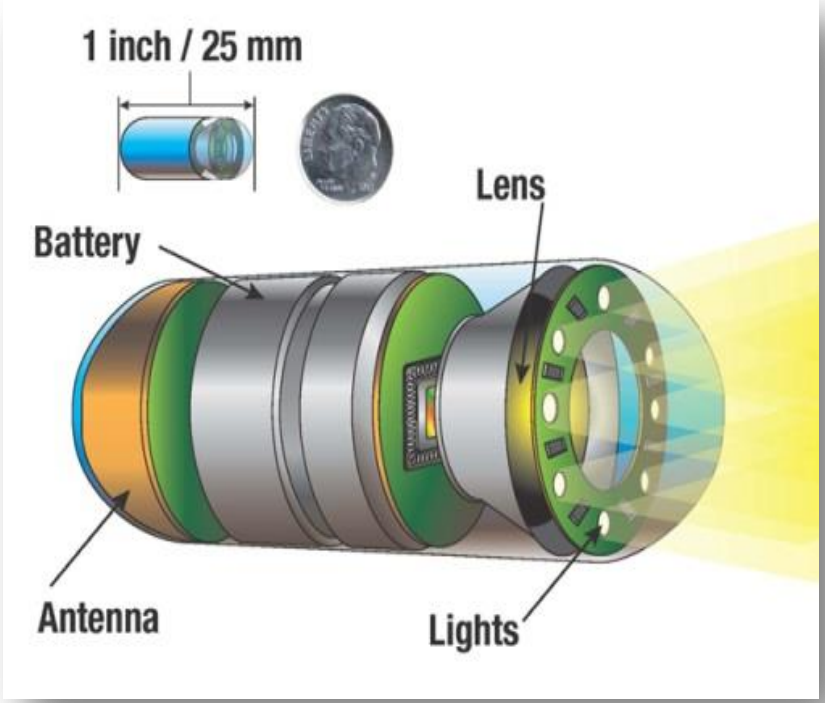

Figure 1: WCE Capsule [1]

[1]https://mydoctor.kaiserpermanente.org/ncal/Images/Endoscopy

Introduction

Wireless Capsule endoscopy is used for the following purposes :

1. Find the cause of gastrointestinal bleeding.
2. Diagnose inflammatory bowel diseases.
3. Diagnose cancer.
4. Diagnose celiac disease.
5. Screen for polyps, etc.

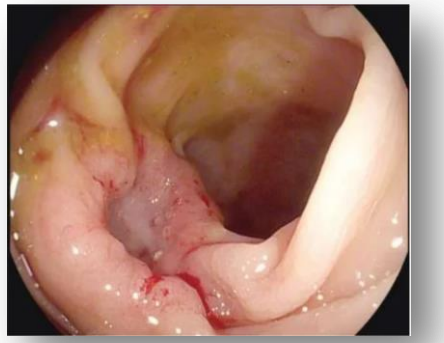

Figure 2: Bleeding

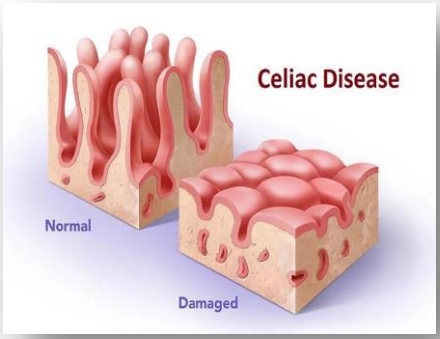

Figure 3: Celiac disease

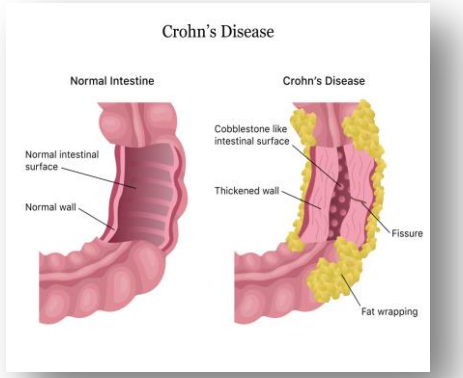

Figure 4: Crohns-Disease

Figure 5: Polyps

Introduction

What am I looking at ?

- The capsule travels at a very slow speed and captures images at the rate of 30 frames per second.
- Capsule endoscopy video is 8 to 12 hours long.
- Slow movement results in huge number of frames, some of which are redundant with high structural similarity.

