SUMMARIZATION OF LARGE-SCALE VIDEOS TO TEXT FORMAT USING SUPERVISED BASED SIMPLE RULE-BASED MACHINE LEARNING MODELS

U.K.H.A. Sugathadasa

199488F

Department of Computational Mathematics,

Faculty of Information Technology

University of Moratuwa

Sri Lanka

March 2022

SUMMARIZATION OF LARGE-SCALE VIDEOS TO TEXT FORMAT USING SUPERVISED BASED SIMPLE RULE-BASED MACHINE LEARNING MODELS

Udage Kankanange Harindu Ashan Sugathadasa

199488F

Thesis/Dissertation submitted in partial fulfilment of the requirements for the degree Master of Science in Artificial Intelligence

Department of Computational Mathematics

University of Moratuwa

Sri Lanka

March 2022

Declaration

I declare that this is my own work, and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Name of the Student:	Signature:	Date:
Sugathadasa U.K.H.A.	……	…….

The above candidate has carried out research for the Masters/MPhil/PhD thesis/ dissertation under my supervision.

Name of the Supervisor:	Signature of the supervisor:	Date:
Dr. Subha Fernando	……	…….

Name of the Co-Supervisor:	Signature of the co-supervisor:	Date:
Dr. Varuna De Silva	……	…….

Abstract

Video Summarization has been one of the most interested research and development field since the late 2000s, thanks to the evolution of social media and the internet, due to the influence to provide a concise and meaningful summary of large-scale video. Even though the video summarization has been elongated through several non-ML and traditional based techniques and ML-based techniques, generation of correct and required summaries from the video is yet a limitation. To overcome this concern, different techniques have been attempted including vision-based approaches and NLP related approaches. With the inspiration of NLP related Transformer networks, researchers are looking to integrate such sequence-based learning algorithm into the video dimension as to apply spatiotemporal extractions. Despite the VS implementations, another extension of VS has been exponentially emphasized, namely TVS which generates the summaries of the video via a text format.

Simply the evolution of VS towards TVS is not a straightforward journey since a lot of blockers have been eliminated using UL, RL, and SL based frameworks. When it comes to the STOA methods in TVS, Transformer based methods are eventually highlighted along the T5 based NLP frameworks. Since this area is still at the ground level, a lot of unknow facts and issues can be explored. Especially the attention-based sequence modelling of the learning algorithm should be carefully imitated to achieve the best accuracy improvements. All the improvements are subjected to apply into a real-time application ulteriorly. To tackle such improvements, a novel standalone method should be introduced with the simplest network layout which can be applicable to the embedded devices. This is where the **Si**mple **Ru**le-based **M**achine **L**earning Network to **T**ext-based **V**ideo **S**ummarization (SiRuML-TVS) has been unveiled.

Though the network contains a single input of large-scale video and a single output of meaningful description for the given video, the high-level network layout compromises three ML modules for Video Recognition, Object Detection, and finally Text Generation. Each module is subjected to different evaluation criterions however, the end-to-end full network is evaluated on a single metric. Different combination of each module can be affected to the performance of the entire pipeline however, the

combination of Transformers and CNNs provide the better tradeoff between accuracy and the computational inferencing. This makes a hope to deploy the proposed method in an edged device thus, the gap between theoretical explanation to practical implementation will be filled.

Dedication

I dedicate this thesis to my parents, my sister, my wife, the university lecturers from undergraduate level who are always withstand in my successes and failures.

Acknowledgements

I would like to pay my greatest appreciation to Dr. Subha Fernando, who motivated my enthusiasm on research and provided guidance and advises through the research journey till the end. Her experience on research was tremendously big support to this work, and I be indebted a great deal of its success for herself.

Every person has its own shadow when at be any place. My second appreciation is to Dr. Varuna De Silva, who is from University of Loughborough, for supporting both me and supervisor when we had a doubtful circumstances and unsolvable stakes.

Next, I would like to elapse my thanks and appreciation to the staff members of the Department of Computation Mathematics, for their supportiveness and consideration for my research work.

Lastly, I would also like to express appreciation through my bottom of heart for all my Lecturers from the Undergraduate program, my fellow colleagues from the University, and all my family members, who supported me through the journey of research program in both verbally and mentally. Even though they might not know about this, the impact from them is substantial, and I might not make it to where I am now at without themselves.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

Abbreviation	Definition
VS	Video Summarization
TVS	Text-based Video Summarization
SL	Supervised Learning
UL	Unsupervised Learning
RL	Reinforcement Learning
WSL	Weakly-Supervised Learning
GAN	Generative Adversarial Network
LSTM	Long-Short Term Memory
NLP	Natural Language Processing
BiLSTM	Bi-Directional Long-Short Term Memory
CNN	Convolutional Neural Network
CSN	Channel Separated Network