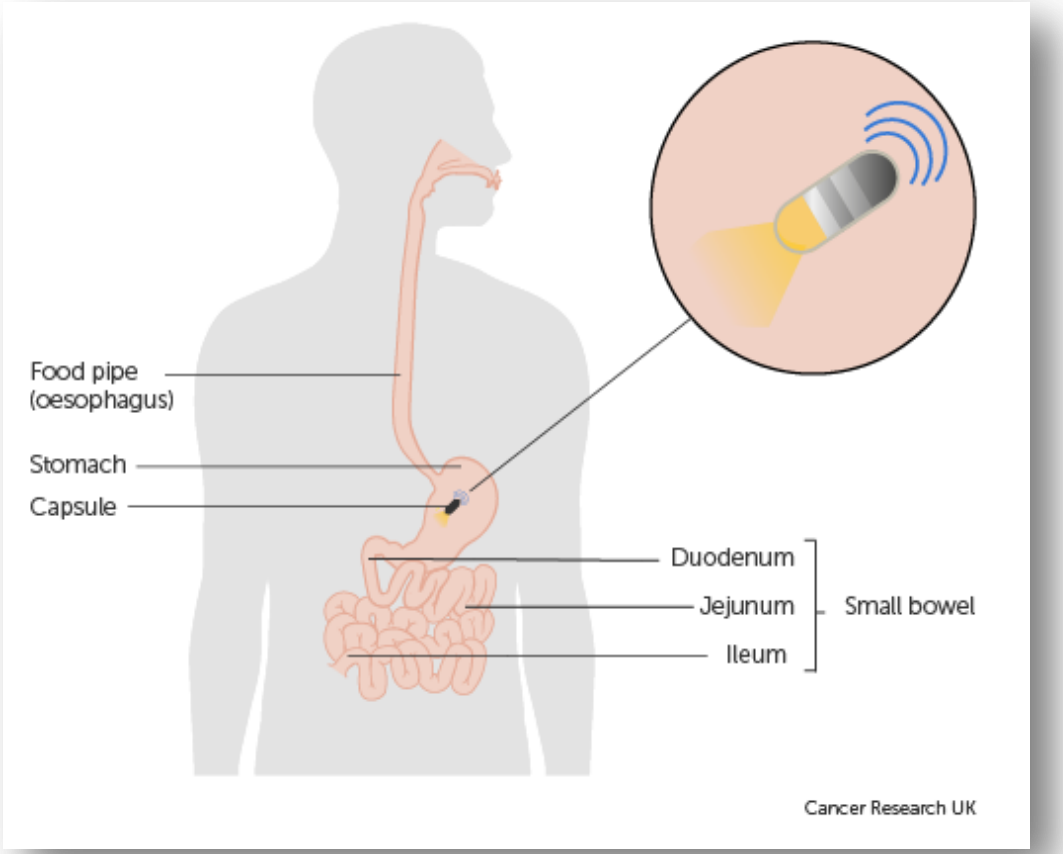

Figure 6: Movement of Capsule into GI tract[2]

[2]B. Sushma and P. Aparna. “Summarization of Wireless Capsule Endoscopy Video Using Deep Feature Matching and Motion Analysis”. In: IEEE Access 9 (2021)

Introduction

The solution to the above problem is:

Video Summarization

Anatomy of a Video:

- Frame: a single still image from a video.
- Shot: sequence of frames recorded in a single camera operation.
- Scene: collection of shots forming a semantic unity (Conceptually, a single time and place)

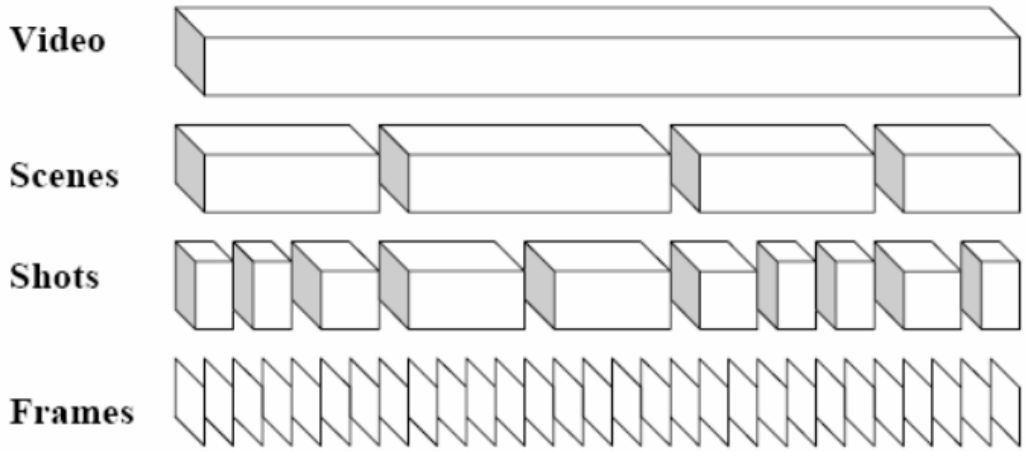

Figure 7: Anatomy of a video[3]

[3]Ben Wing CS 395T Ben Wing. “Video Summarization”. In: Video Summarization. 2008, p.

Video Summarization (VS)

Selecting a small batch of frames from the video data, consisting of large number of video frames, to describe the whole content of original video.

- Summarized video frames may be less than or equal to the number of frame present in the input video.

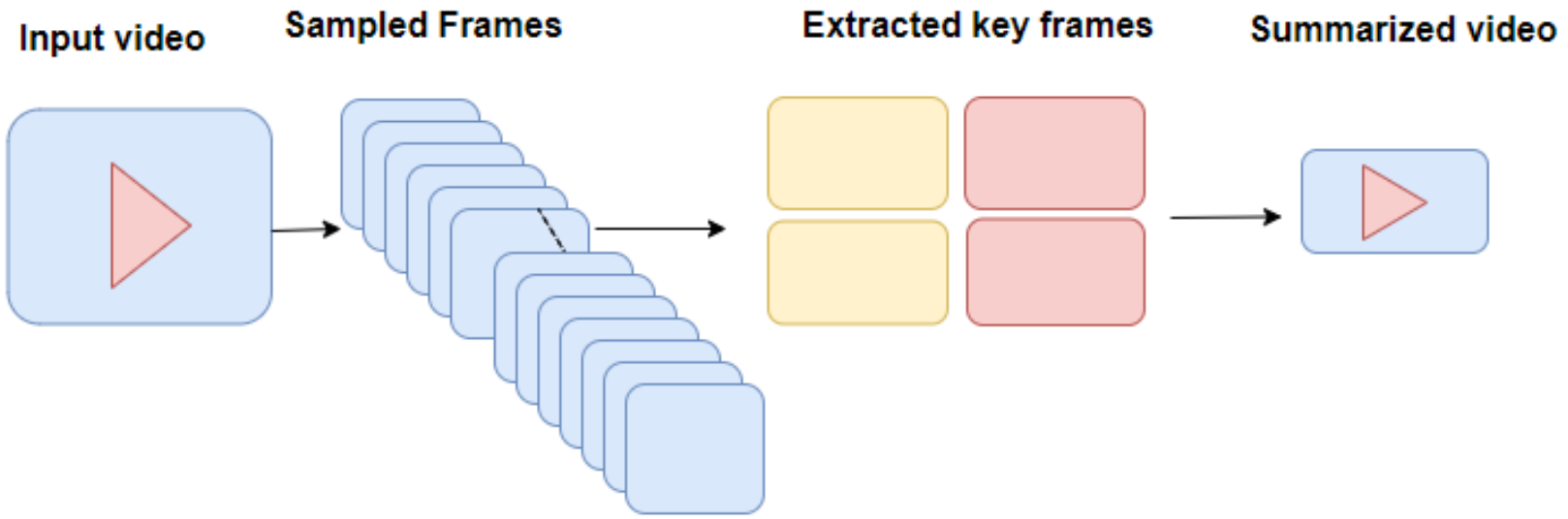

Figure 8: Video Summarization[4]

[4]Evlampios Apostolidis et al. “Video Summarization Using Deep Neural Networks: A Survey”. In: Proceedings of the IEEE 109 (2021), pp. 1838-1863.

Generalized Approach of VS

- The below shown approach is for both supervised machine learning and unsupervised machine learning.

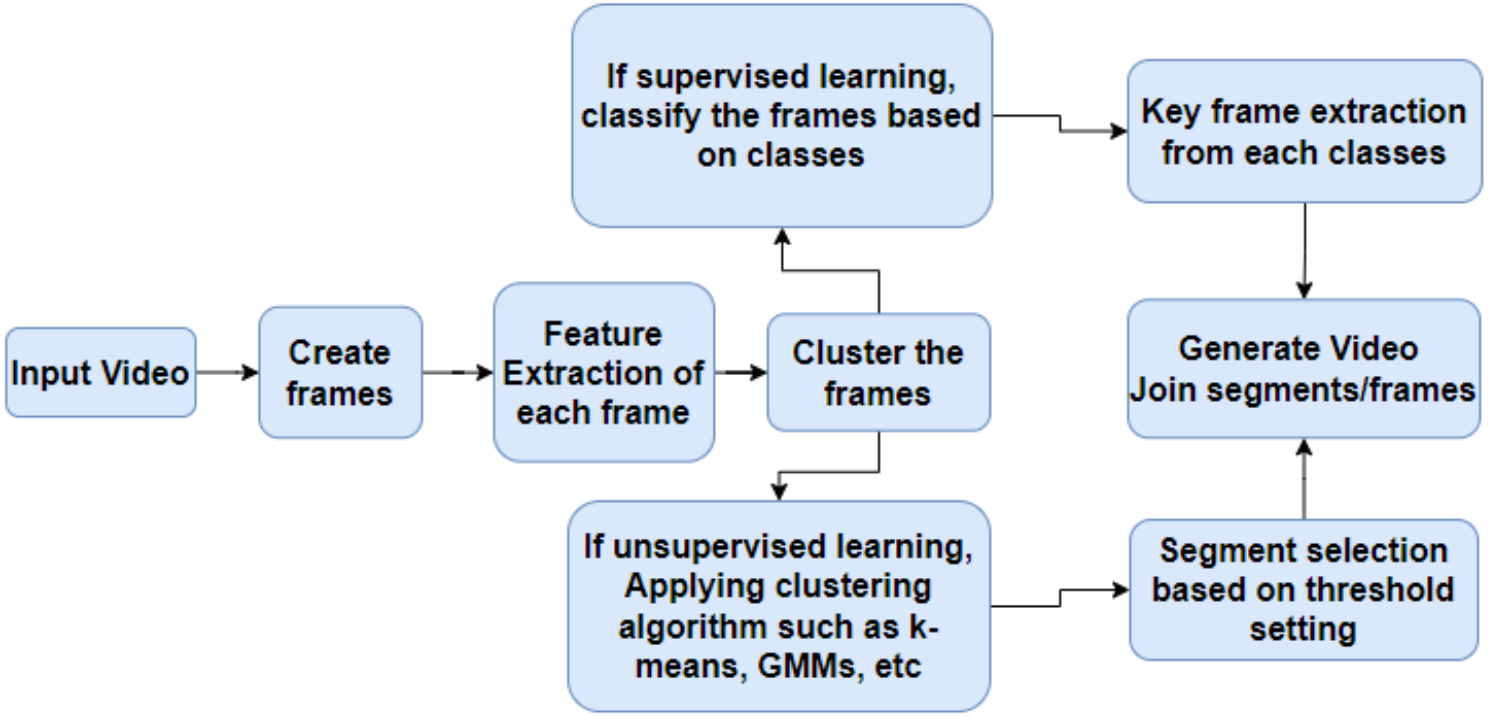

Figure 9: Block diagram of VS[5]

[5]Sushma and P. Aparna. “Summarization of Wireless Capsule Endoscopy Video Using Deep Feature Matching and Motion Analysis”. In: IEEE Access 9 (2021), pp. 13691–13703

Overview of Existing VS Approaches

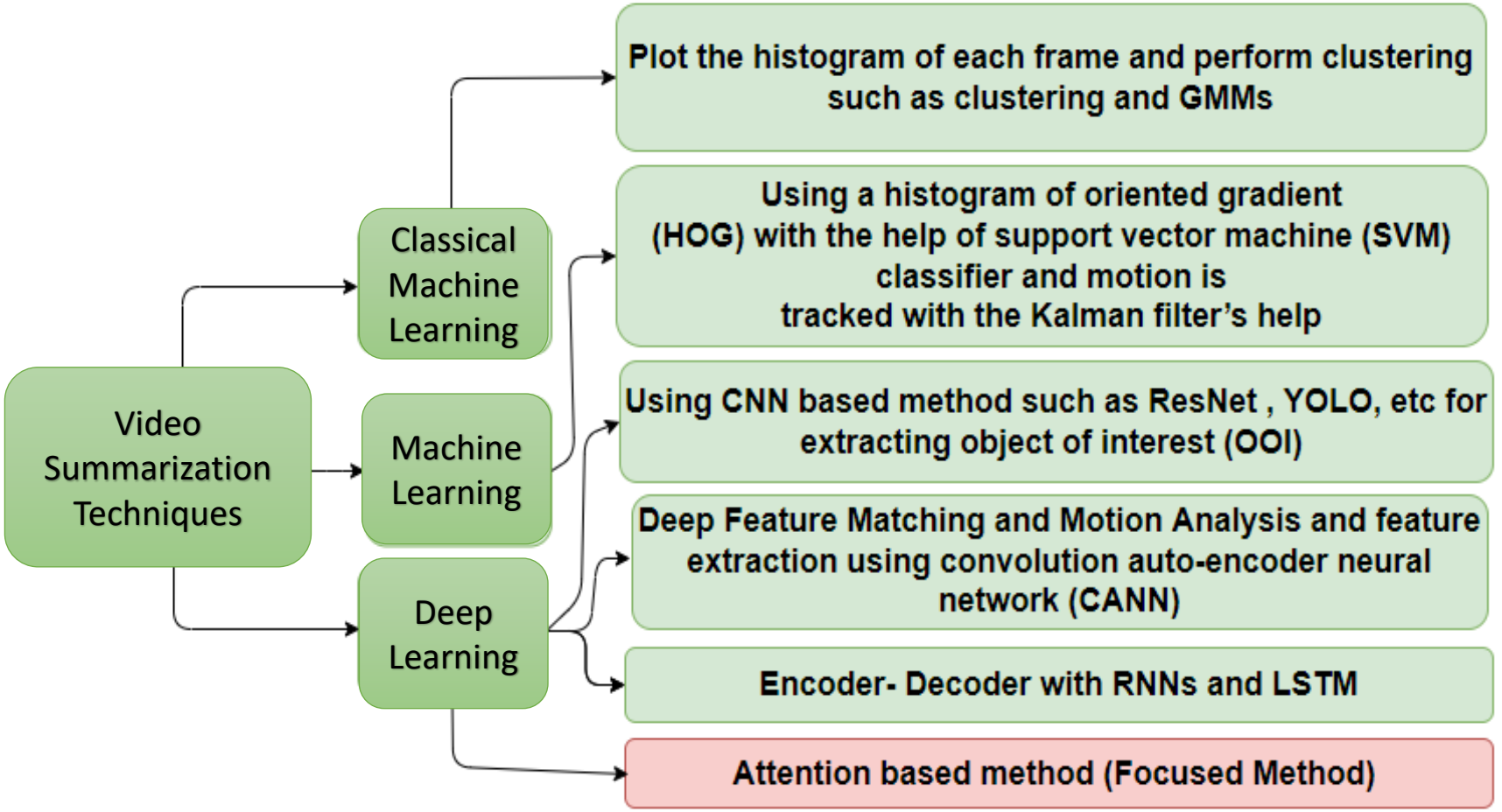

Figure 10: Block diagram of existing methods[6]

[6]Evlampios Apostolidis et al. "Video Summarization Using Deep Neural Networks: A Survey". In: Proceed- ings of the IEEE 109 (2021), pp.1838–1863.

Motivation

- In WCE, the patient can carry out their daily activities without being hospitalized.
- Capsule gives about 12 hours of video, which needs to be analyzed.
- Video summarization provides the summary of the entire content of the video into a few keyFrames. As a result, the doctor’s examination time can be reduced.
- Existing methods of video summarization have the drawbacks such as existing models being complex, and results may not be human semantic in case of unlabeled data.
- The aim of this work is to develop an attention-based model which can overcome these drawbacks.

Problem Definition

- The problem definition of this work is to design and develop an efficient attention-based deep learning model to summarize the video sequence of wireless capsule endoscopy data by selecting its most informative and important frames.

Methodology

Introduction of Attention Network:

To understand proposed Attention Network, we need to discuss below topics:

1. What is Attention ?
2. Long Short-Term Memory (LSTM) Network
3. Why BiLSTM Network ?
4. Encoder with BiLSTM
5. Decoder with Attention Mechanism
6. KeyShot Selection

What is Attention ?

- In psychology, attention is the cognitive process of selectively concentrating on one or a few things while ignoring others.
- It is also an attempt to implement the same action of selectively concentrating on a few relevant things while ignoring others in deep neural networks.
- This is the kind of “Attention” that our brain is exceptionally good at executing.

Figure 11: Group of children with their teacher[7]

[7]https://www.123rf.com/stock-photo/teacherstudentc artoon.html.

Methodology

Long Short-Term Memory (LSTM) Network:

- Long Short-Term Memory networks – usually just called “LSTMs” – are a special kind of RNN.
- LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time.
- It is also used to address vanishing gradient problem occurs in RNNs.

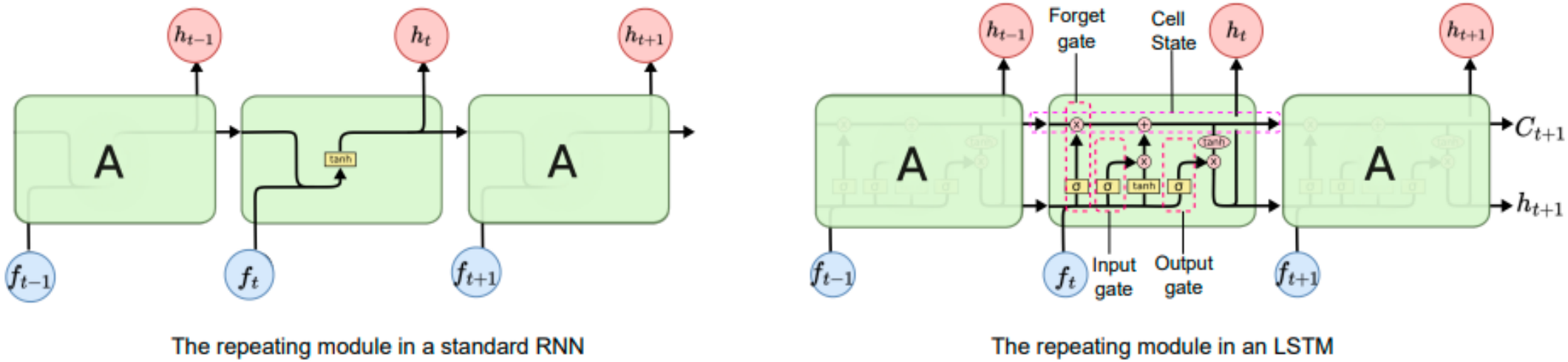

Figure 12: RNN Networks[8]

[8]https://colah.github.io/posts/2015-08-Understanding-LSTMs/fn1

Methodology

Encoder with BiLSTM:

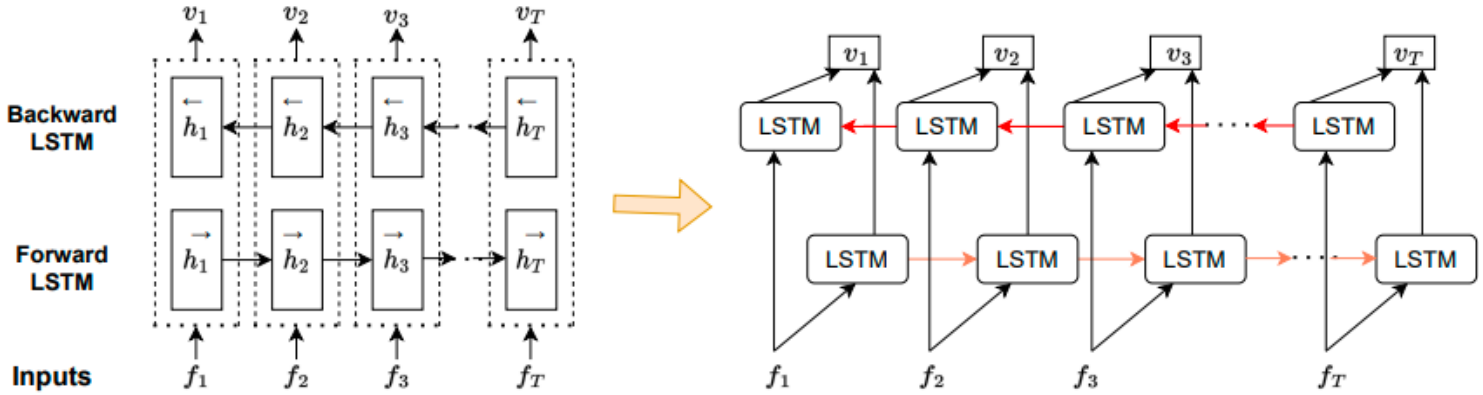

Figure 13: BiLSTM Network[9]

- An encoder converts the input sequence for encoder F = {f_1, f_2, f_3, ..., f_T } into a representative vector V = {v_1, v_2, v_3, ..., v_T }.
- Where $h_t \in R^D$ is a hidden state at time t with dimensions D, extracted for each video frame.
- The forward LSTM reads input frames in a forward direction from f_1 to f_T and calculates its forward hidden states h_1 to h_T.
- Similarly Backward LSTM reads input from f_T to f_1 and then it gives an annotation v_t for each f_t by concatenating the $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$.
- The annotation v_t incorporates the information of both before the f_t frames and after the f_t frames.
- Due to this time tendency of LSTM, then annotation can focus on the frames around f_t.

[9]Zhong Ji et al. “Video Summarization With Attention-Based Encoder–Decoder Networks”. In: IEEE Trans- actions on Circuits and Systems for Video Technology 30.6 (2020), pp. 1709–1717. doi: 10.1109/TCSVT.2019.2904996

Methodology

Decoder with Attention Mechanism: The main components used in architecture are the following

- The decoder is responsible for generating the output sequence Y , which consists of elements Y = $\{y_1, y_2, y_3, ..., y_T\}$ using the representation vector obtained from the encoder.

- The design of the decoder, denoted as ψ and an LSTM decoder can be formulated as :

$$\begin{bmatrix} p(yt|\{y_i|i < t\}, v) \\ S_t \end{bmatrix} = \psi(S_{t\text{-}1}, y_t, v) \qquad (1)$$

Decoder with Attention Mechanism:

- The LSTM decoder with the output of the encoder at time step i. The computation of the relevance score α_t^i can be expressed using a score function

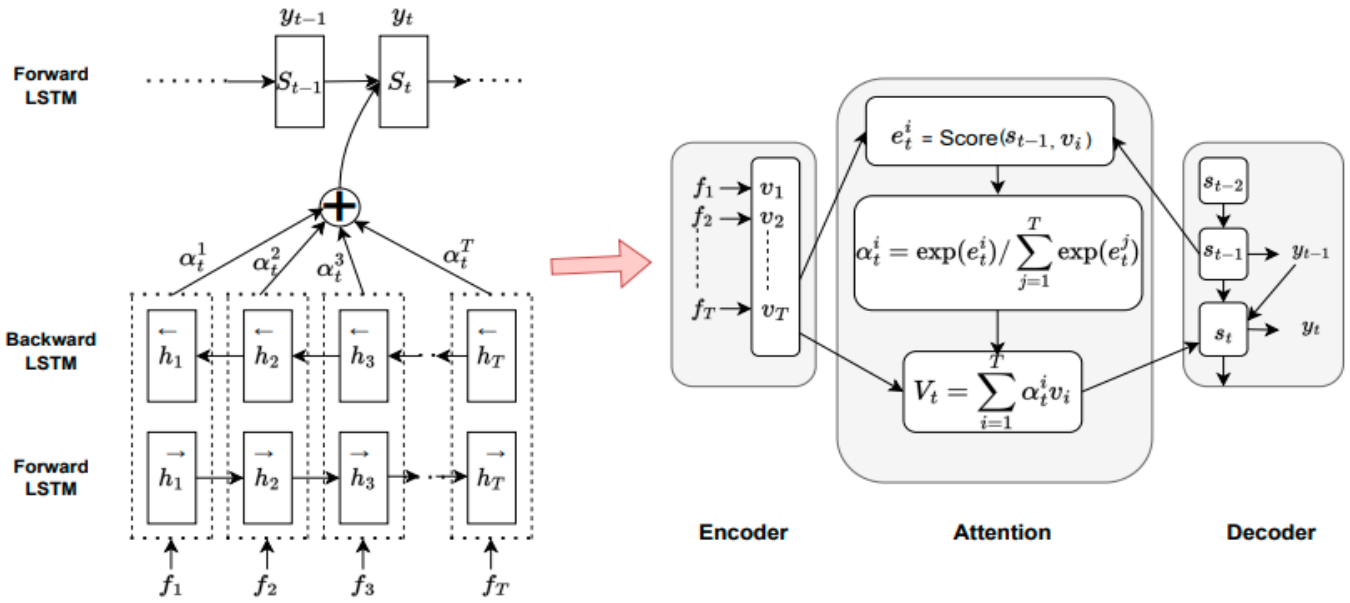

Figure 14: BiLSTM With Attention and its Block Diagram[10]

[10]https://machinelearningmastery.com/the-luong-attention-mechanism/

Methodology

Why BiLSTM Network ?

In Capsule Endoscopy the frame specificity is based on neighborhood such as frames before f_t and frames after f_t:

This is inspired by BiLSTM

- The principle of BiLSTM is to split the neurons of regular LSTM, into two directions.
- One in a positive time direction called a forward state and the other in a negative time direction called a backward state.

Methodology

Decoder with Attention Mechanism:

- Once the relevance scores $e_t^i = \text{score}(s_{t-1}, v_i)$ for all frames $i = 1, \ldots, \dot{T}$ are computed, we normalize them to obtain the α_t^i by:

$$\alpha^i{}_t = \exp(e^i{}_t) \,/\, \sum_{j=1}^{T} \exp(e^i{}_t) \qquad (2)$$

- The decoder decides which parts of the source frames to pay attention. Then the importance score of each frame can be computed.
- Representative vector for entire sequence can be represented as a weighted sum of the annotations:

$$V_t = \sum_{i=1}^{T} \alpha^i{}_t \, v_i \qquad (3)$$

Using the concepts of attention, the decoder can be represented as:

$$\begin{bmatrix} p(yt|\{y_i|i < t\}, Vt) \\ S_t \end{bmatrix} = \psi(S_{t\text{-}1}, y_t, V_t) \qquad (4)$$

- Where, v_t stands for the attention vector at moment $\dot{t}$ and due introduction of attention in decoder output y_t depends on entire input sequence.

- **The Encoder**: The role of the encoder is to generate an annotation v_i for every frame f_i in an input frame of length $\dot{T}$ sequence.
- **The Decoder**: The role of the decoder is to produce the target frame by focusing on the most relevant information contained in the source sequence for this purpose, it makes use of an attention mechanism.

Methodology

KeyShots Selection:

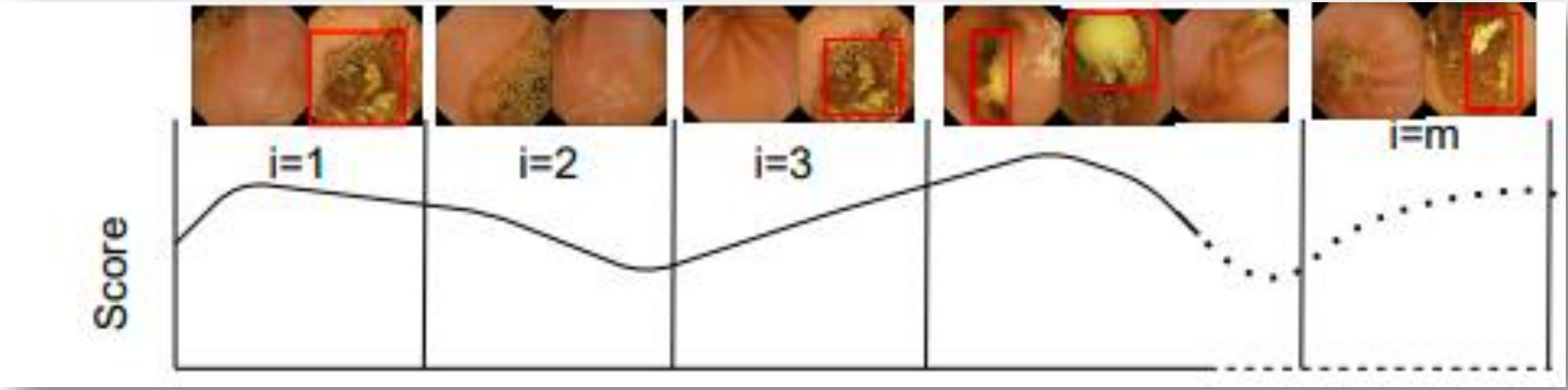

Figure 15: Temporal Segmentation With KTS

- Once obtained the predicted importance scores for all frames, the remaining work is to select the keyShots to generate the VS.
- Apply Kernel Temporal Segmentation (KTS) to segment the visually coherent frame into shots.
- Then it computes shot-level importance scores by taking an average of the frame importance scores within each shot to generate keyShot based summary.

Work Done and Results

Dataset Overview

- Kvasir Capsule is the largest gastrointestinal PillCAM dataset. The dataset contains a total of 43 videos.
- This corresponds to approximately 19 hours of video and 1,955,675 video frames. Each video has been manually assessed by a medical professional working in the field of gastroenterology and resulted in a total of 47,238 annotated frames.
- In this work, the dataset used contains 25 edited videos from available Kvasir Capsule PillCAM dataset.

Data-set link: https://datasets.simula.no/kvasir-capsule.

Ground Truth Preparation:

- This model is trained using frame-level scores, these frame-level scores are obtained by extracting frames from an input video.
- And giving annotation scores by 5 different users on a scale of 1 to 5.
- If the score is 1, the frame is considered less important, and the higher scores are considered as more important.
- The average of annotation scores by 5 different users is used as an importance score. So, each frame in a video is assigned an importance score.

Table 1: Overview of the Kvasir Capsule dataset

Dataset	Videos	Frame Rate	User Annotation Type	Max Length	Min Length
Kvasir Capsule	25	30/Sec	Frame-level importance scores	288 Sec	80 Sec

Work Done and Results

Evaluation Metric:

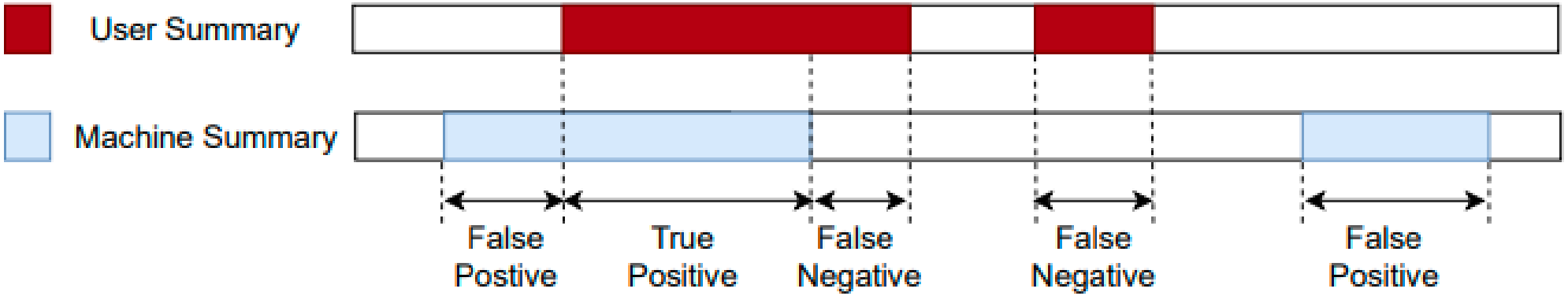

Figure 16: Pictorial Representation of Confusion Matrix

- To ensure a fair comparison with the state-of-the-art, we employ an evaluation protocol.
- This work utilize the harmonic mean of precision and recall, represented as the F-score, to determine how closely the user and machine summaries resemble each other as shown in Fig 16.

$$\text{F-Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

Work Done and Results

Evaluation Metric:

- True positives, False positives and False negatives are calculated per-frame between the ground truth and machine binary keyShot summaries. Recall and Precision defined as:

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative} \qquad (6)$$

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive} \qquad (7)$$

- Average F-score over videos in the dataset is then reported. On the Kvasir Capsule, for each video, a user summary most like to the machine summary is selected.

Work Done and Results

Results:

- The results obtained using proposed approach is presented in Table 2 which included Precision, Recall and F- Score with standard deviation.
- Also, the mean percentage reduction of videos achieved by this work on Kvasir capsule dataset.
- The mean percentage reduction achieved by this work is 81.63%.

Table 2: Mean Performance obtained in VS using attention and without attention mechanism

Approach	Precision	Recall	F-Score
BiLSTM without Attention	84.16 ± 4.97	77.55 ± 4.91	80.63 ± 4.93
BiLSTM with Attention	**87.02 ± 1.35**	**82.37 ± 3.33**	**84.57 ± 2.23**

Summarization Result			
Kvasir Capsule dataset	**Duration**		**Percentage**
	Before VS	**After VS**	**Video Reduced**
With Attention (Proposed Approach)	147 sec	27 sec	**81.63%**

Work Done and Results

Proposed Evaluation Metric:

- In this work, a new concept called the "$\dot{N}$ off " strategy is introduced for calculating accuracy in the context of VS for WCE data.
- The $\dot{N}$ off strategy considers an acceptable error margin of $\pm\dot{N}$ frames when evaluating accuracy.
- This means that for each frame, $\dot{N}$ frames before and after $\dot{N}$ off are considered acceptable frames for pathology detection.
- The accuracy is defined as:

$$\text{Accuracy} = \frac{Predicted \cap \sum_{i=\pm n}^{\dot{N}} Groundtruth}{\sum_{i=\pm n}^{\dot{N}} Groundtruth} \qquad (8)$$

Result for proposed strategy:

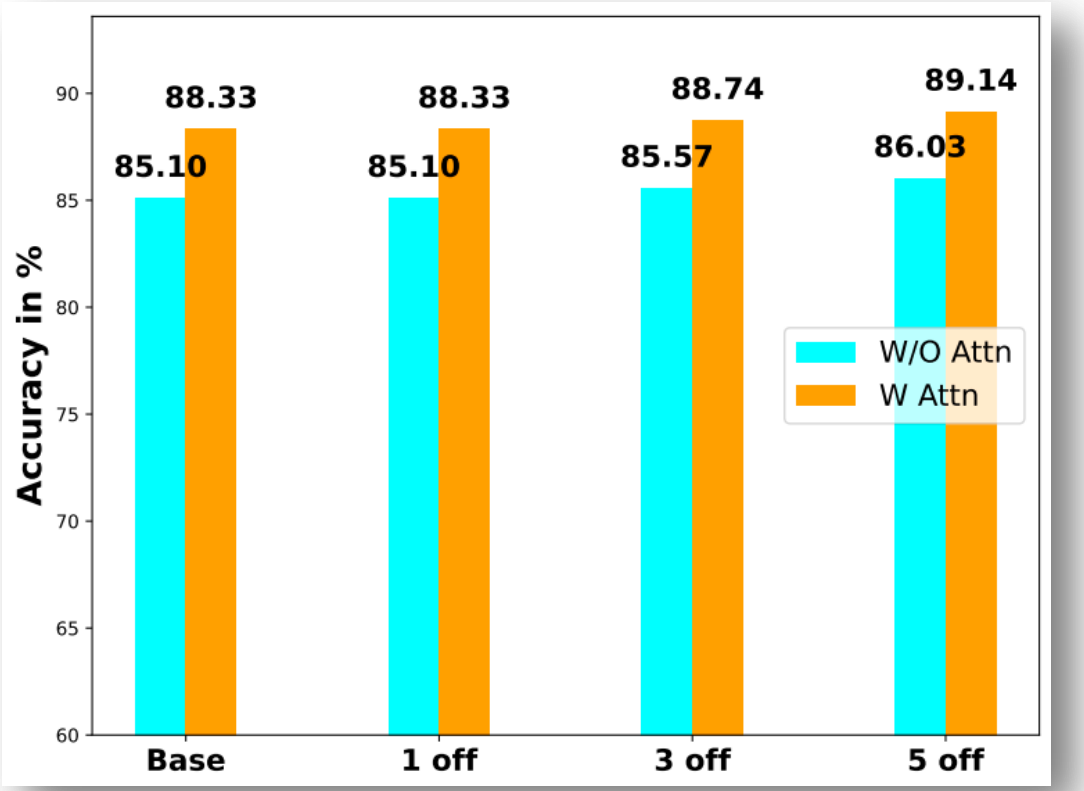

Figure 17: Accuracy for proposed N off decision strategy

Work Done and Results

Input Video

Video 1: Input Vidéo duration 106 sec

Output Video

Video 2: Output Vidéo duration 19 sec.

.

Conclusions and Future Scope

- Unlike natural scene videos, WCE videos pose a unique challenge as they lack significant texture differences between frames and videos.
- These videos typically span a duration of **8-12 hours**, with a frame rate of **30 (fps)** frames per second and a resolution of **336 × 336** pixels.
- In capsule endoscopy, the frames surrounding a diseased frame are also significant. To effectively tackle this challenge, the incorporation of Bidirectional LSTM (BiLSTM) with an attention mechanism is crucial.
- The achieved results show the proposed method's effectiveness in handling WCE data, underscoring its efficacy within this domain. The proposed approach results in an average **F-Score of 84.57%** showing promising performance.
- There is scope of introducing transformer-based attention module and compare the same with this work and observe if there are improvements in the summarization performance.

References I

1. Evlampios Apostolidis et al. “Video Summarization Using Deep Neural Networks: A Survey”. In: Proceedings of the IEEE 109 (2021), pp. 1838–1863
2. B. Sushma and P. Aparna. “Summarization of Wireless Capsule Endoscopy Video Using Deep Feature Matching and Motion Analysis”. In: IEEE Access 9 (2021), pp. 13691–13703.
3. Abbas Biniaz, Reza Aghaeizadeh Zoroofi, and Masoud Reza Sohrabi. “Automatic reduction of wireless capsule endoscopy reviewing time based on factorization analysis”. In: Biomedical Signal Processing and Control 59 (2020), p. 101897. issn: 1746-8094.
4. Yu-Fei Ma et al. “A User Attention Model for Video Summarization”. In: Proceedings of the Tenth ACM International Conference on Multimedia. Association for Computing Machinery, 2002, pp. 533–542. isbn: 158113620X.
5. M. Maher Ben Ismail, Ouiem Bchir, and Ahmed Z. Emam. “Endoscopy video summarization based on unsupervised learning and feature discrimination”. In: 2013 Visual Communications and Image Processing (VCIP). 2013, pp. 1–6.
6. Zhong Ji et al. “Video Summarization With Attention-Based Encoder–Decoder Networks”. In: IEEE Trans-actions on Circuits and Systems for Video Technology 30.6 (2020), pp. 1709–1717. doi: 10.1109/TCSVT.2019.2904996.
7. Ben Wing CS 395T Ben Wing. “Video Summarization”. In: Video Summarization. 2008.
8. Stefania Cristina. The Bahdanau Attention Mechanism. Sept. 2022

References II

9. Xingquan Zhu et al. “Insight Video: toward hierarchical video content organization for efficient browsing, summarization and retrieval”. In: IEEE Transactions on Multimedia 7.4 (2005), pp. 648–666. doi: 10.1109/TMM.2005.850977.
10. Yixuan Yuan and Max Q.-H. Meng. “Hierarchical key frames extraction for WCE video”. In: 2013 IEEE International Conference on Mechatronics and Automation. 2013, pp. 225–229. doi: 10.1109/ICMA.2013.6617922.
11. Jia Sen Huo, Yue Xian Zou, and Lei Li. “An advanced WCE video summary using relation matrix rank”. In: Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics. 2012, pp. 675–678. doi: 10.1109/BHI.2012.6211673.
12. Xuming Feng, Lei Wang, and Yaping Zhu. “Video Summarization with Self-Attention Based Encoder-Decoder Framework”. In: 2020 International Conference on Culture-oriented Science Technology (ICCST). 2020, pp. 208–214. doi: 10.1109/ICCST50977.2020.00046.
13. Gokhan Yalınız and Nazli Ikizler-Cinbis. “Unsupervised Video Summarization with Independently Recurrent Neural Networks”. In: 2019 27th Signal Processing and Communications Applications Conference (SIU). 2019, pp. 1–4. doi: 10.1109/SIU.2019.8806603.
14. Amin Zollanvari et al. “Transformer Fault Prognosis Using Deep Recurrent Neural Network Over Vibration Signals”. In: IEEE Transactions on Instrumentation and Measurement 70 (2021), pp. 1–11. doi: 10.1109/TIM.2020.3026497.